Smart Innovation, Systems and Technologies 382

Kazumi Nakamatsu
Srikanta Patnaik
Roumen Kountchev *Editors*

# AI Technologies and Virtual Reality

## Proceedings of 7th International Conference on Artificial Intelligence and Virtual Reality (AIVR 2023)

KES International

Springer

# Smart Innovation, Systems and Technologies

Volume 382

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST), SCImago, DBLP.

All books published in the series are submitted for consideration in Web of Science.

Kazumi Nakamatsu · Srikanta Patnaik ·
Roumen Kountchev
Editors

# AI Technologies and Virtual Reality

## Proceedings of 7th International Conference on Artificial Intelligence and Virtual Reality (AIVR 2023)

Springer

*Editors*
Kazumi Nakamatsu
University of Hyogo
Kobe, Japan

Roumen Kountchev
Technical University of Sofia
Sofia, Bulgaria

Srikanta Patnaik
Interscience Institute of Management
and Technology
Bhubaneswar, India

ISSN 2190-3018 ISSN 2190-3026 (electronic)
Smart Innovation, Systems and Technologies
ISBN 978-981-99-9017-7 ISBN 978-981-99-9018-4 (eBook)
https://doi.org/10.1007/978-981-99-9018-4

# AIVR2023 Organization

## Honorary Chair

Lakhmi C. Jain, KES International, Australia

## General Chairs

Ruidong Li, Kanazawa University, Japan
Kazumi Nakamatsu, University of Hyogo, Japan

## Conference Chair

Ari Aharari, Sojo University, Japan

## International Advisory Board Chair

Srikanta Patnaik, Interscience Research Network (IRNet), India

## International Advisory Board

Roumen Kountchev, Technical University of Sofia, Bulgaria
Xiang-Gen Xia, University of Delaware, USA
Shrikanth (Shri) Narayanan, University of Southern California, USA

Hossam Gaber, Ontario Tech University, Canada
Jair Minoro Abe, Paulista University, Brazil
Mario Divan, National University de la Pampa, Argentina
Chip Hong Chang, Nanyang Technological University, Singapore
Aboul Ela Hassanien, Cairo University, Egypt
Ari Aharari, Sojo University, Japan

## Program Chairs

Mohd Zaid Abdullah, Universiti Sains Malaysia, Malaysia
Minghui Li, The University of Glasgow, Singapore
Letian Huang, University of Electronic Science and Technology of China, China

## Technical Program Committee

Michael R. M. Jenkin, York University, Canada
Georgios Albanis, Centre for Research and Technology, Greece
Nourddine Bouhmala, Buskerud and Vestfold University College, Norway
Pamela Guevara, University of Concepción, Chile
Joao Manuel R. S. Tavares, University of Porto, Portugal
Punam Bedi, University of Delhi, India
Der-Chyuan Lou, Chang Gung University, Taiwan
Chang Hong Lin, National Taiwan University of Science and Technology, Taiwan
Tsai-Yen Li, National Chengchi University, Taiwan
Yi-Jao Chen, National University of Kaohsiung, Taiwan
Zhang Yu, Harbin Institute of Technology, China
Yew Kee Wong, Jiangxi Normal University, China
Lili Nurliyana Abdullah, University Putra Malaysia, Malaysia
S. Nagarani, Sri Ramakrishna Institute of Technology, India
Liu Huaqun, Beijing Institute of Graphic Communication, China
Jun Lin, Nanjing University, China
Hasan Kadhem, American University of Bahrain, USA
Gennaro Vessio, University of Bari, Italy
Romana Rust, ITA Institute of Technology in Architecture, Switzerland
S. Anne Susan Georgena, Sri Ramakrishna Institute of Technology, India
Juan Gutiérrez-Cárdenas, Universidad de Lima, Peru
Devendra Kumar R. N., Sri Ramakrishna Institute of Technology, Coimbatore, India
Shilei Li, Naval University of Engineering, China
Jinglu Liu, The Open University of China, China
Aiman Darwiche, Instructor and Software Developer, USA
Alexander Arntz, University of Applied Sciences Ruhr West, Germany

Mariella Farella, University of Palermo, Italy
Daniele Schicchi, University of Palermo, Italy
Liviu Octavian Mafteiu-Scai, West University of Timisoara, Romania
Shivaram, Tata Consultancy Services, India
Niket Shastri, Sarvajanik College of Engineering and Technology, India
Gbolahan Olasina, University of KwaZulu-Natal, South Africa
Amar Faiz Zainal Abidin, Universiti Teknikal Malaysia Melaka, Malaysia
Muhammad Naufal Bin Mansor, Universiti Malaysia Perlis (UniMAP), Malaysia
Le Nguyen Quoc Khanh, Nanyang Technological University, Singapore
Pavlo Maruschak, Ternopil Ivan Puluj National Technical University, Ukraine
Dario Cazzato, Uniphore Acquires Emotion Research Lab, Luxembourg
Wanwan Li, University of South Florida, United States
Teodoro A. Macaraeg Jr., University of Caloocan City, Philippines
Lei Qi, lowa State University, USA
Teodoro A. Macaraeg Jr., University of Caloocan City, Philippines
Mohd Saberi Mohamad, United Arab Emirates University, United Arab Emirates
Attilio Sbrana, Instituto Tecnologico de Aeronautica, Brazil
Pengfei Han, Xi'an University of Posts and Telecommunications, China
Achintya K. Bhowmik, Chief Technology Officer and Executive Vice President of Engineering at Starkey, USA
Yahya Abdullah Alzahrani, Makkah College of Technology, Saudi Arabia
Revant Kumar, Apple, USA
Wen Tang, Bournemouth University, UK
Suliman A. Alsuhibany, Qassim University, Saudi Arabia
Chi-Wei Lin, Feng Chia University, Taiwan
Farhad Mehdipour, Otago Polytechnic—Auckland International Campus (OPAIC) and Future Skills Academy, New Zealand
Raha binti Sulaiman, Universiti Malaya, Wilayah Persekutuan, Malaysia
Hui-Wen Huang, Shaoguan University, Guangdong, China
Somkiat Tangjitsitcharoen, Chulalongkorn University, Thailand
Jinyong Hu, Tufts University, USA

## Organizers

Sojo University, Kumamoto, Japan
Changun University, Taoyuan, Taiwan
Universiti Sains Malaysia, Penang, Malaysia
Beijing Huaxia Rongzhi Blockchain Technology Institute, Beijing, China

# Preface

The international conference series, Artificial Intelligence and Virtual Reality (AIVR), has been bringing together researchers and scientists, both industrial and academic, to develop novel Artificial Intelligence and Virtual Reality outcomes. Research in Virtual Reality (VR) is concerned with computing technologies that allow humans to see, hear, talk, think, learn, and solve problems in virtual and augmented environments. Research in Artificial Intelligence (AI) addresses technologies that allow computing machines to mimic these same human abilities. Although these two fields evolved separately, they share an interest in human senses, skills, and knowledge production. Thus, bringing them together will enable us to create more natural and realistic virtual worlds and develop better, more effective applications.

2023 7th International Conference on Artificial Intelligence and Virtual Reality (AIVR2023) was successfully held in Kumamoto, Japan, during July 21–23, 2023. Past AIVR conferences were held in Nagoya (2018), Singapore (2019), and as virtual conferences (2020, 2021, and 2022), respectively. AIVR2023 in the successful AIVR conference series provided an ideal opportunity for reflection on developments over the last two decades and to focus on future developments.

The topics of AIVR2023 cover all aspects of theories and applications of AI technologies, virtual environment design, and virtual reality.

We accepted 4 invited and 28 regular papers among submitted 72 papers from China, Thailand, India, Indonesia, Japan, Malaysia, South Korea, Germany, Egypt, Brazil, France, Austria, New Zealand, the Netherlands, UK, USA, etc. at AIVR2023. This volume is devoted to presenting all those accepted papers of AIVR2023. We hope this volume provides valuable academic insights and the future prospects of the conference topics for the readers.

We wish to express our sincere appreciation to all participants and the technical program committee for their review of all the submissions, which is vital to the success of AIVR2023, and also to the members of the organizer who had dedicated their time and efforts in planning, promoting, organizing, and helping the conference. Special appreciation is extended to our keynote and invited speakers: Prof. Nikola Kasabov, Auckland University of Technology, New Zealand; Prof. Srikanta Patnaik,

Director IIMT, India; Prof. Jair Minoro Abe, Paulista University, Brazil; and Prof. Fabrice Labeau, McGill University, Canada, who made very beneficial speeches for the conference audience.

Kobe, Japan — Kazumi Nakamatsu
Bhubaneswar, India — Srikanta Patnaik
Sofia, Bulgaria — Roumen Kountchev
July 2023

# Contents

# About the Editors

**Kazumi Nakamatsu** received the Ms.Eng. and Dr.Sci. from Shizuoka University and Kyushu University, Japan, respectively. His research interests encompass various kinds of logic and their applications to Computer Science, especially paraconsistent annotated logic programs and their applications. He has developed some paraconsistent annotated logic programs called ALPSN (Annotated Logic Program with Strong Negation), VALPSN (Vector ALPSN), EVALPSN (Extended VALPSN) and bf-EVALPSN (before-after EVALPSN) recently, and applied them to various intelligent systems such as a safety verification based railway interlocking control system and process order control. He is an author of over 180 papers and 30 book chapters and 20 edited books published by prominent publishers. Kazumi Nakamatsu has chaired various international conferences, workshops, and invited sessions, and he has been a member of numerous international program committees of workshops and conferences in the area of Computer Science. He has served as the editor-in-chief of the International *Journal of Reasoning-based Intelligent Systems* (IJRIS); he is now the founding editor of IJRIS and an editorial board member of many international journals. He has contributed numerous invited lectures at international workshops, conferences, and academic organizations. He also is a recipient of numerous research paper awards.

**Prof. Srikanta Patnaik** is now working as the Director of Interscience Institute of Management and Technology, Bhubaneswar, Odisha, India. He has supervised more than 30 Ph.D. Theses and 100 Master theses in the area of Computational Intelligence, Machine Learning, Soft Computing Applications and Re-Engineering. Dr. Patnaik has published more than 100 research papers in international journals and conference proceedings. He is author of 3 text books and edited more than 100 books and few invited book chapters, published by leading international publisher like IEEE, Elsevier, Springer-Verlag, Kluwer Academic, IOS Press and SPIE. Dr. SrikantaPatnaik is the Editors-in-Chief of *International Journal of Information and Communication Technology* and *International Journal of Computational Vision and Robotics* published from Inderscience Publishing House, England and, Editor of *Journal of Information and Communication Convegence Engineering* and Associate

Editor of *Journal of Intelligent and Fuzzy Systems* (JIFS). He is also Editors-in-Chief of Book Series on "*Modeling and Optimization in Science and Technology*" published from Springer, Germany. He is also Fellow of IETE, Life Member of ISTE, and CSI.

**Prof. Roumen Kountchev** Ph.D., D.Sc. is a professor at the Faculty of Telecommunications, Department of Radio Communications and Video Technologies, Technical University of Sofia, Bulgaria. His areas of interests are digital signal and image processing, image compression, multimedia watermarking, video communications, pattern recognition and neural networks. Prof. Kountchev has 350 papers published in magazines and proceedings of conferences; 20 books; 47 book chapters; and 21 patents. He had been a principle investigator of 38 research projects. At present, he is a member of Euro Mediterranean Academy of Arts and Sciences and President of Bulgarian Association for Pattern Recognition (member of Intern. Association for Pattern Recognition). He is an editorial board member of: *International Journal of Reasoning-Based Intelligent Systems*; *International Journal of Broad Research in Artificial Intelligence and Neuroscience*; *KES Focus Group on Intelligent Decision Technologies*; *Egyptian Computer Science Journal*; *International Journal of Bio-Medical Informatics and e-Health*, and *International Journal of Intelligent Decision Technologies*. He has been a plenary speaker at: WSEAS International Conference on Signal Processing, 2009, Istanbul, Turkey; WSEAS International Conference on Signal Processing, Robotics and Automation, University of Cambridge 2010, UK; WSEAS International Conference on Signal Processing, Computational Geometry and Artificial Vision 2012, Istanbul, Turkey; Intern. Workshop on Bioinformatics, Medical Informatics and e-Health 2013, Ain Shams University, Cairo, Egypt; Workshop SCCIBOV 2015, Djillali Liabes University, Sidi Bel Abbes, Algeria; International Conference on Information Technology 2015 and 2017, Al Zayatoonah University, Amman, Jordan; WSEAS European Conference of Computer Science 2016, Rome, Italy; The 9th International Conference on Circuits, Systems and Signals, London, UK, 2017; IEEE International Conference on High Technology for Sustainable Development 2018 and 2019, Sofia, Bulgaria; The 8th International Congress of Information and Communication Technology, Xiamen, China, 2018; General chair of the Intern. Workshop New Approaches for Multidimensional Signal Processing, July 2020, Sofia, Bulgaria.

# Part I
# Invited Papers

# Chapter 1
# A Model for Predicting Crime Risk


**Farhad Mehdipour, U. H. W. A. Hewage, Wisanu Boonrat, April Love Naviza, Vimita Vidhya, and Ari Aharari**

**Abstract** Efficient resource allocation and effective risk management are paramount for governments and police forces. This article presents a comprehensive model designed to predict crime risk levels, with the goal of optimising resource allocation and reducing the associated time, cost, and effort in risk management. By considering crucial factors such as location, time, and crime type, our model endeavours to provide accurate and actionable insights. To ensure the reliability of our predictions, we implemented various data wrangling techniques, including feature selection, data validation, and the creation of new measures. These steps were instrumental in preparing the data for analysis and generating reasonably accurate results. Additionally, we explored a range of machine learning algorithms, namely Logistic Regression, Gaussian Naive Bayes, Decision Tree, XG Boost, and Random Forest, to predict crime risk levels. The models were meticulously validated using cross-validation techniques and evaluated based on diverse performance metrics. In a bid to further advance our predictive capabilities, we leveraged deep learning techniques with TensorFlow, enabling a performance comparison against traditional machine learning models. Notably, the Random Forest algorithm has emerged as the most effective, yielding an impressive accuracy of 90%. The culmination of our efforts is a successful software application, complete with a user interface integrated with our cutting-edge prediction model.

## 1.1 Introduction

New Zealand, ranked as the world's second safest country out of 162 countries in terms of personal violence risk according to the 2021 Global Peace Index [1], has maintained a significantly low crime rate despite its growing population, which has

F. Mehdipour (✉) · U. H. W. A. Hewage · W. Boonrat · A. L. Naviza · V. Vidhya
Department of Information Technology, Otago Polytechnic—Auckland International Campus, Auckland, New Zealand
e-mail: farhadm@op.ac.nz

A. Aharari
Department of Computer and Information Science, SOJO University, Kumamoto, Japan



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_1

been increasing by 1.4–2.0% per year from 2016 to 2018. While the overall victimisation recorded by the police in the year ending December 2020 has decreased by 6.6%, there has been a concerning increase of 12.4% in the number of assault victims compared to the previous year [2]. Considering the ongoing societal challenges, such as inflation, it is reasonable to assume that these numbers may have further escalated up to the present [3]. Consequently, the accurate prediction of crime risk becomes crucial for implementing proactive measures that minimise criminal activities while optimising resource allocation. Moreover, it can empower the police force to effectively prevent crimes.

To address the aforementioned concern, we have developed and implemented an innovative solution that employs machine learning algorithms to predict crime risk levels across three categories: low, medium, and high. This comprehensive approach utilises various machine learning algorithms, including Logistic Regression [4], Gaussian Naive Bayes [5], Decision Tree [6], XG Boost [7], and Random Forest [8]. By integrating a user interface with our prediction model, users are empowered to assess the risk level associated with a specific location, date, and crime type. The application allows users to select from a range of machine learning algorithms, providing flexibility and tailored predictions. Notably, our model incorporates standard features such as crime type, location, and time to ensure the development of an unbiased predictive model. This proposed solution offers significant benefits to the government, particularly the police department, as it facilitates the identification of crime risk levels across different locations and days. However, it is important to note that the proposed application is not intended for public use due to ethical considerations.

The remainder of the paper has been organised as follows. Section 1.2 summarises the literature related to crime prediction. Section 1.3 explains the proposed methodology, including datasets, algorithms, and other operations. In Sect. 1.4, we discuss the results of the experimentations, while Sect. 1.5 concludes the paper and suggests future directions.

## 1.2 Previous Work

Nowadays police forces have been enhancing their traditional methods of crime reporting with new technological advancements to enhance their crime prediction and prevention methods [9]. Various techniques and tools have been introduced, while there are some of commercial tools used by police forces in some regions around the world.

In the field of crime prediction, researchers have highlighted two distinct approaches: qualitative methods that involve identifying the future nature of criminal activity, and quantitative methods that focus on predicting the future scope of crime and crime rates [10]. It is suggested that quantitative analysis is well-suited for observing crime trends, while qualitative analysis is more effective in identifying the underlying factors influencing crime rates. For instance, Berrada et al. [11]

conducted a study that aimed to predict criminal offenses, including location, perpetrator identities, offenders, and victims. They utilised datasets such as Boston local crimes, Population Census Data, and Weather Data. In their analysis, they employed Logistic Regression and Random Forest algorithmic models to generate predictions. Similarly, Almanie et al. [12] concentrated on identifying spatial and temporal criminal hotspots. They utilised the Denver neighbourhood demographics dataset and applied various techniques, including Apriori [13] for association rule learning over relational databases, as well as Naive Bayesian [5] and Decision Tree [6] classifiers. These studies demonstrate the diverse range of methodologies employed to predict and analyse crime, with researchers utilising both traditional statistical models and machine learning algorithms to uncover valuable insights.

Almaw et al. [14] have contributed to the field of crime prediction by focusing on the correlation between population density and crime rates within a specific area. Their work resulted in the development of CrimeTracer and a Crime Prevention Decision Support System. Through a performance analysis of classification algorithms, they found that Naive Bayes demonstrated reasonable accuracy in crime prediction. However, they also discovered that ensemble learning algorithms, which combine predictions from multiple models, yielded even more precise results.

In a related study, Rumi et al. [15] delved into the impact of dynamic features on crime events. They compared these dynamic features with geographic and demographic variables, incorporating historical crime patterns into their analysis. Additionally, they explored the use of Location-Based Social Networks and social media data, as well as human mobility information, for crime prediction. Their study employed four prediction models: Random Forest (RF) [8], Neural Network (NN) [16], Kernel Support Vector Machine (SVM) [17], and Logistic Regression Model (LR) [4], alongside an ensemble-based learning framework [18]. The findings revealed that integrating human mobility data from social media enhanced the accuracy of crime prediction. Furthermore, the combination of dynamic and static features improved the prediction performance across various types of crime events.

Pradhan [19] conducted a comprehensive analysis of various crime types in San Francisco, aiming to understand how different attributes, such as seasons, contribute to specific crimes. The study has delved into the nuanced relationships between these attributes and crime occurrences. In a similar vein, Yuki et al. [20] focused on analysing the crime patterns in Chicago. Their research has aimed to identify the specific crime types that are likely to occur at particular times and locations. To achieve this, they utilised the Chicago crime events dataset from the police and the department's CLEAR system. By employing algorithms such as Random Forest, Decision Tree, and ensemble methods, they gained insights into the dynamics of crime in the city. Kang et al. [21] proposed a novel approach to predicting crime occurrences by leveraging a feature-level data fusion method with an environmental context. Their work incorporated a deep neural network (DNN) to integrate environmental context information. The utilisation of concepts such as the Broken Windows Theory and Crime Prevention Through Environmental Design (CPTED) highlighted the influence of appearance and environmental factors on criminal activity. Collectively, these studies contribute to a deeper understanding of crime

patterns by analysing specific attributes, leveraging advanced algorithms, and considering environmental factors. Their findings shed light on the complex dynamics of crime occurrence and provide valuable insights for crime prediction and prevention efforts.

The factors of the historical records such as Demographic (e.g. age, gender, and population) and Macro-economic features (e.g. unemployment rates) have been the determining factors of crime rates [22]. A study conducted in New Zealand has examined that weather temperature and precipitation showed a significant effect on violence and property crimes [23].

Expanding the scope of the literature review, we explore studies cantered around predictive policing and associated software applications. Mann [24] introduced a predictive crime software called Crimescan, which leverages historical crime data alongside information from 911 hotline calls and police reports pertaining to minor offenses like disorderly conduct, narcotics trafficking, and loitering. This software aims to identify the locations where violent crimes are most likely to occur, providing valuable insights for proactive law enforcement efforts. In a related study, Degeling et al. [25] examined four different predictive policing software applications: PredPol, Hunchlab (Risk Terrain Modelling), Chicago's Heat List, and Beware (Threat Scores). These applications employ various techniques, including Near Repeat theory, to predict crime patterns and assess risks across different locations. Furthermore, these tools aid in identifying potential offenders and victims. The Near Repeat theory posits that when a crime transpires in a particular area, the surrounding vicinity may experience an increased likelihood of similar crimes occurring within a distinct period of time [26]. These studies demonstrate the utilisation of predictive policing software applications that leverage historical data, advanced modelling techniques, and theories such as Near Repeat to enhance crime prediction and prevention efforts. These tools provide law enforcement agencies with valuable insights for resource allocation, targeted interventions, and proactive measures to mitigate crime risks.

Chammah et al. [27, 28] examined the software Auror, initially developed by Dickinson in 2017. Formerly known as Evedentify, this Auckland-based software is a crime intelligence platform designed to assist retailers in preventing crime within their stores. The software facilitates the reporting of incidents swiftly and efficiently, enabling users to record incident reports from any device. It also securely stores and allows for the review of all digital evidence, including CCTV footage, while tracking the outcomes of each criminal report across the company. Auror goes beyond reactive measures by proactively preventing crime. It achieves this by providing real-time alerts when individuals or vehicles of interest enter the store or its vicinity. This proactive approach enhances the retailer's ability to intervene and take appropriate action before a crime occurs. In summary, Auror serves as a comprehensive crime prevention and management solution for retailers, offering features such as efficient incident reporting, secure storage and review of digital evidence, and real-time alerts for individuals or vehicles of interest. By leveraging this software, retailers can strengthen their security measures and mitigate the risks associated with criminal activities within their premises.

It is evident from the literature that the problem of crime prediction is not new and there are various techniques and software tools introduced. However, our focus in on predicting crime based on a new crime risk model which involves common features including geographical and time attributes. The model is integrated with a user interface which facilitates the use of the prediction model for non-technical users.

## 1.3 The Proposed Methodology

### *1.3.1 Datasets and Algorithms*

For the preliminary studies, we sourced six crime datasets for different locations, namely Boston, London, San Francisco [29], Denver [30], and New Zealand [31]. The variables of each dataset were analysed and compared to identify the behaviours of these datasets. From the preliminary study, we selected two NZ-related datasets for further study as our focus is on New Zealand context. We name the selected datasets as NZ version 1 (641,641 records) and NZ version 2 (1,194,764 records). Both the datasets were acquired from publicly available sources as .CSV files. These two datasets have following list of features.

***NZ Version 1 (NZ_1):*** *Crime type and Sub-type, Person/Organisation, Date, Ethnicity, Age type, Method of Proceeding, Number of Records, Police District.*
***NZ Version 2 (NZ_2):*** *Crime type and sub-type, Area Unit, Location Type, Mesh-block, Number of Records, Occurrence Day Of Week, Occurrence Hour Of Day, Territorial Authority, Victimisations, Weapon, Month Year.*

### *1.3.2 Data Preprocessing and Transformations*

Data wrangling, transformation, and feature selection are the most critical processes before building a model. These processes improve the quality of data which can lead to higher model accuracy and less possibility of bias and overfitting [32]. In this part, different wrangling techniques were applied to detect and treat the invalid/incorrect data for all variables.

We explored four feature selection techniques, namely Extra tree classifier, SelectKBest, Chi2, and Mutual Info Classifier [33] to select most important features from the dataset. The normalised results from feature selection from NZ version 2 are summarised in Table 1.1.

**Table 1.1** Features selected in the order of their importance/significance

| Rank | Selected feature (in the order of importance) |
|---|---|
| 1 | Crime type |
| 2 | Crime sub-type |
| 3 | Location type |
| 4 | Location—district |
| 5 | Weapon |
| 6 | Occurrence time (hour of the day) |
| 7 | Territorial authority |

### *1.3.3 Crime Risk Factor—A New Measure*

The attributes of the selected dataset do not directly reflect the risk level of crime; therefore, we have formulated a new variable called crime risk factor based on the selected features. The crime risk factor is calculated by considering the total crime in specific times and locations. Since the original dataset has provided the crime type, day, month, area unit, and territorial authority, we were able to calculate the ratio, rank, and risk level of crimes in certain area and period of interest. Based on the new measures, risk level was identified as low, medium, and high and encoded to 0, l, and 2, respectively. The following section describes the formulation of the crime risk factor. The calculation of the crime risk involves features related to the location/area, time/day, and the crime type.

**Definitions**

*MACC*—Month Area Crime Count: the number of crimes in a certain area and month of the year (Fig. 1.1a).
*DACC*—Day Area Crime Count: the number of crimes in a certain area and day (Fig. 1.1b).
*MACTC*—Month Area Crime-Type Count: the number of crime types in certain area and month of the year (Fig. 1.1c).

The ratio of the crime occurring in a certain day and area is calculated as follows:

$$\text{Day Area Crime Ratio: } \text{DACR} = \text{DACC}/\text{MACC}, \tag{1.1}$$

where the ratio of the crime occurring in a certain day and area unit is formulated as:

$$\text{Month Area Crime Ratio: } \text{MACR} = \text{MACTC}/\text{MACC}. \tag{1.2}$$

Consequently, the ratio of crime (CR) is formulated based on the two above factors as follows:

$$\text{Crime Ratio: } \text{CR} = \text{DACR} \times \text{MACR}. \tag{1.3}$$

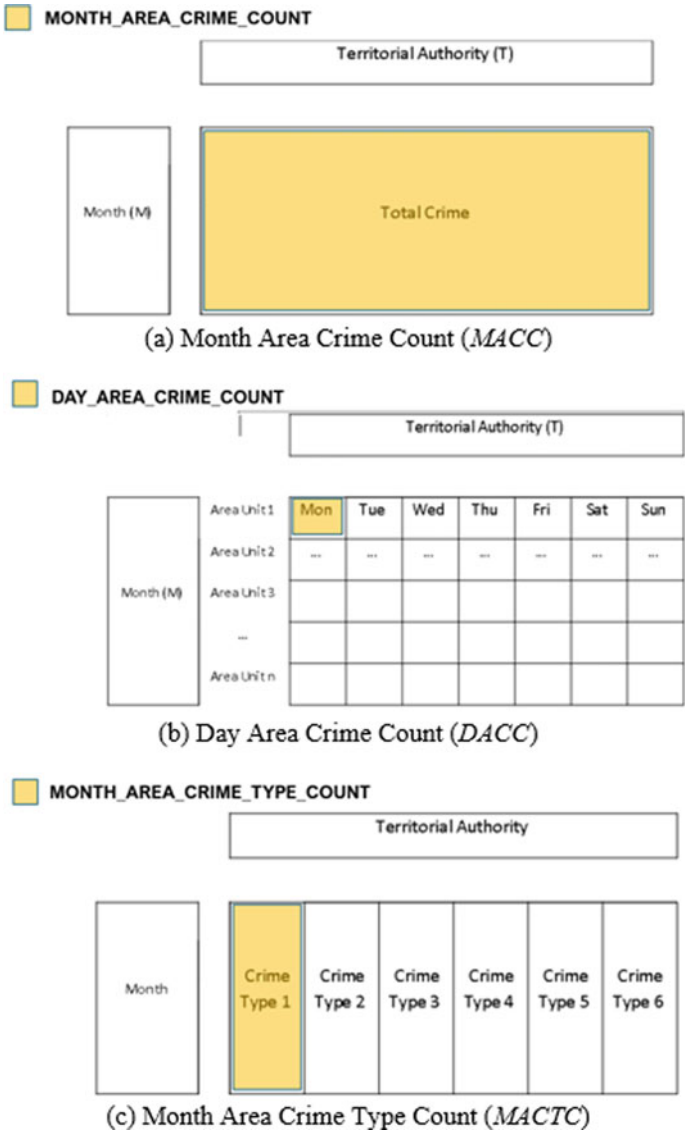



**Fig. 1.1** Definition of the terms used in building the crime risk factor

The crime ratio for a certain area unit can be used to decide a rank based on its value (percentile) range. We allocate 0, 1, and 2 corresponding to low, medium, and high to the crime risk for the crime ratios below 0.3, between 0.3 and 0.7 and above 0.7, respectively (Fig. 1.2). Therefore, the newly introduced crime risk factor is 0 or 1 or 2 based on the crime ratio.

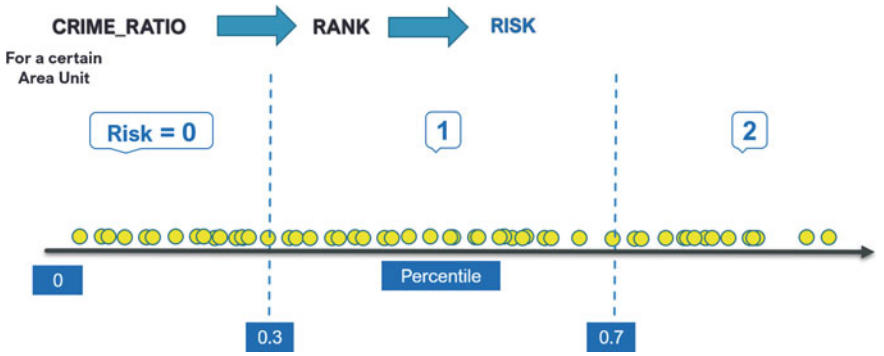



**Fig. 1.2** Conversation of crime ratio to crime rank and crime risk level

**Table 1.2** Baseline scores for different classifiers

| Datasets | Logistic Regression | Gaussian NB | Decision Tree | XGB | Random Forest |
|---|---|---|---|---|---|
| *NZ_1* | 0.16 | 0.16 | 0.10 | 0.18 | 0.08 |
| *NZ_2* | 0.57 | 0.59 | 0.54 | 0.63 | 0.58 |

### *1.3.4 Baseline Scores*

Five classification algorithms were utilised for modelling including Logistic Regression, Gaussian NB, Decision Tree, XGB, and Random Forest classifier. The two datasets were trained using these classifiers and results were compared to identify the best performing algorithm. Baseline scores for both datasets were calculated before feature selection. Predictors and target variables were identified for each dataset and trained on five different classifiers. The trained models have generated the baseline scores from the original datasets as shown in the following table (Table 1.2).

From the results of Table 1.2, it can be seen that *NZ_1* dataset has produced very low baseline scores for all the tested classifiers which are considerably lower than the scores of *NZ_2*. Therefore, the second dataset (*NZ_2*) is suitable to be used for further development because the models can return better scores comparing to the other dataset. Considering the baseline scores and other factors such as high number of records and importance and the relevance of the features, it was decided to use *NZ_2* for building the prediction model.

### *1.3.5 Predictive Modelling and Cross-validation*

After preprocessing and feature engineering process, the newly generated dataset including the crime risk factor attribute was used for predictive modelling. The

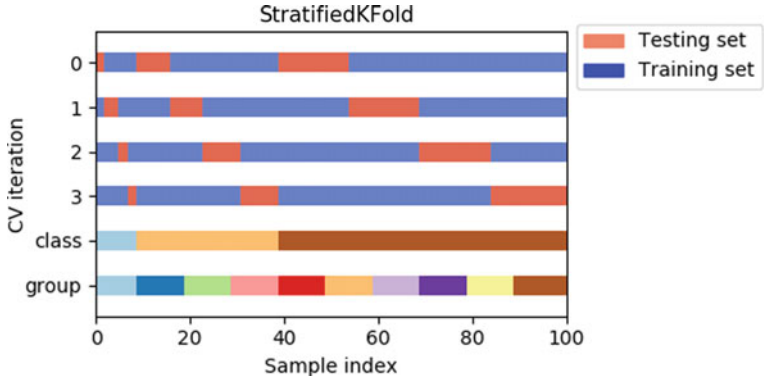



**Fig. 1.3** Tenfold cross-validation using *StratifiedKFold*

processed dataset was used for training classification models and confusion matrices were generated for the five classifiers.

After we had the acceptable accuracy scores from the trained models, cross-validation was used to verify and evaluate results. This step assures that the prediction model performs well with unseen data without overfitting issues. We applied tenfold cross-validation using *StratifiedKFold* (from *sklearn* library in Python) to deal with the overfitting (Fig. 1.3).

### *1.3.6 Deep Learning with TensorFlow*

We explored Deep Learning with TensorFlow [34] and compared the accuracy with machine learning models that we utilised for this experiment. TensorFlow is an open-source platform with easy debugging of nodes which reduces the overhead of debugging the entire code. Additionally, it has a better data visualisation library which supports in visualising results in a more understandable way. We have applied 32 and 64 epochs to the new dataset that undergoes data transformation. For our model to learn more details and relationships within the data, we have applied five layers. *TensorBoard* [34] has been used to visualise our model. *TensorBoard* shows and compares the train and validation movements in every iteration.

**Table 1.3** Accuracy of the prediction for different classifiers

| Algorithm | Train = 70/test = 30 | | *K*-fold CV. ($k = 5$) avg. scores | | *K*-fold CV. ($k = 10$) avg. scores | |
|---|---|---|---|---|---|---|
| | Accuracy | Log_loss | Accuracy | Log_loss | Accuracy | Log_loss |
| Logistic Regression | 0.4 | 1.09 | 0.39 | 1.09 | 0.37 | 1.09 |
| Gaussian NB | 0.37 | 1.12 | 0.36 | 1.12 | 0.34 | 1.12 |
| *K* neighbours ($k = 12$) | 0.77 | 1.03 | 0.78 | 1.04 | 0.79 | 1.04 |
| Decision Tree (max_depth = 20) | 0.76 | 1.81 | 0.75 | 1.81 | 0.75 | 1.7 |
| XG Boost (max_depth = 12) | 0.84 | 0.49 | 0.84 | 0.49 | 0.83 | 0.5 |
| Random Forest (max_depth) = 20 | 0.84 | 0.51 | 0.83 | 0.52 | 0.83 | 0.52 |

## 1.4 Results and Discussion

We obtained the scores of models for predicting crime type and risk. The best baseline score for predicting crime type was around 60% for *NZ_2* dataset; however, after transforming data and creating the new measure (crime risk factor), the accuracy of prediction is increased to 80–90%.

### *1.4.1 Performance Comparison*

The predictive models were trained with different scenarios to predict crime risk. The uneven number of records associated with different types of crime could result in biased results. Since the number of crime types in the dataset is distributed unevenly, we used oversampling techniques to balance the data. This resulted in improved classification accuracy for the top six crime types (the crimes with highest number of incidents).

The results shown in Table 1.3 from Decision Tree, XG Boost, and Random Forest were generated by specifying the *max_depth* parameter, which limits the number of maximum depths of the tree from the root node to a leaf.

### *1.4.2 Evaluating Results Through Visualisation*

Accuracy is an important factor to evaluate the performance of a classifier. However, there are alternative metrics which can provide further insights on the performance of models from different points of views including the following.

**Confusion Matrix**. The following confusion matrices (Fig. 1.4) show the actual and predicted results from ML models. Logistic Regression shows the highest number of predictions in the risk level 2 and the rest of the prediction is sparse. A similar result can be seen in the Gaussian NB, but the highest number happens in the risk level 1. The other models such as K-Nearest Neighbours, Decision Tree, Random Forest, and XG Boost show high accuracy scores and their confusion matrices also show the same trends of the results.

**Receiver Operating Characteristic Curve (ROC)**. The perfect performance of the ROC curve is when the area under the ROC curve (AUC) equals 1. In the plots below, the dashed line in the plots mean to 50% probability that the prediction result will be 0 or 1. We used *OneVsRestClassifier* function (from *scikit-learn* library in Python) to plot the curve of each risk level by comparing the probability of the target class to the other two classes. The graphs indicate that the K-Nearest Neighbour and Decision Tree perform better with a larger area under ROC curve (Fig. 1.5).

The tree-based models such as Decision Tree, XG Boost, and Random Forest result in 80–90% accuracy. This is because the mechanism of tree-based algorithms is designed to solve the problem by creating the rules and splitting the nodes based on the features, which will show a great performance in the large dataset. The K-Nearest Neighbour model also achieved a fair accuracy at 76% when the number of nearest neighbours equals 12. Both Logistic Regression and Gaussian Naive Bayes returned poorest results. Figure 1.6 shows the precision recall curve representing the tradeoff between precision and recall for different threshold.

### *1.4.3 Deep Learning Results*

Upon comparing the outcomes of deep learning and traditional machine learning, we observed highly similar performance achieved with a Decision Tree—specifically, a 76% accuracy was attained after 32 epochs (Fig. 1.7). We believe that the score can be improved; however, it would be less worth to build a deep learning model when we can have the same scores in machine learning models. Training a deep learning model involves more computational resources and time compared to machine learning counterparts. This increased demand stems from the complex architecture and large-scale computations required to optimize the multitude of parameters within a deep neural network.

### *1.4.4 Crime Prediction Application*

We have designed a user-friendly interface (refer to Fig. 1.8) that enables users to predict the level of crime risk by selecting various attributes, including location, time, and the choice of classifier. This interface utilises the machine learning models

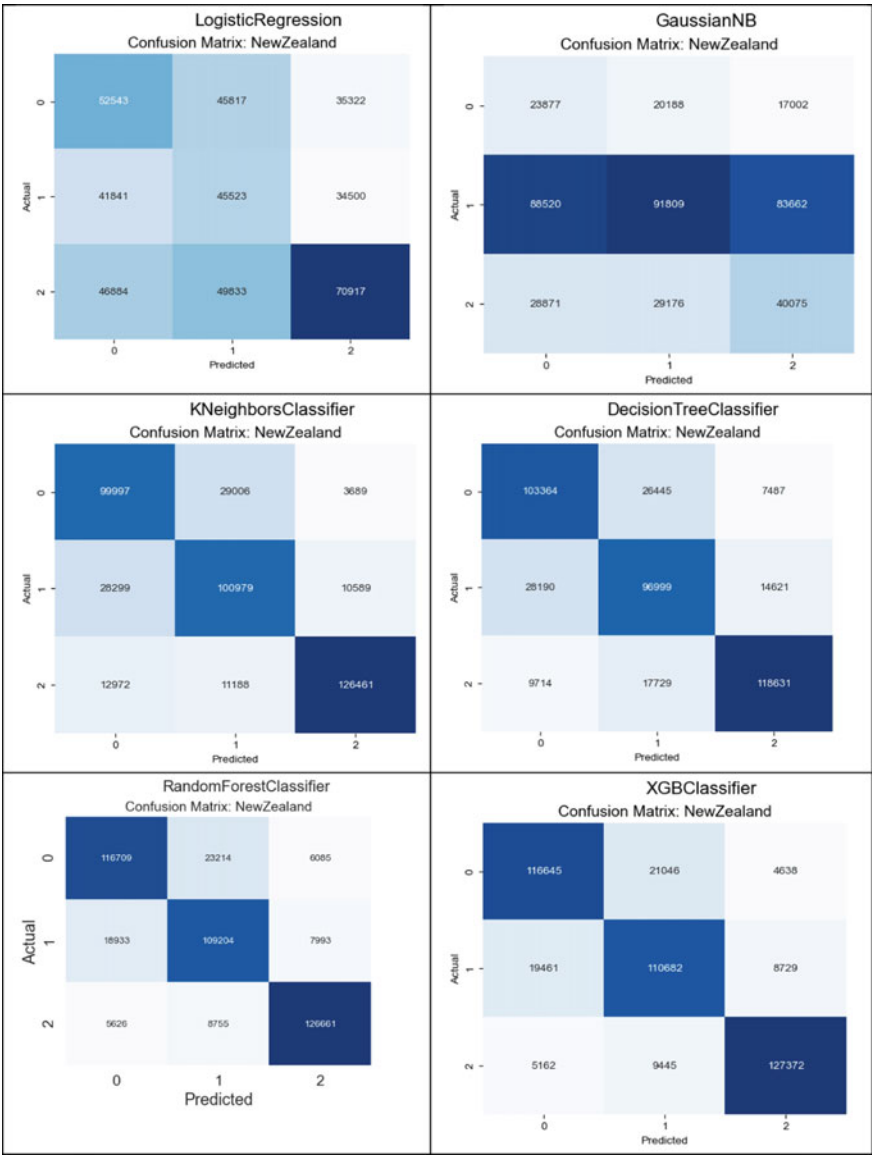



**Fig. 1.4** Confusion matrices

we have developed, along with the incorporation of the new crime risk factor we introduced in our research. Within the interface, users can choose a specific machine learning algorithm from the available options. They can also specify the type of crime, the day, month, hour, and territorial authority for which they want to predict the risk level of crime.

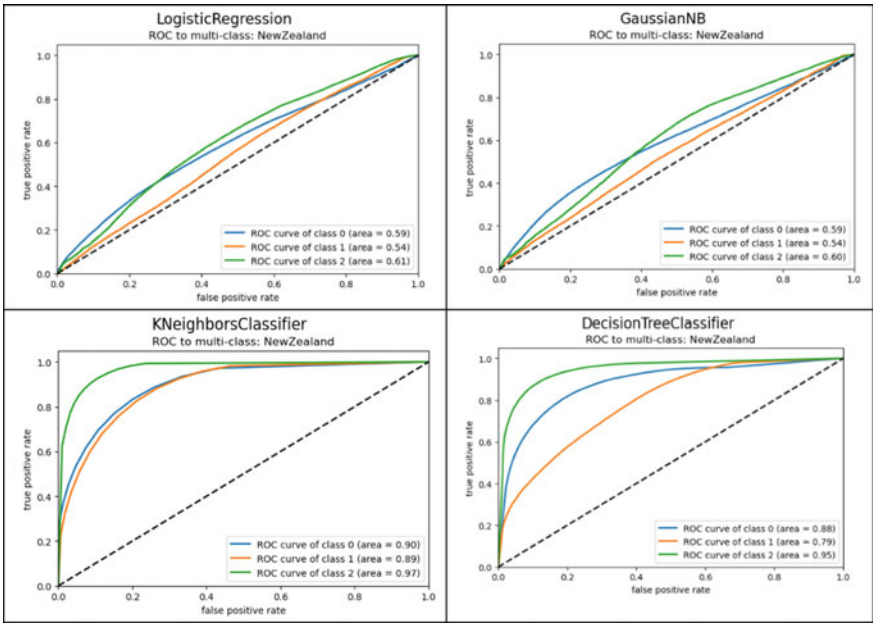



**Fig. 1.5** ROC curves for the classifiers

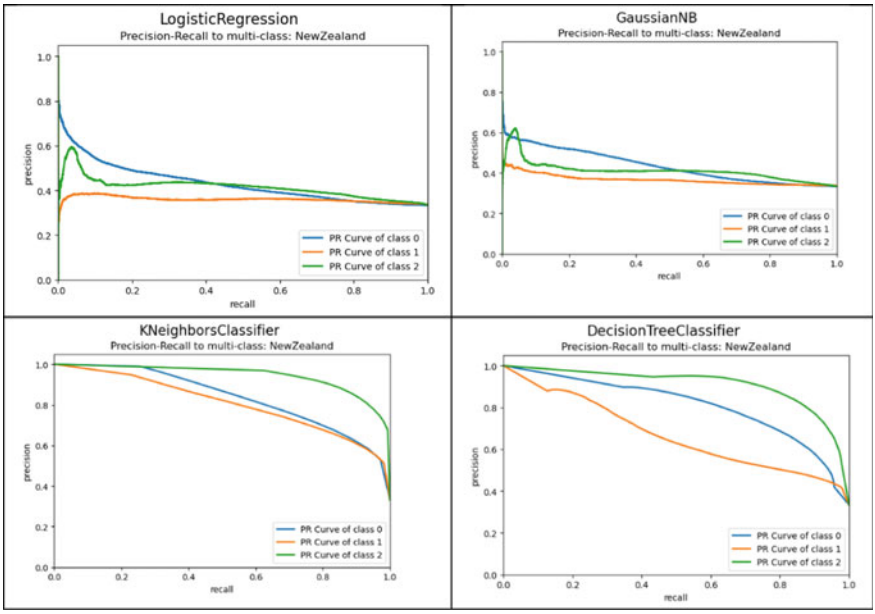



**Fig. 1.6** PR curves for the classifiers


**Fig. 1.7** Epoch accuracy for the deep learning model

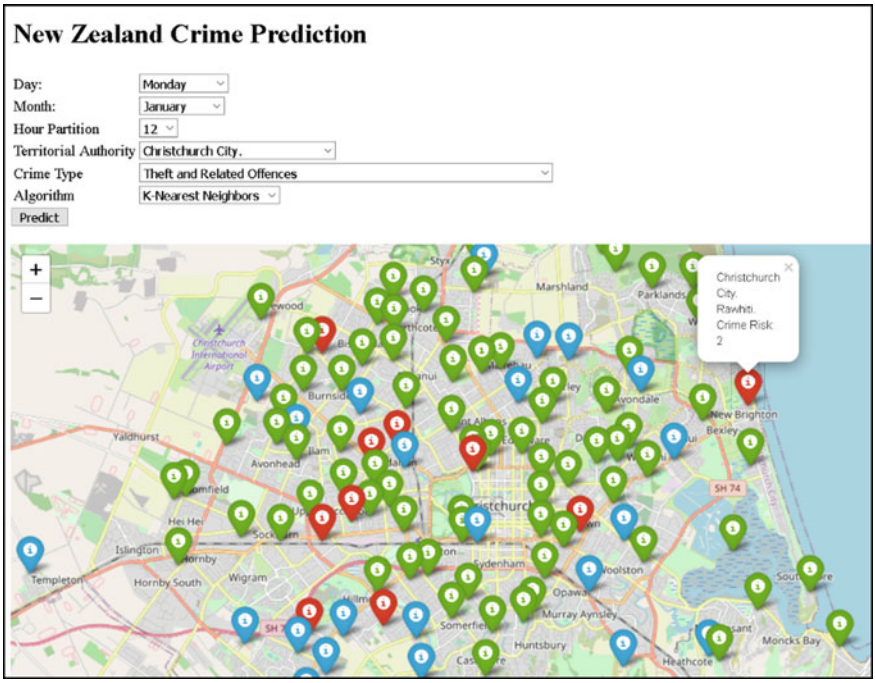



**Fig. 1.8** View of the user interface for the predictive model

By inputting these attributes and utilising the selected machine learning algorithm, the tool will generate a prediction for the crime risk level. This interactive interface allows users, including non-technical individuals, to easily access and utilise our crime prediction models.

## 1.5 Conclusion and Future Directions

The objective of the study is to predict the risk level of a crime. We defined a new crime risk factor to formulate the risk of occurring a crime in certain location and time. We used five different machine learning algorithms: Logistic Regression, Gaussian Naive Bayes, Decision Tree, XG Boost, and Random Forest for crime risk prediction. Out of these, tree-based (Decision Tree, XG Boost and Random Forest) algorithms produced satisfying accuracy results. The Random Forest model is 90% accurate in predicting the likelihood of crime. We used the cross-validation to evaluate the results and, confusion matrix, and ROC curve to test the model results. The deep learning model with TensorFlow provided similar results. The outcome of this work is a prediction tool including machine learning models underneath a user-friendly interface which can be used by the government and police departments to predict the crime risk in advance and allocate their resources more efficiently.

Further studies based on this project could focus on both improving the predictive models and integrating the models with the existing commercialised systems for reporting and visualisation of data.

## References

1. Live and Work New Zealand, 14 07 2022 [online]. Available https://www.live-work.immigration.govt.nz/choose-new-zealand/safe-secure. Accessed 03 2023
2. Crime at a glance, 03 2021 [online]. Available https://www.police.govt.nz/sites/default/files/publications/crime-at-a-glance-dec2020.pdf. Accessed 03 2023
3. The peculiar relationship between inflation and theft, 2023 [online]. Available https://www.firstsecurity.co.nz/blog/the-peculiar-relationship-between-inflation-and-theft/. Accessed 03 2023
4. Cramer, J.S.: The Origins of Logistic Regression (PDF) (Technical Report), vol. 119, pp. 167–178. Tinbergen Institute (2002). https://doi.org/10.2139/ssrn.360300
5. McCallum, A.: Graphical Models, Lecture2: Bayesian Network Representation (PDF). Archived (PDF) from the original on 09 Oct 2022. Retrieved 22 Oct 2019
6. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA (1984). ISBN 978-0-412-04841-8
7. Story and lessons behind the evolution of XGBoost. Retrieved 01 Aug 2016
8. Ho, T.K.: Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, pp. 278–282. Montreal, QC, 14–16 Aug 1995. Archived from the original (PDF) on 17 Apr 2016. Retrieved 5 June 2016
9. Grover, V., Adderley, R., Bramer, M.: Review of Current Crime Prediction Techniques (2007). https://doi.org/10.1007/978-1-84628-666-7_19

10. Schneide, S.: Predicting Crime: A Review (2002)
11. Martegiani, G.: Crime Prediction Using Data Analytics: The Case of the City of Boston
12. Tahani, A., Rsha, M., Elizabeth, L.: Crime prediction based on crime types and using spatial and temporal criminal hotspots. Int. J. Data Min. Knowl. Manage. Process. **5**
13. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pp. 487–499. Santiago, Chile, Sept 1994
14. Almaw, A., Kadam, K.: Crime data analysis and prediction using ensemble learning. In: Second International Conference on Intelligent Computing and Control Systems (ICICCS) (2018)
15. Shakila, K.R., Ke, D., Flora, D.S.: Crime event prediction with dynamic features. EPJ Data Sci. 1027 (2018)
16. Kleene, S.C.: Representation of events in nerve nets and finite automata. Ann. Math. Stud. **34**, 3–41 (1956). Retrieved 17 June 2017
17. Cortes, C., Vapnik, V.: Support-vector networks (PDF). Mach. Learn. **20**(3), 273–297 (1995). CiteSeerX 10.1.1.15.9362. https://doi.org/10.1007/BF00994018.S2CID206787478
18. Opitz, D., Maclin, R.: Popular ensemble methods: an empirical study. J. Artif. Intell. Res.Artif. Intell. Res. **11**, 169–198 (1999). https://doi.org/10.1613/jair.614
19. Pradhan, I.: Exploratory Data Analysis and Crime Prediction In San Francisco. SJSU Scholar Works (2018)
20. Jesia, Y., Sakib, M.Q., Zamal, Z., Habibullah, K.M.: Predicting crime using time and location data. In: 7th International Conference on Computer and Communications Management (2019)
21. Hyeon, W.K., Hang, B.K.: Prediction of crime occurrence from multi-modal data using deep learning. PLoS ONE **24**, 04 (2017)
22. Deadman, D.: Forecasting residential burglary. Int. J. Forecast. **19**(4) (2003)
23. Horrock, J., Menclova, K.: The effects of weather on crime. New Zealand Econ. Pap. **45** (2011)
24. Mann, A.: How Science Is Helping Stop Crime Before It Occurs (2017) [online]. Available https://www.nbcnews.com/mach/science/how-science-helping-stop-crime-it-occurs-ncna805176. Accessed 03 2023
25. Degeling, M., Berendt, B.: What is wrong about Robocops as consultants? A technology-centric critique of predictive policing. AI Soc. **33**(3), 347–356 (2018)
26. Ratcliffe, J.H., Rengert, G.F.: Near-repeat patterns in Philadelphia shootings. Secur. J.. J. **21**(1–2), 58–76 (2008)
27. O'Neill, R.: NZ police widen use of auror crime-fighting tools, 13 Jan 2016 [online]. Available https://www.zdnet.com/article/nz-police-widen-use-of-auror-crime-fighting-tools/. Accessed 09 Mar 2023
28. Chammah, M.: Policing the Future, 02 Mar 2016 [online]. Available https://www.themarshallproject.org/2016/02/03/policing-the-future
29. Kaggle, https://www.kaggle.com/. Accessed 08 2022
30. Denver, https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime. Accessed 08 2022
31. NZ Government, https://catalogue.data.govt.nz/dataset/recorded-crime-victims-and-offenders-statistics-rcvs-and-rcos. Accessed 08 2022
32. Kaushik, S.: Introduction to feature selection methods with an example (or how to select the right variables?), 07 Feb 2023 [online]. Available https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/#:~:text=Top%20reasons%20to%20use%20feature,the%20right%20subset%20is%20chosen. Accessed 09 Mar 2023
33. Scikit Learn, https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection
34. TensorFlow, https://www.tensorflow.org

# Chapter 2
# Early Detection of Red Palm Weevil in Date Palm Trees Using Machine Learning Approaches


**Gehad Ismail Sayed, Fatema Samir, Mariam M. Abdellatif, and Aboul Ella Hassanien**

**Abstract** The destructive pest known as the red palm weevil (RPW) has destroyed several palm tree farms all over the world. It can be difficult to identify RPW early, particularly on large-scale farms. Therefore, this paper introduces a strategy for the early identification of RPW in sizable farms using machine learning approaches. The proposed approach consists of four main phases, namely data pre-processing, feature extraction, classification, and evaluation phases. A total of 483 red palm weevil images are used to evaluate the performance of the proposed approach. It obtained an overall accuracy of 91%, precision of 92%, sensitivity of 90%, and $f$-score of 91%.

## 2.1 Introduction

Millions of people all over the world benefit from the high-value fruit harvest produced by date palms [1]. Furthermore, it is regarded as a significant source of export income for rural smallholders globally. Unfortunately, the red palm weevil (RPW), also known as Rhynchophorus ferruginous, is a threat to commerce and the date production and commerce [2]. The single most devastating pest of palm

G. I. Sayed (✉) · M. M. Abdellatif
School of Computer Science, Canadian International College (CIC), Cairo, Egypt
e-mail: Gehad_Sayed@cic-cairo.com

F. Samir
Faculty of Science, Cairo University, Giza, Egypt

M. M. Abdellatif
Mathematics Department, Faculty of Science, Al-Azhar University (Girls), Cairo, Egypt

A. E. Hassanien
Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt

G. I. Sayed · F. Samir · M. M. Abdellatif · A. E. Hassanien
Scientific Research Group in Egypt (SRGE), Cairo, Egypt



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_2

palms is a Coleopteran snout bug called RPW. Since RPW primarily targets young, tender plants under 20 years old—which make up around 50% of all farmed date palm trees—they are particularly vulnerable [3]. In addition to date palms, RPW also damages decorative, oil, and coconut palms [4].

Chemical treatments can help palm plants recover from an infestation in its early stages [5]. Furthermore, a palm tree only exhibits obvious symptoms of distress when the infestation is well along and it is too late to rescue the tree. For RPW early detection, several methods have been documented in the literature [6, 7]. Although certain detection techniques, such as using trained dogs [8] and x-ray-based tomography [9], are effective, their sluggish scanning times prevent them from being practical in big farms. One of the most effective early detection techniques relies on red palm weevil recognition from images.

There are several domains where machine learning algorithms have been effectively used, including healthcare [10, 11], education [12], and business [13]. In [14], the authors provided one of the early assessments on the application of machine learning algorithms to agricultural issues in domains relevant to agriculture. These algorithms are further used to identify animal noises, predict wine fermentation, estimate soil water parameters, and identify meat and bone meal. The authors employed four machine learning algorithms, namely support vector machine (SVM), artificial neural networks (ANNs), k-nearest neighbor (KNN), and k-means. In New Zealand, kiwifruit plant protection strategies were also aided by machine learning. The machine learning algorithms such as Naive Bayes, decision tree, AdaBoost, random forest, logistic regression, and SVM were utilized in the proposed model that researchers in [15] presented. These machine learning algorithms were used on datasets from the pest and the spray diary monitoring. Recall and Precision metrics were used to assess these models' performance. The outcomes showed that the models with a limited number of features produced accurate predictions. Additionally, AdaBoost outperformed the other classifiers, while the Naive Bayes algorithm came in second.

Termite infestation was discovered using acoustic signal extraction and SVMs with different kernels [16]. According to the experimental outcomes, the SVM using the polynomial kernel function has a high level of classification accuracy. Machine learning has recently been applied by researchers in [17] to solve the issue of pepper Fusarium disease identification. The plant was divided into four groups based on using light reflections from the pepper leaves. They are fusarium-diseased mycorrhizal fungus, healthy, and fusarium-diseased and mycorrhizal fungus. Naive Bayes, KNN, and ANNs were used in experiments to classify data. The usefulness of machine learning algorithms was proved by the findings with high accuracy values.

Despite prior attempts, there is currently a lack of study on the use of machine learning algorithms for RPW infestation identification and prediction. This is due to the difficulty in locating representative datasets of infestation. An automated approach based on ANNs for recognizing RPW was created in [18] to help with RPW prediction. A dataset comprising 326 RPW photos and 93 additional insect images was used to train and test the proposed ANN model. The authors concluded that the best feed-forward supervised learning model for recognizing red palm weevils is the

Powell-Beale restarts and a conjugate gradient with three-layer ANN algorithms. The ability of ten cutting-edge data mining classification algorithms, including KSTAR, Naive Bayes (NB), PART, J48 Decision tree, AdaBoost, bagging, logistic regression, SVM, multilayer perceptron (MLP), and random forest, was evaluated by the authors in [19]. These algorithms are used to predict RPW infestation in its early stages just before the tree sustains major damage. Using a genuine RPW dataset, the recall, accuracy, *F*-measure, and precision of the classification methods were assessed. According to the experimental findings, data mining may forecast RPW infections with a precision of above 87%, an accuracy of up to 93%, an *F*-measure of more than 93%, and a recall of 100%.

The organization of rest of this paper is structured as follows. Section 2.2 briefly discusses the role of machine learning. Section 2.3 describes the adopted dataset. Then, Sect. 2.4 shows the proposed red palm weevil detection model in detail. Section 2.5 discusses the obtained results. Finally, conclusions and future work are proposed in Sect. 2.6.

## 2.2 Machine Learning

Machine learning has recently grown in popularity as a research area in the field of artificial intelligence. Wide-ranging artificial intelligence research and application is now being conducted by well-known Internet corporations, and machine learning, which includes image and speech recognition, is one of these technologies [20]. In particular, given the fast-paced development of big data, machine learning combined with big data can efficiently combine systems and algorithms, enabling machine learning algorithms to operate concurrently across multiple cores and process large amounts of data, which is also the current research direction in the study of artificial intelligence. One of the most well-known and simple machine learning techniques is the decision tree (DT). By using ensemble learning and boosting approaches, DT for some problems can be enhanced. Examples of that include the random forest classifier (RFC) and gradient boosting classifier (GBC). Another branch of machine learning called deep learning has been demonstrated to be among the most effective approaches currently available, particularly for classification issues involving big datasets. Deep learning makes it possible to train deep neural networks with several hidden layers. Although the idea of training neural networks with numerous hidden layers is not new, the lack of processing power and data in the past has severely hindered the development of this discipline.

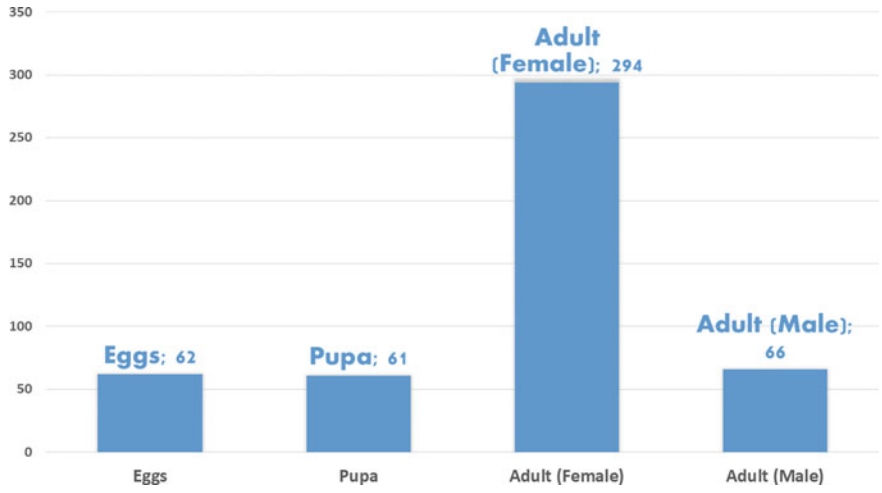



**Fig. 2.1** Class distribution of the collected dataset

## 2.3 Dataset Description

In this study, 483 images of the palm weevil were downloaded using the Google image search engine. The three phases of red palm weevil growth are taken into account in this dataset; Eggs, pupae, and adults make up these life phases. There are two classes of adults: males and females. The adopted dataset's class distribution is shown in Fig. 2.1.

## 2.4 The Proposed Red Palm Weevil Detection Model

In this section, the overall proposed red palm weevil detection based on machine learning algorithms is described in detail. The proposed model consists of four phases: data pre-processing, features extraction, classification, and finally evaluation phases as shown in Fig. 2.2.

### *2.4.1 Data Pre-processing Phase*

In this phase, weevils are extracted from the images using the Rembg tool [21]. Then different oversampling methods are applied to the weevils extracted images. Oversampling is a technique used to balance class distribution in a dataset by increasing the number of samples in the minority class. This is often necessary when working

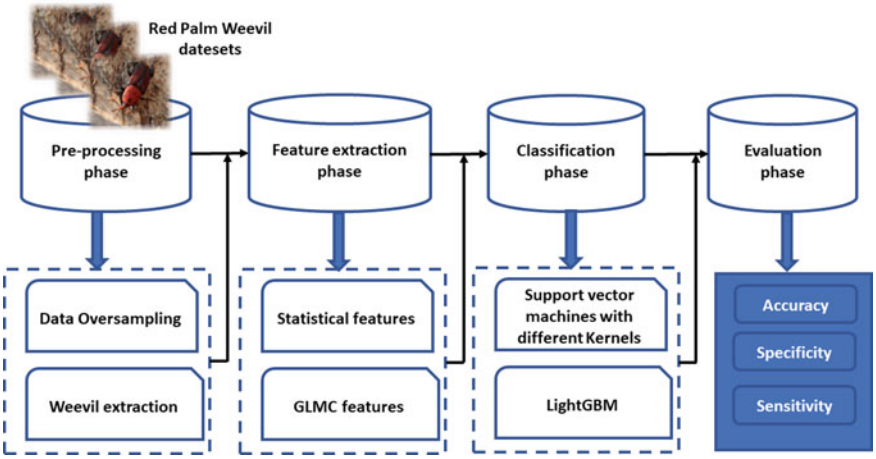



**Fig. 2.2** Proposed red palm weevil detection model

with imbalanced datasets, which can occur when one class significantly outnumbers the other class. There are two popular methods for oversampling imbalanced datasets: random oversampling (ROS) and synthetic minority oversampling technique (SMOTE). In ROS, additional samples are randomly selected from the minority class and added to the dataset, resulting in an equal distribution of classes. However, this method can result in overfitting, as it can introduce duplication of data points [22].

SMOTE, on the other hand, works by generating synthetic minority class samples rather than simply replicating existing minority class samples. To generate synthetic samples, SMOTE selects a minority class sample and its nearest neighbors and interpolates new samples between them. The number of synthetic samples to be generated can be specified by the user. This process is repeated until the desired balance between the classes is achieved. SMOTE has been widely used in various fields, including but not limited to, credit risk assessment, fraud detection, and cancer diagnosis. It has been shown to improve the performance of classification models on imbalanced datasets by reducing bias toward the majority class. The choice of oversampling method will depend on the specific needs and constraints of the problem at hand. It is important to evaluate the results of oversampling and determine whether it has improved the performance of the model on the imbalanced dataset. In this paper, SMOTE oversampling methods are applied to the collected dataset.

### *2.4.2 Feature Extraction Phase*

A dimensionality reduction technique called feature extraction divides a large amount of raw data into smaller, easier-to-process groupings [23]. These huge datasets have

the trait of having many variables that demand a lot of computational power to process. In this phase, several features are extracted from the pre-processed image. These features can be divided into main categories: statistical-based features and gray-level co-occurrence matrix-based features. In this paper, the median statistical-based feature is calculated from the extracted weevil image. One of the earliest techniques for extracting texture features was the grey-level co-occurrence matrices (GLCM), which was first put out by Haralick et al. in [24]. Since that time, it has been extensively utilized in several texture analysis applications and has continued to be a crucial feature extraction technique in the texture analysis field. In this paper, energy, correlation, dissimilarity, homogeneity, and contrast are extracted from the pre-processed images.

### *2.4.3 Classification Phase*

In this phase, the dataset is divided into 70% training and 30% testing. Then, the extracted features from the previous phase are used to feed supervised machine learning algorithms. These algorithms are support vector machine (SVM) and LightGBM. SVM is a type of supervised machine learning algorithm that can be used for classification tasks. The idea behind SVM is to find the hyperplane in a high-dimensional space that maximally separates the different classes [20]. Data points that are closest to the hyperplane are called support vectors and have the most influence on the position of the hyperplane. Once the hyperplane is determined, new data points can be easily classified based on which side of the hyperplane they fall on. SVM can also be used for regression tasks, but they are more commonly used for classification. One advantage of SVM is that it can perform well even when the number of dimensions is much greater than the number of samples. They are also relatively robust to overfitting, which makes them a good choice for many classification tasks [19].

LightGBM is a gradient boosting, type of ensemble learning method that combines multiple weak learners to make a strong prediction [25]. One popular GBM algorithm is LightGBM (LGBM), which is designed for efficient training on large datasets and can handle imbalanced datasets well. LightGBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithms. It is designed to be efficient and scalable and has gained popularity in a variety of machine learning tasks. One of the key features of LightGBM is that it uses a highly efficient implementation of the gradient boosting algorithm, which allows it to train models significantly faster than other implementations, such as XGBoost. Additionally, LightGBM allows users to specify various parameters to control the training process, such as the learning rate, the number of decision trees to include in the model, and the maximum depth of each tree. In terms of performance, LightGBM has been shown to achieve strong results in a variety of benchmarks and real-world applications. The performance of both classifiers is compared and the best classifier is reported.

**Table 2.1** Accuracy result before and after applying SMOTE

| | Accuracy (%) |
|---|---|
| Before | 0.88 |
| After | 0.91 |

## 2.5 Results and Discussions

In this section, there are two main experiments are conducted. The first experiment aims to handle the imbalance problem. The second experiment aims to evaluate the performance of different classifiers, where four classes are considered. These classes are eggs, pupa, male adult, and female adult. It should be noted all the conducted results are implemented on GoogleColab with Python Programming Language.

### *2.5.1 Pre-processing Phase Results*

In this phase, multiple experiments were conducted on the dataset to show the significance of each part in the pre-processing phase. In this phase, SMOTE oversampling method is applied. Table 2.1 compares the performance of the proposed model before applying an oversampling method and after applying it. It should be noted that in this experiment, the full sequence of the proposed model is considered without including SMOTE. Additionally, the LightGBM classifier is used. As can be seen, using an oversampling method such as SMOTE can significantly improve the obtained accuracy. Additionally, it can be observed that the collected dataset suffers from an imbalance problem.

### *2.5.2 Classification Phase Results*

In this subsection, multiple experiments were applied to the dataset to choose the best classifier with the optimal parameters setting that can classify female adults, male adults, eggs, and pupa. Table 2.2 compares the performance of using different kernels of SVM. As can be observed, linear kernel is the optimal kernel, as it obtained the highest accuracy.

**Table 2.2** Comparison of different kernels of SVM in terms of accuracy

| Kernel | Accuracy (%) |
|---|---|
| Linear | 0.72 |
| RBF | 0.65 |
| Polynomial | 0.66 |

**Table 2.3** Comparison between different learning rate of LightGBM in terms of accuracy

| Learning rate | Accuracy (%) |
|---|---|
| 0.9 | 0.73 |
| 0.5 | 0.83 |
| 0.2 | 0.86 |
| 0.1 | 0.91 |

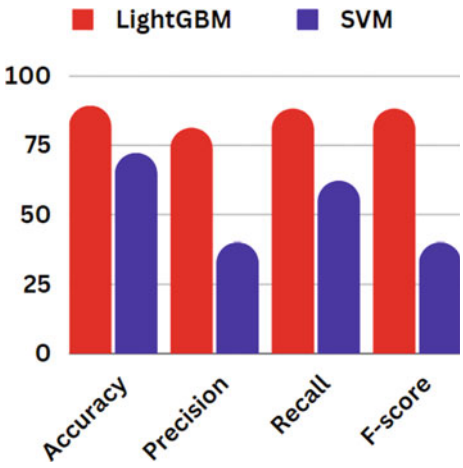



**Fig. 2.3** LightGBM versus SVM in terms of accuracy, precision, recall, and $f$-score

There are many parameters associated with LightGBM; however, we found that the learning rate value can significantly affect its performance. Table 2.3 compares the performance of LightGBM when different learning rate values are considered. As can be observed, when the learning rate value decreases, the obtained accuracy is increased as well. Also, it can be observed that the best learning rate value is 0.1.

Figure 2.3 compares the performance of LightGBM and SVM in terms of accuracy, precision, recall, and $f$-score. As can be observed, LightGBM obtained the best results. Figure 2.4 compares the execution time in seconds for LightGBM and SVM. As it can be observed, LightGBM obtained the minimum execution time. From all results, it can be concluded that LightGBM with a 0.1 learning rate can achieve the highest accuracy with the minimum time.

## 2.6 Conclusions

Red palm weevils have proven to be the world's most devastating pest of palm trees during the past 20 years, especially in the Middle East. Several palm species have suffered significant harm as a result of the RPW. The early detection of the RPW is a difficult challenge for the production of good dates Additionally, early

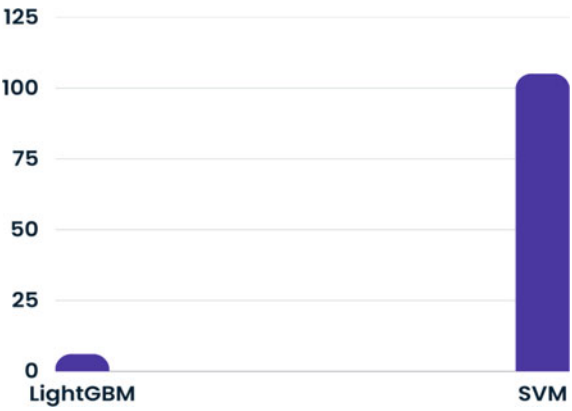



**Fig. 2.4** LightGBM versus SVM in terms of execution time in seconds

detection can prevent the RPW from harming palm plants. This paper proposed a new methodology for the early detection of red palm weevils during their stages. The proposed model starts with data pre-processing followed by feature extraction and finally classification. In the pre-processing phase, weevils are extracted from the images. Then SMOTE oversampling method is applied. Energy, correlation, dissimilarity, homogeneity, and contrast features are extracted from the pre-processed images in the feature extraction phase. In the classification phase, the support vector machine (SVM) and the gradient boosting algorithm (LightGBM), two well-known data mining classification algorithms, were applied and evaluated for their performance. The experimental results showed that the LightGBM algorithm outperforms SVM. It obtained an overall 91% accuracy.

## References

1. Al-Shahib, W., Marshall, R.: The fruit of the date palm: its possible use as the best food for the future? Int. J. Food Sci. Nutr.Nutr. **54**(4), 247–259 (2003)
2. Al-Dosary, N., Al-Dobai, S., Faleiro, J.: Review on the management of red palm weevil rhynchophorus ferrugineus olivier in date palm phoenix dactylifera l. Emirates J. Food Agric. 34–44 (2016)
3. Wahizatul, A., Zazali, C., Abdul, R., Nurul'Izzah, A., et al.: A new invasive coconut pest in Malaysia: the red palm weevil (curculionidae: *Rhynchophorus ferrugineus*). Planter **89**(1043), 97–110 (2013)
4. Ferry, M., Gomez, S., et al.: The red palm weevil in the mediterranean area. Palms **46**(4), 172–178 (2002)
5. Llacer, E., Jacas, J.: Efficacy of phosphine as a fumigant against *Rhynchophorus ferrugineus* (coleoptera: Curculionidae) in palms. Span. J. Agric. Res. **8**(3), 775–779 (2010)
6. Rach, M., Gomis, H., Granado, O., Malumbres, M., Campoy, A., Martin, J.: On the design of a bioacoustic sensor for the early detection of the red palm weevil. Sensors **13**(2), 1706–1729 (2013)
7. Wang, B., Mao, Y., Ashry, I., Al-Fehaid, Y., Al-Shawaf, A., Ng, T., Yu, C., Ooi, B.: Towards detecting red palm weevil using machine learning and fiber optic distributed acoustic sensing. Sensors **21**(5), 1592 (2021)

8. Suma, P., La Pergola, A., Longo, S., Soroker, V.: The use of sniffing dogs for the detection of *Rhynchophorus ferrugineus*. Phytoparasitica **42**(2), 269–274 (2014)
9. Ha, R., Slaughter, D.: Real-time x-ray inspection of wheat for infestation by the granary weevil, *Sitophilus granarius* (l.). Trans. ASAE **47**(2), 531 (2004)
10. Sharma, D., Chakravarthi, D., Boddu, R., Madduri, A., Ayyagari, M., Khaja Mohiddin, M.: Effectiveness of machine learning technology in detecting patterns of certain diseases within patient electronic healthcare records. In: Proceedings of Second International Conference in Mechanical and Energy Technology, pp. 73–81. Springer (2023)
11. Sayed, G., Khoriba, G., Haggag, M.: The novel multi-swarm coyote optimization algorithm for automatic skin lesion segmentation. Evolutionary Intelligence, 1–31 (2020)
12. Alsariera, Y., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A., Ali, N.: Assessment and evaluation of different machine learning algorithms for predicting student performance. Comput. Intell. Neurosci. **2022** (2022)
13. Patriarca, R., Di Gravio, G., Cioponea, R., Licu, A.: Democratizing business intelligence and machine learning for air traffic management safety. Saf. Sci.. Sci. **146**, 105530 (2022)
14. Mucherino, A., Papajorgji, P., Pardalos, P.: A survey of data mining techniques applied to agriculture. Oper. Res. Int. J. **9**(2), 121–140 (2009)
15. Hill, M., Connolly, P., Reutemann, P., Fletcher, D.: The use of data mining to assist crop protection decisions on kiwifruit in New Zealand. Comput. Electron. Agric.. Electron. Agric. **108**, 250–257 (2014)
16. Achirul Nanda, M., Boro Seminar, K., Nandika, D., Maddu, A.: A comparison study of kernel functions in the support vector machine and its application for termite detection. Information **9**(1), 5 (2018)
17. Karadg, K., Tenekeci, M., Taşaltın, R., Bilgili, A.: Detection of pepper fusarium disease using machine learning algorithms based on spectral reflectance. Sustain. Comput. Inform. Syst. **28**, 100299 (2020)
18. Al-Saqer, S., Hassan, G.: Artificial neural networks based red palm weevil (*Rynchophorus ferrugineous*, olivier) recognition system. Am. J. Agric. Biol. Sci. **6**, 356–364 (2011)
19. Kurdi, H., Al-Aldawsari, A., Al-Turaiki, I., Aldawood, A.: Early detection of red palm weevil, *Rhynchophorus ferrugineus* (olivier), infestation using data mining. Plants **10**(1), 95 (2021)
20. Sayed, G., Khoriba, G., Haggag, M.: Parameters optimisation of support vector machine using modified grasshopper optimisation algorithm-based levy-flight method. Int. J. Comput. Aided Eng. Technol. **15**(1), 120–147 (2021)
21. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M.: U2-net: going deeper with nested u-structure for salient object detection. Pattern Recogn.Recogn. **106**, 107404 (2020)
22. Hayaty, M., Muthmainah, S., Ghufran, S.: Random and synthetic over-sampling approach to resolve data imbalance in classification. Int. J. Artif. Intell. Res. **4**(2), 86–94 (2020)
23. Sayed, G., Hassanien, A.: An improved wild horse optimizer for traveling salesman problem. In: The 5th International Conference on Computing and Informatics (ICCI), pp. 274–279 (2022)
24. Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man Cybern.Cybern. **6**, 610–621 (1973)
25. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.: Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural. Inf. Process. Syst. **30**, 3149–3157 (2017)

# Chapter 3
# Unstructured Text Classification Using NLP and LSTM Algorithms


**Sashikanta Prusty, Srikanta Patnaik, Ghanashyam Sahoo, Jyotirmayee Rautaray, and Sushree Gayatri Priyadarsini Prusty**

**Abstract** In the last two decades, cancer has found as the most severe disease that causes a large number of deaths worldwide. However, it seems that the medical data are in the unstructured form, which makes it challenging for the pathologist to classify the disease in the beginning stages. This article provides a novel technique using natural language processing (NLP) for the classification of three common cancers thyroid, lung, and colon from the unstructured medical data, removing at-stop words, and articles such as 'a,' 'an,' and 'the' to form tokens. Tokenization is performed on larger text to divide it into smaller parts with the help of the TensorFlow technique. Besides that, long short-term memory networks (LSTM) model has been proposed to handle long sentences, especially in sequence prediction problems. This article also discusses the basic workflow of model design and prediction of three cancers. Finally, a confusion_matrix has been drawn to evaluate our model performance and found 99.41% at classifying three individual cancers.

S. Prusty (✉) · S. G. P. Prusty
Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be University, Bhubaneswar 751030, India
e-mail: sashi.prusty79@gmail.com

S. Patnaik
Interscience Institute of Management and Technology, Kantabada, Bhubaneswar 752024, India

G. Sahoo
Department of Computer Science and Engineering, GITA Autonomous College, Bhubaneswar 751017, India

J. Rautaray
Department of Computer Science and Engineering, Odisha University of Technology and Research, Bhubaneswar 751003, India



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_3

## 3.1 Introduction

Cancer is a disease that occurs when a few of the body's cells grow out of control and spread across the body. It can begin anywhere in the body of a person. Tumors can develop when this systematic process fails, allowing abnormal or damaged cells to proliferate. Cancerous (malignant) or non-cancerous (benign) tumors are both possible. Cancerous tumors can move to remote regions of the body to establish new tumors. They can also infiltrate neighboring tissues [1]. In the year 2023, there were 19,58,310 new cancer cases and 6,09,820 cancer-related deaths found by the US government [2]. The cancer death rate decreased from 2019 to 2020 (by 1.5%) in Covid-19 pandemic period, helping to contribute to a 33% total drop since 1991 and probably reducing 3.8 million deaths worldwide [3].

The clinical information in pathology reports is primarily presented as unstructured free text, making it difficult to read or search for. The development of NLP has encouraged the inclusion of precise textual information in EHRs to aid patient care and boost cancer research [4, 5]. Several natural language processing (NLP) methods have been developed to streamline the text classification of pathology reports. Unstructured text may be the only way to capture some information that is essential for cancer research and patient care, such as whether and when cancer gets better or worse following a particular therapy [6].

Long short-term memory (LSTM) was first put forth by Hochreiter and Schmidhuber in 1997 to address the issue of vanishing and exploding gradients. However, the biggest advantage is protecting its hidden activation using three gates. An LSTM model has automatic control over whether to keep important properties in the cell state or toss out unimportant ones, and it can recall past long-term time-series data [7, 8]. The input gate regulates how new information enters the cell state. The forget gate purges the cell state of old, irrelevant data. The output gate controls the information that has been taken from the cell state and then chooses the hidden state.

## 3.2 Material and Methods

### *3.2.1 Material*

In this article, the *cancer dataset* has been collected from the Kaggle repository, containing unstructured text about three cancers like thyroid, colon, and lung, which are more common cancer around the globe. Figure 3.1a represents the pie chart of each cancer percentage with three different colors in the dataset, where green is 'thyroid_cancer,' blue is 'colon_cancer,' and orange is 'lung_cancer.' Meanwhile, Fig. 3.1b shows the density distribution of three individual cancers concerning several words present in the dataset.

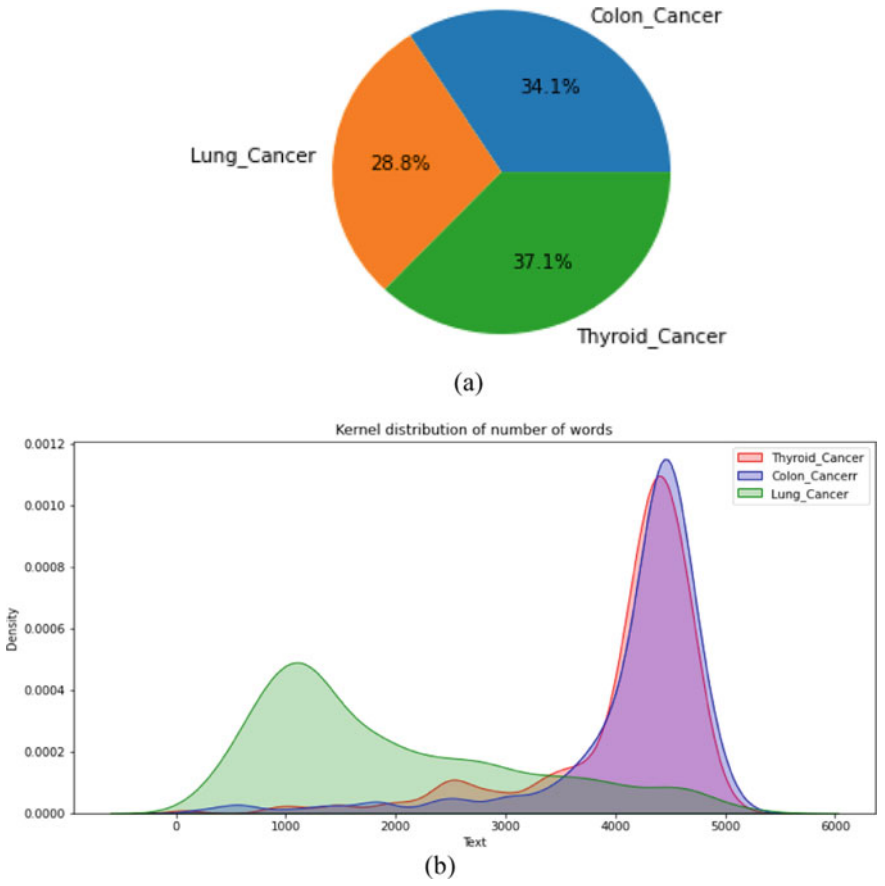



**Fig. 3.1** Dataset representation for three cancers containing **a** classification percentages and **b** density of words

### *3.2.2 Method*

As we have discussed, medical data are generally in unstructured form and need to emphasize related terms while ignoring stop words. More importantly, finding such words from Cancer data makes it challenging for doctors to identify the relevant disease [9]. Thus, there is the need to synchronize the data properly so that it will be suitable for diagnosing the disease at the earlier stages. Tokenization is the initial step in turning the information we provide into numerical values that a machine-learning model can understand. According to the author, choosing the right amount of tokens is crucial since it influences how accurately the Sentence Piece algorithm classifies sentences [10]. Furthermore, to handle long-term dependencies in cancer datasets, the novel LSTM technique is found as the best model these days. They can retain knowledge for long periods, which accounts for this. Second, LSTMs are

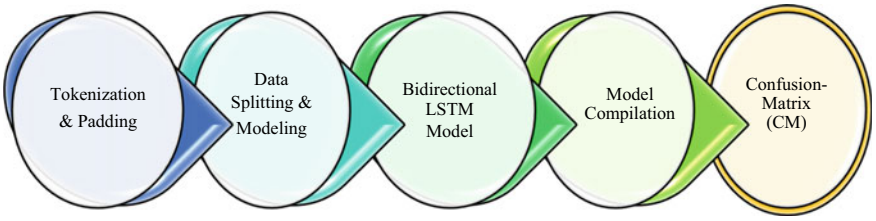



**Fig. 3.2** Proposed method to classify three cancers

significantly less prone to the vanishing gradient issue. Last but not least, LSTMs are excellent at simulating complex sequential data. Therefore, we have proposed a method to classify thyroid, lung, and colon cancers from unstructured cancer datasets using NLP and LSTM techniques as shown in Fig. 3.2.

**Tokenization and Padding**. In general, every textual dataset contains sentences having stop words, symbols, prepositions, and articles like 'the,' 'a,' and 'an.' However, to train a DL model, we require a sequence of tokens with the help of the TensorFlow technique. To do so, we apply tokenization on larger text to divide it into smaller parts, named tokens. This helps in finding patterns which is an initial step before being fed into a model. Sensitive data items can be replaced with impermeable data elements with the aid of tokenization. In Python, the word_tokenize class has been used to form tokens by removing stop words. Figure 3.3, displays the text in the form of tokens from the long sentences in the cancer dataset.

TensorFlow delivers numerical values associated with each token after tokenizing a text. This is sometimes referred to as word_index and is a dictionary composed of the words 'word: index.' Every word that is encountered has a unique number, which is used to identify that word. Any deep learning model frequently requires input with constant size. This implies that our model will have issues with phrases of various lengths. Padding is useful in this situation. In this experiment, cancer data has been tokenized into integer sequences by tokenize_pad_sequences (text), and each sequence has been padded to the same size as shown in Fig. 3.4. This function decides to place zeros before or after the short sequences to make them equal to higher sequences. Set padding as 'post' defines that the zeros are inserted as post-sequences.

**Data Splitting**. Another aspect is to validate the data before being fed into the DL model. Therefore, we used train_test_split () with test_size = 0.3 to get the

```
Out[15]: 0    [thyroid, surgery, children, single, instituti...
         1    [adopted, strategy, used, prior, years, based,...
         2    [coronary, arterybypass, grafting, thrombosis,...
         3    [solitary, plasmacytoma, sp, skull, uncommon, ...
         4    [study, aimed, investigate, serum, matrix, met...
         Name: final_text, dtype: object
```

**Fig. 3.3** Representation of final text after tokenization

Out[21]:

| | Target | Text | final_text |
|---|---|---|---|
| 0 | 0 | Thyroid surgery in children in a single insti... | thyroid surgery children single institution os... |
| 1 | 0 | " The adopted strategy was the same as that us... | adopted strategy used prior years based four e... |
| 2 | 0 | coronary arterybypass grafting thrombosis ï¬b... | coronary arterybypass grafting thrombosis brin... |
| 3 | 0 | Solitary plasmacytoma SP of the skull is an u... | solitary plasmacytoma sp skull uncommon clinic... |
| 4 | 0 | This study aimed to investigate serum matrix ... | study aimed investigate serum matrix metallopr... |
| ... | ... | ... | ... |
| 7565 | 1 | we report the case of a 24yearold man who pres... | report case yearold man presented chief compla... |
| 7566 | 1 | among synchronous colorectal cancers scrcs rep... | among synchronous colorectal cancers scrcs rep... |
| 7567 | 1 | the heterogeneity of cancer cells is generally... | heterogeneity cancer cells generally accepted ... |
| 7568 | 1 | "adipogenesis is the process through which mes... | adipogenesis process mesenchymalstem cells msc... |
| 7569 | 1 | the periparturient period is one of the most c... | periparturient period one challenging periods ... |

7570 rows × 3 columns

**Fig. 3.4** Final tokenization of cancer data containing 7570 rows * 3 columns

**Table 3.1** Classification of texts for three individual sets

| | X_trn.shape | y_trn.shape |
|---|---|---|
| Train set | 4239, 300 | 4239, 3 |
| Validation set | 1817, 300 | 1817, 3 |
| Test set | 1514, 300 | 1514, 3 |

train, validation, and test set. X_trn.shape, X_vld.shape, and X_tst.shape have been performed to classify the shape for each set as follows in Table 3.1.

**TF-IDF and Logistic Regression**. The statistical technique known as term frequency–inverse document frequency (tf-idf) is frequently applied in information retrieval and NLP. It measures a term's significance within a document concerning a group of documents. However, the term frequency (tf) can be classified as the proportion of the number of times the terms appeared to the total number of terms in the document.

$$\text{tf} = \frac{t}{n} \tag{3.1}$$

where $t$ = number of times a term is identified in a document and $n$ = total number of terms.

To find the percentage of phrases from the documents in the cancer dataset, we have calculated the Inverse document frequency (IDF). Words that are specific to a small number of documents are given greater relevance values than words that appear in all documents. So, IDF can be evaluated as:

$$\text{idf} = \log \frac{\text{no. of documents in cacer dataset}}{\text{number of documents containg phrases in the dataset}}) \tag{3.2}$$

From Eqs. (3.1) and (3.2), tf-idf can be calculated as:

$$\text{tf} - \text{idf} = \text{tf} * \text{idf} \tag{3.3}$$

For this experiment, we first required to import sklearn's TfidfVectorizer() to transform an entire set of words and fit them into features. Although, it can be applied for the training and testing of the dataset, we implemented the logistic regression (LR) model using the 'liblinear' solver after creating an 80/20 train-test split in the dataset. The input features are combined with weights before being sent via a sigmoid function by the LR classifier. Any real value entered into the sigmoid function is converted to a number between 0 and 1. However, the confusion_matrix for TF-IDF and logistic regression is shown in Fig. 3.5.

**Bidirectional LSTM**. This technique is capable of learning long-term dependencies, such as for sequence prediction problems. In contrast to single data points like text sequences, LSTM exhibits tremendous performance on a wide range of problems. Figure 3.6, describes the workflow of the LSTM model using three basic gates input, forget, and output gates. In LSTMs, gates regulate the addition and deletion of data from the cell state. Information may be able to enter and leave the cell through these gates. At the output end, there is a pointwise multiplication operation and a layer of sigmoid neural networks. The sigmoid layer outputs values between 0 and 1, where 0 denotes 'nothing' and 1 denotes 'everything should be let through.'

A bidirectional LSTM is a sequence processing model that comprises two LSTMs, one of which receives input forward and the other of which receives it backward.

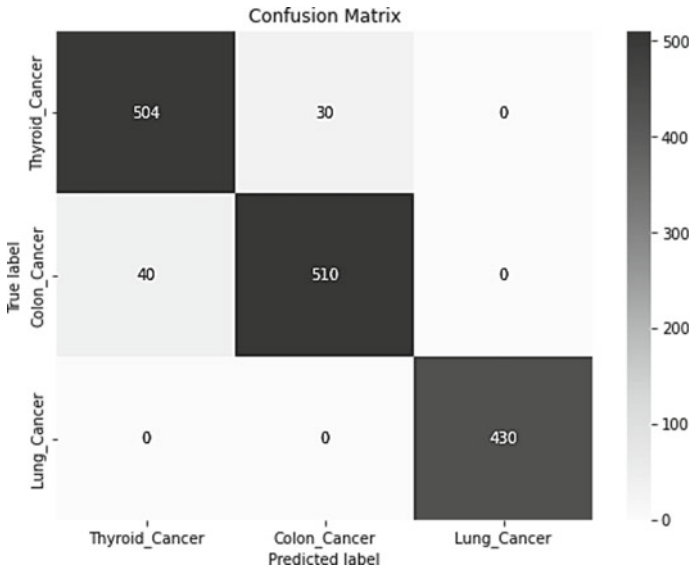



**Fig. 3.5** Representing confusion_matrix for TF-IDF and logistic regression model

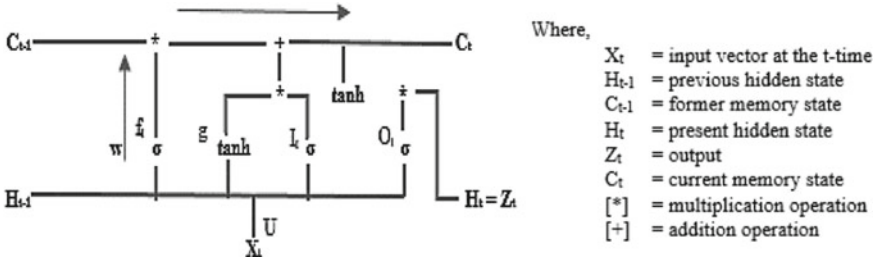



**Fig. 3.6** Structural block diagram of LSTM model

It essentially increases the amount of information that the network has access to, improving the context that the algorithm has access to, for example, by letting it know what words in a sentence come right after and before a given word. Here, Bidirectional () has called with parameter LSTM (32) specifying 64 outputs, as shown in Fig. 3.7. Moreover, an activation function as 'sigmoid,' Mapoolinng1D with 'pool_size as 2,' and embedding_size as '32' have been used for the model building. All these experiments are done on the Jupyter Notebook 6.4.3 platform using the Python programming language.

**Model's Compilation**. After building a model, the next is to compile the model using an optimizer. Creating a model is a step that comes before training. It also defines the metrics, optimizer or learning rate, and loss function, and checks for format issues. For training, a compiled model is required. Here, we have taken loss as 'categorical_ crossentropy,' optimizer as 'adam,' and metrics as 'accuracy' as parameters for the model compilation. The model has been trained using the model.fit() function that takes epochs as 50, and batch_size as 64, as arguments. After successful training, the model performance is shown in Fig. 3.8.

**Model's Performance**. However, a model can be called as good when it is evaluated successfully and achieves better accuracy. To make this happen, we have implemented the model.evaluate (X_tst, y_tst) function over here. This resulted in a 99.41% accuracy, which is much closer to 1. Furthermore, a history plot has been drawn to classify the cancers concerning 50 epochs, resulting in higher accuracy as shown in Fig. 3.9.

**Confusion-Matrix (CM)**. An evaluation of a classification algorithm's performance is done using a confusion matrix. It depicts and summarizes the performance of a classification model and provides a comparison between actual and predicted classes. True positive (TP), false positive (FP), true negative (TN), and false negative (FN) are four common parameters used in CM to find their class. In Fig. 3.10, the *x*-axis and *y*-axis denote the actual and predicted level of three cancers, where their level of classes is shown in dark color.

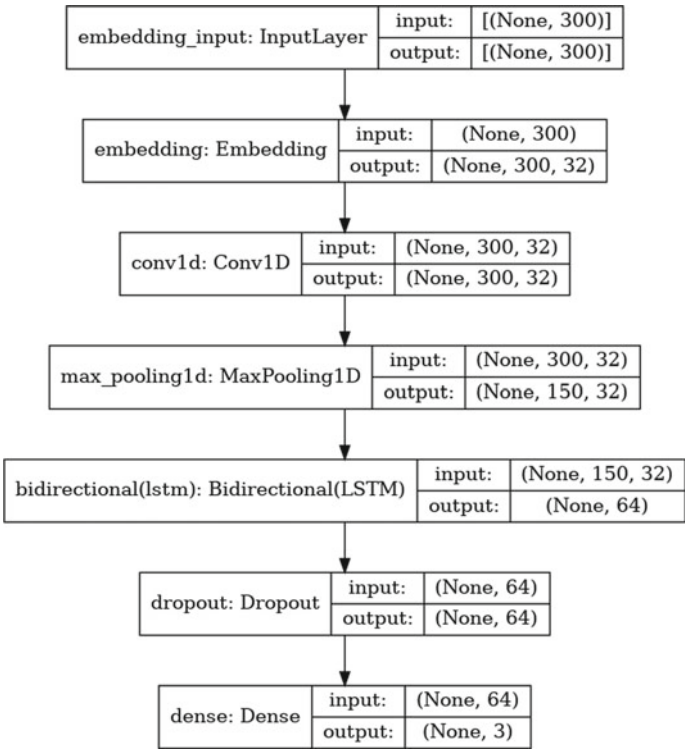



**Fig. 3.7** Process workflow of bidirectional LSTM using NLP technique

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 300, 32)           160000
_________________________________________________________________
conv1d (Conv1D)              (None, 300, 32)           3104
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 150, 32)           0
_________________________________________________________________
bidirectional (Bidirectional (None, 64)                16640
_________________________________________________________________
dropout (Dropout)            (None, 64)                0
_________________________________________________________________
dense (Dense)                (None, 3)                 195
=================================================================
Total params: 179,939
Trainable params: 179,939
Non-trainable params: 0
```

**Fig. 3.8** Showing compilation result for the LSTM model

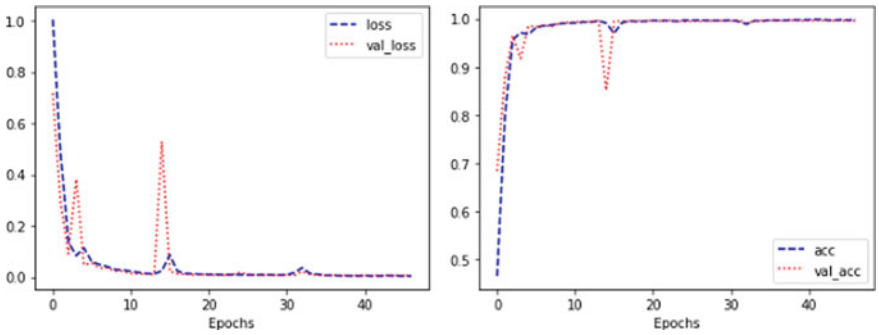



**Fig. 3.9** Representation of LSTM model performance concerning 50 epochs

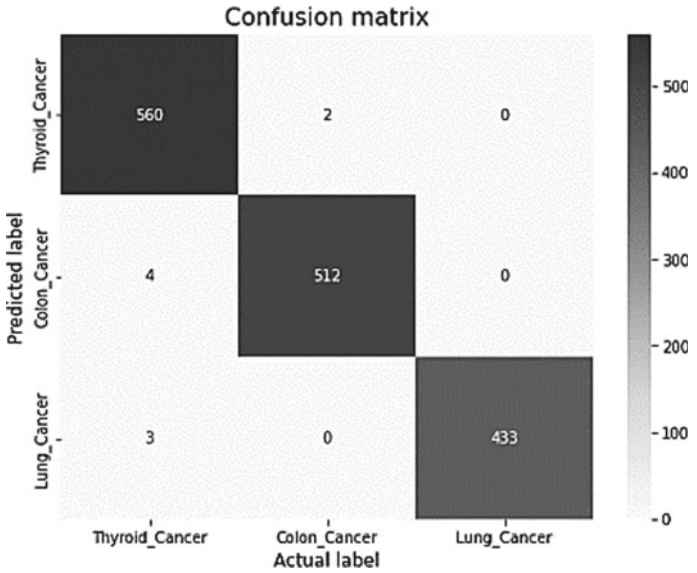



**Fig. 3.10** Representing CM for thyroid, lung, and colon cancers

## 3.3 Conclusion and Future Scope

So far to date, there has been lots of research done worldwide to classify texts that are in unstructured form. Some of them worked, and some of them have not worked properly. To get rid of this, in this research, we have tried to design a novel technique to classify cancer from unstructured text data by removing stop words, articles, and many more. In this research, we tried to design a novel technique to classify cancer from unstructured text data by removing stop words, articles, and many more. Thus, we have applied tokenization to convert long sentences into tokens and a bidirectional LSTM model to classify three cancers lung, thyroid, and colon. As the name suggests,

bidirectional LSTM comprises two LSTMs, one of which receives input forward and the other of which receives it backward. For this reason, we have taken this model for our experiment. However, LSTM forgets the previous modifications that happened in the data and remembers only the selective areas. That is why in future work, we will try to implement the RNN technique for better improvement over the model.

## References

1. Prusty, S., Dash, S.K., Patnaik, S.: A novel transfer learning technique for detecting breast cancer mammograms using VGG16 bottleneck feature. ECS Trans. **107**(1), 733 (2022)
2. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA Cancer J. Clin. **73**(1), 17–48 (2023)
3. Islami, F., Guerra, C.E., Minihan, A., Yabroff, K.R., Fedewa, S.A., Sloan, K., Jemal, A.: American Cancer Society's report on the status of cancer disparities in the United States, 2021. CA Cancer J. Clin. **72**(2), 112–143 (2022)
4. Wang, L., Fu, S., Wen, A., Ruan, X., He, H., Liu, S., Liu, H.: Assessment of electronic health record for cancer research and patient care through a scoping review of cancer natural language processing. JCO Clin. Cancer Inform. **6**, e2200006 (2022)
5. Botsis, T., Murray, J., Alessandro, L.E.A.L., Palsgrove, D., Wei, W.A.N.G., White, J.R., Johns Hopkins Molecular Tumor Board Investigators: Natural language processing approaches for retrieval of clinically relevant genomic information in cancer. Stud. Health Technol. Inform. **295**, 350 (2022)
6. Kehl, K.L., Xu, W., Lepisto, E., et al.: Natural language processing to ascertain cancer outcomes from medical oncologist notes. JCO Clin. Cancer Inform. **4**, 680–690 (2020)
7. Prusty, S., Patnaik, S., Dash, S.K.: Differentiating S1, S2 noises from abnormal heart sounds generated in closure of atrioventricular and semilunar valves using MFCC and LSTM. In: 2022 1st IEEE International Conference on Industrial Electronics: Developments & Applications (ICIDeA), pp. 208–213. IEEE (2022)
8. Mahima, S., Kezia, S., Grace Mary Kanaga, E.: Deep learning-based lung cancer detection. In: Disruptive Technologies for Big Data and Cloud Applications: Proceedings of ICBDCC 2021, pp. 633–641. Springer Nature Singapore, Singapore (2022)
9. Prusty, S., Patnaik, S., Dash, S.K.: SKCV: stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Front. Nanotechnol. **4**, 972421 (2022)
10. Taniguchi, Y., Konomi, S.I., Goda, Y.: Examining language-agnostic methods of automatic coding in the community of inquiry framework. In: 16th International Conference on Cognition and Exploratory Learning in Digital Age IADIS Press, Cagliari, Italy, 19–26 (2019)

# Chapter 4
# Regression-Based Model for Prediction of Road Traffic Congestion: A Case Study of Janpath Segment in Bhubaneswar City


**Sarita Mahapatra, Srikanta Patnaik, Krishna C. Rath, Kabir M. Sethy, and Satya R. Das**

**Abstract** With the integration of communication and information technology in the traffic management system, it is becoming possible to make much-needed changes in the way people commute in big cities. By introducing different modes of commuting, advanced infrastructure, solution related to traffic and mobility management, it is getting feasible to enhance the overall efficacy of the system in terms of overall performance and expenditure. With the help of advanced and emerging communication technologies, a smart and safe mode of traffic management can be ensured. The present work has been inspired by the potential problems arising due to random urban growth. This study is carried out for the capital city of Odisha, that is Bhubaneswar, which is growing spatially and population-wise at exponential rate. The present work has identified one of the important aspects related to the urban traffic management such as analysis and monitoring of traffic congestion. A regression-based model has been proposed to identify and predict the locations for traffic congestion. The model is validated considering a particular road segment in the study area using a geospatial analysis based on geographic information system (GIS). With the increasing significance of urban traffic problems, the application of geographic information system (GIS) can greatly improve the operational efficiency of urban traffic management system. Urban traffic management system based on GIS has become an important part of Intelligent Traffic System (ITS). ITS based on GIS is an open and comprehensive system engaged in control, management and decision-making.

S. Mahapatra (✉)
Department of CSIT, ITER, SOA University, Bhubaneswar, Odisha, India
e-mail: saritamahapatra@soa.ac.in

K. C. Rath · K. M. Sethy
Department of Geography, Utkal University, Bhubaneswar, Odisha, India

S. Patnaik · S. R. Das
Department of CSE, ITER, SOA University, Bhubaneswar, Odisha, India



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_4

## 4.1 Introduction

In the coming decade or so, globally around 812.2 million people will live in urban areas with a city population of ten million or more. With the steep increase in urban population, the city municipalities are facing tough challenges to provide a better living environment for its digital citizens by exploiting technology. In the making of a smart city, among several issues, road traffic management has a prominent role to play. Since billions of people are opting for one among the several public modes for communication, an inefficient traffic management system can certainly make the life difficult in urban areas. Thus, it is crucial to have suitable ways for management of traffic channels, which ultimately defines the living standard of people in the modern, hi-tech cities. Urban traffic management includes some typical issues such as road accidents, traffic congestion, vehicle parking, planned construction works, etc. Among these, traffic congestion is one issue that affects the other issues directly, hence identified for further study.

Congestion in road traffic in India in the last 10–15 years has been a major concern. Issues arising due to acute traffic congestion influence adversely the quality of day-to-day life as well as the financial stability of the country. To face the various challenges due to the traffic congestion, we need to design smart solutions which includes prediction of traffic congestion levels. Efficient analysis of traffic congestion will lead to fast identification of some potential issues related to traffic management and thus will help in reducing the traffic congestion.

Many cities in world including Bhubaneswar in India are experiencing significant growth in road traffic congestion. Due to that, Govt. Of India and transport officials are encouraging use of public and active transport options to ameliorating the congestion. In this scenario, road infrastructure and transportation system is increasingly being shared by more modes which is resulting in complex intersection interactions, and as there is no scope for further expansion available physical infrastructure in many areas, it is needed to manage the available resources and make optimum use of it. So, we need a help of science and technology to design different models to predict the traffic-related issues and identify special traffic-generated activities. We can observe that the road traffic scenario of Bhubaneswar city is mainly due to the population gathering from all parts of the country primarily for education facilities and opportunities in job and business. An unwelcome feature has been the concentration of population in this city, having a population of more than a million which gradually increases the number of two-wheelers, personal vehicles in roads. This increasing population is responsible for increasing vehicle population in form of two-wheelers, personal vehicles in roads which gradually lead to a congested road, slower speeds, longer delays, more fuel consumption and accidents in roads. This unexpected growth of vehicles sometime create dilemma to take decisions on road issues. The important reasons behind these problems are the prevailing imbalance between road infrastructural developments and growth of vehicles in roads along with inadequate public transport system which is not able to keep pace with the demands. In the last few years, many researchers have been encouraged to address these problems so as to reduce the wrong impacts of

traffic congestion. Unfortunately, there is no way to handle when congestion occurs in real in roads. To fill this gap, we are proposing a predictive model in order to make prediction the congestion occurrences by analyzing the congestion trends.

In this paper, a critical study has been conducted on the traffic data of the city of Bhubaneswar. A regression-based neural network model is proposed for an efficient prediction of traffic congestion considering a particular road segment of Janpath road of Bhubaneswar, that is from Vani Vihar Square to Sishu Bhawan Square, as Janpath road has been found to be one of the major roads with high traffic congestion over the years.

By at looking the recent situation and by a recent national survey, delays caused by traffic congestion are more concern for peoples when they are attending any event like official, school, personal events. According to Ali et al. [1], traffic congestion causes nearly millions of price investment every day in terms of opportunity cost and fuel consumption. Also, it has a lot of impacts in our personal life such as (i) loss of time especially during crowded hours increases mental stress and (ii) the added pollution severely affects our environment. Congestion has a serious threat to the lives of vehicle drivers and also serious impact on every inhabitant of the city. Due to congestion, fuel consumption is more [2], and harmful pollutants are added to the air and cause significant slower the economic growth rate [3]. According to a laboratory study [4], pollutant which is emitted due to road congestion increases the chances of health issues like allergies and aggravates the symptoms for those who are sensitive to them. Further researches have given that traffic congestion also enhances the risk of heart attacks [5, 6]. For a developing country like India where smooth running of traffic in roads without congestion is not possible, both comfort and economic growth of road users are very crucial for development. Ensuring these above two requirements for development transportation sectors, the emphasis is on monitoring traffic congestion and its solutions. Accurate traffic congestion prediction is depending on the accurate data collection on roads. Somehow accurate real-life data is required time-to-time plan and availability of modern resources. Traffic congestion prediction provides the authorities to take right steps to avoid more congestions and remedial plan for future in the resource allocations to have a smooth road journey [7]. Keeping in view the traffic congestion issues in different areas of the capital city of Odisha motivated us to focus in these issues and developing a solution.

**Reasons for Congestion Occurrence**: Traffic congestion in most of the cities in India occurs due to various reasons, such as day-to-day work life, huge transport demand, sudden incidents, issues related to weather, or arrangement of special events in crowded places, etc. Based on the reasons for congestion, road traffic congestions can be categorized into the following two types: (1) recurring congestions and (2) nonrecurring congestions [8].

- **Recurring Congestion** In most of the cities, traffic congestions occur during peak hours every day. As per govt. data, approximately 50% of the congestion cases are recurring as experienced by users [5]. The regular causes of recurring congestions are: (1) higher supply and lesser capacity, (2) insufficient traffic controllers, (3) inadequate infrastructure and (4) uncertain and inconsistent traffic flow. [6]

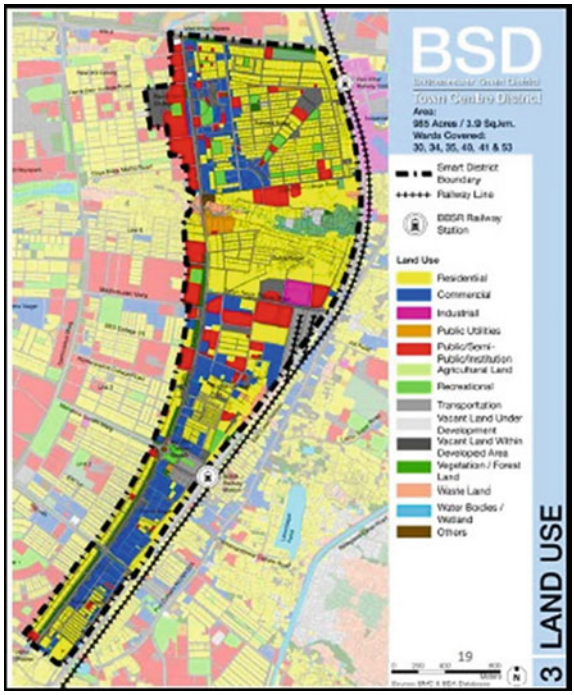



**Fig. 4.1** Janpath road segment

- **Nonrecurring Congestion** The major reasons for nonrecurring congestions are occurrence of unpredictable events, such as traffic accidents, official gatherings, adverse weather or other specific events. Nonrecurring congestion may lead to new congestion in the off periods, along with leading to the delay occurring due to recurring congestion [9–11].

**Case Study on Janpath Smart Corridor of Bhubaneswar City**: In the process of developing Bhubaneswar as a smart city, the pilot project has been initiated during year 2016 to make Janpath (from Vani Vihar sq. to Sishu Bhawan sq.) as a Smart Corridor. As the land use map around Janpath represents Fig. 4.1, the Janpath predominantly encompasses commercial areas on its eastern side and residential and institutional areas on its western side. This characterizes Janpath as a crowded corridor of the city. Hence, this has been taken as a case study for the proposed model here for analysis and validation.

## 4.2 Related Works

Often in intelligent traffic management systems, traffic-related data with missing values and outliers adversely affect the prediction of traffic congestion. Methods proposed in [12, 13] describe historical imputation methods (HIMs) which give

more than one predicted values for the missing value. In HIMs, the mean value of more than one neighboring data points received from same date and location replaces the missing value. But in [14, 15], authors have given another method that is nearest neighbor imputation (NIM) which takes data from the neighboring roads. These methods are not efficient when data from the neighboring roads are not available. However, various machine learning and deep learning models are available which can be applied for prediction of traffic congestion with complicated dataset. In particular, deep learning models are more suitable than the other prediction models due to their architectural advantages. A strategy for predication of traffic congestion based on GPS trajectory data using recurrent neural network (RNN) is proposed by authors in [16] where the model estimates the average speeds from existing data on road extensions with GPS trajectory data. But the RNN-based model faced issues related to long-term dependencies since it could not keep the earlier data. Hence, this led to the evolution of LSTM model. In [17], authors have given a temporary information improvement-based LSTM model that estimates the flow of traffic on a single stretch of road and got advantages over prediction accuracy by emphasizing on the special correlation among the time and flow of traffic. Authors in [18] have developed a spatio-temporal graph convolution network model to highlight on the temporal and spatial features in the congestion prediction. This model has a faster training speed using lesser number of attributes and a full convolution architecture. Thus, there are some substantial efforts made for the prediction of traffic intensity. But looking at the increased complexities in the city of Bhubaneswar, very little attempts have been made for the congestion prediction.

## 4.3 Methodology Used

In the process of developing Bhubaneswar as a smart city, the pilot project has been initiated during year 2016 to make Janpath (from Vani Vihar sq. to Sishu Bhawan sq.) as a Smart Corridor. As the land use map around Janpath represents Fig. 4.1, the Janpath predominantly encompasses commercial areas on its eastern side and residential and institutional areas on its western side. This characterizes Janpath as a crowded corridor of the city. Hence, this has been taken as a case study for the proposed model here for analysis and validation. The Bhubaneswar city has witnessed both planned and unplanned rapid urbanization across all directions such as toward Khurda, Cuttack and Puri. In the last few years, Bhubaneswar city has become the educational as well as business hub of the state along with the state-of-the-art medical facilities. Hence, people from all parts of the state and the country are moving to Bhubaneswar for their livelihood. It has substantially affected the urban environment, population growth and most importantly traffic congestion in the city. Presently, the traffic-related issues in the city have become a major concern. In Fig. 4.1, we have given the road map of Janpath, Bhubaneswar, prepared using QGIS, an open-source GIS software.

### *4.3.1 Data Collection, Processing and Analysis*

Traffic congestion-related data have been collected from both primary and secondary source. The primary sources included field surveys and observations aided with observation schedule, field books and GPS. We have collected traffic information of different road segments of Janpath, a crowded road of Bhubaneswar with subdivided segments like Vani Vihar sq. to Rupali sq., Rupali sq. to Ram Mandir sq., Ram Mandir sq. to Master canteen sq., Master canteen sq. to Rajmahal sq., Rajmahal sq. to Shishu Bhawan sq. Questionnaire surveys have also been conducted through online (google form) and offline mode for capturing respondent perceptions on traffic congestion-related parameters. The proposed model has been developed using linear regression implemented using Python libraries. In this regard, the data from master database integrated with primary sources (field observations) for Janpath smart corridor have been used. Traffic data studies are conducted to determine the traffic-related attributes like number, movements and classification of road vehicles at a given location. These data can help to find out traffic volume, traffic density, traffic speed, and critical flow time period, traffic volume patterns and determine the different vehicle impacts in traffic issues. We have mostly collected the data like volume of traffic, count of different types of vehicles, direction to travel and vehicle occupancy. Traffic congestion prediction has two important and basic steps such as: (1) data collection and (2) model development for prediction. Both of these steps need to be executed methodically to get higher degree of accuracy. After collecting data, data processing needs to be done for preparing the training and testing datasets.

Data cleaning plays vital role in the accuracy and better performance of our model. Missing data handling is a deceptively tricky issue so we cannot ignore it by removing from our dataset. They must need extra care as they can reveal some important information about the dataset. We have dropped the observations with missing values to handle our dataset, hence being able to predict even with new data if some of the features are missing.

Traffic data studies are conducted to determine the traffic-related attributes like number, movements and classification of road vehicles at a given location. These data can help to find out traffic volume, traffic density, traffic speed, critical flow time period, traffic volume patterns and determine the different vehicle impacts in traffic issues. There are two types of traffic survey such as: (1) automatic count survey and (2) manual traffic count. We selected manual count method to study the road traffic characteristics of capital city Bhubaneswar. By using manual count method, we have mostly collected the data like volume of traffic, count of different types of vehicles, direction to travel and vehicle occupancy. Manual count of survey method has disadvantages like time consuming, duplication of data entry and human errors, but it has lots of advantages also like scope for error correction, detailed desired information can be collected, simple to conduct, flexibility and not required much cost to perform the survey. We have collected traffic information of different road segments of a crowded road of Bhubaneswar like Janpath Road: from Vani Vihar to

**Fig. 4.2** View of Janpath from flyover on Vani Vihar Sq.

Rupali, Rupali to Ram mandir, Ram mandir to Master canteen, Master canteen to Rajmahal, Rajmahal to Shishu Bhawan Sq. etc. (Fig. 4.2).

### *4.3.2 Observations for Janpath Smart Corridor*

Tables 4.1, 4.2, 4.3, 4.4 and 4.5 represent derived data from master database developed from field observations. Analyzed information for the five sub-segments of Janpath smart corridor for four time zones including two peak hours and two lean hours has been described in the table. It can be observed from the table that in the peak hours, the sub-segment Ram Mandir Sq. to Master Canteen Sq. has the highest value of PCU which indicates highest traffic density.

**Table 4.1** Traffic status from Vani Vihar to Rupali Square of Janpath Road in Bhubaneswar

| S. No. | Vehicle type | Vehicle count | % share in congestion |
|---|---|---|---|
| 1 | Car | 380 | 22.32 |
| 2 | Truck/bus | 44 | 2.58 |
| 3 | 3-wheeler | 258 | 15.16 |
| 4 | 2-wheeler | 418 | 24.55 |
| 5 | Van | 25 | 1.46 |
| 6 | Others | 577 | 33.90 |
| Total vehicle count = 1702 | | | |
| PCU = 1817.5 | | | |

**Table 4.2** Traffic status from Rupali Sq. to Rammandir Sq. of Janpath road in Bhubaneswar

| S. No. | Vehicle type | Vehicle count | % share in congestion |
|---|---|---|---|
| 1 | Car | 425 | 24.955 |
| 2 | Truck/bus | 45 | 2.64 |
| 3 | 3-wheeler | 390 | 22.90 |
| 4 | 2-wheeler | 529 | 31.06 |
| 5 | Van | 36 | 2.11 |
| 6 | Others | 278 | 16.324 |
| Total vehicle count = 1703 | | | |
| PCU = 1588.0 | | | |

**Table 4.3** Traffic status from Rammandir Sq. to master canteen Sq. of Janpath road in Bhubaneswar

| S. No. | Vehicle type | Vehicle count | % share in congestion |
|---|---|---|---|
| 1 | Car | 423 | 27.02 |
| 2 | Truck/bus | 46 | 2.93 |
| 3 | 3-wheeler | 345 | 22.04 |
| 4 | 2-wheeler | 449 | 28.69 |
| 5 | Van | 25 | 1.59 |
| 6 | Others | 277 | 17.69 |
| Total vehicle count = 1565 | | | |
| PCU = 1497.25 | | | |

**Table 4.4** Traffic status from master canteen Sq. to Rajmahal Sq. of Janpath road in Bhubaneswar

| S. No. | Vehicle type | Vehicle count | % share in congestion |
|---|---|---|---|
| 1 | Car | 410 | 27.29 |
| 2 | Truck/bus | 45 | 2.99 |
| 3 | 3-wheeler | 310 | 20.63 |
| 4 | 2-wheeler | 436 | 29.02 |
| 5 | Van | 26 | 1.73 |
| 6 | Others | 275 | 18.30 |
| Total vehicle count = 1502 | | | |
| PCU = 1447.0 | | | |

### *4.3.3 Proposed Regression-Based Model*

Traffic congestion prediction is not so straightforward using models. Most common factors for traffic congestion prediction are the study area, data collection methodology, selection of predicted parameters, prediction intervals and validation procedures. Data collection for traffic congestion prediction varied from year to year. Traffic congestion level prediction is done by predicting time of observation, traffic

**Table 4.5** Traffic status from Rajmahal Sq. to Sishu Bhawan Sq. of Janpath road in Bhubaneswar

| S. No. | Vehicle type | Vehicle count | % share in congestion |
|---|---|---|---|
| 1 | Car | 278 | 34.66 |
| 2 | Truck/bus | 14 | 1.74 |
| 3 | 3-wheeler | 208 | 25.93 |
| 4 | 2-wheeler | 223 | 27.80 |
| 5 | Van | 12 | 1.49 |
| 6 | Others | 67 | 8.35 |
| Total vehicle count = 802 | | | |
| PCU = 706 | | | |

speed, flow of traffic or vehicle count, traffic density. Our regression-based model for the prediction of congestion level is shown in Fig. 4.3.

Simple linear regression is one of the widely used regression algorithms in machine learning where a significant variable from the data set is selected to predict the future values in terms of output variables. Regression analysis is a predictive modeling strategy which is mainly used to find the degree of relationships among two or more variables and also helps to design a predictive model. In this analysis, the dependent variable and the independent variable are called as the target and predictor, respectively. The values of predictors are required to predict the possible value of the target variable.

The entire process of regression analysis includes important steps such as (1) identifying the predictors and target, (2) finding the relationship between the target and predictor, (3) estimating the coefficients, (4) finding the estimated values of the target and (5) eventually calculating the model accuracy of the fitted relationship.

Our regression-based model considers the traffic congestion data from geospatial traffic database, which primarily contains vehicle count data for different types of vehicle such as motorcycle, auto rickshaw, car, truck/bus and others for seven different one-hour time zones (including both peak and lean time zones for traffic congestion). The data have been collected for the five sub-segments of Janpath road segment as mentioned in Sect. 3.1. The total vehicle count ($\text{Total}_{\text{VC}}$), passenger car unit (PCU) and traffic density (TD) have been calculated for each row in the table which means for a particular time zone and sub-segment of the road by the following formula: $\text{Total}_{\text{VC}}$ = sum of vehicle count for each vehicle type

$$\text{PCU} = \sum_{i=1}^{n} \text{vehiclecount}_i \times \mu,$$

where $\mu = 0.5, 1.0, 0.75, 3.0, 1.5$ and 1.5 for 2-wheelers, car, 3-wheelers, truck-/bus, vans and others, respectively, and each value of $i$ represents a particular vehicle type with $n$ as the number of different types of vehicle.

$$\text{TD} = \text{PCU}/(\text{length of road sub - segment}).$$

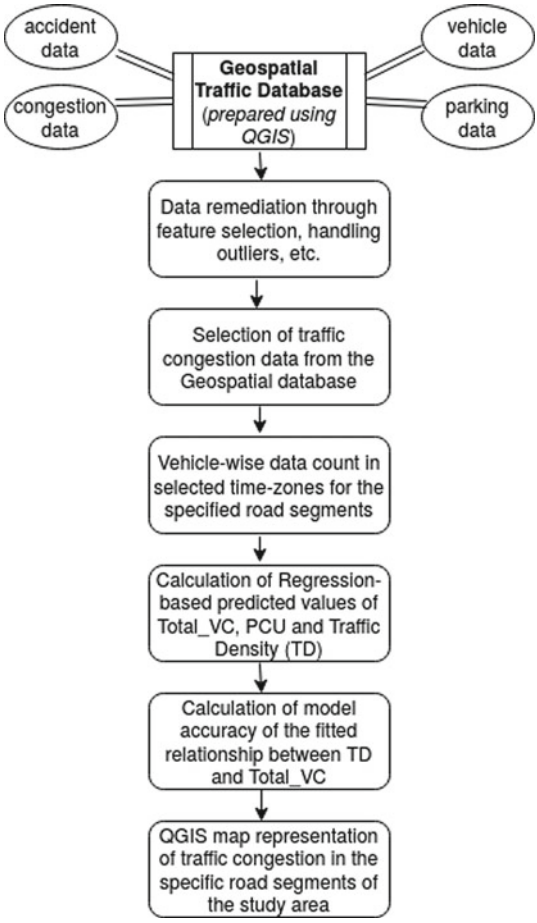



**Fig. 4.3** Flow diagram of the model for congestion prediction

**Algorithm 1**: Regression-Based-Prediction ($R$, $n$, Pos, $k$)

Step1: Identifying the predictors and target from the traffic database
Step2: Finding the relationship between the target and predictor
Step3: $\text{Total}_{\text{VC}}$ = sum of vehicle count for each vehicle type
Step4: for $j = 1$ to $l$
Step5: for $i = 1$ to $k$
Step6: PCU = $\mu[i]$ * $\text{Total}_{\text{VC}}$ $[i]$

where $\mu[i] = 0.5, 1.0, 0.75, 3.0, 1.5$ and $1.5$ for 2-wheelers, car, 3-wheelers, truck/bus, vans and others, respectively, and each value of $i$ represents a particular vehicle type with $n$ as the number of different types of vehicle.

Step7: TD = PCU/(length of road sub-segment)
Step8: Estimating the coefficients
Step9: Finding the estimated values of the target

Step10: Calculating the model accuracy of the fitted relationship
Step11: Computing residue of each instance by the difference of $TD_{observed}$ and $TD_{predicted}$.

## 4.4 Analysis and Discussion

There is a random sample of 104 observations in total for regression-based prediction of the traffic congestion based on the different time zones and total vehicle count. The primary goal of the model is to determine the best fitting line through the data points using a scatter plot of ($x = \text{TD}$, $y = \text{Total}_{\text{VC}}$) and the straight line drawn among the data points indicates the fitting with the experimental data 4. Hence, the scatter plot with the straight line in between shows the causal relationship between traffic density, the dependent variable and total vehicle count, the independent variable. Here, traffic density is considered as the target variable since it is directly proportional to traffic congestion since a high traffic density can lead to a high degree of traffic congestion.

**Evaluating the Linear Regression Model**: In the following discussions, we have presented some methods for the evaluation of the accuracy of the regression-based model based on its predictive power. After fitting our linear regression model, we have shown the accuracy of the model based on its predictive ability with the following test results.

**Correlation Coefficient**: The Multiple *R* test value as shown in Table 4.6 known as the correlation coefficient, comes out to be 0.899685713 which means the dependent variable (traffic density) and the independent variable (total vehicle count) are **highly correlated**.

The Multiple *R* test value as shown in 5.6 known as the correlation coefficient comes out to be 0.899685713 which means the dependent variable (traffic density) and the independent variable (total vehicle count) are highly correlated. The degree of correlation is also shown in the scatter plot in Fig. 4.4 which indicates the target and predictor are linearly and highly correlated as the data points are mostly scattered around the predicted regression line. The line fit plot shown in Fig. 4.5 is a special type of scatter plot that presents the data points across the fitted regression line. Both of these graphs indicate how better the model fits the data. The normal probability plot

**Table 4.6** Regression statistics

| Test type | Test value |
|---|---|
| Multiple *R* | 0.899685713 |
| *R*-square | 0.809434382 |
| Adjusted *R*-square | 0.807566091 |
| Standard error | 290.9545091 |
| Observations | 104 |

shown in Fig. 4.6 almost has a straight line pattern which provides the information that the errors or residuals are normally distributed.

**Coefficient of Determination**: The *R*-square test value of the model as shown in Table 4.6, which is also known as the coefficient of determination, indicates the number of data points falling on the regression line means the percentage of accuracy of the model. Here, this *R*-square value comes out to be 0.809434382 means the **accuracy of our model is nearly 80.94%**. *R*-square value is always between 0 and 100%. As per the guideline, the higher the *R*-square value, the better the model. But,

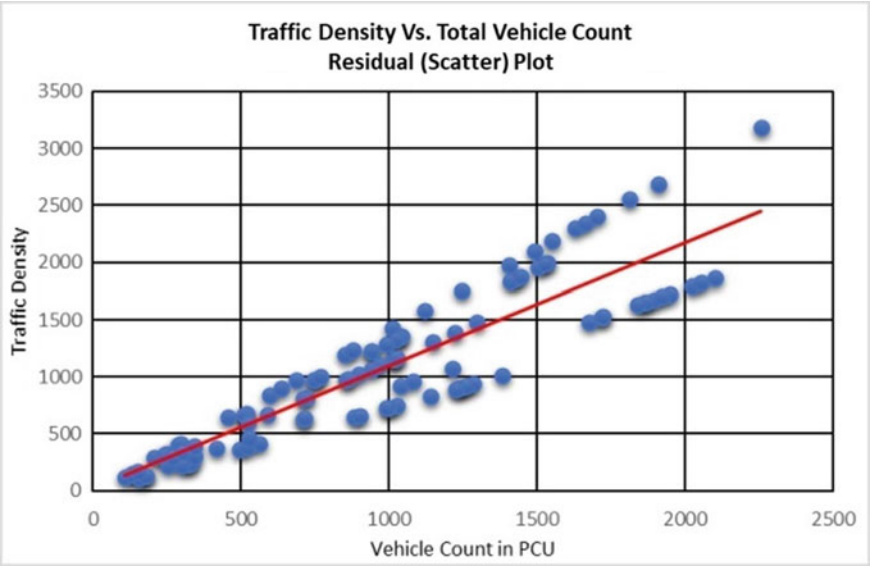



**Fig. 4.4** Result1

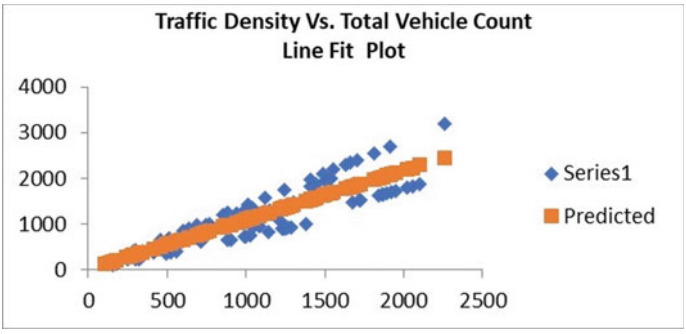



**Fig. 4.5** Result2

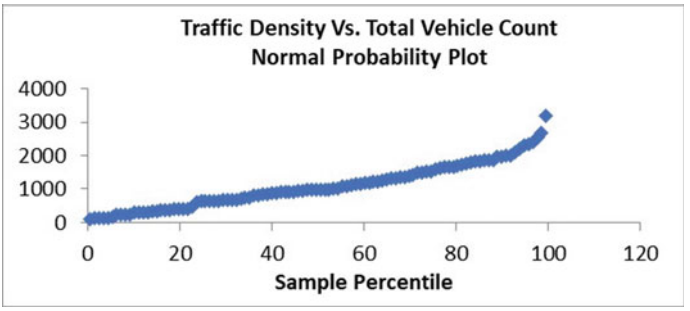



**Fig. 4.6** Result3

**Table 4.7** Length and traffic densities of the five selected road segments in Janpath road, Bhubaneswar

| Road segment no. | From Sq. | To Sq. | Length in km | Traffic density |
|---|---|---|---|---|
| 1 | Vani Vihar | Rupali | 0.71 | 1833.80 |
| 2 | Rupali | Ram Mandir | 1.37 | 1243.06 |
| 3 | RamMandir | Master Canteen | 1.13 | 2032.46 |
| 4 | Master canteen | Rajmahal | 0.77 | 1950.64 |
| 5 | Rajmahal | Sishu Bhawan | 0.88 | 911.36 |

at the same time, the goal is not to maximize the *R*-square value as a higher *R*-square value may also affect the applicability and stability of the model.

The adjusted *R*-square value or the modified version of the *R*-square value comes out to be 0.807566091 which is very close to the result in *R*-square test value, and hence, it can be inferred that our regression-based **model has obtained a best fit**.

The estimated traffic density derived from the model has been presented in Table 4.7. The results have also been presented in GIS map as given in Fig. 4.7. The results indicate that the sub-segment from RamMandir Sq. to Master Canteen Sq. is characterized by highest traffic density and thereby highest degree of congestion. Analysis of the model results in line with the observed data from the master database and map analysis provides us the information that the said sub-segment from RamMandir Sq. to Master Canteen Sq. is congested mostly due to the following reasons:

- Presence of number of crowd-pulling centers such as temples, shopping malls, hotels, bus stand, railway station, cinema halls etc.
- Lack of enough subways and intersections.
- Unauthorized parking.
- Unplanned construction work.
- Rally, crowd gathering, etc. without any prior notice.

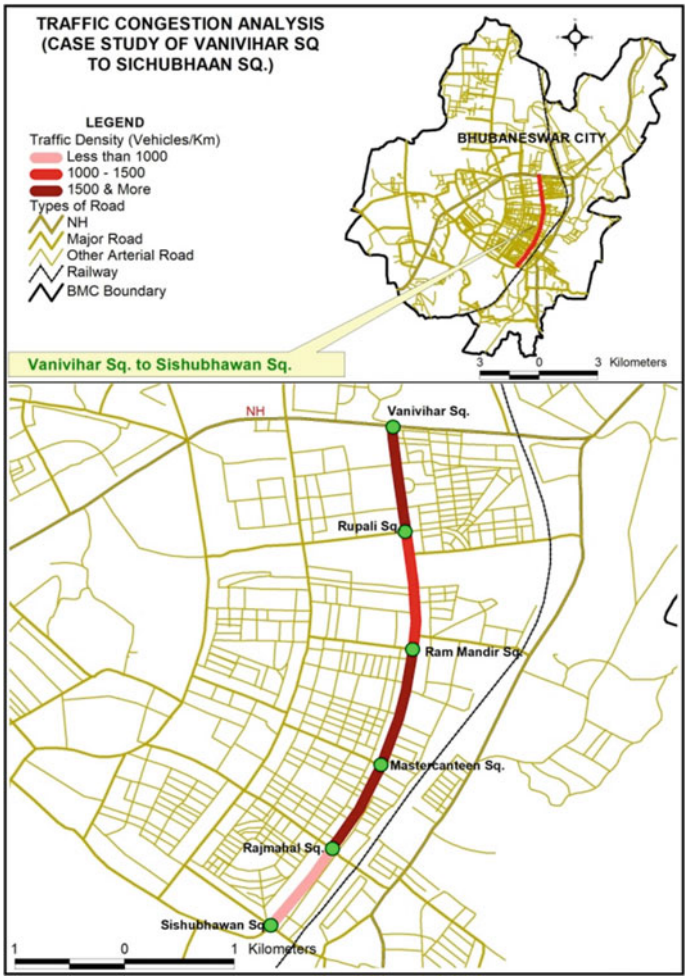



**Fig. 4.7** Traffic congestion analysis (Vani Vihar Sq. to Sishu Bhawan Sq.)

## 4.5 Conclusion

Accurate data collection and huge data availability help us to get more accurate prediction results. Correct and accurate prediction of traffic congestion can provide safe traffic environments; reduce delay time which is spent in the congestion, cost saving and also avoid traffic accidents. So, it is most important to prevent and predict traffic congestion accurately. The working procedure of this model is very simple and has a very less computation time. The final result of this prediction model shows that it is not only effective to predict traffic congestion but also helps to find out different related parameters associated with traffic congestion. This paper presents

a strategy for prediction of traffic congestion levels and provides basis ideas for managing issues related to road traffic congestions.

## References


1. Aftabuzzaman, M.: Measuring traffic congestion: a critical review. In: Proceedings of the 30th Australasian Transport Research Forum (ATRF), Melbourne, Australia, 25–27 Sept 2007
2. Falcocchio, J.C., Levinson, H.S.: Managing nonrecurring congestion. In: Road Traffic Congestion: A Concise Guide, pp. 197–211. Springer, Berlin, Heidelberg, Germany (2015)
3. Ghosh, B.: Predicting the Duration and Impact of the Nonrecurring Road Incidents on the Transportation Network. Ph.D. Thesis, Nanyang Technological University, Singapore, May 2019
4. Fonseca, D.J., Moynihan, G.P., Fernandes, H.: The role of nonrecurring congestion in massive hurricane evacuation events. In: Recent Hurricane Research Climate, Dynamics, and Societal Impacts, pp 441–458. In Tech, London, UK (2011)
5. Tonne, C., Beevers, S., Armstrong, B., Kelly, F., Wilkinson, P.: Air pollution and mortality benefits of the London congestion charge: spatial and socioeconomic inequalities. Occup. Environ. Med.. Environ. Med. **65**, 620–627 (2008)
6. Falcocchio, J.C., Levinson, H.S.: Road Traffic Congestion: A Concise Guide, vol. 7. Springer, Berlin/Heidelberg, Germany
7. Robinson, R.M., Collins, A.J., Jordan, C.A., Foytik, P., Khattak, A.J.: Modeling the impact of traffic incidents during hurricane evacuations using a large scale micro simulation. Int. J. Disaster Risk Reduct. **31**, 1159–1165 (2018)
8. Haselkorn, M., Yancey, S., Savelli, S.: Coordinated Traffic Incident and Congestion Management (TIM-CM): Mitigating Regional Impacts of Major Traffic Incidents in the Seattle I-5 Corridor, Department of Transportation. Office of Research and Library: Washington, DC, USA, 2018
9. Mahmassani, H.S., Dong, J., Kim, J., Chen, R.B., Park, B.B.: Incorporating Weather Impacts in Traffic Estimation and Prediction Systems. Joint Program Office for Intelligent Transportation Systems, Washington, DC, USA, 2009
10. He, F., Yan, X., Liu, Y., Ma, L.: A traffic congestion assessment method for urban road networks based on speed performance index. Procedia Eng. **137**, 425–433 (2016)
11. Documentation and Definitions Urban Congestion Reports Operations Performance Measurement FHWA Operations. Available online https://ops.fhwa.dot.gov/perfmeasurement/ucr/documentation.htm
12. Ni, D., Leonard, J.D., Guin, A., Feng, C.: Multiple imputation scheme—for overcoming the missing values and variability issues in ITS data. J. Transp. Eng. **131**(12), 931938 (2005)
13. Luo, X., Meng, X., Gan, W., Chen, Y.: Traffic data imputation algorithm based on improved low-rank matrix decomposition. J. Sensors **2019**, 111 (2019)
14. Chen, J., Shao, J.: Nearest neighbor imputation for survey data. J. Official Statist. **16**(2), 113–131 (2000)
15. Beretta, L., Santaniello, A.: Nearest neighbor imputation algorithms: a critical evaluation. BMC Med. Inform. Decis. Making **16**(S3), 74 (2016)
16. Sun, S., Chen, J., Sun, J.: Traffic congestion prediction based on GPS trajectory data. Int. J. Distrib. Sensor Netw. **15**(5), 155014771984744 (2019)
17. Mou, L., Zhao, P., Xie, H., Chen, Y.: T-LSTM: a long short-term memory neural network enhanced by temporal information for traffic flow prediction. IEEE Access **7**, 98053–98060 (2019)
18. Yu, H.Y., Zhu, Z.: Spatio-temporal graph convolution networks: a deep learning framework for traffic forecasting. In: Proceedings of International Joint Conference on Artificial Intelligence, 2018, pp. 3634–3640

# Part II
# Regular Papers

# Chapter 5
# Automated Landmark Detection for AR-Based Craniofacial Surgical Assistance System

**Sanghyun Byun, Muhammad Twaha Ibrahim, M. Gopi, Aditi Majumder, Lohrasb R. Sayadi, Usama S. Hamdan, and Raj M. Vyas**

**Abstract** The shape of the face of cleft lip patients varies significantly from a regular face due to the unique form and differing levels of severity of their condition. The first step in cleft lip repair requires surgeons to mark anthropometric landmarks that are used as a guide to conduct surgical incisions. These landmarks are different from the ones that are deemed important in a regular face and cannot be detected by existing facial landmark detection frameworks. We propose a AI/ML-based assistive tool that can automatically mark the anthropometric landmarks for cleft repair on the image of the cleft lip patient. We use a novel method for training a convolutional neural network that detects the anthropometric landmarks for patients with cleft lip without requiring a large number of images for training. By utilizing image region of interest (ROI) warp and direct regression, the proposed approach is able to accurately detect landmarks despite variation in the appearance of the cleft. Further, we show the significant improvement ROI warp has on the prediction of anthropometric landmarks used for cleft surgeries. We collaborate closely with reputed craniofacial surgeons to build our training datasets and validate the accuracy of our automated markings. This tool is anticipated to have a tremendous impact on building surgical capacity for cleft repair surgeries, which has a huge shortage, in particular in rural areas, especially in emerging global areas of South America, Africa, and India.

## 5.1 Introduction

Every year, around 195,000 babies globally are born with oral or facial clefts. 4.62 million people in the world today are living with an unrepaired cleft, which increases their chances of suffering from life-threatening problems like malnutrition, or death

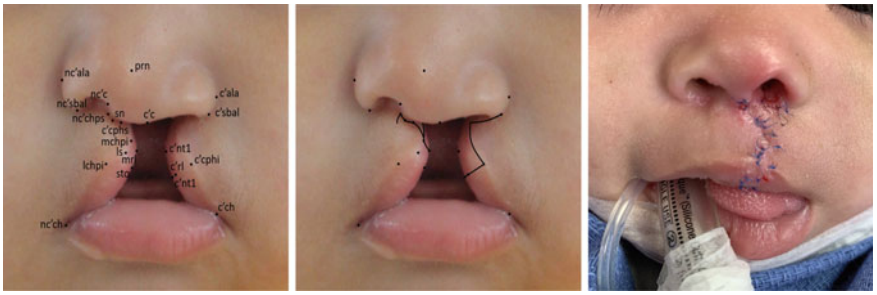
S. Byun (✉) · M. T. Ibrahim · M. Gopi · A. Majumder
Department of Computer Science,University of California—Irvine, Irvine, USA
e-mail: sanghyub@uci.edu

L. R. Sayadi · R. M. Vyas
Department of Plastic Surgery, University of California—Irvine, Irvine, USA

U. S. Hamdan
Global Smile Foundation, Norwood, USA



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_5




(a) cleft anthropometric landmark (b) Millard repair (c) post-surgery

**Fig. 5.1** 21 cleft anthropometric landmarks from anterior view (**a**), surgery guideline used for Millard repair (**b**), and post-surgery incision marks with blue stitches outlining the incision (**c**). *c*' stands for cleft side and *nc*' stands for non-cleft side

due to choking, by 2.15 times. Other life-impacting effects include speech impediment, deafness, malocclusion, gross facial deformity, and severe psychological problems [1]. Therefore, children born with a cleft lip must undergo reconstructive surgeries that aim to stitch the lip and palette together. The first reconstructive surgery must happen before the 18th month when the facial tissue is soft, malleable, and therefore, more amenable to repair. As the child grows older and the face shape changes with age, they may require follow-up corrective surgeries up until the age of 14–15 years [2].

Cleft surgery is one of the most challenging surgeries. Most of the time, surgery is performed on infants 3–6 months old where the surgical area is less than $4 \times 4$ cm. in size. Cleft lip deformations come in several levels of complexities and severities. Around 85% of surgeries utilize the rotation technique, named so due to the skin flap moving in a curved path during surgery. Successful planning and execution of the different kinds of incisions (e.g., Millard, Tennison-Randall, Mulliken) all start with a precision marking of 21 anatomical points of reference [3], called anthropometric landmarks, that are used to plan the incision (see Fig. 5.1). Surgeons use these points to make measurements and guide the cleft repair surgery. They mark incisions using these points, and then during the surgery, they try to align the landmarks on either side of the cleft to reconstruct a balanced, symmetric, and aesthetically pleasing lip. Figure 5.1 shows an example of techniques used.

Accurate markings of these points directly impact the quality and accuracy of the repair, which in turn determines the number of corrective surgeries required in the future. Cleft surgery is very sensitive to anthropometric landmarks. Incorrect marking can lead to an asymmetric reconstructed lip that requires further corrective surgeries, thereby increasing costs and discomfort to the patient. Despite years of experience, surgeons put in tremendous effort and time to mark the 21 keypoints precisely due to their short-term and long-term impact on the surgical outcome. Getting to a level

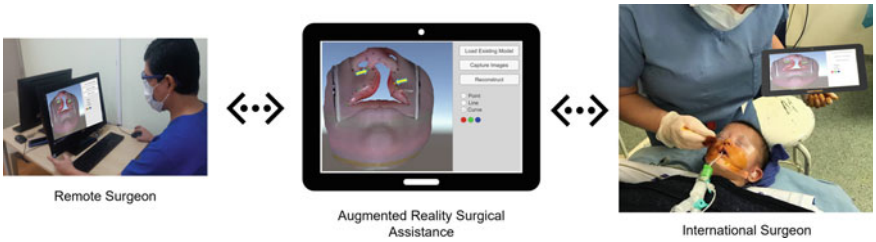



**Fig. 5.2** Example of a remote surgery done through augmented reality proctoring on a tablet

of accuracy that assures good outcomes takes a lot of repetitive practice and is a skill built by the surgeon with years of experience.

Imagine an assistive tool that can collect data from expert surgeons during surgical planning and use an AI/ML-based method to predict the 21 anthropometric landmarks automatically. This assistive tool can be used in the following scenarios:

- Practicing surgeons can use this tool for skill building. Perform repetitive hands-on training on a large database of images even when they do not have access to an in-person trainer with direct feedback. They can even practice on a 3D mannequin, and the markings on its pictures can provide feedback on the accuracy they achieved.
- AR-based expert surgical assistance is becoming common in remote areas with low surgical capacity (see Fig. 5.2). When a novice surgeon is performing a complex surgery with expert guidance using an AR application on a tablet, the predicted markings provide a great starting point. The accuracy of these predicted points provides a great foundation and reference for the remote expert to guide the surgeon. We are working closely with Dr. Raj Vyas and Dr. Ross Sayadi of the University of California, Irvine (UCI), and Children's Hospital of Orange County (CHOC) to provide such an assistive tool as part of a bigger effort on novel AR systems for remote surgical guidance [4, 5].

### *5.1.1 Our Contributions*

In this paper, we propose a novel method that uses convolutional neural networks (CNNs) to develop an assistive tool that automatically detects and labels the anthropometric landmarks used in cleft surgeries. The input to our tool is a database of images of cleft lip patients with anthropometric marks. Such data can easily be collected during the planning phase of any cleft repair surgery. Our main contributions are as follows:

- An end-to-end network for accurately detecting all 21 anthropometric landmarks in an image of a face with cleft lip deformity. We make use of transfer learning to avoid building an entirely new solution for the cleft lip. Instead, we leverage

the large amount of work done on detecting general facial features and the large training datasets thereof. This allows us to achieve high-accuracy detection very efficiently with training datasets that are an order of magnitude smaller in size.
- We show that warping the input image to a pre-computed template face and extracting the predicted landmark coordinates from this region of interest (ROI)-warped space significantly improves precision.
- We also build an efficient GUI that can help surgeons mark the 21 anthropometric landmarks on a cleft lip image which is subsequently used for training.

Both the AI/ML-based automatic landmark prediction method and the GUI can be adapted for different kinds of surgeries in the future. We find, talking to our surgeon collaborators, that almost all surgeries involve such marking of key landmarks before planning the incisions.

## 5.2 Related Work

Recent advances in artificial intelligence and machine learning, especially in the subfield of face detection and recognition, can be used to assist surgeons with the marking process by automatically detecting the keypoints from a single capture of a face with a cleft lip. However, current state-of-the-art methods would not be able to detect landmarks accurately. This is because a cleft face has unique features that are not present in normal faces (see Fig. 5.4). This makes it harder for convolutional models to extract correct feature maps. However, they can be leveraged to help us in detecting landmark points for faces with cleft lip deformities.

### *5.2.1 Regression-Based Facial Landmark Detection*

Regression-based methods are a more traditional approach to landmark detection and can be further divided into *direct regression* and *cascaded regression* models. Facial landmark detection through *direct regression* [6–11] predicts landmarks by passing an image through a backbone, then putting the output into a fully connected network to predict the coordinates. Here, the backbone network can be any network (e.g., ResNet or HRNet) that can extract necessary features from a given image to predict the coordinates. *Cascaded regression* models [12–17] take regression models a step further by using coarse-to-fine methods [17] to predict coordinates. These models use the previously detected landmarks to iteratively update the coordinates, generating predicted landmark coordinates after a certain number of iterations. Though cascaded regression models tend to be more effective as they use iterative steps for prediction, they need a large training set to attain prior knowledge of predefined face shapes.

### 5.2.2 Heatmap-Based Facial Landmark Detection

Inspired by fully convolutional networks [18], heatmap-based methods aim to predict landmark coordinates without the use of a fully connected layer. Therefore, most heatmap facial landmark detection models use coordinated prediction to generate a semantic map, i.e., a heatmap that has Gaussian distributions around the predicted landmark coordinates [19–24]. Softmax function is then used to force the sum of elements to one. The coordinate is then extracted by the argmax function. As information gathered from a local view of an image does not give a full understanding of the image, Wei et al. [25] proposed an alternative *convolutional pose model* where the heatmap is generated in stages, slowly increasing the effective receptive field, an area of the image that the regressor focuses on. Although designed to be used for human pose estimation, it is often altered and trained to detect facial landmarks as well. When compared to regression models, heatmap facial landmark detection often shows superior results in terms of accuracy. However, as heatmaps are highly vulnerable to correlated features, in cases where there are only a few training images, heatmap-based methods will often exhibit noise, reducing the robustness, and making it harder to assess the usefulness of the model.

### 5.2.3 One-Hot Facial Landmark Detection

Region-based convolutional neural network (R-CNN) [26], a network built to detect objects in an image through region proposals, is also used for facial landmark detection. An initial CNN is used to extract rectangular region proposals, which are then used to classify each region using a separate classifier. To enhance this method to an even finer level of detection and classification, *Mask R-CNN* [27] was developed on top of Faster R-CNN [28]. This replaces the classification head with a region of interest pooling, reducing the overhead as well as allowing the model to give pixel-wise classification of the image.

### 5.2.4 Other Techniques for Facial Landmark Detection

Although most prior works put a heavy focus on the model architecture for the performance of their methods, other aspects of learning such as *loss function* and *image preprocessing* can affect the final results. While not studied as much, there are a few cases where such methods have shown to have a significant impact on the results. As all training sequences aim to reduce the value of their loss functions [29–31], choosing the right function is crucial to any machine learning model. In the domain of facial landmark detection using regression techniques, Feng et al. [29] proposed a wing-shaped loss function to improve the accuracy of facial detection models by

increasing the impact of small to medium errors using a combination of linear and nonlinear parts, while Fard et al. [30] proposed an adaptive loss function using a difference between predicted landmarks and ground truth. In heatmap-based techniques, Yan et al. [31] calculated the difference between two probability distributions using Wasserstein distance to output its value. In the domain of image preprocessing, Zhao et al. [32] used a correction network in the image preprocessing stage to enhance the result of the detection network. It corrects an image that has been warped by a fisheye lens (e.g., the ones used with doorbell cameras for wider view angle). This is done by using two networks in sequence. Correction networks predict coefficients for their radial transformation equation, and alignment networks generate a projective transformation matrix.

### *5.2.5 Medical Domain*

Facial landmark detection is used widely in the medical field, especially in the field of plastic surgery, to analyze facial structures before suggesting treatment plans [33–35]. Freitas et al. [34] proposed a facial feature detection network extracting facial contour, contour simplification, and point localization from a side face profile image to be used for general facial plastic surgery (e.g., reconstructive nose surgery). AI/ML-based landmark prediction for landmark detection in cleft lip faces was suggested first by our collaborator Sayadi et al. [36] in a medical journal which is developed into a comprehensive method and system in this work.

## 5.3 Method

Our method seeks to detect 21 anthropometric cleft lip landmarks (CLL) that surgeons use to guide the surgery (see Fig. 5.1). These landmarks are located around the cleft and are unique to the cleft side, denoted by the prefix *c*', as well as the non-cleft side, denoted by the prefix *nc*'. Five of these 21 landmarks that are least subjective to the surgeon preferences are: (a) *prn*: the tip of the nose; (b) *c'ala, nc'ala*: the wings of the nostrils on the cleft and non-cleft side, respectively; and (c) *c'ch, nc'ch*: the junction of the upper and lower lips on the cleft and non-cleft side, respectively.

Our method uses a network that consists of four main components: (a) detection of landmark points in the face outline of the input image to cut out a region of interest (ROI) that focuses on the cleft lip deformity, (b) creating a rectangular input image from just the ROI via a warp-and-crop, (c) detection of cleft landmarks in the warped and cropped image, and finally, (d) inverse cropping and warping on the detected landmarks to find their location in the original image. Figure 5.3 shows the outline of our network.

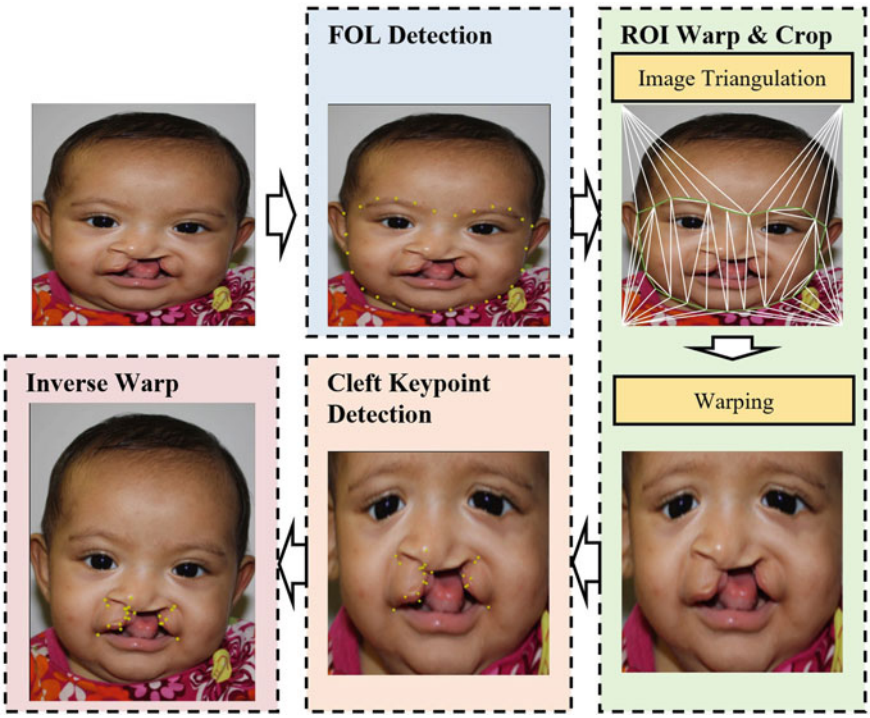
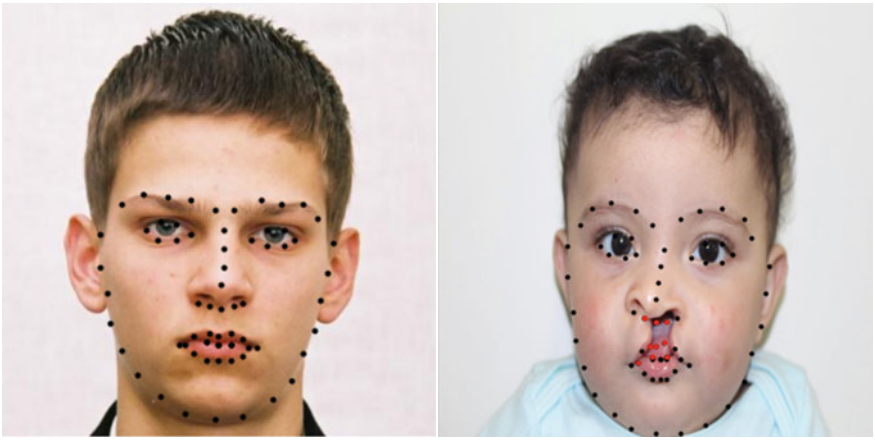



**Fig. 5.3** Flowchart of the proposed method. *FOL Detection*: the facial outline landmarks (FOLs) are shown in yellow. *ROI Warp-and-Crop*: The ROI, drawn in green, is triangulated and is used to warp-and-crop the image. *Cleft Keypoint Detection*: The cleft landmarks are then detected on the warped image. *Inverse Warp*: Finally, the keypoints are inverse-warped to determine their correct locations in the original image

### *5.3.1 Facial Outline Landmark (FOL) Detection*

In order to detect the face outline that focuses on the cleft deformity, we use 27 landmarks along the mandible and the eyebrows of the face. These facial outline landmark (FOL) points were computed by taking the average of all training images' reference points. The face was cropped and resized into a $1024 \times 1024$ image, which was then passed through HRNet [37] to detect the landmarks along the mandible and eyebrows for each image. For each of these FOL points, the mean coordinate was calculated after excluding the outliers that were unusually far from the mean.

We start by detecting FOL points since the presence of a cleft lip has almost no effect on the landmarks on the facial outline (see Fig. 5.4). The deformity affects the nasolabial region most significantly, and therefore any heatmap-based FOL point detection method can still accurately detect landmarks along the mandible and eyebrows for cleft faces. Heatmap-based methods work better for this purpose than regression-based methods since they deliberately focus on regions near the landmark



(a) normal face (b) cleft patient

**Fig. 5.4** Comparison of detected FOL points on a sample face from the 300 W dataset and cleft lip patient. Incorrectly detected landmarks with large errors on nasolabial region of cleft lip patients are marked in red

for prediction, allowing the FOL points with sufficient distance from the nasolabial region to be detected correctly.

We use HRNet [37] that has been trained over the 300 W dataset to detect the FOL of the cleft face. This dataset consists of images of faces with varying poses, expressions, skin tones, and lighting conditions and has been labeled with 68 facial landmarks along the mandible, eyes, nose, and mouth. In our work, we detect all 68 facial landmarks but only retain the 27 points comprising the facial outline. Figure 5.4 compares the detected FOL points for a normal face and a cleft lip face. Note how the points on the cleft are incorrectly detected. However, the outline is still detected accurately.

### *5.3.2 ROI Warp and Crop*

The detected FOL cut out a region of interest (ROI) in the image that focuses only on the cleft deformity. However, though we use front-face images, they may be somewhat different in scale and orientation. Therefore, we want to cast all images to a general template to make the subsequent detection of CLL points more robust. Therefore, we use piecewise affine transformation to warp the ROI to a template ROI. The template ROI is generated by taking the average of facial outlines of 50 images, following the concept presented by Felzenswalb et al. [38], where intended detection objects are given deformable templates in the form of triangular meshes. Piecewise affine transformation separates the image into triangular segments for warping, allowing the warping of images using multiple reference points as shown

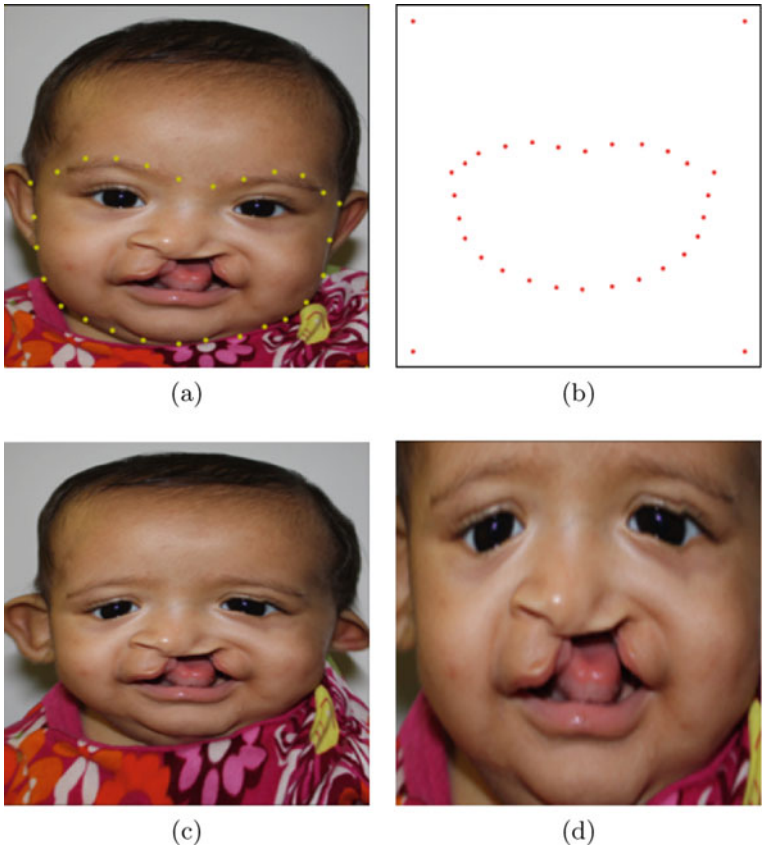



**Fig. 5.5** **a** Input image of a cleft patient with the FOL points (in yellow) carving out the ROI. **b** The template ROI is in red. **c** Warped image where the ROI of the input image is warped to the template ROI. **d** The warped image is cropped to retain only the minimal rectangle enclosing the warped ROI

in Fig. 5.5. As shown by Ye et al. [39], due to the use of multiple anchor points for warping, piecewise affine transformation results in a much more controlled transformation compared to a general error-minimizing nonlinear transformation. It also preserves the local shape of sub-regions in the area of cleft deformity. Finally, we crop the warped image to retain the minimal rectangle enclosing the warped ROI. Since our cleft lip dataset is small, and a large variation in the appearance occurs due to the severity and shape of the cleft as well as their scale and orientation, the previously mentioned warp-and-crop applied to each input image standardizes the appearance of the ROI for more accurate predictions. Figure 5.6 shows the triangulations used for warping for different images in greater detail.

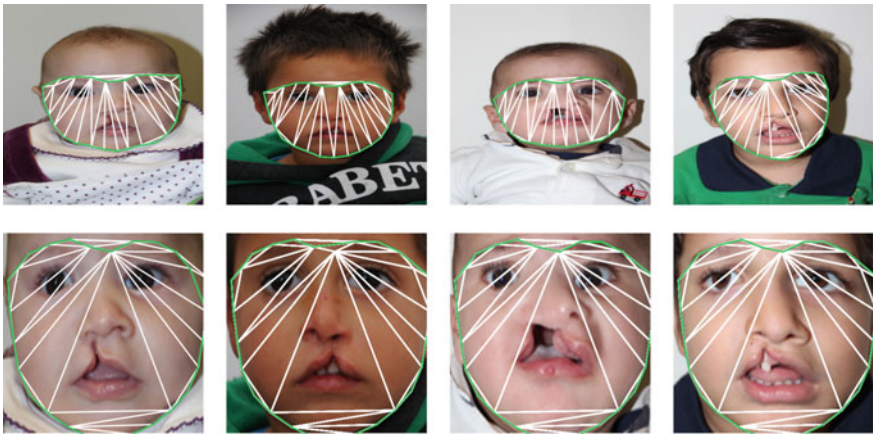


**Fig. 5.6** Triangulations used for piecewise affine transform for different kinds of cleft lip patients. This demonstrates the generality of our method on multiple images with different attributes (e.g., skin tone color, illumination, type of cleft)

### *5.3.3 Cleft Lip Landmark (CLL) Detection*

Following the warp, we proceed to detect the cleft lip landmarks (CLL) using a small dataset of 200 images. As noted in Sect. 5.2, although heatmap regression has proven to produce slightly better results in facial landmark detection, our dataset is not sufficiently large for the model to reach convergence. Therefore, we use ResNet-50 [40] with direct regression trained over the Wingloss [29] function.

ResNet [40] proposes a solution to the vanishing gradient problem by introducing a residual block, which merges previous input with outputs to prevent harmful layers from affecting the result. Wingloss [29] is a loss function used in network training in which a wing-shaped function is used to control the linearity of the training. Feng et al. [41] have shown Wingloss to significantly improve the accuracy of facial landmark detection models using direct regression.

### *5.3.4 Inverse Image Warp*

Finally, the predicted CLL points need to be inverse-warped to determine their locations in the original image domain. Figure 5.7 shows the results and compares the predictions with ground truth hand-marked by collaborating surgeons. We employ the same technique used for transforming ground truth to the warped image domain for creating an inverse warp.

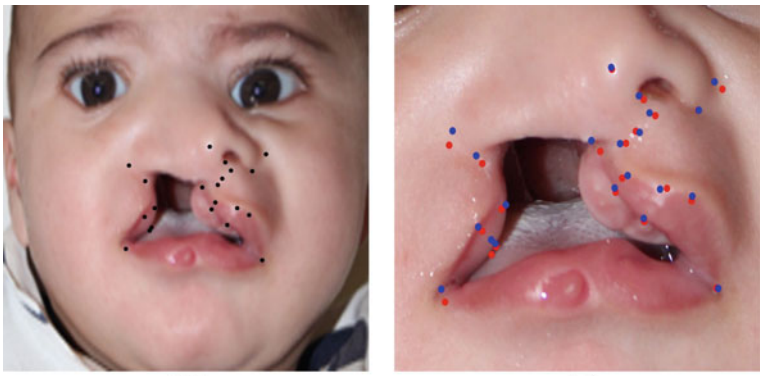



**Fig. 5.7** Detected landmarks on warped ROI (**a**). In a view zoomed in on the cleft lip deformity (**b**), we show the predictions in red and the ground truth in blue. Note the precision alignment of red and blue points

## 5.4 Implementation and Results

### *5.4.1 Dataset*

Our dataset consists of 500 unilateral cleft lip photos provided to us by the Global Smile Foundation and is approved by the Institutional Review Boards, a group that has been formally designated to review and monitor biomedical research involving human subjects. Of the 500 images, approximately 200 frontal images of patients with cleft lips were labeled with 21 keypoints that are used for cleft surgery techniques, such as one shown in Fig. 5.1. These images capture a frontal view of the entire face of patients and were carefully labeled by our surgeon collaborators who have deep experience in performing cleft lip repair. Almost all patients are less than 2 years which is the typical age range for cleft lip repair surgery. Excluding incompletely labeled images, we train our model on 150 images and evaluate our approach on the remaining 50.

**Table 5.1** Interocular NME for cleft lip dataset

| Method | Test | Full | 90% | 80% |
|---|---|---|---|---|
| Wingloss [29] | 2.84e − 2 | 2.81e − 2 | 2.38e − 2 | 2.12e − 2 |
| Wingloss-warped | **2.35e − 2** | **2.35e − 2** | **1.84e − 2** | **1.71e − 2** |
| [1]Sayadi et al. [36] | N/A | 3.87e − 2 | N/A | N/A |

Only full NME is provided by Sayadi et al. [36]

### 5.4.2 Evaluation Metric

In this work, we measure the accuracy of the detected keypoints by reporting the interocular normalized mean error (NME), $\epsilon_i$, computed as:

$$\epsilon_i = \frac{1}{V_i} \sum_{j \in K} v_{ij} \frac{d_{ij}}{L_i}, \tag{5.1}$$

where $K$ is the set of all 21 keypoints, $d_{ij}$ is the Euclidean distance (in pixels) between the $j$-th detected keypoint in image $i$ and its ground truth location, $v_{ij}$ is a binary variable that is 1 if the keypoint is visible in image $i$ and 0 otherwise, $L_i$ is the interocular distance, i.e., the normalized Euclidean distance between the centers of the pupils and $V_i = \sum_{j \in K} v_{ij}$, the number of unoccluded points in image $i$.

### 5.4.3 Training

We train two networks separately: one where the input images have been ROI-warped and one where the input images are not ROI-warped. Both networks are trained on 150 training images.

### 5.4.4 Results

The test dataset consists of 50 cleft images of various face shapes, skin tones, and severity of cleft lip to best analyze the accuracy of the networks with minimal bias. The full set is evaluated using all 200 marked images. Figure 5.8 shows our predicted points compared with ground truth for some of these images. Note that despite having a large variation in the cleft lip deformity, our method is able to predict the CLLs accurately. Table 5.1 lists the average NMEs for our dataset both with and without the warp-and-crop. Figure 5.9 compares our method to one that does not perform the warp-and-crop and also with those reported by Sayadi et al. [36]. Our network provides superior performance with warp-and-crop and has a higher NME without warp-and-crop. Additionally, both versions of our network perform better than Sayadi et al. [36] on the full dataset (see Table 5.1).

Table 5.2 shows the NME for all 21 keypoints detected by the proposed method with and without warp-and-crop and compares them to the results reported by Sayadi et al. [36]. Our algorithm accurately and precisely predicts the anthropometric landmarks well within the boundaries set forth by benchmarks [42, 43]. The landmarks *c'ala*, *nc'ala*, and *prn*, all on the nose, have the lowest NME, whereas *c'ch*, the

**Fig. 5.8** Cleft anthropometric landmark predictions. Original image (left), predictions on warped ROI (left-middle), predictions on original image (right-middle), zoomed in predictions on original image (right)

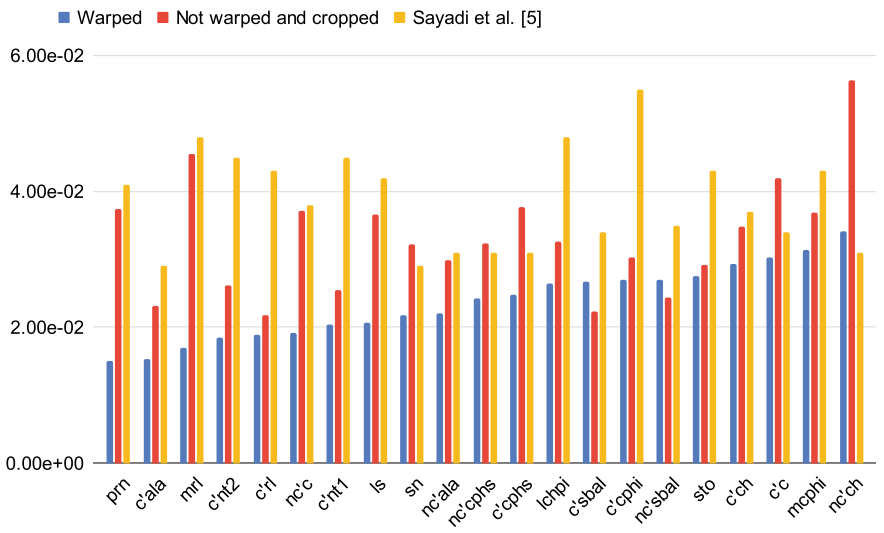



**Fig. 5.9** NME for all 21 cleft anthropological landmark predictions with warped Wingloss, without warp-and-crop Wingloss [29], and Sayadi et al. [36]

**Table 5.2** Interocular NME comparison for All 21 landmarks

| Name | Wingloss [29] | Warped Wingloss | [1] Sayadi et al. [36] |
|---|---|---|---|
| prn | 3.75e − 02 | **1.51e − 02** | 4.10e − 02 |
| c'ala | 2.32e − 02 | **1.53e − 02** | 2.90e − 02 |
| mrl | 4.55e − 02 | **1.70e − 02** | 4.80e − 02 |
| c'nt2 | 2.61e − 02 | **1.85e − 02** | 4.50e − 02 |
| c'rl | 2.18e − 02 | **1.89e − 02** | 4.30e − 02 |
| nc'c | 3.71e − 02 | **1.92e − 02** | 3.80e − 02 |
| c'nt1 | 2.55e − 02 | **2.04e − 02** | 4.50e − 02 |
| ls | 3.66e − 02 | **2.06e − 02** | 4.20e − 02 |
| sn | 3.22e − 02 | **2.18e − 02** | 2.90e − 02 |
| nc'ala | 2.99e − 02 | **2.21e − 02** | 3.10e − 02 |
| nc'cphs | 3.23e − 02 | **2.42e − 02** | 3.10e − 02 |
| c'cphs | 3.76e − 02 | **2.48e − 02** | 3.10e − 02 |
| lchpi | 3.26e − 02 | **2.64e − 02** | 4.80e − 02 |
| c'sbal | **2.24e − 02** | 2.67e − 02 | 3.40e − 02 |
| c'cphi | 3.03e − 02 | **2.70e − 02** | 5.50e − 02 |
| nc'sbal | **2.44e − 02** | 2.70e − 02 | 3.50e − 02 |
| sto | 2.92e − 02 | **2.75e − 02** | 4.30e − 02 |
| c'ch | 3.48e − 02 | **2.93e − 02** | 3.70e − 02 |
| c'c | 4.19e − 02 | **3.02e − 02** | 3.40e − 02 |
| nc'cphi | 3.69e − 02 | **3.14e − 02** | 4.30e − 02 |
| nc'ch | 5.64e − 02 | 3.41e − 02 | **3.10e − 02** |

[1] Sayadi et al. [36] values are approximated from provided graph

**Table 5.3** Computation time on network components

| FOL detection (ms) | ROI warp (ms) | Cleft keypoint detection (ms) |
|---|---|---|
| 358 | 661 | 950 |

Computation time calculated on 11,900 K with RTX 4090
Inverse warp not shown as computation takes less than 1 ms

cleft-side lip corner, was close to the median NME. *nc'ch*, the lip corner on the non-cleft side, showed the highest NME. Overall, our network, both with and without warp-and-crop, detects all keypoints, except *nc'ch*, with a lower NME than Sayadi et al. [36].

Table 5.3 notes the average computation time for the major components of the network: FOL detection, ROI warp, and cleft keypoint detection. Our machine used for computation was equipped with 11,900 K processor, RTX 4090 GPU, and 128 GB 3600 MHz DDR4. Note that inverse warp is omitted from Table 5.3 as it takes less than 1 ms.

### *5.4.5 Discussion and Limitations*

The results show that our network has a lower error for certain landmarks but a higher error for others. For example, *c'ch*, *c'ala*, and *nc'ala* have the lowest NME, whereas *c'ch* and *nc'ch* have the highest NME. *c'ala*, *nc'ala*, and *prn*, all on the nose, have the lowest NME as they are anchor points that are the least subjective to surgeon preferences. In comparison, *c'ch* and *nc'ch*, the corners of the mouth, have a larger region of acceptable marking, and therefore, it is harder for our model to pinpoint the exact location. In marking the anthropometric landmarks, *c'cphi*, *c'nt1*, *c'nt2*, and *c'rl* are most subjective to surgeon preferences. However, as all images were marked by Dr. Ross Sayadi, they did not show high NME in our study.

Benefiting from ROI warp, our network can accurately detect cleft landmarks on images at a camera angle between $-10°$ and $10°$. Any camera angle beyond this range results in incorrect detections, as our data does not contain enough example images to train the network. Despite thorough training, certain image features have been shown to cause performance drops and failure (see Fig. 5.10). Such cases are (i) facial occlusion by apparatus, e.g., due to breathing tubes, band-aids, etc., (ii) non-frontal views of the cleft, (iii) adult patients with cleft lip, and (iv) patients with eyes closed. We suspect the small size and variety of our dataset to be the root cause of most of the failure cases. As our dataset consists mainly of full frontal face view images of patients less than two years of age with eyes open, our network is more likely to fail on images not meeting these criteria.

The proposed method initially detects FOL for warp-and-crop. This means that the network would fail to detect keypoints in an image where the facial boundary is not visible. The trained model also shows significant performance drops in subnasal

(a) with apparatus (b) tilted face (c) adult face

**Fig. 5.10** Examples of failure cases. **a** FOLs could not be detected for images of patients with surgical apparatus, **b** images where the camera is oriented more than 10° from the frontal view, and **c** cleft lip deformity in adult faces

views of the cleft. This is because the dataset used for training does not have any subnasal images. However, as long as the angle is not severely skewed, the warp-and-crop step of the proposed method corrects the image enough for the detection to operate correctly.

## 5.5 Conclusion

In summary, we have presented the first method to accurately detect cleft lip landmarks using an AI/ML-based training method. In order to enhance the accuracy of this detection, we have used a warp-and-crop method that standardizes the image to counterbalance the facial deformations caused by the cleft lip condition. Not only does it improve the performance of detection, but it also allows an end-to-end detection of cleft landmarks. With the help of ResNet-50 and Wingloss function, we optimize the network for use in cleft lip surgeries. The empirical results show that the proposed method outperforms prior methods significantly. In the future, we would like to enhance this technique in the following ways.

- First, we would like our technique to be robust to different camera angles by expanding our training dataset to non-frontal views.
- In addition, we would like to use temporal coherence and GPUs to enhance the performance of the prediction to get it near real-time so that the predicted markings stick to the face in AR-based video surgical assistance sessions. Finally, we are also building a spatially augmented reality (SAR) system that

- Spatially augmented reality (SAR)-based systems exist today that use a projector and RGBD camera (e.g., Azure Kinect) to illuminate surgical guidance marks directly on the surgical site. A remote expert marks the landmark points or lines on the 3D model of the surgical site (captured by a structured light scan) using a GUI that shows up on-site in the surgical area in real-time [44]. We would like to extend our predictions to 3D models (instead of 2D images) to be used as a starting point on SAR systems. VR headset-based systems can also benefit from this.
- We would like to forge new directions by adapting the same technique for critical landmark detection for other surgeries as well.

**Acknowledgements** This work was supported in part by The American Society of Maxillofacial Surgeons. We thank Dr. Raj Vyas and Dr. Ross Sayadi for the large amount of time spent with us on numerous discussions, marking of cleft lip images for creating datasets, and helping us understand the key importance of the anthropometric landmarks. We thank the Global Smile Foundation for providing us with the valuable cleft lip datasets.

## References

1. Vyas, T., Gupta, P., Kumar, S., Gupta, R., Gupta, T., Singh, H.: Cleft of lip and palate: a review. J. Family Med. Primary Care **9**, 2621 (2020). https://doi.org/10.4103/jfmpc.jfmpc_472_20
2. Guerrero, C.: Cleft lip and palate surgery: 30 years follow-up. Ann. Maxillofacial Surg. **2**, 153–157 (2012). https://doi.org/10.4103/2231-0746.101342
3. Rossell Perry, P.: A 20-year experience in unilateral cleft lip repair: from millard to the triple unilimb z-plasty technique. Indian J. Plastic Surg. **49**, 340 (2016). https://doi.org/10.4103/0970-0358.197226
4. Sayadi, L., Chopan, M., Sayadi, J., Samai, A., Arora, J., Anand, S., Evans, G., Widgerow, A., Vyas, R.: Operating room stencil: a novel mobile application for surgical planning. Plastic Reconstr. Surg. Glob. Open **9**, e3807 (2021). https://doi.org/10.1097/GOX.0000000000003807
5. Vyas, R., Sayadi, L., Bendit, D., Hamdan, U.: Using virtual augmented reality to remotely proctor overseas surgical outreach: building long-term international capacity and sustainability. Plastic Reconstr. Surg. **146**, 622e–629e (2020). https://doi.org/10.1097/PRS.0000000000007293
6. Bulat, A., Tzimiropoulos, G.: Two-Stage Convolutional Part Heatmap Regression for the 1st 3D Face Alignment in the Wild (3Dfaw) Challenge, vol. 9914 (2016). https://doi.org/10.1007/978-3-319-48881-3_43
7. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at Boundary: A Boundary-Aware Face Alignment Algorithm, pp. 2129–2138 (2018). https://doi.org/10.1109/CVPR.2018.00227
8. Wu, Y., Hassner, T., Kim, K., Medioni, G., Natarajan, P.: Facial landmark detection with tweaked convolutional neural networks (2015). https://doi.org/10.1109/TPAMI.2017.2787130
9. Yang, J., Liu, Q., Zhang, K.: Stacked Hourglass Network for Robust Facial Landmark Localisation, pp. 2025–2033 (2017). https://doi.org/10.1109/CVPRW.2017.253
10. Zadeh, A., Lim, Y., Baltrusaitis, T., Morency, L.P.: Convolutional Experts Constrained Local Model for 3D Facial Landmark Detection, pp. 2519–2528 (2017). https://doi.org/10.1109/ICCVW.2017.296
11. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Trans. Pattern Anal. Mach. Intell. **38**, 1–1 (2015). https://doi.org/10.1109/TPAMI.2015.2469286

12. Kowalski, M., Naruniec, J., Trzcinski, T.: Deep Alignment Network: A Convolutional Neural Network for Robust Face Alignment, pp. 2034–2043 (2017). https://doi.org/10.1109/CVPRW.2017.254
13. Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A Deep Regression Architecture with Two-Stage Re-Initialization for High Performance Facial Landmark Detection, pp. 3691–3700 (2017). https://doi.org/10.1109/CVPR.2017.393
14. Ranjan, R., Patel, V., Chellappa, R.: Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2016). https://doi.org/10.1109/TPAMI.2017.2781233
15. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment, pp. 1–16 (2014). https://doi.org/10.1007/978-3-319-10605-2_1
16. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23** (2016). https://doi.org/10.1109/LSP.2016.2603342
17. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive Facial Landmark Localization With Coarse-to-Fine Convolutional Network Cascade, pp. 386–391 (2013). https://doi.org/10.1109/ICCVW.2013.58
18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. **79** (2014)
19. Bulat, A., Sanchez, E., Tzimiropoulos, G.: Subpixel heatmap regression for facial landmark localization. https://arxiv.org/abs/2111.02360 (2021)
20. Bulat, A., Tzimiropoulos, G.: Super-fan: integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans (2017). https://doi.org/10.1109/CVPR.2018.00019
21. Jackson, A., Valstar, M., Tzimiropoulos, G.: A cnn cascade for landmark guided semantic part segmentation (2016)
22. Lan, X., Hu, Q., Cheng, J.: HIH: towards more accurate face alignment via heatmap in heatmap. https://arxiv.org/abs/2104.03100 (2021)
23. Robinson, J.P., Li, Y., Zhang, N., Fu, Y., Tulyakov, S.: Laplace landmark localization. https://arxiv.org/abs/1903.11633 (2019)
24. Yin, S., Wang, S., Chen, X., Chen, E., Liang, C.: Attentive One-Dimensional Heatmap Regression for Facial Landmark Detection and Tracking, pp. 538–546 (2020). https://doi.org/10.1145/3394171.3413509
25. Wei, S., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. https://arxiv.org/abs/1602.00134 (2016)
26. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2013). https://doi.org/10.1109/CVPR.2014.81
27. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. 1 (2018). https://doi.org/10.1109/TPAMI.2018.2844175
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39** (2015). https://doi.org/10.1109/TPAMI.2016.2577031
29. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks (2018). https://doi.org/10.1109/CVPR.2018.00238
30. Pourramezan Fard, A., Mahoor, M.: ACR Loss: Adaptive Coordinate-Based Regression Loss for Face Alignment, pp. 1807–1814 (2022). https://doi.org/10.1109/ICPR56361.2022.9956683
31. Yan, Y., Duffner, S., Phutane, P., Berthelier, A., Blanc, C., Garcia, C., Chateau, T.: 2D Wasserstein loss for robust facial landmark detection. Pattern Recogn. **116** (2021). https://doi.org/10.1016/j.patcog.2021.107945

32. Zhao, H., Ying, X., Shi, Y., Tong, X., Wen, J., Zha, H.: Rdcface: Radial Distortion Correction for Face Recognition, pp. 7718–7727 (2020). https://doi.org/10.1109/CVPR42600.2020.00774
33. Chandaliya, P., Nain, N.: Plasticgan: holistic generative adversarial network on face plastic and aesthetic surgery. Multimedia Tools Appl. **81**, 1–22 (2022). https://doi.org/10.1007/s11042-022-12865-5
34. Freitas, R., Aires, K., Campelo, V.: Automatic location of facial landmarks for plastic surgery procedures. Conf. Proc. IEEE Int. Conf. Syst. Man Cybern. **2014**, 1444–1449 (2014). https://doi.org/10.1109/smc.2014.6974118
35. Freitas, R.T., Aires, K.R.T., Campelo, V.E.S.: Locating facial landmarks towards plastic surgery. In: 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, pp. 219–225 (2015). https://doi.org/10.1109/SIBGRAPI.2015.40
36. Sayadi, L., Hamdan, U., Zhangli, Q., Hu, J., Vyas, R.: Harnessing the power of artificial intelligence to teach cleft lip surgery. Plastic Reconstr. Surg. Glob. Open **10**, e4451 (2022). https://doi.org/10.1097/GOX.0000000000004451
37. Ke, S., Xiao, B., Liu, D., Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation, pp. 5686–5696 (2019). https://doi.org/10.1109/CVPR.2019.00584
38. Felzenszwalb, P.: Representation and detection of deformable shapes. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 208–20 (2005). https://doi.org/10.1109/TPAMI.2005.35
39. Ye, Y., Shan, J., Bruzzone, L., Shen, L.: Robust registration of multimodal remote sensing images based on structural similarity. IEEE Trans. Geosci. Remote Sens. 1–18 (2017). https://doi.org/10.1109/TGRS.2017.2656380
40. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
41. Feng, Z.H., Kittler, J., Awais, M., Wu, X.J.: Rectified wing loss for efficient and robust facial landmark localisation with convolutional neural networks. Int. J. Comput. Vis. **128** (2020). https://doi.org/10.1007/s11263-019-01275-0
42. Belhumeur, P., Jacobs, D., Kriegman, D., Kumar, N.: Localizing Parts of Faces Using a Consensus of Exemplars, pp. 545–552 (2011). https://doi.org/10.1109/CVPR.2011.5995602
43. Çeliktutan, O., Ulukaya, S., Sankur, B.: A comparative study of face landmarking techniques. EURASIP J. Image Video Process. **2013** (2013). https://doi.org/10.1186/1687-5281-2013-13
44. Ibrahim, M.T., Gopi, M., Vyas, R., Sayadi, L.R., Majumder, A.: Projector illuminated precise stencils on surgical sites. In: IEEE Conference on Virtual Reality and 3D User Interfaces (2023)

# Chapter 6
# The Influence of Eye-Height and Body Posture on Size Perception in Virtual Reality

**Ayumu Mitsuzumi and Saori Aida**

**Abstract** This paper investigates the impact of avatar eye-level variation on the perceived size of visual objects in virtual reality (VR) environments. The study utilized an HTC Vive Pro Eye head-mounted display (HMD) and observers assumed three postures: upright, supine, and prone. We investigated the relationship between eye-height and posture difference based on eye-height estimation using two environments, background without texture and background with texture. The experimental results showed that the results were in line with previous studies on the differences in eye-height. As in the previous study, the differences in the background without texture environment and background with texture were also influenced by perception, and it was confirmed that the background with texture environment was greatly influenced by the eye-height estimation. In terms of posture, upright posture was perceived as the largest at 70 and 120 cm, followed by the supine and prone posture, and upright posture was perceived as the smallest at 220 and 270 cm, followed by the supine and prone posture. This suggests that the upright posture is the most sensitive to eye-height information in size estimation, followed by the supine and prone posture.

## 6.1 Introduction

In recent years, the progress of virtual reality (VR) has been remarkable, garnering attention as a technology that provides an immersive experience. VR gained widespread recognition subsequent to Facebook's announcement on October 28, 2021, wherein they declared their rebranding as Meta and their focus on the development of the Metaverse. Furthermore, VR is increasingly being adopted not only for corporate or research purposes but also for home entertainment, with Meta's Oculus Quest2 and HTC's Vive series leading the way. The demand for VR is anticipated to

A. Mitsuzumi · S. Aida (✉)
Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Ube 755-8611, Japan
e-mail: saoaida@yamaguchi-u.ac.jp



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_6

escalate in the future due to the impact of Covid-19, which has necessitated a shift toward online meetings and activities, replacing traditional face-to-face interactions. Consequently, a myriad of VR games and social VR services are currently under development. In VR games, individuals assume the role of a character and engage in battles against adversaries, while in social VR services, users can interact closely with people from all corners of the globe, immersing themselves in an ideal avatar [1]. People can not only embody their own ideal persona but also easily assume a different gender. Furthermore, it is effortless to disregard physical constraints and adopt the behavior associated with another gender, which proves challenging in the real world. In addition to gaming and social services, VR finds applications in the realm of research. For instance, by establishing virtual laboratories, experiments can be conducted on observers located remotely [2]. In such scenarios, players assume the role of avatars and explore the virtual world from the avatar's perspective. Consequently, the eye-level often corresponds to the avatar's height. If the avatar is tall, the eye-level might exceed the actual player's eye-level, whereas a small avatar may result in a lower eye-level. Moreover, players may engage in gameplay while standing, sitting in a chair, or even in a reclined position. Such postural variations can lead to disparities in visibility when observing the same object within the game. The purpose of this study is to examine the impact of avatar eye-level variation on the apparent size of the visual object, with a particular emphasis on the avatar's eye-height and body posture within the virtual environment.

It is known that the observer's own eye-height is involved in the size perception of the visibility of the visual object. When engaging in object observation, one may adopt various postures such as standing or sitting, which in turn affects the eye-height. When the same object was observed in different postures and at varying eye-heights, it appeared larger than its actual size in an upright or seated position and smaller in a supine posture [3]. However, consistent perception was observed across different postures when the eye-height remained constant, suggesting that discrepancies in perception arise primarily from variations in eye-height rather than posture. Furthermore, manipulating floor texture information was found to influence size estimation when the eye-height is in proximity to the ground. Size estimation outcomes were similar across different postures as long as the eye-height remained unchanged.

When eye-height was manipulated within the virtual environment using a head-mounted display (HMD), the object was perceived to possess a larger size when observed with a lower eye-height compared to the actual eye-height [4]. This indicates that eye-height serves as the most prominent perceptual cue in size estimation.

Eye-height can also be manipulated by adjusting the height of the floor [5]. Decreasing the eye-height leads to an observer perceiving the object as larger. Furthermore, it has been confirmed that the observer's perception can be altered by manipulating the eye-height in the environment without directly manipulating the observer's eyes.

Perception within a virtual domain employing an HMD deviated from the perceptual experience encountered in the real world [6]. The perceptual impact of manipulating the eye-heights also differed between observing while projecting on a screen

and manipulating with an HMD [4]. Furthermore, irrespective of the specific HMD device employed, these effects were inclined to be perceived as smaller than reality [7].

Through the utilization of a virtual environment, an object can be observed under consistent conditions, regardless of the actual physical posture. However, a discrepancy arises between the observed posture in reality and within the virtual environment. In the upright, supine, and prone postures, observers displayed a tendency to perceive the object as smaller when in the supine and prone postures compared to the upright posture [8]. Furthermore, the object tended to be consistently perceived as smaller than its true size across all postures [8]. A discrepancy in perception exists between the real and virtual environments, with a correlation suggesting a propensity for smaller perceptual experiences during the lying posture in the real world.

In comparison to the upright posture, the supine and prone postures are inclined to elicit a perception of reduced size [7], while a lower eye-height tends to evoke a perception of increased size [5]. This study aims to examine the correlation between eye-height and posture, with a specific focus on the eye-height estimation [4], and investigates the effects of eye-height and posture on size perception in virtual environment in three postures: upright, supine, and prone.

## 6.2 Method

### *6.2.1 Stimuli and Apparatus*

We used HTC Vive Pro Eye HMD (1440 × 1600 pixel per monocular, 90 Hz refresh rate) and graphics displayed in the Vive were generated on a Windows 10 computer (CPU: Intel Core i9-10900 K 3.70 GHz, memory: 32.0 GB, GPU: NVIDIA GeForce GTX 3080, SSD: 500 GB). All observers used HMD and held HTC Vive controllers (2018) in their left and right hands for the experiment. The 3DCG software Blender (Ver. 3.3.1) from Blender Foundation and the game engine Unity from Unity Technologies were used to create the virtual environment for the experiment. While the creation of the virtual environment, experimental program was created in C# to perform the necessary controls to conduct the experiments. The program software Visual Studio Code (VScode), a code editor made by Microsoft, was used to make C# program script.

The experiment utilized Unity to construct the virtual environment, and observers assumed three real-life postures: upright, supine, and prone (see Fig. 6.1). Observers used a mini folding bed (OTB-MN) in the prone and supine posture conditions.

We used two virtual environments for the experiment: background without texture and background with texture. Background without texture: The gray floor, walls, and sky were also unified light gray (#808,080). Background with texture: A pattern consisting of two colors, gray (#424,345) and light gray (#808,080), was used on the floor and walls, and the sky was the same light gray color as the background without

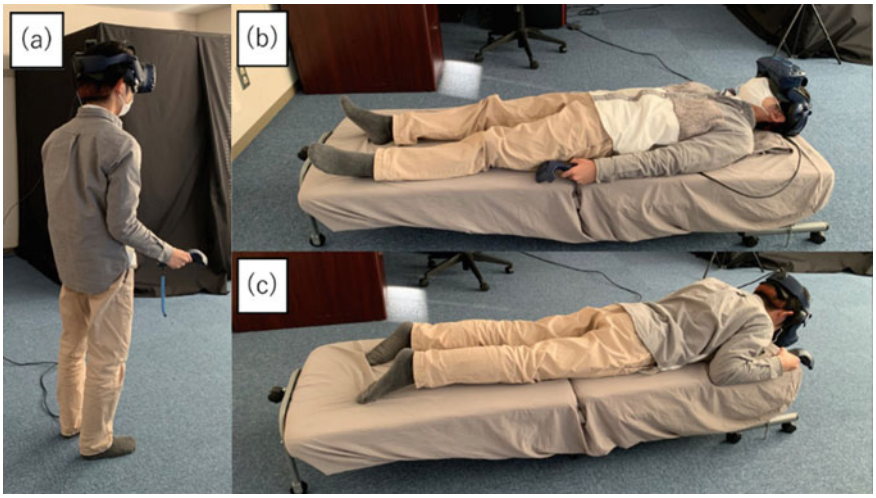



**Fig. 6.1** Experiment posture: **a** upright, **b** supine, **c** prone

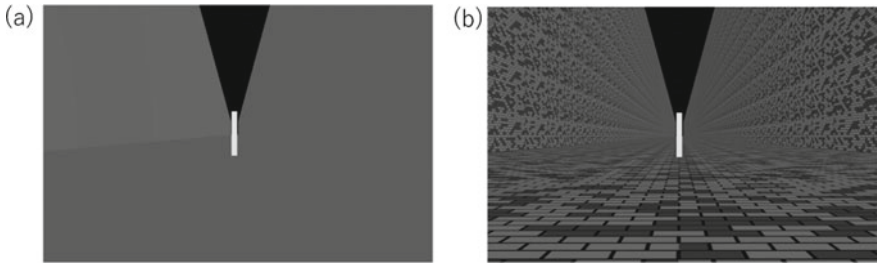



**Fig. 6.2** Virtual environment: **a** background without texture. **b** Background with texture

texture environment (see Fig. 6.2). A white stick (long: 10 cm, wide: 10 cm, height: 150 cm, color: #FFFFFF) was placed on the ground 10 m away from observer's position in virtual environment. Using this stick as a reference, we prepared six different sticks (120, 130, 140, 160, 170, 180 cm) in addition to one with the same height (150 cm) and the same length and wide (see Fig. 6.3).

### *6.2.2 Observers*

Eleven observers (11 male, ages 21–25, average height 169.4 cm, average eye-height 158.6 cm) participated in the experiment. All of observers had normal or corrected-to-normal vision. One observer was an author of this study, and the others were naive as to the purpose of the experiment. Informed consent was obtained from all observers

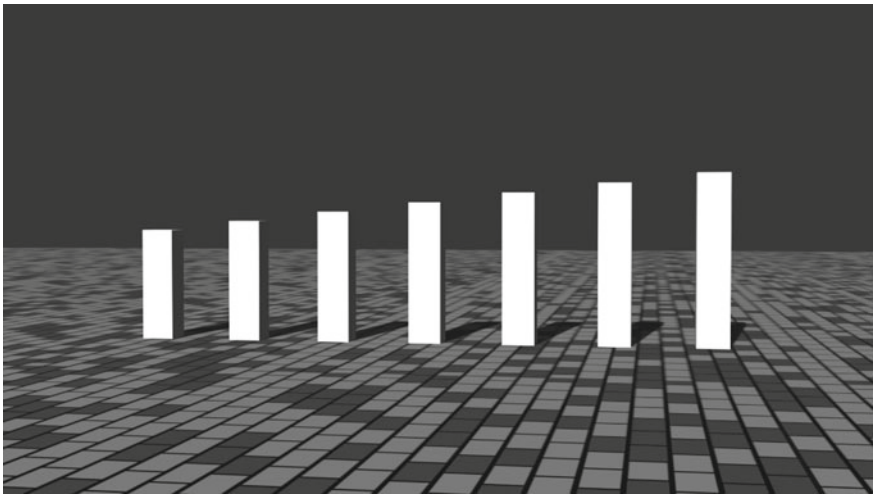


**Fig. 6.3** Stimuli (from left to right: 120, 130, 140, 150, 160, 170, 180 cm)

involved in the study. The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Yamaguchi University (protocol code 2022-003-01 and date of approval May 16, 2022).

### *6.2.3 Procedure*

Observers wore HTC Vive Pro Eye HMD on their head and held HTC Vive controller in each hand. Once observers were fitted with HMD, they participated in a virtual environment. Then, they received explanations and instructions about the experiment. After observers understood and consented to the experiment, the experiment began. After the experiment started, observers moved to a round marker at their feet and checked whether they could see the stimuli from there or not. After confirming that they could see the stimuli sufficiently, observers answered the questions at their own timing. The questions were in the form of two stimuli. Observers answered which stimuli were perceived larger. Two seconds after the first stimuli were presented, the second stimuli replaced on the same position. And after three seconds, the stimuli disappeared and a sound was heard, indicating the observers answer time. Observers were given an immediate break if they requested it, and observers were also given a break when postures or when their posture changed.

The experiment consisted of six sessions. In each session, there were six combinations of three types of body postures (i.e., upright, supine, and prone), and two types of environment (i.e., without texture and with texture). In each session, there were four blocks, in which the height on standard stimuli consisted of 70,120, 220, and 270 cm; the presentation order of the four blocks differed among observers. In

each block, the height on the comparison was randomly selected from seven different heights, with ten repetitions. Consequently, each observer was presented with stimuli for a total of 1680 times (3 body postures × 2 environment × 4 standard height × 7 height on the comparison × 10 repetitions).

### *6.2.4 Data Analysis*

In this study, the constancy method was employed and a sigmoid function (Eq. 6.1) was applied to the answers. (The value was set to "0" if the comparison stimuli were small and "1" if it was large.)

$$p = \frac{a}{1 + e^{-k(x-x0)}} + c \tag{6.1}$$

$p$ is the probability, $x$ is the probability that the observer judged the stimuli to be larger for comparison, and $a$, $x0$, and $c$ are parameters for adjusting the graph depiction.

The results of the observer's answers for each posture were analyzed. Each data was curve-fitted to a sigmoid function to determine the difference between the height of the bar and the median (150 cm) when the response was 50%, and the mean value was obtained. Figure 6.4 shows the results of curve fitting to the sigmoid function for observer 1 in the environment with background with texture at 120–170 cm condition.

## 6.3 Results

The results of the observer's answers for each posture were analyzed. Each data was curve-fitted to a sigmoid function to determine the difference between the height of the bar and the median (150 cm) when the response was 50% and the mean value was obtained. The relationship between no background with texture and with background with texture in the upright posture is shown in Fig. 6.5.

The vertical axis of the Fig. 6.5 shows the mean value calculated from the difference between the observer's bar height at 50% (horizontal axis) and the median (150 cm), and the horizontal axis shows the eye-height for each comparison (average of 50% difference). Orange is the background without texture, and blue is the background with texture. N represents the number of observers. Error bars indicate 95% confidence intervals. The 95% confidence interval is used to evaluate significant differences between the two environments, and the 95% confidence interval is also used to evaluate significant differences between the two environments and a median difference line was 0.

We conducted a three-way repeated measures ANOVA (4 eye-height × 3 body posture × 2 environment) on the average of 50% difference. The results showed that

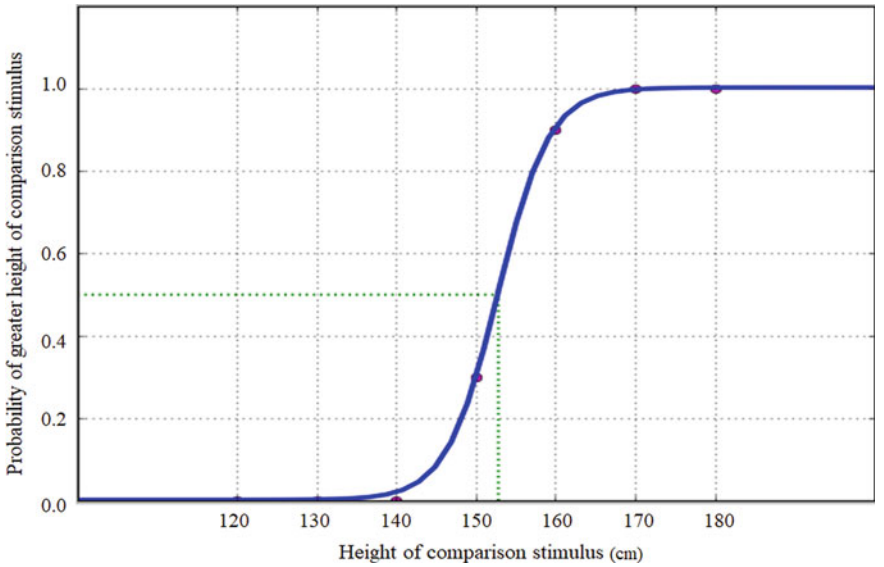



**Fig. 6.4** Curve fitting results to the sigmoid function for observer 1 in the environment with background with texture at 120–170 cm condition

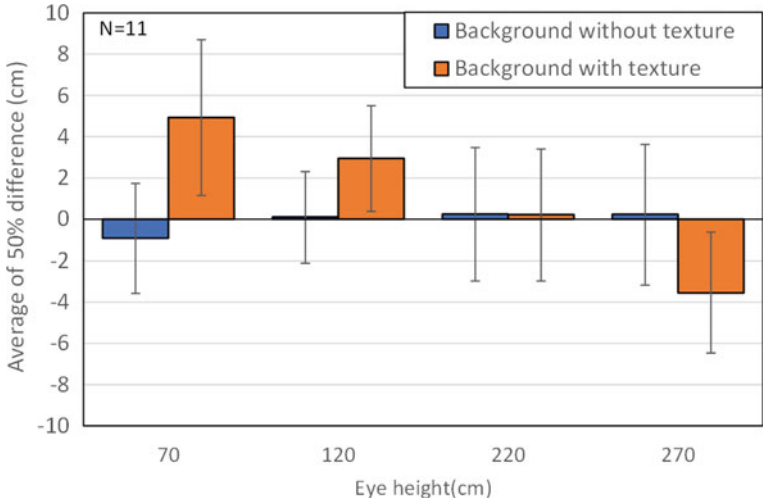



**Fig. 6.5** Relationship between background without texture and with background with texture environment in upright posture. Error bars are 95% confidence intervals

the main effect of environmental conditions [$F(1, 9) = 4.711$, $p < 0.01$] was significant. The main effects of eye-height [$F(3, 27) = 2.024$, $0.1 < p < 0.5$] and posture [$F(2, 18) = 0.821$, $0.1 < p < 0.5$] each had a significant difference trends. There was also a significant interaction between eye-height and environmental conditions [$F(3, 27) = 12.754$, $p < 0.001$]. There was a significant tread interaction between eye-height and posture [$F(6, 54) = 1.292$, $0.1 < p < 0.5$]. There was no significant difference between posture and environmental condition [$F(2, 18) = 0.506$, $p > 0.5$]. The subtest showed that there was a significant difference in the main effect [$F(3, 57) = 8.209$, $p < 0.001$] if the environmental condition on the condition with background with texture.

Figure 6.5 shows that the difference of eye-height effects in the background without texture environment is smaller than that in the background with texture environment, this indication that the eye-height information is not used very much in the background without texture environment. However, despite the lack of use of eye-height information, there was a slight tendency for the perception to be small at low eye-height and large at high eye-height; however, we could not confirm a significant difference by eye-height because the median difference line was 0 and the 95% confidence interval was crossed. In the background without texture environment, the perception tended to be larger at lower eye-heights. Significant differences were confirmed when the median difference line was 0 and the 95% confidence interval was not crossed when the eye-height was 70, 120, and 270 cm, thus indicating that there was an effect due to eye-height. There was a significant trend in the main effect of eye-height [$F(3, 27) = 2.024$, $0.1 < p < 0.5$]. The result showed that the perception of larger than actual size when eye-height was low and smaller than actual size when eye-height was high. This shows difference in perception from eye-height [3], which is supported by the results of this study. The difference in perception by the environment was confirmed by ANOVA as described above. However, the 95% confidence interval did not confirm a significant difference between the background without texture and the background with texture environment.

Figures 6.6 and 6.7 show the relationship between background without texture environment and background with texture environment in the supine and prone posture. Figures 6.6 and 6.7 show that in both the supine and prone postures, as in the upright posture, the difference eye-height in the background without texture environment was almost no different from the upright posture. This indicates that there was no effect from the eye-height information. Since the 95% confidence interval crossed 0 for all postures except for the 270 cm eye-level in the prone posture, we could not confirm the effect of eye-height in the background without texture environment. The vertical axis of the Figs. 6.6 and 6.7 shows the mean value calculated from the difference between the observer's bar height at 50% (horizontal axis) and the median (150 cm), and the horizontal axis shows the eye-height for each comparison (Average of 50% difference). Orange is the background without texture, and blue is the background with texture.

Figure 6.7 shows that in the background with texture environment, as in the upright posture, the perception tended to be larger when the eye-height was lower and smaller when the eye-height was higher. Figures 6.6 and 6.7 show that in the

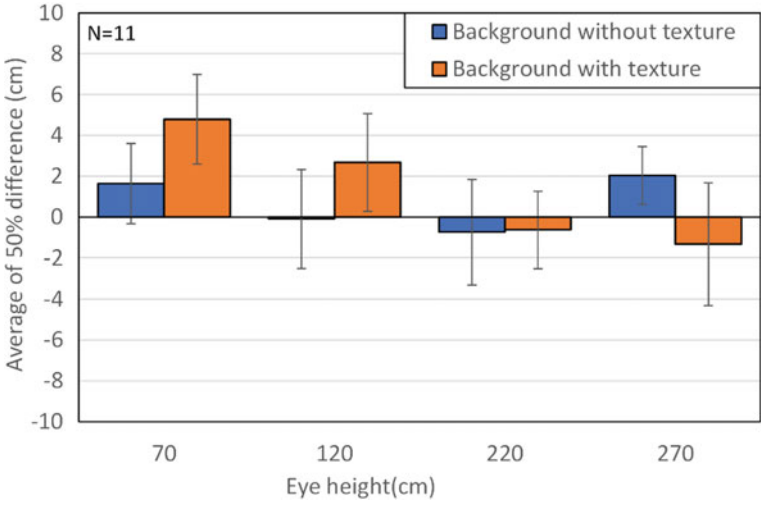



**Fig. 6.6** Relationship between background without texture and with background with texture environment in supine posture. Error bars are 95% confidence intervals

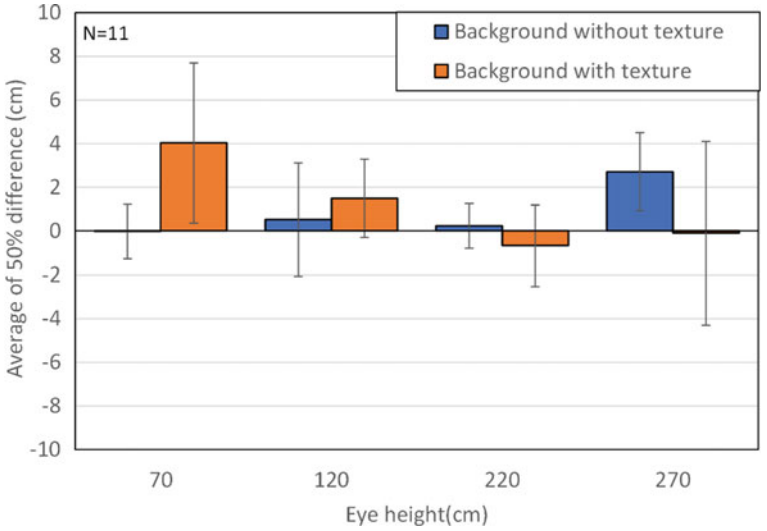



**Fig. 6.7** Relationship between background without texture and background with texture environment in prone posture. Error bars are 95% confidence intervals

background with texture environment the 70 and 120 cm eye-height in the supine posture, 70 cm eye-height in prone posture, the median difference line was 0 and the 95% confidence interval was not crossed, thus significant differences were confirmed. However, the 95% confidence intervals for the 220 and 270 cm in both postures were

not significant because the difference between the median and the median crossed the 0 line. The main effect of eye level [$F(3, 27) = 2.024$, $0.1 < p < 0.5$] had a significant difference trend, supporting the result of Wraga [3] that differences in perception are generated from eye-height rather than differences in posture in the supine and prone posture. However, the 95% confidence interval did not confirm a significant difference between the environment with background without texture and background with texture environment, and there was no significant difference between the posture and the environment [$F(2, 18) = 0.506$, $p > 0.5$].

Next, Figs. 6.8 and 6.9 show the relationship between the three postures of upright, supine, and prone in the background without texture environment and the three postures of upright, supine, and prone in the environment with background with texture environment. The vertical axis of the Figs. 6.8 and 6.9 shows the mean value calculated from the difference between the observer's bar height at 50% (horizontal axis) and the median (150 cm), and the horizontal axis shows the eye-height for each comparison (Average of 50% difference). Orange is the upright posture, gray is supine posture, and blue is prone posture.

The results of analysis of variance by ANOVA showed that the main effect of posture [$F(2, 18) = 0.821, 0.1 < p < 0.5$] had a significant difference trend. Figures 6.8. and 6.9. shows that in the background without texture environment, when the eye-height was 120 and 220 cm, the perception was relatively accurate in any postures; however, when the eye-height was 70 and 270 cm, a variety of perceptions appeared depending on the posture, the perception tended to be small when the eye-height was low and large when the eye-height was high, while in the supine posture, the

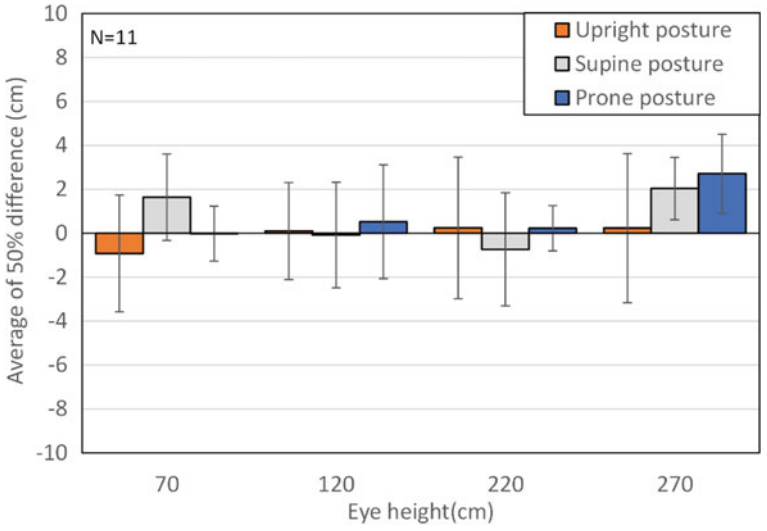



**Fig. 6.8** Relationship between the three postures of upright, supine, and prone in the background without texture. Error bars are 95% confidence intervals

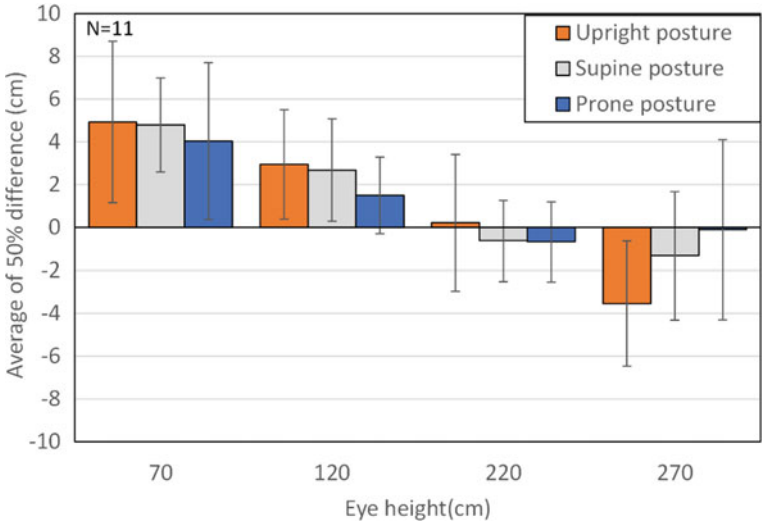



**Fig. 6.9** Relationship between the three postures of upright, supine, and prone in the background with texture. Error bars are 95% confidence intervals

perception tended to be small when the eye-height was low and large when the eye-height was high except when the eye-height was 270 cm. When we checked the 95% confidence interval and the difference between the median and 0 line, we could not confirm significant differences in all of the upright and other postures except eye-height was 270 cm in supine and prone posture. This indicates that we were unable to confirm the effect of eye-height on perception in the background without texture environment.

In the background with texture environment, the perception tended to be larger when the eye-height was low and smaller when the eye-height was high, regardless of the posture. The effect of eye-height on perception was most pronounced in the upright position, followed by supine, and then prone posture. At 70 and 120 cm, the 95% confidence interval tended to be greater than 0 for any posture, and at 220 and 270 cm, the 95% confidence interval tended to be less than 0 for any posture. Thus, it was confirmed that eye-height was affected in the background with texture environment in all postures. The results of ANOVA analysis of variance showed that there was a significant difference trend for eye-height and posture [$F(6, 54) = 1.292$, $0.1 < p < 0.5$]. Therefore, we use eye-height information in the size estimation, and we found that its effect is greatest in the upright and decreases as move to the supine and prone posture.

## 6.4 Discussion

This study involved the construction of an experimental virtual environment using Unity to explore the correlation between eye-height and body posture. Three postures, namely upright, supine, and prone, were examined. The perception of four different eye-heights (70, 120, 220, and 270 cm) was investigated under two environmental conditions: a background without texture and a background with texture. The experimental results were subjected to analysis of variance, revealing significant differences in the main effect of the environment. Additionally, the subtests provided further evidence of significant differences in the main effect under the condition with a textured background. These findings are consistent with the results reported by Wraga [3].

Figures 6.5, 6.6 and 6.7 accurately replicated the observed differences in perception based on eye-height, as reported by Dixon et al. [4]. In size perception using eye-height estimation with virtual environment, it has also been confirmed that when the eye-height is low, the perception is large [4], and in this study in the background with texture environment, no matter which posture was used, when the eye-height was low, the perception was large, and in this study, in the background with texture environment, the perception tended to be larger when the eye-height was low and smaller when the eye-height was high, regardless of the postural. This indicates that eye-height information is used for size estimation in all postures. In the background without texture environment, there was no difference in perception depending on eye-height as in the background with texture environment, and no significant difference could be confirmed from the 95% confidence interval. In the background without texture environment, cues for size estimation were removed as much as possible, making eye-height estimation more difficult, which may have resulted in the perception being less affected by the height of the eyes.

Figures 6.8 and 6.9 show that in the background without texture environment, the perception was relatively accurate at 120 and 220 cm eye-height in all postures, but at 70 and 270 cm eye-height, the perception varied depending on the posture. The fact that differences in perception were observed indicates that the observers are not accustomed to observing from a height higher than their own eye-height unless they stretch or jump [9] which is partially supported by the results. We also suspect that observers were not accustomed to estimating size in the absence of perceptual cues, which may have resulted in a variety of perceptions depending on the posture in the case of eye-height that was more distant from the normal eye-height. The fact that the 95% confidence interval did not confirm a significant difference suggests that there may be other factors besides the effect from eye-height.

In the background with texture environment, for all postures, Fig. 6.9 showed that the perception tended to be larger when eye-height was low and smaller when the eye-height was high. The results support the findings of [4]. However, the three postures (upright, supine, and prone) were not perceived the same, with the upright posture being perceived as the largest when the eye-height was 70 cm or 120 cm, followed by the supine and the prone posture, and when the eye-height was 220 cm

or 270 cm, the upright posture was perceived as the smallest, followed by the supine, and prone posture was perceived as the largest when the eye-height was 220 cm or 270 cm. The results support the findings of [7]. The ANOVA analysis showed that there was a significant difference trend between eye-height and posture [$F(6, 54) = 1.292$, $0.1 < p < 0.5$]. This shows that there is a relationship between eye-height and posture in the size estimation. Considering Figs. 6.8 and 6.9 and the ANOVA results, it is suggested that the upright posture may be more sensitive to eye-height information in size estimation than other postures.

During the experiment, several observers commented that they felt as if their own bodies were buried in the ground in the virtual environment. It is possible that the observers do not recognize the eye-height in the virtual environment and are unconsciously using the eye-height their own body in the real world. Therefore, it is possible that the eye-height used in the real world is also used in the virtual environment. In this study, size estimation was measured in virtual environment but size estimation by the actual eye-height of the observer was not measured. Since size estimation is usually done from the actual eye-height in real world in size estimation. There is a possibility that some estimation from the eye-height in the real world is made even if the experiment is conducted using virtual environment. The sample size was relatively small, and all participants were male and within a narrow age range. These may limit the generalizability of our results.

## 6.5 Conclusion

In this study, we investigated the relationship between eye-height and posture difference based on eye-height estimation using two environments, background without texture and background with texture. The experimental results showed that the results were in line with previous studies on the differences in eye-height. As in the previous study, the differences in the background without texture environment and background with texture were also influenced by perception, and it was confirmed that the background with texture environment was greatly influenced by the eye-height estimation. In terms of posture, upright posture was perceived as the largest at 70 and 120 cm, followed by the supine and prone posture, and upright posture was perceived as the smallest at 220 and 270 cm, followed by the supine and prone posture. This suggests that the upright posture is the most sensitive to eye-height information in size estimation, followed by the supine and prone posture.

**Acknowledgements** This research was supported by JSPS KAKENHI (Grant-in-Aid for Early-Career Scientists), grant number 21K18027.

## References

1. Virtual Girl Nem: Metaverse Evolution Theory, Gijutsu-Hyoron Co., Ltd., Tokyo (2022)
2. Ishimoto, K., Kato, Y., Kitazaki, M.: Development and validation of experimental environment for body ownership research in virtual-reality social network service VRChat. Jpn. J. Psychonomic Sci. **40**(2), 121–134 (2022)
3. Wraga, M.: Using eye height in different postures to scale the heights of objects. J. Exp. Psychol. Hum. Percept. Perform. **25**(2), 518 (1999)
4. Dixon, M.W., Wraga, M., Proffitt, D.R., Williams, G.C.: Eye height scaling of absolute size in immersive and nonimmersive displays. J. Exp. Psychol. Hum. Percept. Perform. **26**(2), 582 (2000)
5. Wraga, M.: The role of eye height in perceiving affordances and object dimensions. Percept. Psychophys.Psychophys. **61**(3), 490–507 (1999)
6. Murakami, M., Yokoda, H., Yamada, S.: Virtual reality design based on comparative verification of perception and impression for height in reality and virtual reality space—a research on space design for virtual reality architectural theory. In: Proceeding of the 42nd Symposium on Computer Technology of Information, 36–39 (2019)
7. Kelly, J., Doty, T., Ambourn, M., Cherep, L.: Distance perception in the oculus quest and oculus quest 2. Front. Virtual Reality **3**, 850471 (2022)
8. Kelly, J.W., Cherep, L.A., Siegel, Z.D.: Perceived space in the HTC vive. ACM Trans. Appl. Perception (TAP) **15**(1), 2, 1–16 (2017)
9. Leyrer, M., Linkenauger, S.A., Bülthoff, H.H., Kloos, U., Mohler, B.: The influence of eye height and avatars on egocentric distance estimates in immersive virtual environments. In: Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization, 67–74 (2011)

# Chapter 7
# The IVE-IEQ Model: A Conceptual Framework for Immersive IEQ Learning

**Fatin Nursyafiqah Khairul Anuar, Raha Sulaiman, Nazli Bin Che Din, and Asrul Sani Razak**

**Abstract** Immersive virtual environments (IVE) have garnered considerable attention in various educational contexts due to their proven efficacy in facilitating enhanced learning performance. However, there is indeed a lack of IVE technology integration in indoor environmental quality (IEQ) education and related fields. To demonstrate such complex multisensory phenomena of IEQ in a learning module, students require a comprehensive depiction of the sensory information. However, the current teaching and learning strategies are inadequate in providing such a detailed representation. In the view of ongoing research, this study sought to construct a conceptual framework for the implementation of IVE within IEQ learning to promote learning experiences in higher education by highlighting the fundamental experience, key elements, and contributing factors of IVE in developing a conceptual framework for multisensory phenomena of IEQ. This literature studies seek to contribute to the educators and instructional designers a comprehensive framework for incorporating IVE technology into higher education curriculums for IEQ learning. The developed framework could offer insights into prospective research areas for the development and refinement of IVE technology with the aim of enhancing the learning outcome within the context of IEQ education. Ultimately, this research paper demonstrates the potential of IVE technology to revolutionise IEQ teaching and provides a road map for its effective adoption in higher education.

## 7.1 Introduction

As the educational capabilities of immersive learning in virtual environment have been widely perceived on a global scale, educators from a variety of fields are considering the educational potential of immersive learning environments for their student's academic pursuits. This advanced educational approach persisted with

F. N. K. Anuar · R. Sulaiman (✉) · N. B. C. Din · A. S. Razak
The Centre for Building, Construction and Tropical Architecture (BuCTA), Faculty of Built Environment, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: rahasulaiman@um.edu.my



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_7

educators attempting to implement better experiential teaching and learning environments which enable students to access knowledge, engage with it, and implement it according to their individual preferences [1]. Chen [2] claimed that virtual learning environments (VLE) are beneficial in three forms: (a) the execution and versatility of distance and online learning; (b) the efficiency of knowledge transfer; as well as (c) the enrichment of learning by action. Immersive learning environments enabled students to facilitate the learning process through active involvement, observation, and collaboration with other users in the immersive medium, thus further enabling learning more enjoyable than conventional teaching methods.

Moreover, the digitised imaging settings in teaching and learning activities nowadays do not facilitate students to perceive the comprehensive phenomenon of indoor environmental quality (IEQ), as they are currently practising face-to-face teaching method, presentation, examination, test, and discussion which remain in the classroom. Thus, these students are unable to convey the scientific data on the impact on building users' health, comfort, and productivity as they correlate with people's sense of perception and multisensory experiences. Reviews by Radianti [3] concluded that the majority of IEQ research concentrates on usability testing, and there are very few studies analysing student comprehension in using VR as an immersive learning method in a particular subject.

There are currently a substantial number of comprehensive reviews VR applications in various academic domains such as healthcare [4, 5], entertainment [6], marketing [7], and education [8, 9]. Studies have proved conclusively that immersive learning technologies improve academic performance, learning engagement, and interest, which promote theoretical, innovative, conceptual, and functional methods of learning [10, 11]. However, the execution of immersive learning in Malaysian curriculum is still experimental and has not been thoroughly adopted or oriented on frameworks [12]. It was determined that virtual reality as an immersive educational instrument in Malaysian education lacks a well-defined framework due to the fact that it is not utilised effectively in actual teaching methods.

Building science educational module primarily covers the study of indoor environmental quality (IEQ) sensory measures, such as thermal comfort, acoustic, and lighting; human science (physiology and psychology); physical sciences (enclosure design and performance); interior components (finishing, colour, composition, and so forth.); energy and building materials. In addition, the typical interpretation of IEQ parameters is centred on two-dimensional (2D) data, often numerical data and complex graphs. Ultimately, by integrating the IEQ phenomenon with the advanced multisensory learning mechanism, these immersive rendering and simulation tools could establish skills transfer from IVE where it can efficiently deliver the scientific data on building users' health and well-being.

Therefore, in this study, immersive virtual environment (IVE) is defined as a virtual environment produced using computer hardware and software. It conveys an authentic environment to the user so that they can experience and engage with VR devices. It enables the creation of a fully immersive multisensory experience of indoor environmental quality (IEQ) with the appropriate level of immersion. Based

on this ongoing research, it is essential to identify the fundamental experience, characteristics, and implementation of these educational advances in constructing the IVE-IEQ conceptual framework.

## 7.2 Fundamental Experience of Immersive Learning

Immersive learning (IL) is said to be a constant development and adaptation of cognitive, affective, and sensorimotor models that develop whenever a user engages with technological interfaces [13]. Immersive learning enables a much more engagement than traditional learning content, as students feel far more immersed while engaging with learning material and learning activities by efficiently leveraging the immersive virtual environment features [14, 15]. The factors and fundamental experiences that contribute to immersive learning experience as an educational tool are highlighted in the following section.

### *7.2.1 Key Elements of IVE*

Virtual reality (VR), multi-user virtual environment (MUVE), and mixed reality (MR) are three main types of digital technology applications that underpin an increasing growing amount of structured and unstructured immersive learning approaches. Therefore, the fundamental experience that VR conveys in an immersive virtual environment consists of the user's internal assessments as a response to the deployment of the technology. Immersion and presence are the two key elements of virtual environment [14], while illusion, flow, situated cognition, and psychological ownership are recognised as the independent element in delivering immersion and presence [16, 17]. Below are the key element of this technology and must be carefully considered when designing and implementing VR experiences for various applications:

**Immersion**. Immersion is a key element that establishes the cognitive response of individuals to immersive technology engagement. Generally, one school of thought affirms that immersion is typically characterised as a technological attribute which could be evaluated objective manner, namely the desktop displays, software, and hardware that are able to generate an all-encompassing, substantial surrounding, and vivid artificial environment [18]. Others define immersion as a psychological condition in which the person experiences a perception of isolation from the real world [19]. A notion that is closely correlated with the degree of immersion delivered by IVE would be ecological validity, where it is determined by the IVE's capacity to replicate a realistic environment in which the environment is intended to represent the actual event [20].

**Presence and Illusion**. Witmer et al. [19] suggest that presence and immersion are correlated but not equivalent. In contrast, presence is predominantly associated with the sensation of being in a different location or environment—a disengagement from reality as well as a sensory connection to a different dimension [21]. Moreover, several experts even suggested that presence is characterised as an illusion of actually existing within that virtual environment regardless of the fact that you are aware of actual surrounding. It is just a visual illusion but not a cognitive one [16]. Consequently, presence is interpreted into two distinct notions: Place Illusion (PI) and Plausibility Illusion (Psi). PI pertains to the perception of being physically present within that virtual environment [22], where it is the initial impression of witnessing an idealised event or scene through a VR device, whereas Plausibility Illusion (Psi) involves the impression that the participant's interactions within virtual reality (VR) are believable and realistic despite of being aware of the actual surroundings. Psi demands that the simulated virtual environment responds to the participant's movement spontaneously. Ecologically, it is relevant when the virtual environment is designed to represent real-world occurrences. Participants are more prone to engage authentically in a virtual space if both PI and Psi are fully functional, thus having much further impact and benefits [22].

**Situated Cognition and Psychological Ownership**. Another aspect that can be generated in a virtual environment is body ownership which is also referred to as psychological ownership. From a first-person viewpoint, the participant perceives a life-sized virtual avatar in place of their own. The participant's original body motions can be synchronised with the designated virtual body or an avatar, establishing the illusion that the virtual body as their own [23]. Thus, participants can cognitively integrate virtual content, offering them the impression of being in a real settings; thus, increasing learning outcomes in an immersive environment [24].

### *7.2.2 Sensory Stimuli and Its System Drivers*

Immersion correlates to a realistic experience that enables users to visualise their own appearance via immersive technology, thereby enabling them to sense and manipulate various components in virtual space, as well as actively interact in various situations and conditions. Consequently, one strategy to promote the sense of presence is to enhance immersion [25]. Immersion is primarily generated by the human perceptual and behavioural systems. Perception system in the realm of human sensory information comprises sight, hearing, touch, smell, and taste, among many others, while behavioural system comprises body position, orientation, mobility, and spatial awareness [17, 26].

As aforementioned, immersion was considered the idea of illusion in which system technologies serve as the drivers [19, 22]. Thus, immersion is acquired by facilitating the participant with virtual devices and systems [26], such as a high-resolution virtual reality headset with real-time motion capture, gloves or controllers with motion

sensors, sound system, as well as any devices that generate and stimulate the sensorial sensory inputs, or sensor systems that enable users to engage with a simulated environment like it was a real environment. In this manner, these systems can be further refined and improved according to the desired degree towards which one system could be used to emulate others, which would then deepen the level of immersion [16].

### 7.2.3 Users Response in IVE

**Immersive Learning Outcome**. In IVE, user response refers to a direct consequence of utilising immersive technology. According to professionals in the field of education, the use of immersive technology encourages active participation, engagement, resulting in content understanding, overall academic performance, efficiency, capabilities and competency [27–29].

Furthermore, in the context of architectural and building science, the application of immersive technology in BIM-VR-based learning has indeed been established in terms of usability, immersion which results in learning outcomes, thus, demonstrating that it can facilitate students' engagement through its high real-time visualisation capacities [28, 30]. Chavez and Bayona's review [31] highlights that virtual reality's collaborative and interactive capability empowers individuals to actively interact with digital representations in a virtual environment. This interaction facilitates improved learning outcomes for students, allowing them to experience a virtual world closely resembling reality. Furthermore, virtual reality fosters interest, motivation, and increased enthusiasm for learning.

### 7.2.4 IVE Application in Higher Education

Considering the IVE characteristic, there are three major methods of immersive learning in architectural and building science education: simulation, visualisation, and exploration. Table 7.1 summarises relevant papers on the main IVE application in architectural and building science education.

Simulation in virtual environment is geared towards delivering a learning experience wherein three-dimensional (3D) models are designed as an immersive virtual environment and therefore are tailored to achieve desired learning objectives such as experiencing the building construction process, understanding construction detailing [35], the architectural design phase, understanding human psychology [43], and spatial experience [47]. Exploration serves as a tool for interacting with environmental data during the experimental phase of the design process [52], whereas visualisation operates as a visual communication platform by interacting with the objects or information in various contexts and for experimenting with course material and technologies [28, 49].

**Table 7.1** Summary of IVE application in architectural and building science education

| Authors | IVE application | Area of study |
|---|---|---|
| [28] | Visualisation | Building construction; BIM |
| [30] | Simulation | Building construction |
| [32] | Simulation | Building evacuation |
| [33] | Exploration | Architectural design |
| [34] | Simulation | Building: fire safety education |
| [35] | Simulation | Building construction |
| [36] | Visualisation | Architecture: CAD/BIM |
| [37] | Exploration | Architectural design |
| [38] | Exploration | Construction safety education |
| [39] | Exploration | Building science: sustainable design |
| [40] | Exploration | Building science: solar and environment |
| [41] | Visualisation | Architectural design: design process |
| [42] | Simulation | Construction education |
| [43] | Simulation | Architectural design: fire safety |
| [44] | Exploration | Architectural design: design process |
| [45] | Simulation | Building construction |
| [46] | Exploration | Architectural design: spatial experience |
| [47] | Simulation | Architectural design: design process |
| [48] | Simulation | Building service |
| [49] | Visualisation | Building construction: BIM-based |
| [50] | Visualisation | Construction education |
| [51] | Simulation | Architectural education: daylighting |
| [52] | Exploration | Architectural design: design process |
| [53] | Visualisation | Building construction education: user interface |

## 7.3 Summary

Virtual reality that functions as an immersive learning technology environment represents the cutting edge of technology advancement and education revolution. Progressively, a significant number of studies were analysed in the context of applying immersive virtual environment (IVE) as an educational approach for building science education, where major attributes of IVE were identified, particularly presence and immersion, as this component plays a crucial role in effectively disseminating the knowledge to students. In order to elevate one's sense of presence in its simulated virtual surroundings, it is crucial to heighten the degree of immersion, thereby creating a more immersive and realistic virtual experience [25]. Accordingly, to successfully deliver technology-based learning content, it is necessary to enhance

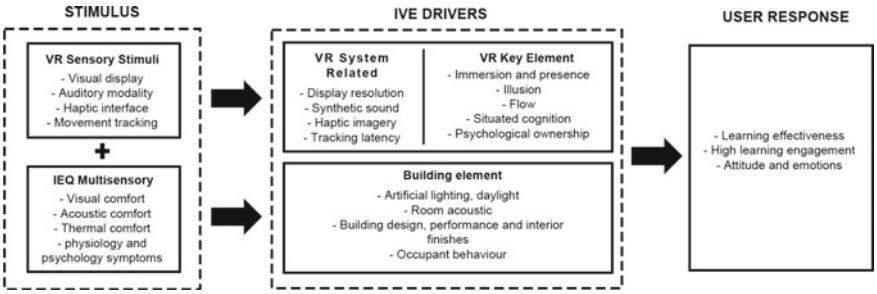



**Fig. 7.1** IVE-IEQ multisensory experience conceptual framework

the technical characteristics (immersion) when implementing a virtual reality technology for education, as these attributes would therefore deliver authentic information. Consequently, this could have a positive impact on students' academic performance, thereby enhancing their learning outcomes, establishing level of motivation, and boosting their degree of excitement in learning as well as emotion. In conclusion, as depicted in Fig. 7.1, this review summarises the variables in building conceptual frameworks centred on key aspects of multisensory experience of immersive virtual environment (IVE) integrated with indoor environmental quality (IEQ) domain.

**Acknowledgements** This research is fully funded by Fundamental Research Grant Scheme (FRGS) FRGS/1/2022/TK06/UM/02/45 and FP092-2022, and partly funded by FP117-2018 and GPF006F-2019.

## References

1. Martín-Gutiérrez, J., Mora, C.E., Añorbe-Díaz, B., González-Marrero, A.: Virtual technologies trends in education. Eurasia J. Math. Sci. Technol. Educ. **13**, 469–486 (2017). https://doi.org/10.12973/eurasia.2017.00626a
2. Chen, Y.: A study on student self-efficacy and technology acceptance model within an online task-based learning environment. **9**, 34–43 (2014). https://doi.org/10.4304/jcp.9.1.34-43
3. Radianti, J., Majchrzak, T.A., Fromm, J., Wohlgenannt, I.: A systematic review of immersive virtual reality applications for higher education: design elements, lessons learned, and research agenda. Comput. Educ.. Educ. **147**, 103778 (2020). https://doi.org/10.1016/j.compedu.2019.103778
4. Halton, C., Cartwright, T.: Walking in a patient's shoes: an evaluation study of immersive learning using a digital training intervention. Front. Psychol. **9**, 1–13 (2018). https://doi.org/10.3389/fpsyg.2018.02124
5. Mohr, D.C., Burns, M.N., Schueller, S.M., Clarke, G., Klinkman, M.: Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen. Hosp. Psychiatry **35**, 332–338 (2013). https://doi.org/10.1016/j.genhosppsych.2013.03.008
6. Ijaz, K., Ahmadpour, N., Wang, Y., Calvo, R.A.: Player experience of needs satisfaction (PENS) in an immersive virtual reality exercise platform describes motivation and enjoyment. Int. J. Hum. Comput. Interact.Comput. Interact. **36**, 1195–1204 (2020). https://doi.org/10.1080/10447318.2020.1726107

7. Guttentag, D.A.: Virtual reality: applications and implications for tourism. Tour. Manag. **31**, 637–651 (2010). https://doi.org/10.1016/j.tourman.2009.07.003
8. Frank, J.A., Kapila, V.: Mixed-reality learning environments: Integrating mobile interfaces with laboratory test-beds. Comput. Educ.. Educ. **110**, 88–104 (2017). https://doi.org/10.1016/j.compedu.2017.02.009
9. Kubra Altun, H., Lee, J.: Immersive learning technologies in English language teaching: a meta-analysis. Int. J. Contents **16**, 155–191 (2020)
10. Hao, K.C., Lee, L.C.: The development and evaluation of an educational game integrating augmented reality, ARCS model, and types of games for English experiment learning: an analysis of learning. Interact. Learn. Environ. **29**, 1101–1114 (2021). https://doi.org/10.1080/10494820.2019.1619590
11. Taskiran, A.: The effect of augmented reality games on English as foreign language motivation. E-Learn. Digital Media **16**, 122–135 (2019). https://doi.org/10.1177/2042753018817541
12. Md Shamsudin, N., Md Yunus, M.: Mirror..mirror on the wall are we real in reality? Virtual reality learning application in malaysian education. Environ.-Behav. Proc. J. **7**, 111–117 (2022). https://doi.org/10.21834/ebpj.v7i19.3245
13. Dengel, A.: What is immersive learning? In: 2022 8th International Conference of the Immersive Learning Research Network (iLRN), pp. 1–5. IEEE (2022). https://doi.org/10.23919/iLRN55037.2022.9815941
14. Jensen, L., Konradsen, F.: A review of the use of virtual reality head-mounted displays in education and training. Educ. Inf. Technol. **23**, 1515–1529 (2018). https://doi.org/10.1007/s10639-017-9676-0
15. Saeidi, S., Rizzuto, T., Zhu, Y., Kooima, R.: Measuring the effectiveness of an immersive virtual environment for the modeling and prediction of occupant behavior. sustainable human-building ecosystems. In: Selected Papers from the 1st International Symposium on Sustainable Human-Building Ecosystems, 159–167 (2015). https://doi.org/10.1061/9780784479681.017
16. Slater, M.: Immersion and the illusion of presence in virtual reality. Br. J. Psychol. **109**, 431–433 (2018). https://doi.org/10.1111/bjop.12305
17. Suh, A., Prophet, J.: The state of immersive technology research: a literature analysis. Comput. Human Behav. **86**, 77–90 (2018). https://doi.org/10.1016/j.chb.2018.04.019
18. Slater, M., Wilbur, S.: A framework for immersive virtual environments (FIVE): speculations on the role of presence in virtual environments. Presence Teleoper. Virtual Environ. **6**, 603–616 (1997). https://doi.org/10.1162/pres.1997.6.6.603
19. Witmer, B.G., Singer, M.J.: Measuring presence in virtual environments: a presence questionnaire. Presence **7**, 225–240 (1998). https://doi.org/10.1007/978-3-030-17287-9_28
20. Alamirah, H., Schweiker, M., Azar, E.: Immersive virtual environments for occupant comfort and adaptive behavior research—a comprehensive review of tools and applications. Build. Environ. **207**, 108396 (2022). https://doi.org/10.1016/j.buildenv.2021.108396
21. Wilkinson, M., Brantley, S., Feng, J.: A mini review of presence and immersion in virtual reality. Proc. Human Factors Ergon. Soc. Annu. Meeting **65**, 1099–1103 (2021). https://doi.org/10.1177/1071181321651148
22. Slater, M.: Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philosoph. Trans. R. Soc. B Biol. Sci. **364**, 3549–3557 (2009). https://doi.org/10.1098/rstb.2009.0138
23. Liu, D., Dede, C., Huang, R., Richards, J.: Virtual, Augmented, and Mixed Realities in Education. Springer Singapore, Singapore (2017). https://doi.org/10.1007/978-981-10-5490-7
24. Cheng, K.H., Tsai, C.C.: Children and parents' reading of an augmented reality picture book: analyses of behavioral patterns and cognitive attainment. Comput. Educ.. Educ. **72**, 302–312 (2014). https://doi.org/10.1016/j.compedu.2013.12.003
25. Wilkerson, M., Maldonado, V., Sivaraman, S., Rao, R.R., Elsaadany, M.: Incorporating immersive learning into biomedical engineering laboratories using virtual reality. J. Biol. Eng. **16**, 1–12 (2022). https://doi.org/10.1186/s13036-022-00300-0
26. Wu, F., Liu, Z., Wang, J., Zhao, Y.: Establishment virtual maintenance environment based on VIRTOOLS to effectively enhance the sense of immersion of teaching equipment. In:

Proceedings of the 2015 International Conference on Education Technology, Management and Humanities Science, 27, 333–337 (2015). https://doi.org/10.2991/etmhs-15.2015.93
27. Liubchak, V.O., Zuban, Y.O., Artyukhov, A.E.: Immersive learning technology for ensuring quality education: Ukrainian university case. In: CTE Workshop Proceedings, 9, 336–354 (2022). https://doi.org/10.55056/cte.124
28. Elgewely, M.H., Nadim, W., ElKassed, A., Yehiah, M., Talaat, M.A., Abdennadher, S.: Immersive construction detailing education: building information modeling (BIM)–based virtual reality (VR). Open House Int. **46**, 359–375 (2021). https://doi.org/10.1108/OHI-02-2021-0032
29. Asad, M.M., Naz, A., Churi, P., Tahanzadeh, M.M.: Virtual reality as pedagogical tool to enhance experiential learning: a systematic literature review. Educ. Res. Int. **2021** (2021). https://doi.org/10.1155/2021/7061623
30. Bashabsheh, A.K., Alzoubi, H.H., Ali, M.Z.: The application of virtual reality technology in architectural pedagogy for building constructions. Alex. Eng. J. **58**, 713–723 (2019). https://doi.org/10.1016/j.aej.2019.06.002
31. Chavez, B., Bayona, S.: Virtual reality in the learning process. In: Advances in Intelligent Systems and Computing, pp. 1345–1356. Springer International Publishing (2018). https://doi.org/10.1007/978-3-319-77712-2_129
32. Feng, Z., González, V.A., Amor, R., Lovreglio, R., Cabrera-Guerrero, G.: Immersive virtual reality serious games for evacuation training and research: a systematic literature review. Comput. Educ.. Educ. **127**, 252–266 (2018). https://doi.org/10.1016/j.compedu.2018.09.002
33. Abdelhameed, W.A.: Creativity in the initial phases of architectural design. Open House Int. **42**, 29–34 (2017). https://doi.org/10.1108/ohi-01-2017-b0005
34. Zhang, K., Suo, J., Chen, J., Liu, X., Gao, L.: Design and implementation of fire safety education system on campus based on virtual reality technology. In: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017, 11, 1297–1300 (2017). https://doi.org/10.15439/2017F376
35. Abdullah, F., Kassim, M.H. Bin, Sanusi, A.N.Z.: Go virtual: Exploring augmented reality application in representation of steel architectural construction for the enhancement of architecture education. Adv. Sci. Lett. **23**, 804–808 (2017). https://doi.org/10.1166/asl.2017.7449
36. Fonseca, D., Redondo, E., Valls, F., Villagrasa, S.: Technological adaptation of the student to the educational density of the course. A case study: 3D architectural visualization. Comput. Human Behav. **72**, 599–611 (2017). https://doi.org/10.1016/j.chb.2016.05.048
37. Abu Alatta, R., Freewan, A.: Investigating the effect of employing immersive virtual environment on enhancing spatial perception within design process. Archnet-IJAR **11**, 219–238 (2017). https://doi.org/10.26687/archnet-ijar.v11i2.1258
38. Peña, A.M., Ragan, E.D.: Contextualizing construction accident reports in virtual environments for safety education. In: Proceedings—IEEE Virtual Reality, 389–390 (2017). https://doi.org/10.1109/VR.2017.7892340
39. Ayer, S.K., Messner, J.I., Anumba, C.J.: Augmented reality gaming in sustainable design education. J. Archit. Eng. **22**, 1–8 (2016). https://doi.org/10.1061/(asce)ae.1943-5568.0000195
40. Bartosh, A., Krietemeyer, B.: Virtual environment for design and analysis (VEDA): interactive and immersive energy data visualizations for architectural design. Technol. Archit. Des. **1**, 50–60 (2017). https://doi.org/10.1080/24751448.2017.1292794
41. González, N.A.A.: Development of spatial skills with virtual reality and augmented reality. Int. J. Interact. Des. Manuf. **12**, 133–144 (2018). https://doi.org/10.1007/s12008-017-0388-x
42. Zhou, Y., Ji, S., Xu, T., Wang, Z.: Promoting knowledge construction: a model for using virtual reality interaction to enhance learning. Procedia Comput. Sci. **130**, 239–246 (2018). https://doi.org/10.1016/j.procs.2018.04.035
43. Pitana, T., Prastowo, H., Mahdali, A.P.: The development of fire safety appliances inspection training using virtual reality (VR) technology. In: IOP Conference Series: Earth and Environmental Science. IOP Publishing Ltd. (2020). https://doi.org/10.1088/1755-1315/557/1/012064
44. Lin, C.H., Hsu, P.H.: Integrating procedural modelling process and immersive VR environment for architectural design education. MATEC Web Conf. **104** (2017). https://doi.org/10.1051/matecconf/201710403007

45. Hossain Maghool, S.A., Moeini, S.H. (Iradj), Arefazar, Y.: An educational application based on virtual reality technology for learning architectural details: challenges and benefits. Archnet-IJAR **12**, 246–272 (2018). https://doi.org/10.26687/archnet-ijar.v12i3.1719
46. Moleta, T.J.: Game on: exploring constructive design behaviors through the use of real-time virtual engines in architectural education. Int. J. Archit. Comput.Comput. **14**, 212–218 (2016). https://doi.org/10.1177/1478077116663341
47. Sapto Pamungkas, L., Meytasari, C., Trieddiantoro, H.: Virtual reality as a spatial experience for architecture design: a study of effectiveness for architecture students. SHS Web Conf. **41**, 05005 (2018). https://doi.org/10.1051/shsconf/20184105005
48. Mai, L.T., Werdin, H.: VRLab4BES—a virtual reality implementation approach of building service simulation for educational purposes. In: International Conference on Virtual Rehabilitation, ICVR. 2022-May, 82–89 (2022). https://doi.org/10.1109/ICVR55215.2022.9848010
49. Seyman Guray, T., Kismet, B.: Applicability of a digitalization model based on augmented reality for building construction education in architecture. Constr. Innov.Innov. **23**, 193–212 (2021). https://doi.org/10.1108/CI-07-2021-0136
50. Ventura, S.M., Castronovo, F., Nikolić, D., Ciribini, A.L.C.: Implementation of virtual reality in construction education: a content-analysis based literature review. J. Inf. Technol. Constr. **27**, 705–731 (2022). https://doi.org/10.36680/j.itcon.2022.035
51. Sabry, H., Sherif, A., Rakha, T., Fekry, A.: Integration of daylighting simulation software in architectural education. In: EG-ICE 2010—17th International Workshop on Intelligent Computing in Engineering (2019)
52. Gomez-Tone, H.C., Chávez, M.A., Samalvides, L.V., Martin-Gutierrez, J.: Introducing immersive virtual reality in the initial phases of the design process—case study: freshmen designing ephemeral architecture. Buildings **12** (2022). https://doi.org/10.3390/buildings12050518
53. Sun, C., Hu, W., Xu, D.: Navigation modes, operation methods, observation scales and background options in UI design for high learning performance in VR-based architectural applications. J. Comput. Des. Eng. **6**, 189–196 (2019). https://doi.org/10.1016/j.jcde.2018.05.006

# Chapter 8
# The Effect of Distance on Audiovisual Temporal Integration in an Indoor Virtual Environment

**Victoria Fucci and Raymond H. Cuijpers**

**Abstract** For several decades, it has been debated whether a distance compensation mechanism exists during audiovisual (AV) synchrony judgements, regardless of the vast difference between the speed of sound and light. Here we aimed to investigate the effect of stimulus distance on the human tolerance for (physical) asynchronies and broaden earlier findings with a state-of-the-art head-mounted display (HMD). In this study, we measured the point of subjective simultaneity (PSS) of visual and auditory stimuli in an indoor virtual environment (VE). The synchrony judgement method was used for 11 stimulus onset asynchronies (SOA) and six egocentric distances up to 30 m. In addition, to obtain higher validity of the dataset, we implemented in our data analysis the results from the previous studies of the egocentric distance perception and the AV hardware latency delay. Our findings displayed positive PSS values that increased with distance showing that in our VE, a distance compensation mechanism is taking its place. However, the gain was smaller than was expected for complete compensation for the slower speed of sound.

## 8.1 Introduction

It is known that the unitary representation of the physical world comes through various sensory modalities such as vision, hearing, touch and smell. In usual circumstances, we hardly imagine a separate perception of auditory or visual events happening around us. By nature, in physical world conditions, they are automatically integrated and best understood in a single multisensory event [1, 2]. However, we still need to understand how multisensory events are integrated into virtual world conditions and the best design decisions for the most ecologically valid integration.

By using information from multiple modalities, an individual can combine the different types of information and interpret the world more accurately. The combination of multisensory modalities can also cause the illusory binding of a multimodal

V. Fucci (✉) · R. H. Cuijpers
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: victoria.k.fucci@gmail.com



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_8

cue in a single event. For example, in the ventriloquist effect, people hear the sound coming from the mouth of the dummy instead of the ventriloquist. They perceive a coherent audiovisual event where the visual source captures the sound source [3]. The dominance of vision causes the brain to interpret the effects caused by spatial disparity as everything coming from the source that provides the visual aspect. This dominance is negatively correlated with the amount of visual noise present. The more visual noise (e.g. blurred stimuli) an individual perceives, the more (s)he relies on the sound source to properly understand the event. Using such phenomena, perception of the world can greatly be influenced by different features such as spatial or temporal differences.

Compared to tactile stimulation, which usually remains limited to the nearby area, visual and auditory modalities can appear on a more extensive range of distances [4–6]. Even though there is a significant difference in the propagation speed for sound and light (343 m/s for sound and about 300,000 km/s for light), the perception of audiovisual (AV) events in our daily environment remains synchronous. However, in a real-world scenario, the auditory component will always arrive later at the observer, and this difference will increase in accordance with the physical distance. Synchronicity in a physical representation of visual or auditory events is not always necessary to be present, as also suggested by examples from daily situations, but tolerance for temporal disparities is needed. Vroomen et al. [7] found one of the most precise explanations of why we still tend to perceive sensorial modalities synchronously despite all neural and physical disparities between light and sound: Our brain functions are ready to determine two stimulations in (a)synchrony as long as they fall into a specific window of temporal disparity. Several studies in AV temporal alignment found that the perception of an audiovisual stimulus is maximally synchronous if the visual event reaches the subject just before the auditory event [8–10].

One of the most cited works that showed a vision-first bias was performed by Sugita and Suzuki [11]. In their experiment, participants were given headphones. They were presented with flashing LEDs from various distances (1, 5, 10, 20, 30, 40 and 50 m) to stimulate visual perception and a sound from the headphones to stimulate auditory perception. The intensity of the flashes increased with distance so that the perceived luminance was constant. The subjects were instructed to think that the sound came from the same LED array as the light. Participants were asked to judge the timing of the two stimuli using a temporal order judgement (TOJ) task. They judged whether the light came before or after the sound. The study reported that the stimulus onset (a)synchrony (SOA) only provided the best perception of synchrony through a positive value which speaks for an audio delay with respect to vision. The best perception of synchrony provided by SOA was a point of subjective simultaneity (PSS) which correlated positively with distance. Sugita and Suzuki [11] concluded that humans rely on data about stimulation distance to compensate for the differences in propagation velocity between light and sound. This would explain why the PSS increases with stimulus distance. They also found that this compensation has limits: humans can compensate for the sound and visual sources up to a delay of 106 ms at a 40 m distance, which shows that the source's distance affects how people perceive it.

In contrast, the research conducted by Lewald and Guski [12] showed that the distance compensation mechanism was absent. They repeated Sugita et al.'s [11] experiment in the outdoors with real speakers and LEDs so that the attenuation of the intensity of the flash and sound stimuli were physically correct. They concluded that it simply works by integrating stimuli in a wide time window rather than an implicit estimation of sound velocity and that the observed distance effect by Sugita and colleagues is an artefact of the unnatural stimuli. Signals falling within that window are perceived as synchronous and may be perceived as part of a single multisensory event. Another study by Arnold et al. [13] had similar conclusions and even proposed that the results of the study by Sugita and Suzuki [11] were more cognitive than perceptual, as they requested their participants *imagine* the auditory and visual stimuli to originate from one source.

Since asynchrony is caused by the physical difference in travel time between light and sound, the signal that arrives at our ear's lags with respect to the visual signal. The problem is that significant differences between the studies made interpretation difficult. For example, Sugita and Suzuki [11] used headphones to deliver white-noise bursts without distance information (constant loudness and no reverberation) and the light intensity was attenuated to compensate for the effect of distance. It was found that up to 20 m sound delays were compensated. On the other hand, Lewald and Guski [12] used an actual speaker setup with a LED in its centre in an outdoor open field to deliver similar white-noise bursts. Thus, sound and light intensity were attenuated according to distance from the observer, but there was no distance information from reverberations as it was an outdoor experiment. They found no evidence of compensation when using these more realistic stimuli that do contain distance information. It is well known that the attenuation between direct and indirect sound is an essential cue for sound localization [14, 15]. Since this auditory cue was absent in both studies, it could have reduced the effects of any distance compensation mechanism.

An important aspect of synchrony perception is that nowadays, AV information arrives from our natural environment and the digital world. Science and technology are continuously searching for advanced forms of sensory reproduction systems. For example, for the past two decades, interest in immersive digital technology has been growing, decreasing and growing again. It has enormously impacted entertainment, arts and several research industries. Various virtual reality (VR) systems could influence human mental states and perceptions differently and affect social behaviour in the real world [16]. Recent technological developments have made HMDs more accurate and smaller, in several instances, hardly distinctive from regular glasses. As a result of these developments, VR technology is now an exciting subject of investigation again [17]. HMD displays incorporate every type of technological innovation which mounts displays on the individual's head. However, despite all the progress, simulator sickness issues consistently become challenging for extensive use of HMDs [18]. In such artificial environments, timing relations among modalities rely entirely on VR technology, which has a hardware latency. Temporal differences can have a negative impact on the perception of AV production components [19] and even reduce the feeling of presence in a virtual environment [20]. Due to these observations,

an important task for developers is to regulate optimally the temporal connection between auditory and visual signals to achieve higher perceived quality.

Silva et al. [21] were among the first to use immersive technology in an AV synchrony perception experiment, created ecologically valid stimuli [22] of biological motion and manipulated the visual depth cues. They used multiple distances (ranging from 10 to 35 m, with 5-step increments) and onset asynchronies to test whether an internal distance compensation mechanism existed for the physical delays of the different types of stimuli. They suggested that the compensation mechanism exists, and it might depend on the depth cues available to the observer. At the furthest distances (30 and 35 m), the synchrony judgement was more uncertain, which might suggest a limit to the compensation mechanism.

This study aims to contribute to the research area focused on the effect of stimulus distance on human tolerance for (physical) (a)synchronies by using a high-fidelity HMD display and a realistic virtual environment simulation. Our interest is to clarify the relation between distance and the perception of synchrony when both visual and auditory depth cues are as ecologically valid as possible. To do so, we simulated realistic impact sounds of boxes falling on a conveyor belt in an indoor virtual environment. These stimuli provided both rich visual and auditory depth cues. Based on the previous work [8, 11, 21, 23, 24], we would expect that simulation of realistic audiovisual depth cues would result in the activation of the distance compensation mechanism when making synchrony judgements and the PSS at the observer will shift with distance increments towards increasing audio delays.

## 8.2 Method

### *8.2.1 Participants*

The participants included nine undergraduate students between 19 and 24 years old ($\mu_{\text{age}} = 21.1$ years, $\sigma_{\text{age}} = 1.69$ years, five females and four males), recruited utilizing convenience sampling. All participants had normal or corrected-to-normal vision and hearing. Each participant received financial compensation of 50 euros after participation in the experiment.

The study by Lewald and Guski [12] on AV synchrony perception on distance in the real world acquired an effect size $f^2 = 1.48$. To calculate the required sample size for our experiment, an ANOVA was used comparing the PSS for the five different distances used in the study and the effect size acquired from the study by Lewald and Guski [12]. Given an $\alpha = 0.05$ and a power of 0.9, this led to a sample size of at least four participants. We analysed the data after every four participants to see whether the results were sufficient for the scope of the study on the presence of a distance compensation mechanism. In addition, a minimum of eight participants was chosen to avoid unknown covariance, such as hidden hearing problems. Ultimately, the experiment had nine participants.

The study was conducted according to the guidelines in the Declaration of Helsinki [25] and was approved by the Ethical Board of the Human Technology Interaction Group (HTI) of the Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology.

### *8.2.2 Apparatus*

The experiment took place at Eindhoven University VR Lab. The VE was administered to participants via Oculus Rift HMD with built-in on-ear headphones. The Oculus Rift had two OLED panels with a 1080 × 1200 resolution running at 90 Hz with a 90° horizontal field-of-view (FOV) and a 110° vertical FOV. The PC ran Windows 10 on an Intel® Core™ i9-9900 K processor, with the NVIDIA® Titan RTX™ graphics card and a Realtek® LC1220P-VB2 sound card.

As a data collection framework during the experiment, we used an open-source package called Unity Experiment Framework for the Unity game engine [26]. This framework allowed us to collect all participants' responses within the VE comfortably. All the Unity scripts for the experimental setup were written in C# programming language.

### *8.2.3 Stimuli*

The experiment used a virtual reality environment to simulate an old brick factory hall with a conveyor belt located in the centre, with several objects around to give a better sensation of realism (see Fig. 8.1). The factory had a surface of 25 m by 40 m and a height of 7 m. The participants' position in the virtual space is 4.2 m to the side (perpendicular to the conveyor belt) and 4.7 m in front of the closest box position (parallel to the conveyor belt), as shown in Fig. 8.1. The distance to the start of the nearest box position (hypotheses of the triangle) amounted to 6.3 m. Figure 8.2 depicts a schematic overview of the room with scales.

This VE was created in the 3D design software SketchUp (see Fig. 8.3). The textures were assigned to the environment and objects for a more ecologically valid visual experience. The finalized 3D scene was imported into the Unity Game Engine and integrated with an Oculus Rift. For faster rendering power, it was chosen to use Baked Light behaviour with lightmap. The participant's position was fixed, but the head movement remained free to experience a more "natural" feeling of being present in the VE. However, during the experiment, participants were seated, which allowed them to focus on a very narrow area of the VE along the conveyor belt.

The fall of a cardboard box has been designed as the only dynamic event happening in the environment. The participants assist a box falling from the metallic tubes at different egocentric simulated distances (6.3, 10.6, 15.3, 20.1, 25.1, 30 m); after the event, the box disappears. In the designed space, no wind or other factors that


**Fig. 8.1** Screen capture of the virtual factory hall. Three stands marked by letters *A* (audio first), *S* (synchronous) and *V* (video first) for participants to respond according to the synchrony judgement method (SJ-3)

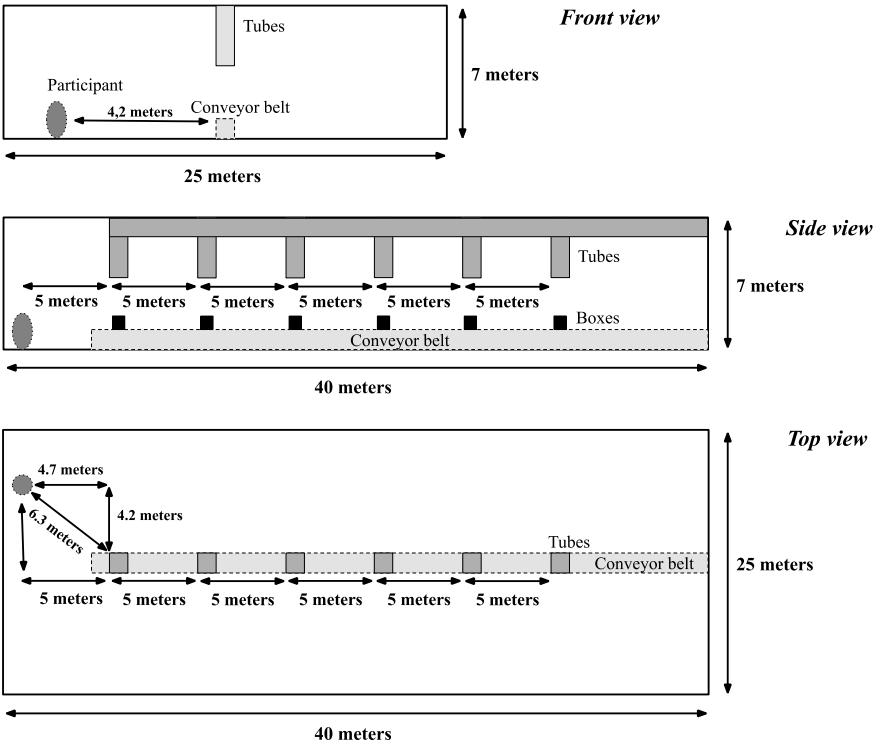



**Fig. 8.2** Layout of the VE, as seen from the front, side and top

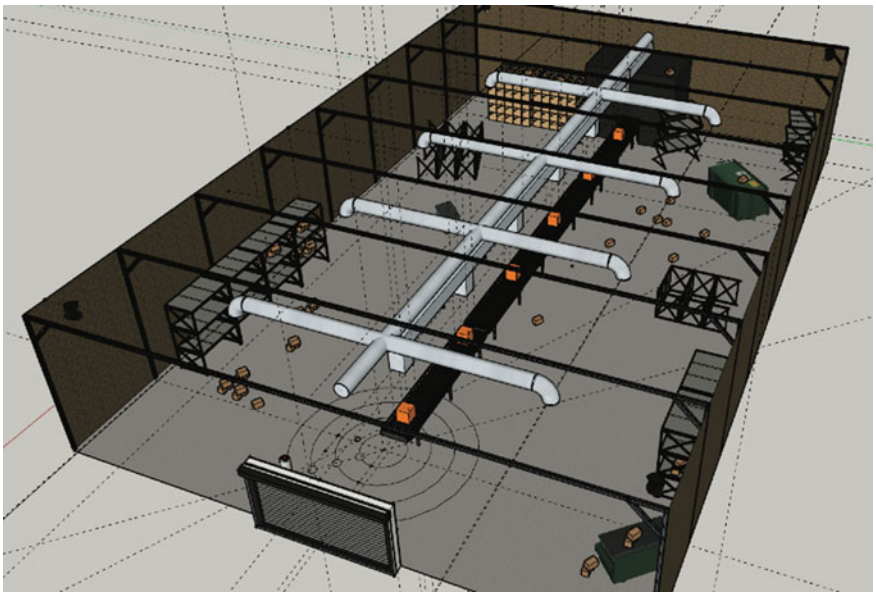


**Fig. 8.3** 3D model sketch of the VE, the top view

could influence the falling of the boxes are simulated. Hence, the falling time and the falling speed of the box have been considered constant. SOAs are defined regarding the falling time of impact (0.892 s) and with the default setting for gravitational acceleration. In the VE, the above calculation was used to specify the time the sound will be presented at the observer's location/headphones.

The environment's geometry was simplified following the room characteristics introduced in the previous section and exported as a geometrical model into Odeon. Sound absorption coefficients were applied to the walls, floor, ceiling and 3D objects to guarantee that the simulated room's binaural impulse response (BRIR) was correctly calculated. This improved the realism of the acoustic properties of the designed space, resembling an experience closer to the real space. Two audio files were used during the experiment: a seamlessly looped soundscape of the environment (ventilation system) and a box hitting the conveyor belt. These sound sources were generated for each distance separately. Figures 8.4 and 8.5 provide an overview of how the sound sources and receiver (observer position) were placed in the Odeon software suite.

Each BRIR was generated for one source-receiver position: eight BRIRs for ventilation soundscape and six BRIRs for each simulated distance. On each generated BRIR, the convolution calculations were performed using a MATLAB script with a mono audio recording of a fan noise or the falling sound of a cardboard box. This would make the audio files sound like they were heard in that specific space and at a specific distance due to the simulated reflections. The sound of a box falling on the conveyor belt lasted 1 s as an impact sound and 2 s as a reverberation tail (3 s

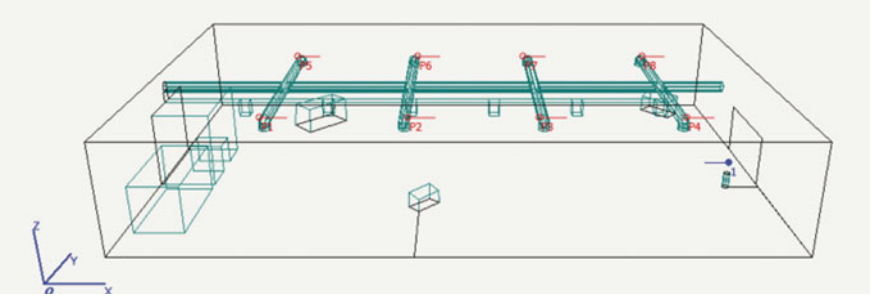


**Fig. 8.4** Simplified 3D model for Odeon software to simulate the ventilation system soundscape. The blue point represents the receiver(listener) position, and the red points are the sound sources from each air tube, eight in total

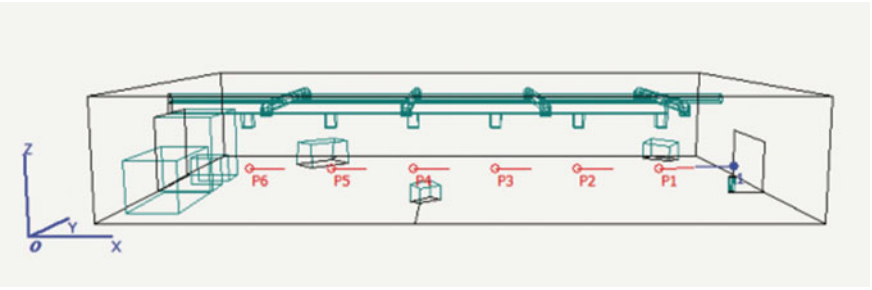



**Fig. 8.5** Simplified 3D model for Odeon software to simulate the collision sound of the box dropping on the conveyor belt. The blue point represents the receiver(listener) position, and the red points are the sound sources for six different distances (6.3, 10.6, 15.3, 20.1, 25.1, 30 m)

in total). All generated audio files of the VE soundscape (ventilation system) were seamlessly mixed at − 36 dB in one waveform audio file format file [27] using audio editing software [28].

### *8.2.4 Design*

The experiment used a within-subjects design which involved two independent variables and one dependent variable. The first independent variable was, for a given event, the time interval between the occurrence of a visual stimulus and the occurrence of its corresponding auditory stimulus. In the context of synchrony experiments, this time interval is referred to as the stimuli onset (a)synchrony (SOA). It is expressed in the number of milliseconds after which the auditory stimulus is introduced relative to the presentation of the corresponding visual stimulus. Negative SOAs, therefore, indicate that the sound stimulus is presented before the visual stimulus.

The second independent variable was six egocentric distances (simulated/perceived) at which the AV event occurred. The simulated and perceived distances obtained in a separate study [29], are listed in Table 8.1. The 11 SOAs ranged from − 250 to 250 ms, with a step size of 50 ms. Every combination of distance and delay was presented to the participant 30 times, leading to a total of 1980 trials. The trials were divided into twelve quasi-random blocks of 165 trials.

The synchrony judgement (SJ3) method [30, 31] was used to measure the perceived AV synchrony and determine the point of subjective simultaneity (PSS) values. The PSS was defined as the midpoint of the range of delays that are usually judged to be synchronous [31]. These PSS values were then compared over distance to find if sound delays caused by distance affected perceived synchrony. The participants responded to the SJ3 method by indicating a perceived stimulus order, expressed by one of the following options: audio first (*A*), synchronous (*S*), or video first (*V*). This perceived order was the dependent variable.

### *8.2.5 Procedure*

To start the experiment, participants signed the informed consent and filled in a short questionnaire about their age, gender, experience using VR and the condition of their hearing and eyesight.

Due to the large number of trials, the experiment was divided into two sessions. Before the start of the actual experiment, the participants first performed a practice trial to understand the meaning of audio first (*A*), synchronous (*S*), video first (*V*), and how to give the corresponding response. In this practice session, no final results were recorded. The participants were observed during the trial session to determine whether they understood the task. If this were true, the actual experiment would start.

The box hitting the conveyor belt was accompanied by an auditory stimulus, presented either shortly before, shortly after, or simultaneously with the hit to the participant's location. The time interval between the visual occurrence of the hit and its corresponding auditory occurrence was randomly selected from a set of predefined SOAs. Immediately after the box would hit the conveyor belt, the participant was asked to judge in which order the stimuli occurred.

The answer was recorded along with the quasi-randomized parameters, including the chosen tube from where the box dropped, the timestamps of the response input, and the time interval between the visual and auditory occurrence of the box hitting the conveyor belt. After each finished block, the participant was allowed to take a break to prevent fatigue and nausea caused by the VR headset. Overall, the experiment consisted of two sessions, each lasting 1.5–2 h, wherefore it could take two consecutive days or one full day with a lunch break to finish the whole experiment.

**Table 8.1** Individual values represented with applied hardware latency compensation for each PSS (+104 ms) in ms per distance in m (mean, SE, SD)

| Simulated dist. | 6.3 m | 10.6 m | 15.3 m | 20.1 m | 25.1 m | 30 m |
|---|---|---|---|---|---|---|
| **Perceived dist.** | 2.8 m ± 0.17 | 5.8 m ± 0.38 | 8.8 m ± 0.52 | 11.5 m ± 0.63 | 14.5 m ± 0.74 | 18 m ± 1.09 |
| Pp 1 | 35.62 | 55 | 45.15 | 46.14 | 67.74 | 34.57 |
| Pp 2 | − 22.03 | − 9.04 | 15.12 | 34.57 | 31.18 | 27.73 |
| Pp 4 | 37.28 | 24.09 | 39.31 | 60.86 | 46.94 | 39.59 |
| Pp 5 | 44.41 | 46.84 | 60.75 | 50.63 | 49.53 | 60.01 |
| Pp 6 | 27.43 | 82.17 | 25.9 | 53.98 | 76.03 | − 3.83 |
| Pp 7 | − 3.8 | − 67.83 | 33.53 | 3.47 | − 28.85 | − 8.97 |
| Pp 8 | 50.32 | 59.34 | 51.85 | 64.32 | 77.84 | 69.62 |
| Pp 9 | 74.86 | 90.14 | 92.66 | 88.9 | 105.71 | 82.86 |
| **Mean** | **30.54** | **33.06** | **51.72** | **57.81** | **63.66** | **49.37** |
| SE | 9.55 | 16.38 | 9.68 | 10.72 | 16.31 | 15.54 |
| SD | 28.66 | 49.15 | 29.04 | 32.15 | 48.93 | 46.63 |

### *8.2.6 Data Analysis*

When virtual reality is used, the underestimation of egocentric distance occurs in various scenarios [32], and the hardware also introduces an additional latency/delay [33]. In our study, the perception of distance could play an important role, and we also report in our data analysis the perception of previously measured egocentric distances in the used VE [29]. We measured these distances using verbal judgement and position adjustment tasks. In a previous study, we found that the average egocentric distance underestimation was 38.5%. The perceived distance values obtained from the study on egocentric distance perception (2.8, 5.8, 8.8, 11.5, 14.5, 18 m) (see Table 8.1) are the mean values of the different groups of participants and therefore should be considered as an additional observation perspective which allowed us to demonstrate the effect of not only the simulated (designed) distance of AV (a)synchrony in VR but also the effect of perceived distance.

The end-to-end hardware latencies were measured for Unity Engine with Oculus Rift, using Arduino Uno, VINT Hub Phidget HUB0000, SparkFunSound Detector, light sensor and TBS1102B Tektronix digital Oscilloscope [34]. Noticeably, as far as we know, previous research on AV timing perception did not conduct or implement such measurements into their design or data analysis [12, 23, 30, 31]. Therefore, our experiment was designed and conducted in accordance with the preceding standards.

The obtained hardware delay values were applied during the data analysis of this study. The SOA values for each participant for each simulated/perceived distance condition were compensated for the hardware end-to-end visual (46 ms) and audio (150 ms) latency delay of the whole Apparatus. This resulted in all SOAs being shifted by + 104 ms, resulting in a final SOA range from − 146 to + 354 ms instead of the original − 250 to + 250 ms at the moment of the experiment. This did not affect our results significantly, as this range was wide enough to capture all PSS values.

The fitting model-based psychometric function and statistical tests such as regression analysis with the repeated measures ANOVA were conducted using MATLAB [35].

## 8.3 Results

As a result of the SJ3 task, we obtained the proportions of responses about when participants felt that audio was either leading, lagging or synchronous. Figure 8.6 shows the proportions of responses that a participant gave for each choice with a stimulus distance of 15.3 m as a function of SOA. To obtain an accurate estimate of the PSS, we used a logistic model that could simultaneously fit three curves. The advantage of this method is that it is very robust against missing data, and the resulting fits will always add up to 1. The video-first and audio-first responses are fitted with logistic sigmoid functions, the equation of which is given in Eq. 8.1.

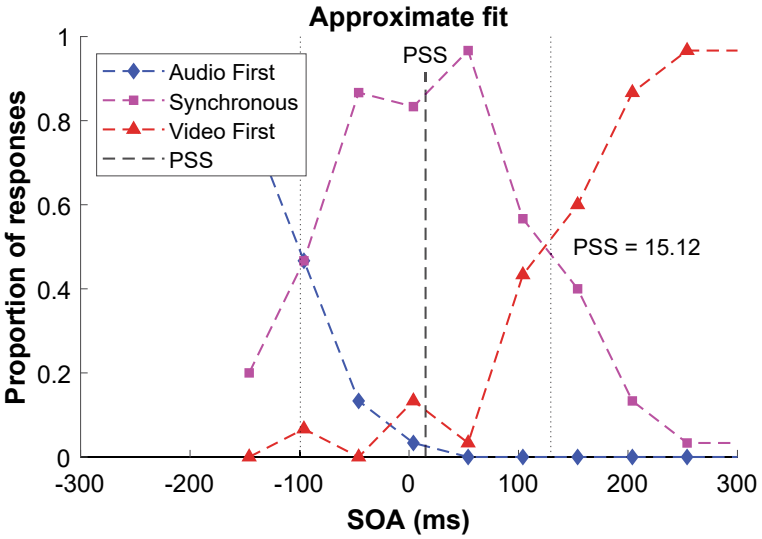



**Fig. 8.6** Example data of one participant for a distance of 15.3 m/8.8 m with hardware latency compensation. Represented are the audio-first (blue), synchronous (pink), and video-first (blue) response proportions for each of the measured SOA

$$\sigma\left(x, \mathscr{k}, x_0\right) = \frac{1}{1 + \mathrm{e}^{-\mathscr{k}(x - x_0)}} \tag{8.1}$$

The slope parameter $\mathscr{k}$ for the audio-first responses was chosen to be negative and opposite to video-first responses. It is also possible to fit them independently, but this did not improve the results. The equation for the synchronous responses follows from the fact that the proportions of responses must always add up to 1. For this reason, it is more elegant to fit the curves simultaneously, as the fit parameters are not independent of one another. The resulting equation for the synchronous responses is:

$$\rho\left(x, \mathscr{k}, x_A, x_V\right) = 1 - \sigma\left(x, -\mathscr{k}, x_A\right) - \sigma\left(x, \mathscr{k}, x_V\right) \tag{8.2}$$

where $\mathscr{k}$ is the steepness, $x_A$ the horizontal shift for audio-first responses and $x_V$ the horizontal shift for video-first responses. We then minimized the total sum of squared error for these parameters. The total sum of squared errors is given by:

$$\begin{aligned} \mathrm{SSE} = & \sum_{\text{audiofirst}} \left(\mathscr{y} - \sigma\left(x, -\mathscr{k}, x_A\right)\right)^2 \\ & + \sum_{\text{videofirst}} \left(\mathscr{y} - \sigma\left(x, -\mathscr{k}, x_\mathrm{V}\right)\right)^2 \end{aligned}$$

$$+ \sum_{\text{synchronous}} (y - p(x, \mathscr{k}, x_A, x_V))^2 \tag{8.3}$$

Here $y$ is the observed proportions $x$ and the SOA. Using MATLAB's *fminsearch,* we then obtained the desired parameters of the fit, as illustrated by Mareschal et al. [36]. The value of PSS is computed as follows $(x_V + x_A/2)$. The synchrony range is given by the range between the left and right intersection points, which is approximately equal to $x_V - x_A$. In Fig. 8.7, the same data are shown together with the fitted curves for one participant and a stimulus distance of 15.3 m.

The raw data were plotted for each participant for each simulated/perceived distance condition. The data was analysed, and the PSS values were defined as the mean of the synchrony boundaries ($L$ and $R$). The PSS values for each simulated/perceived distance condition of each participant can be found in Table 8.1. The obtained results supported the hypothesis about the existence of an internal compensation mechanism for audiovisual delays being in agreement with Sugita and Suzuki [11] and (to some extent) with Silva et al. [21] but in clear opposition with Lewald and Guski [12]. In addition, the average reaction time (RT) during each trial was calculated, RT = 335 ms ± 0.014, which is within the range of the studied literature on SOAs in AV reaction time tasks [37].

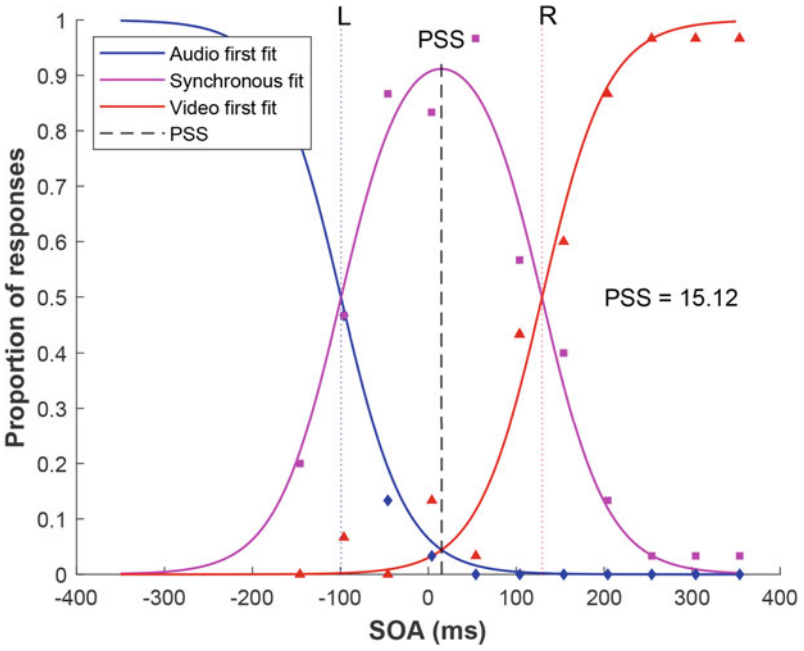



**Fig. 8.7** Logistic model fit of the data demonstrated in Fig. 8.6. Solid lines show the fitted audio-first (blue), synchronous (pink) and video-first (red) response curves. The synchrony range is formed by intersection points $L$ and $R$ representing the left and right synchrony boundaries. The dashed grey line indicates the PSS at the midpoint of this range

The linear regression model of PSS values with an averaged slope of 1.14 ± 0.48 ms/m and a shift of 27.31 ms on the *y*-intercept, plotted as a function of simulated distances (6.3, 10.6, 15.3, 20.1, 25.1 and 30.0 m) established by the Unity game engine (1 Unity unit = 1 m), presented in Fig. 8.8, showed a trend effect of distance on PSS ($F$ (1.4) = 5.596, $t$ (14) = 0.708, $p = 0.077$). It turns out there was considerable variation between participants (see below). Despite this result, there is an $R^2 = 0.583$ and a substantial effect size $f^2 = 1.398$. We also did a repeated measures ANOVA after verifying that sphericity could be assumed (*Mauchly's* $W = 0.010$, d$f = 14$, $p = 0.072$), showing a significant main effect of distance on the PSS ($F$ (5.35) = 3.554, $p = 0.011$). The distance compensation did not scale with the same amount as one would expect based on the speed of sound, and on average, only 1.1/2.9 = 38% was compensated.

Previously, we found that distance in this VE was strongly underestimated [29] by an average of 38,5%. Assuming that people use the perceived distance to judge asynchronies, it is useful to plot the PSS judgements as a function of perceived distance, as shown in Fig. 8.9. The linear regression model of PSS values with an averaged slope of 1.79 ± 0.78 ms/m, and a shift of 29.39 ms on the *y*-intercept, was plotted as a function of perceived distances (2.8, 5.8, 8.8, 11.5, 14.5, 18 m), presented on Fig. 8.9, also showed that distance positively predicts PSS ($F$ (1.4) = 5.23, $p$ =

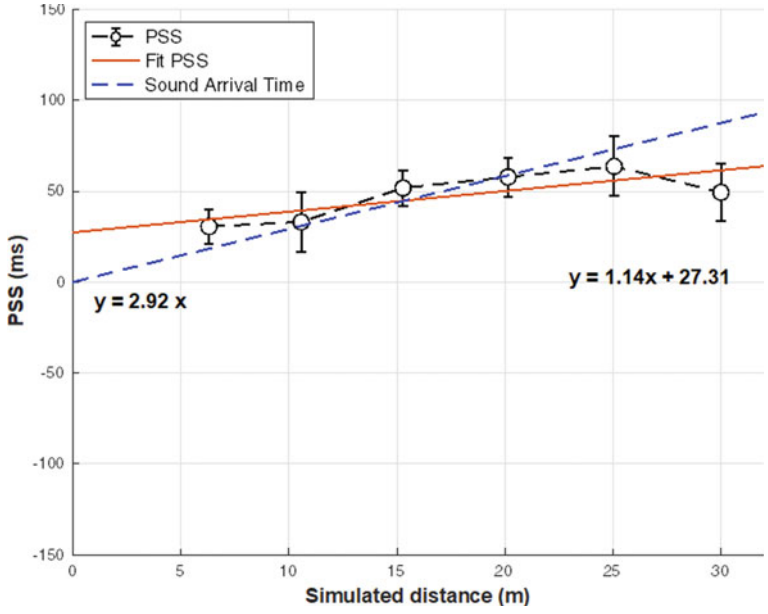



**Fig. 8.8** Linear regression model (solid red line) of mean across PSS values (white circles) at the observer position with hardware latency compensation, plotted as a function of simulated egocentric distance. The dashed blue line represents the theoretical distance compensation mechanism at the observer position for differences in sound arrival time with the speed of sound 344 m/s. The error bars indicate the standard error (SE) across the individual subject's mean

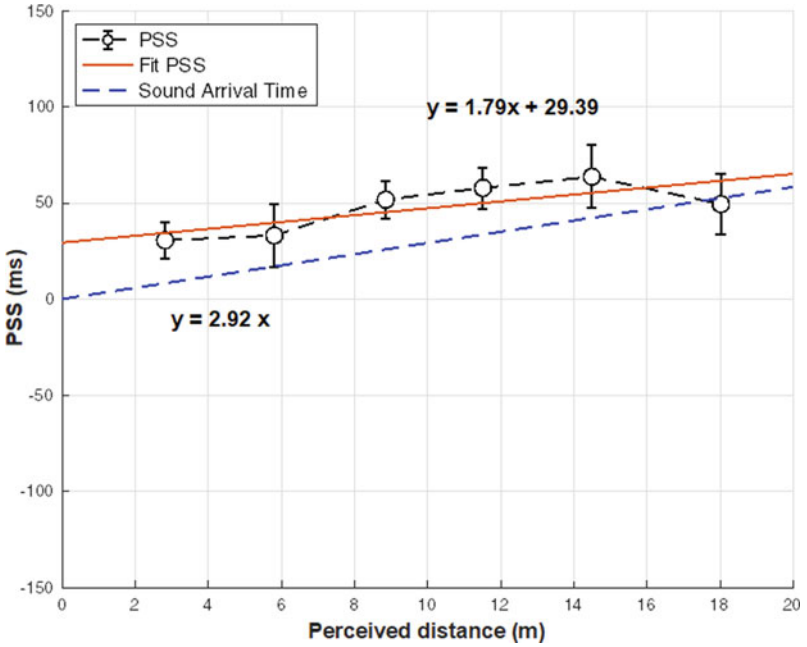



**Fig. 8.9** Linear regression model (solid red line) of mean across PSS values (white circles) at the observer position with hardware latency compensation, plotted as a function of perceived egocentric distance. The dashed blue line represents the theoretical distance compensation mechanism at the observer position for differences in sound arrival time with the speed of sound 344 m/s. The error bars indicate the standard error (SE) across the individual subject's mean

0.084), with an $R^2 = 0.567$ and a sufficiently large effect size $f^2 = 1.309$. Similarly, the distance compensation did not scale strictly based on the speed of sound, and on average, only 1.7/2.9 = 58% was compensated. It is 20% higher in comparison to the results from Fig. 8.8.

The analyses of the individual fit of a linear function for simulated distances per participant shown in Fig. 8.11 resulted in an average slope of $1.139 \pm 0.82$ ms/m. From the graph, it is clear that there is a significant variation between subjects.

We plotted each participant's individual slopes in ascending order to investigate this further. Figure 8.11 represents the individual slopes of the participants with the standard error of the mean. Again, we observe a wide range of individual differences, but all are positive, indicating distance compensation. In particular, for Participant 5 ($p = 0.005$)**, Participant 2 ($p = 0.02$)* and Participant 8 ($p = 0.041$)*, the slopes are significantly larger than zero.

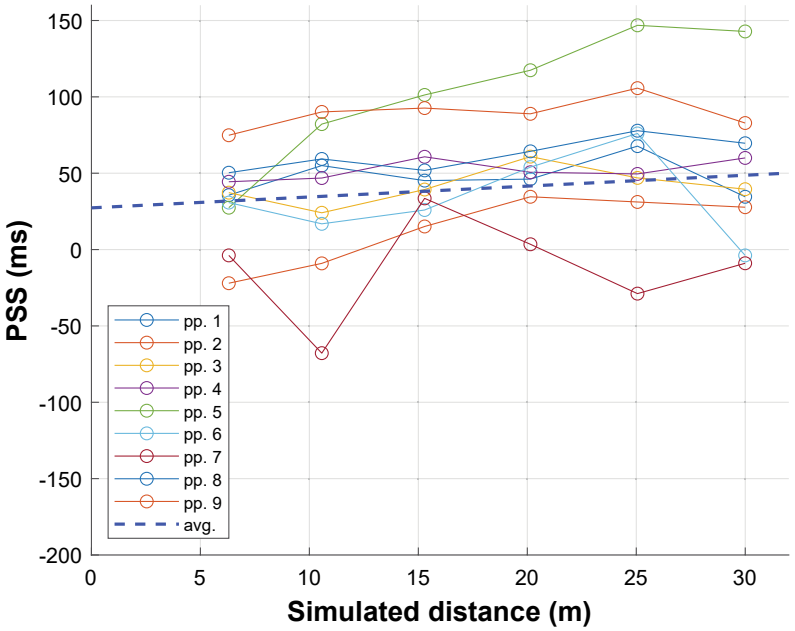



**Fig. 8.10** Individual fits of a linear function for simulated distances per participant. The coloured lines represent individual data, and the dashed blue line is the average fit

## 8.4 Discussion

This study investigated the effect of distance on AV (a)synchrony perception in a high-fidelity indoor VE, considering the egocentric distance underestimation and the end-to-end hardware latency. Based on the previous research by Silva et al. [21], Sugita and Suzuki [11], it was hypothesized that the point of subjective simultaneity (PSS) at the observer would shift with distance increments towards increasing audio delays.

The results obtained in the chosen VE supported the proposed hypothesis, saying: "The PSS at the observer shifts with distance increments towards increasing audio delays", being in agreement with Sugita and Suzuki [11] and (to some extent) with Silva et al. [21], but in clear opposition with Lewald and Guski [12]. Furthermore, data showed that PSS values increased with distance, showing a distance compensation mechanism for the designed virtual environment.

The statistical results on averaged data showed a significant main effect of distance on the PSS. Still, there was a significantly positive slope when fitting a straight line only for some participants. In particular, we found strong individual differences: six participants had a positive slope below 1 ms/m, which was not significantly different from zero, and three participants had a larger, significantly positive slope

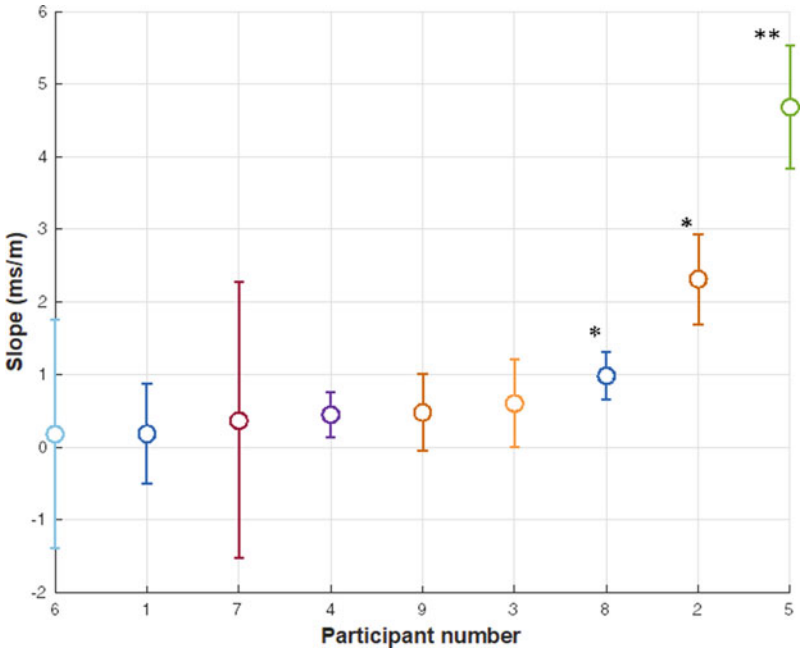



**Fig. 8.11** Slopes and SE of the coefficients. The y-axis values are the individual slopes of the participants, while the bars are the SE of the participants (* $= p \leq 0.05$, ** $= p \leq 0.01$)

(pp. 5**, pp. 2*, pp. 8*). This shows that the synchrony judgement task was quite challenging for some participants. It does appear, however, that a distance compensation mechanism is in effect, although the quantitative effects vary across participants.

The Oculus Rift is not the best HMD available and was discontinued in 2021. Still, the specifications are decent and marginally worse than more state-of-the-art systems. In any case, they are sufficient for our perceptual experiment. The most relevant parameter for measuring PSS is the hardware latencies. They turn out to be quite similar between various devices. However, we could remove any hardware delays due to our calibration procedure. The results obtained from the latency measurement showed lower latency performance of Oculus Rift compared to Oculus Quest 2 (newer version) [34]. In addition, Kelly [38] recently analysed 131 studies on distance perception and determined that despite the improvement among modern HMDs, the noticeable underestimation continues. Moreover, there is insufficient evidence for any positive effect on distance perception, irrespective of the resolution. This means that human perception is a more complex construct and might not purely depend on technological advancements.

A possibly interesting perceptual phenomenon occurred at a simulated distance of 30 m. When analysing Fig. 8.8, it seems that the PSS at 30 m is relatively low compared to the linear fit. The data of the five closer distances lie much better on

a straight line without the 30 m distance included. Although this effect is insignificant, it aligns with other research observations suggesting a limit to how much audiovisual latency can be compensated. Noticeably, if the furthest distance were invalidated, the other points' slope would become steeper, showing a more substantial distance compensation effect. One explanation of this behaviour might be depth perception. In the study by Silva et al. [21], they found that the results from the mean of the Gaussian curves at the 30 and 35 m could not reach a value for the 100% of synchrony judgement responses. This might mean that the synchrony judgements at farther distances become more complicated, and it somehow affects the linearity of the distance compensation mechanism. However, it might also be the result of the experimental artefact, which should be investigated further in future studies.

In line with Silva et al. [21], future studies should also focus on the presence and quality of visual depth cues and their effect on the distance compensation mechanism for sound propagation velocity. In addition, it is interesting to implement and simulate the real-time effect of physical sound propagation for different VE conditions (indoor/outdoor; farther distances) and the effect of the direct sound-only (no reflections) condition.

## 8.5 Conclusion

This study aimed to test the effect of distance on the perception of audiovisual (a)synchrony in an indoor virtual environment. Our results demonstrated that as the sound transmission time at the observer increased with distance, the PSS values measured at the observer also increased at an approximately similar rate (shifting towards visual leads). This provides evidence that people implicitly estimate the sound transmission time when judging the synchrony of an audiovisual stimulus in the chosen VE. These findings are in line with previous research [8, 11, 23, 24] and to some extent with Silva et al. [21], but in contrast to other research findings where the distance compensation mechanism was missing [12, 13].

In conclusion, using the state-of-the-art HMD, our study contributes to the continuous dispute on whether the effect of distance compensation mechanism on audiovisual (a)synchrony perception exists. Furthermore, before running similar experiments in VE, we want to remind you about the necessary quantification of egocentric distance perception and the hardware end-to-end AV latency measurements. Both act as valuable design parameters and independent variables. These practices can provide a higher validity and understanding of the data.

**Acknowledgements** We want to thank the late Armin Kohlrausch for his supervision and considerable inspiration that made this work possible. We also thank the Building Acoustics research group at the Eindhoven University of Technology for the advice and help with the audio auralization process. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 812719.

## References


1. King, A.J.: Multisensory integration: strategies for synchronization. Curr. Biol. **15**(9), 339–341 (2005). https://doi.org/10.1016/j.cub.2005.04.022
2. Spence, C., Squire, S.: Multisensory integration: maintaining the perception of synchrony. Curr. Biol. **13**(13), 519–521 (2003). https://doi.org/10.1016/s0960-9822(03)00445-7
3. Alais, D., Burr, D.: The ventriloquist effect results from near-optimal bimodal integration. Curr. Biol. **14**(3), 257–262 (2004). https://doi.org/10.1016/j.cub.2004.01.029
4. Gepshtein, S., Burge, J., Ernst, M.O., Banks, M.S.: The combination of vision and touch depends on spatial proximity. J. Vision **5**(7) (2005). https://doi.org/10.1167/5.11.7
5. Hillis, J.M., Ernst, M.O., Banks, M.S., Landy, M.S.: Combining sensory information: mandatory fusion within, but not between. Senses Sci. **298**, 1627–1630 (2002). https://doi.org/10.1126/science.1075396
6. Miyazaki, M., Yamamoto, S., Uchida, S., Kitazawa, S.: Bayesian calibration of simultaneity in tactile temporal order judgment. Nat. Neurosci. **9**, 875–877 (2006). https://doi.org/10.1038/nn1712
7. Vroomen, J., Keetels, M.: The spatial constraint in intersensory pairing: no role in temporal ventriloquism. J. Exp. Psychol. Hum. Percept. Perform. **32**(4), 1063–1071 (2006). https://doi.org/10.1037/0096-1523.32.4.1063
8. Alais, D., Carlile, S.: Synchronizing to real events: subjective audiovisual alignment scales with perceived auditory depth and speed of sound. Proc. Natl. Acad. Sci. **102**, 2244–2247 (2005). https://doi.org/10.1073/pnas.0407034102
9. Arrighi, R., Alais, D., Burr, D.: Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. J. Vision **6**(6) (2006). https://doi.org/10.1167/6.3.6
10. Keetels, M., Vroomen, J.: The role of spatial disparity and hemifields in audio-visual temporal order judgments. Exp. Brain Res. **167**, 635–640 (2005). https://doi.org/10.1007/s00221-005-0067-1
11. Sugita, Y., Suzuki, Y.: Implicit estimation of sound-arrival time. Nature **421**, 911–911 (2003). https://doi.org/10.1038/421911a
12. Lewald, J., Guski, R.: Auditory-visual temporal integration as a function of distance: no compensation for sound-transmission time in human perception. Neurosci. Lett. **357**, 119–122 (2004). https://doi.org/10.1016/j.neulet.2003.12.045
13. Arnold, D.H., Johnston, A., Nishida, S.: Timing sight and sound. Vision. Res. **45**, 1275–1284 (2005). https://doi.org/10.1016/j.visres.2004.11.014
14. Burijngame, J.A., Butler, R.A.: The effects of attenuation of frequency segments on binaural localization of sound. Percept. Psychophys. **60**, 1374–1383 (1998). https://doi.org/10.3758/BF03207999
15. Middlebrooks, J.C., Makous, J.C., Green, D.M.: Directional sensitivity of sound-pressure levels in the human ear canal. J. Acoust. Soc. Am. **86**, 89–108 (1989). https://doi.org/10.1121/1.398224
16. Mazuryk, T., Gervautz, M.: Virtual reality-history, applications, technology and future. Comput. Sci. **12** (1996)
17. Rummukainen, O.: Reproducing reality: perception and quality in immersive audiovisual environments. Doctoral thesis, Aalto University (2016). https://urn.fi/URN:ISBN:978-952-60-7115-2
18. Moss, J.D., Muth, E.R.: Characteristics of head-mounted displays and their effects on simulator sickness. Human Factors J. Human Factors Ergon. Soc. **53**, 308–319 (2011). https://doi.org/10.1177/0018720811405196
19. Rihs, S.: The influence of audio on perceived picture quality and subjective audio-video delay tolerance. In: Proceeding of the MOSAIC Workshop Advanced Methods for the Evaluation of Television Picture Quality. J. Audio Eng. Soc. **47**(5) (1995). https://research.tue.nl/en/publications/d6a8c289-73e5-49e0-8261-e3faf7e55f64

20. Kohlrausch, A., van de Par, S.: Audio-visual interaction in the context of multi-media applications. Commun. Acoust. pp. 109–138 (2005). https://doi.org/10.1007/3-540-27437-5_5
21. Silva, C., Mendonça, C., Mouta, S., Silva, R., Campos, J.C., Santos, J.: Depth cues and perceived audiovisual synchrony of biological motion. PLoS One **9** (2014). https://doi.org/10.1371%2Fjournal.pone.0080096. Erratum in: PLoS One **9**(1) 2014. https://doi.org/10.1371/annotation/d0e27a68-ad6d-452f-bfca-337487fc933c. PMID: 24244617; PMCID: PMC3828238
22. Kohlrausch, A., van Eijk, R., Juola, J.F., Brandt, I., van de Par, S.: Apparent causality affects perceived simultaneity. Atten. Percept. Psychophys. **75**, 1366–1373 (2013). https://doi.org/10.3758/s13414-013-0531-0
23. Engel, G.R., Dougherty, W.G.: Visual-auditory distance constancy. Nature **234**, 308–308 (1971). https://doi.org/10.1038/234308a0
24. Kopinska, A., Harris, L.R.: Simultaneity constancy. Perception **33**, 1049–1060 (2004). https://doi.org/10.1068/p5169
25. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA **310**(20), 2191–2194 (2013). https://doi.org/10.1001/jama.2013.281053
26. Brookes, J., Warburton, M., Alghadier, M., Mon-Williams, M., Mushtaq, F.: Studying human behavior with virtual reality: the unity experiment framework. Behav. Res. Methods. **52**, 455–463 (2019). https://doi.org/10.3758/s13428-019-01242-0
27. Bhatnagar, G., Mehta, S., Mitra, S.: Chapter 7. The WAV File Format. Introduction to Multimedia Systems. Academic Press, San Diego (2007)
28. Mazzoni, D., Dannenberg, R.B.: A fast data structure for disk-based audio editing. Comput. Music. J. **26**(2), 62–76 (2002)
29. Korshunova-Fucci, V., van Himbergen, F.F., Fan, H.M., Kohlrausch, A., Cuijpers, R.H.: Quantifying egocentric distance perception in virtual reality environment. Int. J. Human Comp. Interact. (2023). https://doi.org/10.1080/10447318.2023.2234117
30. van Eijk, R.L., Kohlrausch, A., Juola, J.F., van de Par, S.: Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type. Percept. Psychophys. **70**, 955–968 (2008)
31. Kuling, I.A., van Eijk, R.L., Juola, J.F., Kohlrausch, A.: Effects of stimulus duration on audio-visual synchrony perception. Exp. Brain Res. **221**, 403–412 (2012). https://doi.org/10.1007/s00221-012-3182-9
32. Feldstein, I.T., Kölsch, F.M., Konrad, R.: Egocentric distance perception: a comparative study investigating differences between real and virtual environments. Perception **49**, 940–967 (2020). https://doi.org/10.1177/0301006620951997
33. Raaen, K., Kjellmo, I.: Measuring latency in virtual reality systems. Entertain. Comput. ICEC 2015, 457–462 (2015). https://doi.org/10.1007/978-3-319-24589-8_40
34. Fucci, V., Liu, J., You, Y., Cuijpers, R.H.: Measuring Audio-Visual Latencies in Virtual Reality Systems. In: Proceedings of 7th International Conference on Artificial Intelligence and Virtual Reality. Smart Innovation, Systems and Technologies, **382**(10), Springer, Singapore (2023)
35. The MathWorks Inc.: Matlab version: 9.12.0, r2022a (2022)
36. Mareschal, I., Calder, A.J., Dadds, M.R., Clifford, C.W.: Gaze categorization under uncertainty: psychophysics and modeling. J. Vision **13**, 18–18 (2013). https://doi.org/10.1167/13.5.18
37. Bazilinskyy, P., de Winter, J.: Crowdsourced measurement of reaction times to audiovisual stimuli with various degrees of asynchrony. Human Factors J. Human Factors Ergon. Soc. **60**, 1192–1206 (2018). https://doi.org/10.1177/0018720818787126
38. Kelly, J.W.: Distance perception in virtual reality: a meta-analysis of the effect of head-mounted display characteristics. IEEE Trans. Visual. Comput. Graph. pp. 1–13 (2022). https://doi.org/10.1109/TVCG.2022.3196606

# Chapter 9
# Inside the Black Box: Modeling a Cybersickness Dose Value Through Built-In Sensors of Head-Mounted Displays

**Judith Josupeit and Fabienne Andrees**

**Abstract** The postural instability theory claims that postural instability precedes visually induced motion sickness (VIMS) or, in the case of virtual reality (VR), cybersickness. If this theory holds, there needs to be a temporal connection between postural instability and the onset of reported cybersickness. Thus, a head-movement-based cybersickness dose value (hmCSDV) is postulated. The hmCSDV uses the individuals' motion patterns before, during, and after VR exposure. For reasons of efficiency head movement is accessed via the built-in sensors of the head-mounted display. In addition, controller input during VR exposure is used to account for individual differences in perceived virtual motion. In total, data from 169 participants were available for modeling. To address the aspect of gamification the experimental task allowed the participant to virtually explore a VR city and collect checkpoints. Multivariate non-normality was respected in all statistical analyses. The feasible generalized least squares regressions with the within effect of time showed significant results for the prediction of cybersickness ratings during VR exposure, but not for the comparison before and after VR. The final hmCSDV suggested that shorter distances, higher mean acceleration, longer duration of for-or-aft motion, more frequent stops, and shorter duration of these stops as a function of total time spent in VR accounted for 5.4% of the total variability in impending cybersickness ratings. Methodological features and limitations of the study are discussed. This finding holds promise for algorithms that can be used to predict individual cybersickness severity and provide potential countermeasures before symptoms occur.

J. Josupeit (✉) · F. Andrees
Technical University Dresden, Dresden, Germany
e-mail: judith.josupeit@tu-dresden.de

F. Andrees
e-mail: fabienne.andrees@tu-dresden.de



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_9

## 9.1 Introduction

Practitioners of yoga will probably know the famous pose "the tree"—standing on one foot with the other foot placed on the inner thigh. For sure, they have perceived the differences in balance and posture by exercising in this posture with their eyes open and closed. This example illustrates the reliance of body posture and balance on visual input.

It follows that the visual input and postural output are interconnected. Yet, not only lacking visual input but also artificial visual information has the potential to perturbate posture. Specifically, posture is affected not only when standing on one foot or a wobbly ground, but also when standing on a flat surface in case the visual input is apt [1]. According to the postural instability theory, the visual movement information induces corrective postural movements—mostly for-and-aft movement—to maintain the momentary upright position [2]. Unfortunately, due to the artificiality, instead of being helpful, these movements lead to a resonance effect. Thus the artificial visual movement information now additionally contains a real physical movement "artifact". Once visual input and postural output got out of hand, the likelihood of somatic responses like dizziness, headache, disorientation, or nausea is increased [2–4].

All of these responses are subsumed under the term *visually induced motion sickness* (VIMS). Depending on the sources for visual input, a variety of *sicknesses* are differentiated, e.g., simulator sickness or cybersickness or cinema sickness, because these different effectors lead to different patterns of somatic responses [5, 6]. Whereas the former and the latter should be self-explanatory, the term *cybersickness* might not be that familiar. In short, cybersickness is VIMS induced by virtual reality (VR). Simulators and VR are characterized by interaction with the environment; in contrast, effectors like 3D movies allocate a passive role to the user. In contrast to simulators, VR applies an immersive technology that achieves a reality or sense of presence. To this end, stereoscopic images with binocular disparities are used to convey a three-dimensional impression, but also head-movement-based rendering to transfer the real physical movement into a virtual shift in perspective in the VR environment. Hardware-wise stereo-lenses and position tracking are used for instance via a head-mounted display (HMD).

Compared to the other effectors of VIMS, VR comes with several advantages like easily accessible head-tracking data, controller input, and a naturalistic environment. Therefore, VR might even solve a chicken-and-egg problem that is discussed in the field of VIMS: Namely, whether postural instability is the precursor or the manifestation of VIMS. In favor of the latter statement is the postural instability theory [2]: Consequently, the experimental data should indicate the onset of corrective movements before the onset of VIMS. This leads to the convenient situation in VR, in which the severity of cybersickness would become predictable with information that is already standard in consumer electronics via the physical movement—by logging the position of the HMD—while controlling for the virtual movement—by logging the users' controller input. Previous small sample studies indicate that changes in

the range and frequencies of head movement, accessed through the built-in sensors of the HMD, do predict cybersickness by applying a deep fusion network [7, 8]. In contrast to this data-driven black box approach, we want to postulate and test a head-movement cybersickness dose value (hmCSDV), which would not necessarily need any data from target users, but be based on theoretical grounds. The hmCSDV defines a metric for the expected cybersickness severity in dependence on time spent in VR and individually different characteristics of head movement. Previously postulated models that focus on the mean expected severity of cybersickness (CSDV) did primarily focus on soft- and hardware-specific characteristics of the setup [9]. As our setup is the same for all participants, these characteristics are experimentally controlled.

To bundle the easily accessible data from the built-in sensors and the individual controller input together, while keeping the soft- and hardware-specific factors at par, two independently conducted unpublished project studies will be used for theory-based inference with a larger sample size. As a prerequisite for modeling the hmCSDV, the VR application has to have an impact on the reported symptoms. Therefore, in comparison with the baseline, post VR symptoms should be significantly higher. Moreover, during the VR exposure, the cybersickness ratings should significantly increase over time. On an individual level, postural instability should significantly increase in the baseline vs. post VR comparison, as well as gradually during VR if the posture is reliant on visual input. Additionally, considering cybersickness, the individual postural instability should predict the difference between the baseline and post VR ratings. Furthermore, during the VR exposure, the individual postural instability should predict the upcoming cybersickness ratings when accounted for the moderating effect of the virtual movement.

## 9.2 Methods

### *9.2.1 Participants*

In total, 189 participants took part in either study. Twelve participants had to be excluded from the final analysis as they reached the previously defined abort criterion, which is covered in more detail in the Procedures section. Moreover, four datasets were missing and two other datasets were incomplete because of a malfunction of the hardware. In the second study, one participant had already participated in the first study and thus needed to be excluded to rectify the sample. This leaves a total of 169 individuals (76 male, 92 female, and one diverse) for the analysis. Participants' age ranged from 18 to 64 years ($M = 23.83$, $SD = 5.27$).

Both studies were run consecutively in a laboratory at the Technical University of Dresden from October to December 2020. Participants were recruited via flyers or mail through the university's participant data bank and received either course credit or 5€ per 1/2 h. Two hours before their session participants were instructed

to refrain from eating. Predefined exclusion criteria included being under-aged ($\leq$ 17 years), having epilepsy, having a history of migraine, being pregnant, and/or having non-corrected visual impairments. Written informed consent was obtained before participation. All data were analyzed in anonymized and aggregated form. The studies were approved by the local ethics committee (SR-EK-315072020 and SR-EK-316072020).

### *9.2.2 Design*

Both studies use repeated measures variables for cybersickness and postural instability with a baseline and post VR comparison and a temporal sequence with six time points. For transparency, it has to be admitted that one of the studies used a between-subject design with a treatment and a control group. The treatment group was additionally presented with a virtual nose which was discussed to have mitigating effects on the self-reported cybersickness [10, 11, but also see 12]. As no mitigating effect between the treatment and the control group was proven, we will use all data from both studies[1] [13].

### *9.2.3 Materials and Measured Variables*

The HTC VIVE (HTC, Taiwan, China, and Valve, Bellevue, WA, USA) was deployed as HMD. The computer that rendered the VR environment was custom-built with an NVIDIA GeForce RTX 2070 GPU, an Intel Core i7-9700 K CPU, and 32 GB (2 $\times$ 16 GB) RAM. The VR environment was created with Unity Professional (v 2019.1.1.1f1). In order to access the wireless motion-tracked controllers, the Steam VR plugin was utilized with custom key-bindings. Prefabs for the VR environment of the Windridge City asset were used. Additionally, some simple prefabs were custom objects made with Blender (v2.92.0).

As Unity meta-data a timestamp, logs for the displayed scenes in VR, and the number of reached virtual checkpoints were collected. Moreover, for the postural instability the participants' position and rotation of the head in *x*-, *y*-, and *z*-Unity-coordinates were recorded. For the interaction in VR the controller trackpad input in *x*- and *y*-Unity-coordinates, the controller position and rotation in *x*-, *y*-, and *z*-Unity-coordinates were added. All Unity meta-data had a mean sampling frequency

[1] We are aware that not finding a statistically significant difference does not necessarily lead to the conclusion that samples are equivalent. Therefore, TOST-equivalence tests were used to compare the treatment with the control group. We defined that a meaningful effect as a decrement of one point for the MISC and of one item for the VRSQ. The results for both sides were significant MISC $\Delta_L\ t_{(107.960)} = 3.179$, $p < 0.001$ $\Delta_U\ t_{(107.960)} = -1.962$, $p = 0.026$; and VRSQ $\Delta_L\ t_{(106.061)} = 2.157$, $p = 0.016$ $\Delta_U\ t_{(106.061)} = -3.496$, $p < 0.001$. In contrast, both NHST were not significant, meaning the effect size bounds do include zero.

of 60 Hz and were stored in a csv file. Demographic data comprising age, gender, duration of any previous experience with VR, and whether visual aid was needed, were asked via a Lime Survey questionnaire. Before and after VR exposure the virtual reality sickness questionnaire (VRSQ) [14] was assessed to represent the multi-dimensionality of the cybersickness symptoms. During the VR exposure the Misery Scale (MISC) [15] was used as a single-item questionnaire with the abort criterion nausea (MISC $\geq$ 6) [16].

### *9.2.4 Procedure*

For a quick overview of the experimental procedure, the reader is referred to Fig. 9.1. After reception, signing the consent form, and completing the demographic questionnaire, the experimenter explained the cybersickness questionnaires in use, and the functionality of the trackpad. A cover story about visuospatial orientation and its relation to eye movements in VR was used to reduce priming effects.

Then, the experimenter adjusted the straps of the HMD and assessed the VRSQ and MISC baseline measures in a dark virtual environment, that was used for calibrating the eye-tracker (eye-tracking data are not reported). During calibration, the participant was instructed to look straight ahead, while standing as still as possible with their feet hip-width apart, for at least 30 s. After that, a VR city environment containing various virtual checkpoints was displayed. The participant's task was to explore this environment freely to find as many virtual checkpoints as they could in 10 min. To navigate in the city, participants were handed the controller to their dominant hand. The longitudinal (forward or backward) acceleration was achieved by pressing the trackpad on its' corresponding edge (front or back). For changing the lateral direction, the participant had to move their head. Rotational movement along the participant's axis was therefore allowed, whereas translational movement was constrained. Every 2 min the MISC was deployed to monitor the participants' subjective cybersickness state. After 10 min, the dark environment was displayed to measure the post VR VRSQ (same instruction as above). Overall, this procedure resulted in six MISC (baseline and five times during VR) and two VRSQ (baseline and post VR) measurement time points. Thereafter, the experimenter stopped the VR application, removed the HMD, and made sure the participant was capable of leaving. After all potential after-effects vanished, the participants were debriefed and compensated. In total, the procedures took roughly 30 min for each participant.

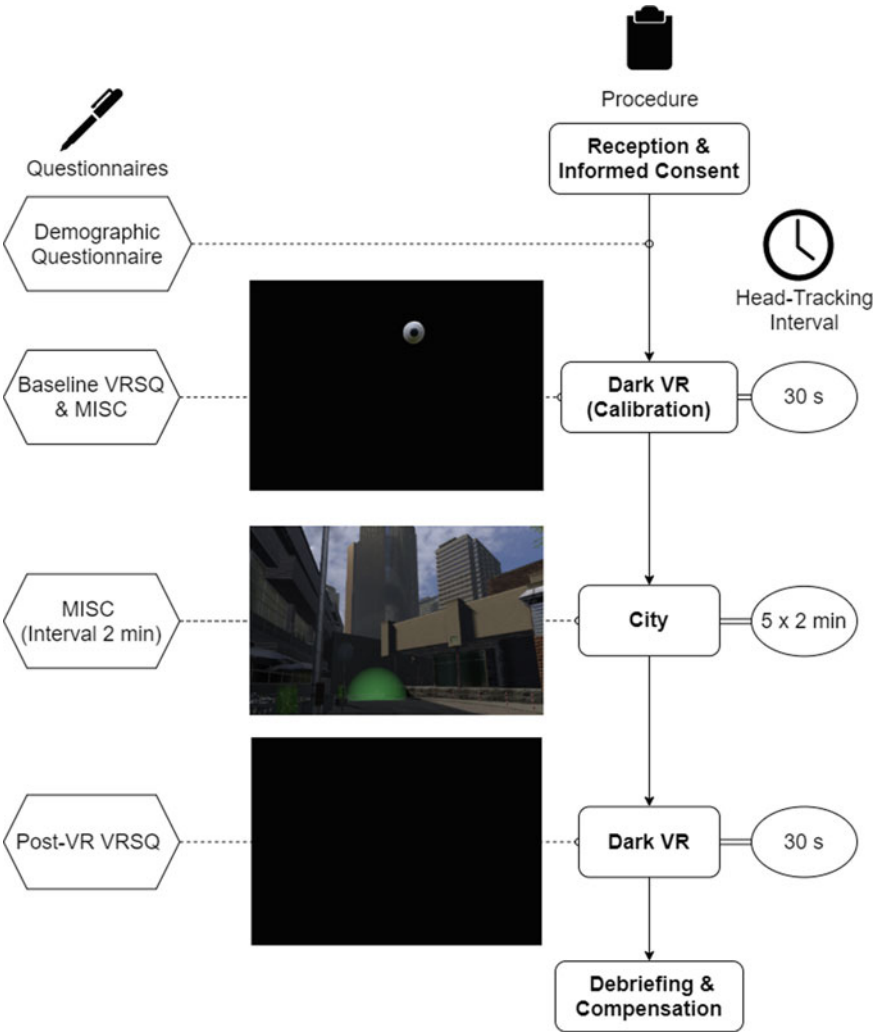



**Fig. 9.1** Illustration and summary of the experimental procedure

## 9.3 Results

### *9.3.1 Data Preprocessing*

Baseline and post VR were defined as the first 30 s after the calibration onset, respectively; the first 30 s after dark VR was displayed for the second time. The other five blocks were defined as 2 min intervals starting with the first controller input, i.e., after

the instruction had been read to the participant, following the experimental procedures for the query of the MISC. For each block, head positions and timestamps were used to calculate the Euclidean distances, accelerations, and changes in *x*-, *y*-, and *z*-directions. Additionally, the frequency and duration of each directional physical movement were summarized. Moreover, controller input and timestamps during the city environment were used to calculate the frequency and duration of the start-stop motion. All data were grouped by the block they were assigned to and summary statistics, namely minima, maxima, median, and mean, were calculated.

For one subject, a missing at-random value in the cybersickness ratings was imputed using linear imputation. Because cybersickness data are prone to be non-normally distributed, we will use robust or nonparametric equivalents of parametric inference statistics, whenever the assumption of multivariate normality is violated [17]. We decided to use the naturalistic non-logarithmized data to have generalizable and interpretable results. The multi-dimensionality of the VRSQ was met by calculating the $\text{VRSQ}_{\text{Total Score}}$ based on the questionnaires' manual.

### *9.3.2 Descriptive Analysis*

The descriptive statistics and tests for multivariate normality revealed that nonparametric tests should be applied for all statistical analyses. Looking at Fig. 9.2, it becomes apparent that the cybersickness ratings are right-skewed and do vary vastly on individual level. Additionally, comparing baseline ($M = 5.486$, $\text{SD} = 6.446$) to Post VR $\text{VRSQ}_{\text{Total Score}}$ ($M = 14.65$, $\text{SD} = 12.057$), ratings seem to increase (Fig. 9.2a). The same pattern, a linear increase over time spent in VR, applies to the MISC ($\text{MISC}_{\text{Baseline}}$ $M = 0.316, \text{SD} = 0.735, \text{MISC}_{10\text{ min VR}}$ $M = 1.577, \text{SD} = 1.842$) with a right skewness and a gradually decreasing kurtosis when the time spend in VR progressed (Fig. 9.2b).

Figure 9.3 illustrates the postural instability development over time spent in VR for the baseline and post VR comparison, as well as during VR exposure. We define postural instability with the following facets: the mean Euclidean distance a participant moved physically in all directions, and the mean acceleration of this movement, the mean duration of movements in *z*-direction, and the frequency of changes in this direction. All data are right-skewed; the kurtosis ranges indicate that the data include a lot of heavy-tailed data. For an overview of some assorted descriptive statistics the reader is referred to Table 9.1, the linear trend suggested by the descriptive statistics illustrated in Fig. 9.3 for the data during VR exposure can be found in the majority of postural instability indicators but is not reported here for clarity reasons.

The mean Euclidean distance gets larger for the baseline and post VR comparison as well as during the VR exposure, whereas the mean acceleration gets smaller. Moreover, the frequency of the changes in movement in *z*-direction decreases for the baseline and post VR comparison as well as during the VR exposure. In parallel, the duration of this movement shows the same reciprocal pattern, meaning an increase over time. The changes of movement in *z*-direction decreased over time spent in

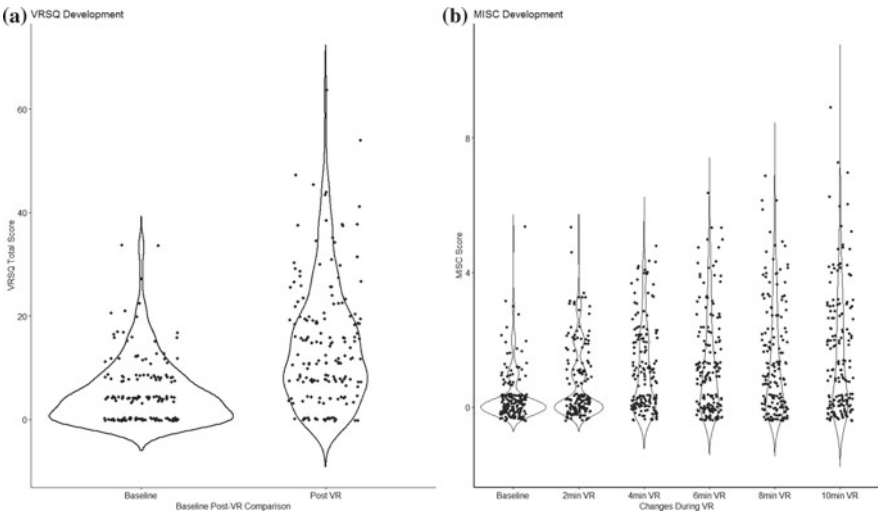



**Fig. 9.2** Cybersickness development over time. **a** Comparison of the $VRSQ_{Total\ Score}$ baseline and post VR. **b** Development of the MISC ratings during the VR exposure query interval was fixed to 2 min, violin plots with added jitter are used to account for the non-normality of the data

VR, while the duration of the movement in one direction reciprocally increased. The $z$-axis was chosen, as for-and-aft movement is often used to systematically provoke VIMS [18]. For the instability during VR, the frequency and the mean duration of stops in the virtual movement were added. With prolonged time spent in VR, less frequent and shorter stops were found (Fig. 9.4).

One must be cautious when comparing the baseline and post VR with the instability during VR exposure as timeframes for these instances differ (30 s vs. 2 min), which also applies for the $VRSQ_{Total\ Score}$ and MISC, due to the different dimensionality of the questionnaires.

### *9.3.3 Inference Statistics*

The nonparametric longitudinal analysis MATS inference for potentially singular and heteroscedastic MANOVA [19, 20] of the baseline and post VR comparison of the $VRSQ_{Total\ Score}$ revealed a significant time effect ($ATS_{(1,166)} = 181.243$, $p < 0.001$). Moreover, the nonparametric longitudinal analysis of the effect of time during VR for the MISC was also significant ($ATS_{(3.477,167)} = 51.017$, $p < 0.001$).

As the longitudinal (or repeated measures) factor time has six levels, sequential post-hoc contrasts were run [21]. The sequential post-hoc contrasts revealed that the $MISC_{Baseline}$ ($\widehat{p}_{\ Baseline} = 0.348$) was smaller than the $MISC_{2\ min\ VR}$ ($\widehat{p}_{\ 2\ min\ VR} = 0.446$) and gained significance ($\chi^2_{(275)} = 2.774$, $p = 0.003$). Furthermore, the

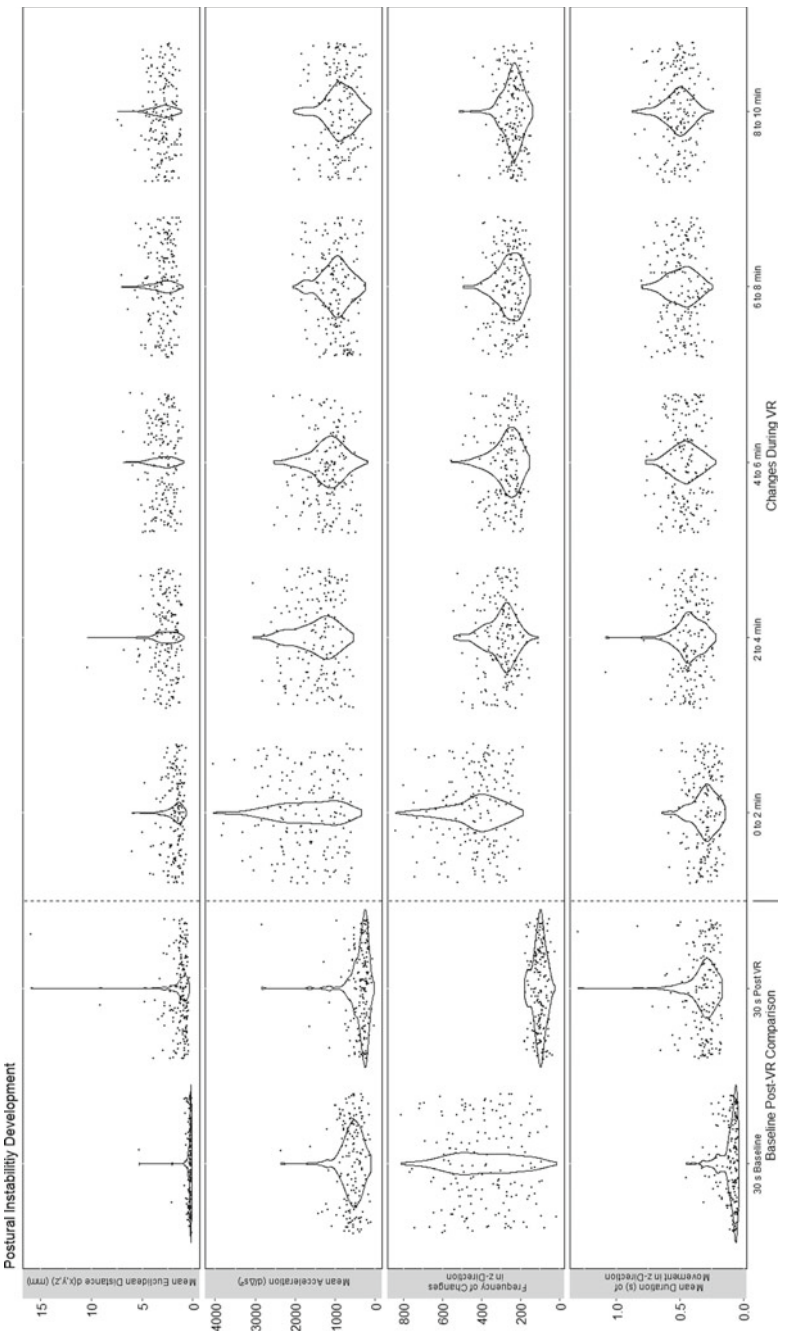



**Fig. 9.3** Postural instability development is defined as the mean Euclidean distance, mean acceleration, frequency of changes in *z*-direction, and duration of movement into *z*-direction for the baseline and post VR comparison and the development during VR exposure. Violin plots with added jitter are used to account for the non-normality of the data. Note the different scales in the facet grid referring to different data types and timeframes (30 s vs. 2 min)

**Table 9.1** Assorted descriptive statistics for postural instability

| | | 30 s baseline | 30 s post VR | 0–2 min | 8–10 min |
|---|---|---|---|---|---|
| Mean Euclidean distance | *M* | 0.0003 | 0.001 | 0.002 | 0.003 |
| | SD | 0.0004 | 0.001 | 0.001 | 0.001 |
| Mean acceleration | *M* | 0.583 | 0.375 | 1.708 | 1.002 |
| | SD | 0.291 | 0.291 | 0.789 | 0.376 |
| Frequency of directional changes *z*-direction | *M* | 387.720 | 102.941 | 430.744 | 241.238 |
| | SD | 175.165 | 31.317 | 134.649 | 54.366 |
| Mean duration of movement *z*-direction | *M* | 0.102 | 0.325 | 0.306 | 0.516 |
| | SD | 0.071 | 0.133 | 0.009 | 0.117 |
| Frequency of stops | *M* | | | 60.560 | 30.762 |
| | SD | | | 32.390 | 30.153 |
| Mean duration of stops | *M* | | | 1.516 | 1.022 |
| | SD | | | 1.345 | 0.833 |

$\text{MISC}_{8\ \text{minVR}}$ ($\widehat{p}_{\ 8\ \text{min VR}} = 0.562$) was estimated significantly smaller ($\chi^2_{(275)} = 8.375$, $p < 0.001$) than the $\text{MISC}_{10\ \text{minVR}}$ ($\widehat{p}_{\ 10\ \text{min VR}} = 0.574$).

Additionally, the baseline and post VR comparison for the postural instability measures was run, via a multivariate-repeated measures MANOVA-type analysis [22]. The main effect of the comparisons of the postural instability measures was significant ($\text{MATS}_{(3,167)} = 3517.622$, $p_{BS} < 0.001$). The same applied to the main effect of the baseline and post VR comparison ($\text{MATS}_{(1,167)} = 477.605$, $p_{BS} < 0.001$) and the interaction between these two effects $\text{MATS}_{(3,167)} = 808.399$, $p_{BS} < 0.001$. Nonparametric univariate post-hoc comparisons with a Bonferroni–Holm-adjusted α-level of 0.05 revealed that all measures significantly differed between baseline and post VR.

To compare the temporal dependency of the postural instability during VR exposure the multivariate-repeated measures MANOVA-type analysis was applied once more. Likewise, the main effect of the comparisons of the postural instability measures was significant ($\text{MATS}_{(3,167)} = 31{,}241.204$, $p_{BS} < 0.001$), as well as the main effect of the time spend in VR ($\text{MATS}_{(1,167)} = 394.513$, $p_{BS} < 0.001$) and the interaction between these two effects $\text{MATS}_{(3,167)} = 1110.248$, $p_{BS} < 0.001$. Nonparametric univariate post-hoc comparisons with a Bonferroni–Holm-adjusted α-level of 0.05 revealed that all measures were significantly different in dependence on time. Sequential post-hoc contrasts indicated that especially the first 2–3 occurrences did differ significantly.

To account for the heteroscedasticity a linear panel model with feasible generalized least squares estimators (FGLS) was used to regress the severity of cybersickness [23], the estimators of the model can be found in Table 9.2. The factor subject was included in the model as a random effect, whereas the effect of time was included as a within effect, as the cybersickness ratings do vary over time. The model included the general mean Euclidean distance, mean acceleration, the frequency of directional

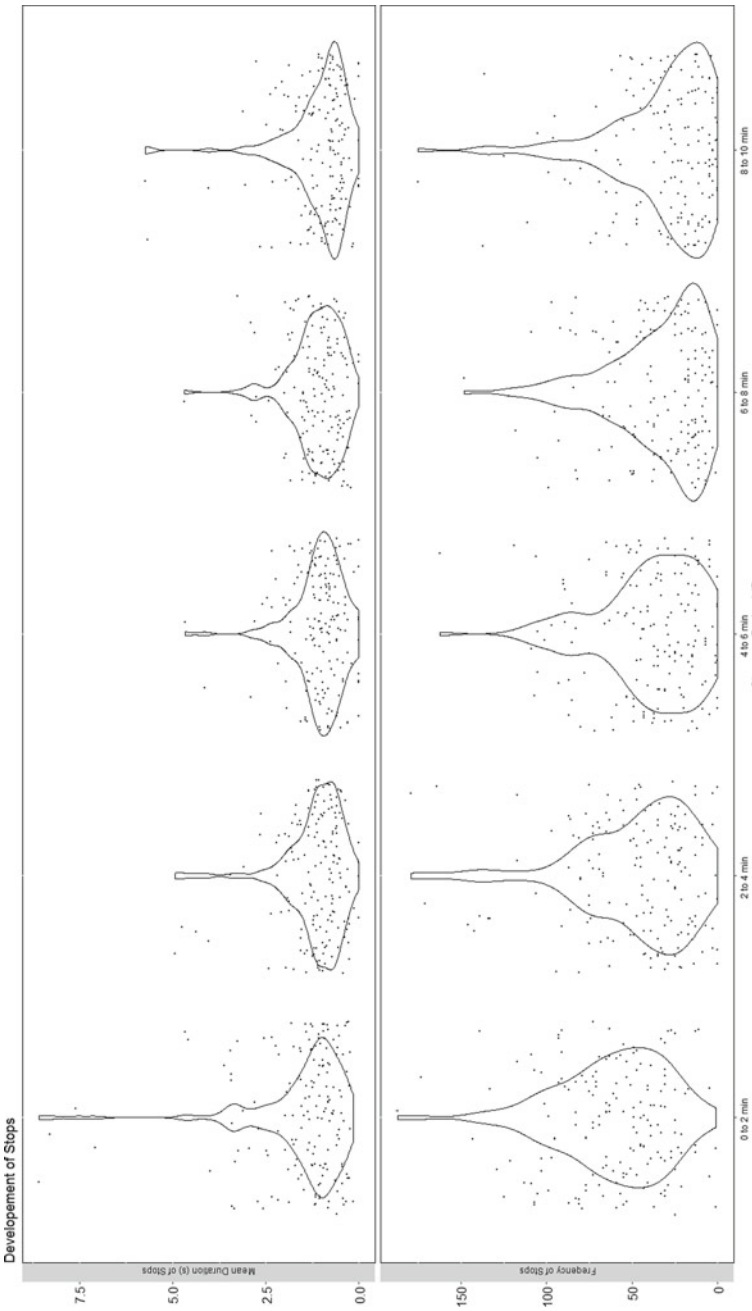



**Fig. 9.4** Mean duration and frequency of stops during VR exposure, violin plots with added jitter are used to account for the non-normality of the data

changes in $z$-direction, and the mean duration of movement in $z$-direction for the baseline post VR comparison. The regressand was the $VRSQ_{\text{Total Score}}$. The implausible results indicated that the model was misspecified ($R^2 = -7.88$ indicating $F_{(4,160)} = -35.496$, $p = 1$, n.s.).

In addition to the predictors used in the baseline and post VR comparison, for the cybersickness ratings during VR exposure the frequency of stops and the mean duration of stops were included as predictors. The regressand was the MISC excluding the $MISC_{\text{Baseline}}$ to enable equal measurement intervals and account for the predictiveness the model was aiming at. The $R^2$ revealed that the model significantly ($F_{(6,806)} = 7.708$, $p < 0.001$) explained 5.4% of the total variability ($R^2 = 0.054$). A summary of the models' regressors and their estimators including the statistics are shown in Table 9.2. The mean Euclidean distance and mean duration of stops are negatively correlated with the upcoming cybersickness rating. In contrast, the acceleration, the duration of movement on the $z$-axis, and the frequency of stops suggest a positive correlation with the upcoming cybersickness rating. It becomes apparent that the interconnections between the regressand and the regressors in the model are not parallel to the descriptive statistics mentioned above.

**Table 9.2** Estimators of FGLS regression models

| | *B* | SE | *z* | *p* |
|---|---|---|---|---|
| *Baseline versus post VR Y: $VRSQ_{Total\ Score}$* | | | | |
| $X_1$: Mean Euclidean distance | 2227.986 | 65,720.897 | 0.034 | 0.973 |
| $X_2$: Mean acceleration | − 94.874 | 142.811 | − 0.664 | 0.507 |
| $X_3$: Frequency of directional changes $z$-direction | 0.127 | 0.200 | 0.634 | 0.526 |
| $X_4$: Mean duration of movement $z$-direction | − 12.044 | 79.419 | − 0.152 | 0.880 |
| *During VR Y: $MISC_{2\ min\ VR - Post\ VR}$* | | | | |
| **$X_1$: Mean Euclidean distance** | **− 210.696** | **18.86** | **− 11.183** | **< 0.001*****|
| **$X_2$: Mean acceleration** | **0.123** | **0.008** | **7.119** | **< 0.001*****|
| $X_3$: Frequency of directional changes $z$-direction | < 0.001 | < 0.001 | 0.903 | 0.366 |
| **$X_4$: Mean duration of movement z-direction** | **1.036** | **0.209** | **4.955** | **< 0.001*****|
| **$X_5$: Frequency of stops** | **< 0.001** | **< 0.001** | **3.610** | **< 0.001*****|
| **$X_6$: Mean duration of stops** | **− 0.002** | **< 0.001** | **− 5.578** | **< 0.001*****|

## 9.4 Discussion

This study aimed at postulating a hmCSDV that predicts the severity of experienced cybersickness based on participants' head-tracking data in combination with a potentially moderating effect of the distinctive controller input. For modeling the data, the non-normality and heteroscedasticity were taken into account. Although the baseline and post VR comparison did not converge to a meaningful model, the predictive model for the time during VR was found to be significant. This model found that shorter mean Euclidean distances, higher mean accelerations, longer mean durations of movement in $z$-direction, and higher frequencies and shorter mean durations of stops were significant predictors.

When applying a fixed interval the reciprocal relation between distance and acceleration can be derived from the formula for acceleration, which is distance divided by time to the power of two. This logic cannot explain the other findings, as they are diametrical to the descriptive statistics. In general, the duration of stops and the frequency of stops significantly decrease over time, but the estimates of the model show that a higher frequency of stops is positively correlated with the upcoming cybersickness rating. It is possible that the gamification of the VR environment was evaluated as an incentive to keep moving and the participants gradually got used to the handling of the equipment, which led to a global effect of less frequent and shorter stops on average. In case of an onset of cybersickness, more frequent stops were seen. From one perspective more stops lead to an experience of a lot of artificial de- and acceleration, but on the other hand being able to stop the virtual movement in your own need, a higher controllability should be experienced, which is found to have a reducing effect on VIMS [24]. Moreover, another subsumption might be an interaction with the duration of the stops: Unless stops are frequent and long, controllability is not experienced. If stops are frequent but also short, it could be argued that these stops are even involuntary and opposed to controllability. Unfortunately, the user experience for the VR application was not assessed, which makes any interpretation highly speculative. Future studies should take an evaluation of the VR application into account.

Looking at the model, the hmCSDV can be seen as a very general approach, as it focuses on the participants' head movement and the start-stop-motion via the distinctive controller input, but overlooks any other influencing factors. Including the users' demographic data could potentially reduce error variance and therefore increase the power of the model. Additionally, it might be useful to merge the model with uncontroversial factors like the soft- and hardware-specific factors of the CSDV [17, 25].

Nevertheless, controversial factors the CSDV takes into account like gender [26, but also see 27], were attentively excluded from the hmCSDV. This decision was based on pre-studies which did not replicate a gender-specific effect of cybersickness, and our data do not support this claim either.[2] Because there is no clear-cut effect

[2] TOST-equivalence tests were run with all individuals that fit into the binary definition of gender. As before, a meaningful difference was defined as a decrement of one point for the MISC and of

of gender on cybersickness susceptibility, we suggest that a predictive model should exclude this factor.

Interestingly the model fit for the hmCSDV was different between the baseline and post VR compared to the during VR exposure model. Both models regress the measures to different operationalizations of cybersickness, it can be criticized that cybersickness is not properly operationalizable. While self-reported cybersickness during VR is only efficiently reportable with a single-item questionnaire, the dimensionality of the symptoms is only accessible via multi-item questionnaires that are inefficient during VR exposure. Although $\text{MISC}_{10\text{ min}}$ and $\text{VRSQ}_{\text{Total Score post VR}}$ do correlate positively to some extent (Spearman's rank correlation $r_{(167)} = 0.53$, $p < 0.001$), it is unsurprising that the MISC is more superficial and tells a slightly shorter story than the VRSQ. However, using the VRSQ during VR exposure would add artificiality to the setup, as the breaks would become longer with more items sampled. This would reduce the immersion and add more variability to the query intervals. Especially if a participant needs to assess the severity of their symptoms, they would need longer to answer compared to a participant with no symptoms, and therefore the total time spent in VR would become incomparable by a systematic error. Additionally, the assessment of the VRSQ can potentially interfere with the experimental task in VR. Particularly studies that focus on physiological indicators and use eye-tracking would—at least for item eight of the VRSQ (dizzy with eyes closed)—add artificial blink events to the data. Nevertheless, the operationalizability argument is just one of the explanations for the differences in model fit.

Moreover, for a prediction of cybersickness the temporal consecutiveness seems necessary. For the VRSQ query, the meta-data was recorded in parallel, therefore the aspect of prediction is lacking. To enable insights into the temporal connection future studies should consider shorter query intervals for a closer-mashed mapping of cybersickness.

Another argument for the differences in model fit is that the baseline and post VR comparison do not representing postural instability due to cybersickness. Instead, it might only be a comparison of postural control without any visual input before and after VR exposure. This brings the measurements known from the center of pressure task to mind [28]. As the key component for cybersickness, the visual input, was missing, it can be argued that not finding any correlation with postural instability and reported cybersickness is in favor of the postural instability theory. Thus, in line with the theory, the visual input is a necessary and sufficient prerequisite for cybersickness, whereas the role of natural postural control for cybersickness is ambiguous [29].

A limitation, which needs to be addressed, is the differing visual input for each participant based on the controller input and head rotation they used. Although it would be possible to replicate the paths taken to explore the city, it would still be

---

one item for the VRSQ. The results for both sides and both cybersickness ratings were significant ($\text{VRSQ}_{\text{post VR}}$ $\Delta_L\, t_{(148.246)} = 4.396, p < 0.001$ $\Delta_U\, t_{(148.246)} = -1.960, p = 0.026$; and $\text{MISC}_{10\text{ min}}$ $\Delta_L\, t_{(154.208)} = 4.914, p < 0.001$ $\Delta_U\, t_{(154.208)} = -2.001, p = 0.023$) which means that no meaningful gender difference has been found.

impossible to have matching visual input for participants as long as there is free exploration. At the same time, this limitation can be seen in the much broader context of naturalistic experimental settings as natural validity in general counteracts controllability. As mentioned in the introduction, we aimed at using a VR environment as application-related as possible. Different studies display highly controlled environments, like a roller coaster ride [30], with almost no interaction—what we regard as the key component of naturalistic VR. In our perception literature on cybersickness often overlooks the characteristics of interaction in VR and is more constrained than the use case. Our setting is a trade against controllability that leads to biased results when applying them in the "real world," but results in a higher error variance and makes random effects more likely.

One big advantage compared to many other VR studies is the decent amount of participants, the naturalistic VR including the use of built-in sensors of the hardware and applied inference statistics that take the multivariate non-normality and right skewness of the data into account. In the future, more advanced algorithms might predict cybersickness via meta-data that can be read out of all current HMD without a big fuss. If implemented as an early-warning mechanism, the hmCSDV could be used to mitigate or even cancel cybersickness out by stopping the VR application before the onset of symptoms. Following up on our method by using only built-in sensors in combination with a cloud connection it would become possible to collect and model the data of users during VR game play at home. Because no sensitive personal data are included in the model, as long as one sticks to the aggregated form of the data, privacy is not threatened. Furthermore, the need for laboratory studies, which are struggling with ecological validity and small sample sizes, could be further reduced.

**Acknowledgements** We acknowledge A. Klingenfuss, I.M. Bundil, K. Holzmeyer, and J. Schöppe for collecting the experimental data.

## References

1. Berencsi, A., Ishihara, M., Imanaka, K.: The functional role of central and peripheral vision in the control of posture. Hum. Mov. Sci. **24**(5–6), 689–709 (2005)
2. Stoffregen, T.A., Smart, L.J.: Postural instability precedes motion sickness. Brain Res. Bull. **47**, 437–448 (1998)
3. LaViola, J.J., Jr.: A discussion of cybersickness in virtual environments. ACM Sigchi. Bull. **32**(1), 47–56 (2000)
4. Reason, J.T., Brand, J.J.: Motion Sickness. Academic Press Inc., 24/28 Oval Road London NW1 (1975)
5. Stanney, K.M., Kennedy, R.S., Drexler, J.M.: Cybersickness is not simulator sickness. Proc. Human Factors Ergon. Soc. Annu. Meet. **41**(2), 1138–1142 (1997)
6. Golding, J.F., Gresty, M.A.: Motion sickness. Curr. Opin. Neurol.Opin. Neurol. **18**, 29–34 (2005)

7. Islam, R., Desai, K., Quarles, J.: Cybersickness prediction from integrated HMD's sensors: a multimodal deep fusion approach using eye-tracking and head-tracking data. In: 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (2021)
8. Jin, W., et al.: Automatic prediction of cybersickness for virtual reality games. In: 2018 IEEE Games, Entertainment, Media Conference (GEM). IEEE (2018)
9. So, R.H., Lo, W., Ho, A.T.: Effects of navigation speed on motion sickness caused by an immersive virtual environment. Hum. Factors **43**(3), 452–461 (2001)
10. Prothero, J., Furness, T.: The role of rest frames in vection, presence and motion sickness (1998)
11. Wienrich, C., et al.: A virtual nose as a rest-frame—the impact on simulator sickness and game experience. In: 2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games) (2018)
12. Servotte, J.-C., et al.: Virtual reality experience: immersion, sense of presence, and cybersickness. Clin. Simul. Nurs.Nurs. **38**, 35–43 (2020)
13. Schuirmann, D.J.: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J. Pharmacokinet. Biopharm.Pharmacokinet. Biopharm. **15**(6), 657–680 (1987)
14. Kim, H.K., et al.: Virtual reality sickness questionnaire (VRSQ): motion sickness measurement index in a virtual reality environment. Appl. Ergon. **69**, 66–73 (2018)
15. Bos, J., Mackinnon, S., Patterson, A.: Motion sickness symptoms in a ship motion simulator: effects of inside, outside and no view. Aviat. Space Environ. Med.. Space Environ. Med. **76**, 1111–1118 (2006)
16. Kuiper, O.X., et al.: Knowing what's coming: unpredictable motion causes more motion sickness. Hum. Factors **62**, 1339–1348 (2020)
17. Rebenitsch, L., Owen, C.: Estimating cybersickness from virtual reality applications. Virtual Reality **25**(1), 165–174 (2021)
18. Diels, C., Howarth, P.A.: Frequency characteristics of visually induced motion sickness. Hum. Factors **55**(3), 595–604 (2013)
19. Noguchi, K., et al.: NparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. J. Stat. Softw.Softw. **50**(12), 1–23 (2012)
20. Friedrich, S., Pauly, M.: MATS: inference for potentially singular and heteroscedastic MANOVA. J. Multivar. Anal.Multivar. Anal. **165**, 166–179 (2018)
21. Konietschke, F., et al.: Nparcomp: an R software package for nonparametric multiple comparisons and simultaneous confidence intervals. J. Stat. Softw.Softw. **64**(9), 1–17 (2015)
22. Friedrich, S., Konietschke, F., Pauly, M.: Resampling-based analysis of multivariate data and repeated measures designs with the R package MANOVA.RM. R J., 11, p. 380 (2019)
23. Croissant, Y., Millo, G.: Panel data econometrics in R: the PLM package. J. Stat. Softw.Softw. **27**(2), 1–43 (2008)
24. Dong, X., Yoshida, K., Stoffregen, T.A.: Control of a virtual vehicle influences postural activity and motion sickness. J. Exp. Psychol. Appl. **17**, 128–138 (2011)
25. Chen, W., Yuen, S.L., So, R.H.Y.: A progress report on the quest to establish a cybersickness dose value. Proc. Human Factors Ergon. Soc. Annu. Meet. **46**(26), 2119–2123 (2002)
26. Munafo, J., Diedrick, M., Stoffregen, T.A.: The virtual reality head-mounted display oculus rift induces motion sickness and is sexist in its effects. Exp. Brain Res. **235**(3), 889–901 (2017)
27. Grassini, S., Laumann, K.: Are modern head-mounted displays sexist? A systematic review on gender differences in HMD-mediated virtual reality. Front. Psychol. **11**(1604) (2020)
28. Doyle, R.J., et al.: Generalizability of center of pressure measures of quiet standing. Gait Posture **25**(2), 166–171 (2007)
29. Petel, A., et al.: Motion sickness susceptibility and visually induced motion sickness as diagnostic signs in Parkinson's disease. Euro. J. Transl. Myol. **32**(4) (2022)
30. Stanney, K., Fidopiastis, C., Foster, L.: Virtual reality is sexist: but it does not have to be. Front. Robot. AI **7**(4) (2020)

# Chapter 10
# Measuring Audio-Visual Latencies in Virtual Reality Systems

**Victoria Fucci, Jinqi Liu, Yunjia You, and Raymond H. Cuijpers**

**Abstract** In virtual reality (VR) systems, delays may occur while the signal passes through hardware and software components, thus causing asynchrony or even cybersickness, as a result. To better understand and control the role of delays in VR experiments, we tested an accessible method for measuring audio and visual end-to-end latency between game engines (Unity and Unreal), head-mounted displays (HMD) (Oculus Rift and Oculus Quest 2) and vertical synchronisation (V-sync) setting (on/off). The measuring setup consisted of the microcontroller, a dedicated serial port, a microphone, a light sensor and an oscilloscope. The measurements showed that Unreal Engine with Oculus Rift had ≈ 16 ms less visual delay and ≈ 33 ms less audio delay than Oculus Quest 2. The Unity Engine with Oculus Rift had ≈ 22 ms less visual delay and ≈ 39 ms less audio delay than Oculus Quest 2. These values may differ between systems, but they are above the discrimination thresholds. No differences were found for the V-sync on/off parameter. Compared to the Unity Engine, the Unreal Engine showed much lower visual latency performance and significantly lower audio latency. In addition, Oculus Rift's audio and visual latency performance had lower delays than Oculus Quest 2; therefore, using Oculus Rift is advisable in VR research where lower latencies are essential, even though Oculus Quest 2 is a newer version of the HMD. Our approach provides a convenient way to measure audio and visual end-to-end latency in VR without a strong engineering background.

## 10.1 Introduction

Over the last decade, virtual reality (VR) technology has rapidly developed and is widely used in research and application areas. In addition, it allowed us to create virtual scenarios that are too difficult or expensive to achieve in the physical world. Although head-mounted displays (HMD) have become more advanced since the first prototype [1], there are still some challenges to address when using HMDs, such as the audio and visual latency of the hardware system.

V. Fucci (✉) · J. Liu · Y. You · R. H. Cuijpers
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: victoria.k.fucci@gmail.com



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_10

Latency generally refers to a lag somewhere in the system. It is determined by measuring the time difference between two locations as the signal passes through the "system" components [2]. Among other things, system latency can be considered as the accumulation of delays from the generation of the signal to the arrival of all components in the visual or auditory system of the person. The term end-to-end latency ("motion-to-photons") is often used to define the time delay between the motion signal and the corresponding display output [2]. Specifically, the motion-to-photon delay is the total time between the movement of the user's head and the display emitting photons for the updated stereoscopic images that reflect that movement [3]. The motion-to-photon latency usually includes sampling the head tracking sensors, combining the sensor fusion, rendering the stereoscopic images, the display reading from the frame buffer and the display emitting photons [4].

While the visual latency is focused on kinematic inputs and visual outputs, another type of delay is more relevant to audio inputs and outputs, which was noted as mouth-to-ear latency by Becher et al. [2] or end-to-end audio latency. This latency occurs when there is a time delay between the generation and reception of audio signals in the system. Since it is based on a human communication framework, end-to-end audio latency is more common in remotely shared virtual environments, where users may find it challenging to get timely feedback from others. Therefore, low audio latency is also a significant concern for designing natural communication in VR systems.

Although the general goal of designing the VR system is to reduce the overall latencies, some are unavoidable as these latencies are inherent features of hardware and software that generate the virtual environment (VE) [5]. For example, to deal with a signal, the system needs time to go through a sequential procedure, including recording, identifying, translating and responding appropriately. Each step requires the joint performance of hardware and software. Furthermore, if the VE is shared online, the network condition will also play a role in contributing to the overall latencies. Thus, whether the overall delay is noticeable depends on the cumulative effect of the multiple delays arising from the VR system. From this point of view, instead of eliminating those latencies embedded in the VR system, mediating controllable latency might be a more appropriate solution.

As a criterion for the quality of the VR experience and an inherent property of the VR system, latency impacts various aspects. A possible impact is a decreasing sense of "presence"—the feeling of "being in the virtual environment". This is contrary to the primary role of VR, which is to build realistic virtual scenes. Thus, minimising VR latency is essential for generating perceptual experiences close to physical reality [6]. Furthermore, latency affects subjective experience and the presence of VR scenarios and task performance in physical interactions or collaborative tasks [7, 8]. In addition, significant delays over 64 ms can provoke cybersickness in VR applications [7]. Cybersickness is closely related to simulator sickness and motion sickness. Generally speaking, cybersickness is defined by specific adverse symptoms induced by VR or augmented reality (AR) applications that do not apply external forces to the user. The negative symptoms include disorientation, apathy, fatigue, dizziness,

headache, increased salivation, dry mouth, difficulty focusing, eye strain, vomiting, pallor, sweating and postural instability [9].

The degree to which latency could be noticed is related to its length. Typically, users appear sensitive to differences in latency of approximately 15 ms [10]. However, there is still discussion about this criterion. The performance might be worse, or at least not better, at very low latency, possibly explained by the behaviour of the human motor system having inherent latencies [7, 11].

No standardised method for measuring audio and visual end-to-end latencies in VR exists. Many methods require expensive experimental equipment and do not have broad applicability. Finally, most methods for measuring VR end-to-end latency are complicated to operate and difficult to apply. Furthermore, they require a high level of expertise in relevant devices, thus preventing researchers without a technical background from replicating the measurements.

Out of the above considerations, this study aims to contribute to the measurement methods of audio and visual end-to-end latencies in VR by adapting and modifying the method by Raaen and Kjellmo [12], specifically, using two HMDs (Oculus Rift and Oculus Quest 2) and two most popular game engines (Unity and Unreal) to explore the difference in latency between several cross configurations, including vertical synchronisation mode (on/off). In addition, we intend to find and recommend the configuration with the best (lowest latencies) based on our results.

## 10.2 Related Work

Depending on the type of latency and lab conditions, there are various ways to measure latency. For example, the end-to-end visual latency in immersive virtual reality (IVR) refers to the time delay between a user's action and when this action is visible on the HMD [13]. The end-to-end visual latency usually includes sampling the head tracking sensors, combining the sensor fusion, rendering the stereoscopic images, the display reading from the frame buffer and the display emitting photons [4]. The high-speed camera is one of the most frequently used equipments for end-to-end latency measurements, which allows direct observation and frame-by-frame comparison. For example, in the study of Gruen et al. [10], two cameras were synchronised to capture multiple pictures of the clock running simultaneously by the Arduino board. While one camera observed the time through the HMD, the other observed the time directly. Therefore, the latency was represented by the difference between the two observations.

Kijima and Miyajima [14] employed a different strategy using two cameras. They placed the optical centres of two vertically offset cameras at the rotation axis of a turntable. The upper camera aimed at the real visual target while the lower camera aimed at the virtual visual target through HMD in front of it. With the turntable's rotation, the trajectories of two targets were obtained. The average visual end-to-end latencies were calculated based on the evaluation of angular trajectories of the real and virtual targets.

While the high-speed camera is a handy tool for latency measurement, researchers may have other alternatives. For example, Becher et al. [2] replaced cameras with photodiodes and a potentiometer as external references when comparing the real motion and the corresponding display output of the object. Noticeably, all components functioned by the microcontroller, which helped to simplify the procedure and avoid unnecessary errors.

In the study of Raaen and Kjellmo [12], latency was aroused by creating an abrupt change in a simulated environment while changing the vertical synchronisation (V-sync) parameter. The main setup consisted of the HMD, laser pen, light sensor and oscilloscope. While one light sensor was attached to the device's screen to capture the change in the VE, the other was attached to pick up the laser. Once the VR device was moved, the light sensor illuminated by the laser pen registered the disappearance of the light. This method measured the actual delay from the beginning of the head movement to the screen update of the VR device. The study found a significant effect of V-sync on a total delay and that V-sync on/off mode will result in different frame rates. Additionally, a recent study by Pape et al. [15] improved this method by replacing the oscilloscope with a microcontroller, thus making the setup more portable, affordable and wirelessly controlled.

We found a limited amount of relevant studies and measurement practices for end-to-end audio latency. As one of the few examples, Becher et al. [2] used a piezo buzzer to generate the sound signal and a microphone attached to the headset to detect an audio signal. The audio latency was then represented as the elapsed time from the activation of the buzzer until a specific sound pressure threshold was reached.

## 10.3 Experiment Setup

### *10.3.1 Apparatus*

The measurements took place at Eindhoven University Game Lab. For latency measurement, a simple setup was introduced, which included the VR HMDs (Oculus Rift and Oculus Quest 2 with Link cable), Arduino Uno, VINT Hub Phidget, SparkFun Sound Detector, light sensor and TBS1102B Tektronix digital oscilloscope as shown in Fig. 10.1. Specifically, the Arduino board was applied as a 5-voltage power supply for the sound detector and the light sensor. Also, a resistor was connected in series to the circuit of a light sensor according to the instruction in Fig. 10.2.

The Oculus Rift had two AMOLED panels with a $1080 \times 1200$ resolution running at 90 Hz with a $\approx 110°$ horizontal Field-of-View (FOV) and a $\approx 90°$ vertical FOV. The Oculus Quest 2 had a single fast switch LCD $1832 \times 1920$ resolution running at 120 Hz with a $\approx 92°$ horizontal Field-of-View (FOV) and an $\approx 89°$ vertical FOV. The PC ran Windows 10 on an Intel® Core™ i9-9900 K processor, with the NVIDIA® Titan RTX™ graphics card and a Realtek® LC1220P-VB2 sound card.

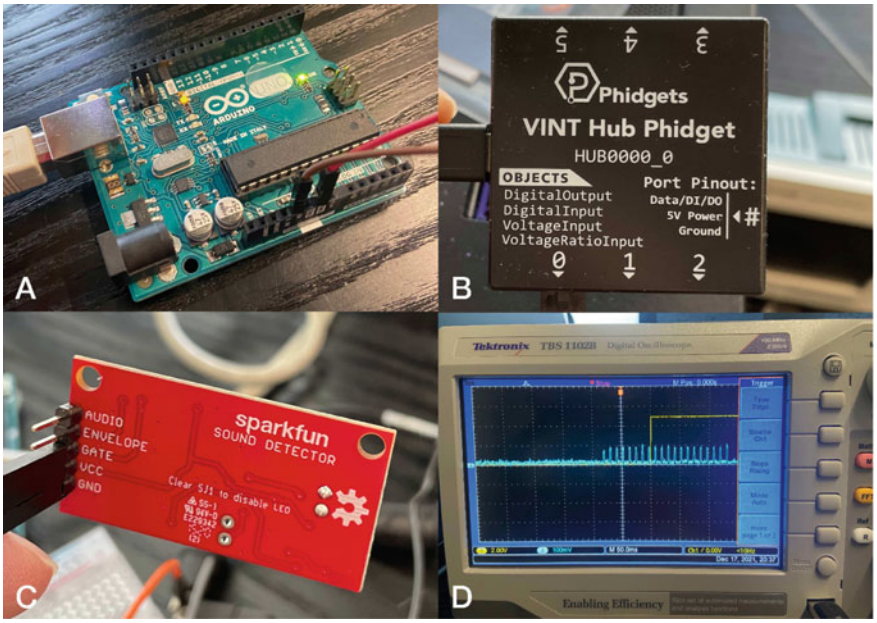



**Fig. 10.1** Apparatus with Arduino Uno board (**a**), VINT Hub Phidget (**b**), SparkFun sound detector (**c**) and oscilloscope (**d**)

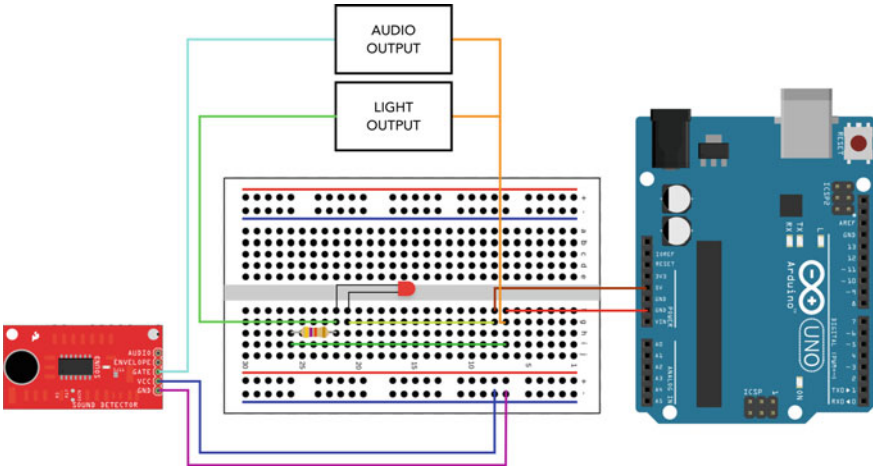



**Fig. 10.2** Breadboard wiring visualisation of Arduino Uno, light sensor and Sparkfun sound detector

Instead of using the camera or the microcontroller as an objective reference to latency, we used VINT Hub Phidget, an I/O board equipped with digital inputs and outputs, thus allowing real-time data exchange with the computer station. Furthermore, the sound detection system was instrumented with multiple ports for different uses, one of which was to detect whether a sound was above a certain threshold (GATE), and the other was to show the varying pitch of the sound (ENVELOPE). In this case, the former port was more applicable as we were only interested in the binary judgement of the audio output.

Once the physical setup was complete, two identical warehouse 3D models were exported into the projects for measurements via Unity and Unreal game engines [16, 17]. The visual representation of the 3D model did not play any significant role in the measurements. However, this 3D model was chosen to measure audio and visual latencies for this specific setup. In addition, the Phidget library software interface was imported as an asset/plugin into each game engine to use the Phidget functions [16]. The Phidgets Inc. provides installer for the software interface on their webpage [18]. Finally, the Unreal Engine plugin was created manually by transforming Unity's C# programming language into the C++ programming language used in Unreal following the plugins documentation [17].

To minimise GPU processing time and to ensure the change is sharp enough for the external sensors to capture, a grounded dark plane in front of a simulated button was placed, which only illuminated itself when the space bar was pressed. This would trigger a short audio burst and the illumination of the plane, both lasting 200 ms. For the consistency between the two models, both models applied the same fixed first-person perspective in VR previewing mode (see Fig. 10.3). Noticeably, the infrared sensor between the two lenses of the HMD was covered with a lens cleaning cloth to remain active without wearing.

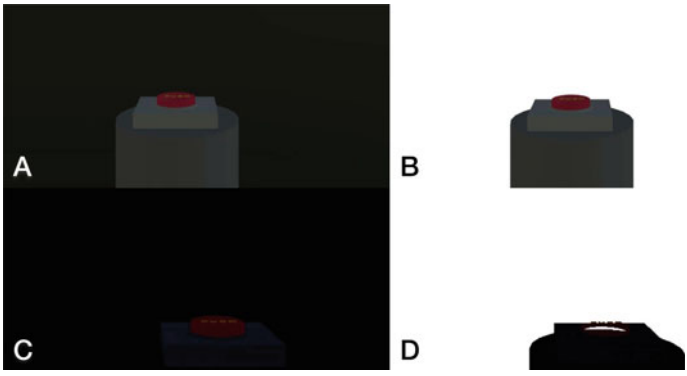



**Fig. 10.3** First-person visual perspective in Unity Engine (**a**, **b**) and Unreal Engine (**c**, **d**) in darkened (**a**, **c**) and illuminated (**b**, **d**) virtual environment

### *10.3.2 Procedure*

The main goal was to stimulate sudden changes in visual (on/off) and audio (on/off) events by pressing the space bar. Once the key was pressed (signal sent), the Phidget would store the signal through its installed library [16], corresponding to the game engine program, which was marked as the beginning of the input-to-output loop. Then the trigger signal was sent to the oscilloscope and graphically displayed as varying signal voltages. Meanwhile, the flash and burst sound following the action of the pressed keyboard would be sensed by the light sensor and the sound detector, which were mounted beside the lens and the speaker of the HMD, as displayed in Fig. 10.4.

Through the connection to the oscilloscope probe, light and sound output were displayed as a 2D waveform with $X$-axis representing time starting from 0 and $Y$-axis representing voltage. Therefore, the end of the input-to-output loop was the $X$-coordinate corresponding to the start of the first peak in the waveform, which was also the moment when the sensors perceived the light and the sound. For each configuration, ten measurements were done, resulting in 80 datasets for visual and audio end-to-end latency. All the data per frame displayed on the oscilloscope could be retrieved from a built-in USB port and exported as 2D coordinates in the CSV file. After, the visual and audio latency could be measured by calculating the distance between the corresponding $X$-coordinates, as shown in Fig. 10.5.

## 10.4 Results

Tables 10.1 and 10.2 represent measurement results (minimum (Min.), maximum (Max.) and average (Avg.) values) of visual and audio end-to-end latency for Oculus Rift and Oculus Quest 2 HMDs, Unreal and Unity game engines and V-sync mode on/off. For the end-to-end visual latency, Oculus Rift had the lowest average latency in the Unreal Engine with both V-sync modes (on/off) $\approx$ 29 ms. In comparison, Oculus

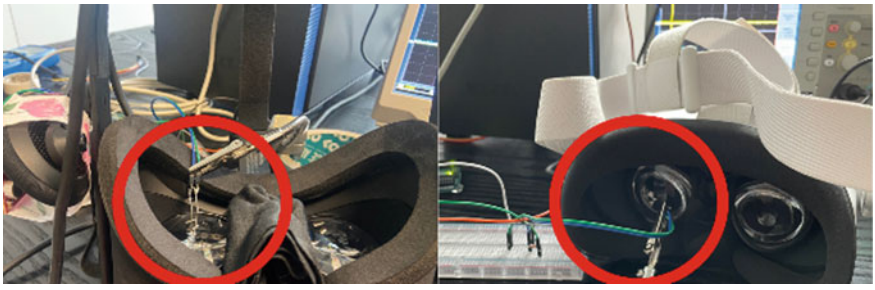


**Fig. 10.4** Measurement setup of visual latency for Oculus Rift (left image) and Oculus Quest 2 (right image). The light sensor is inside the red circle

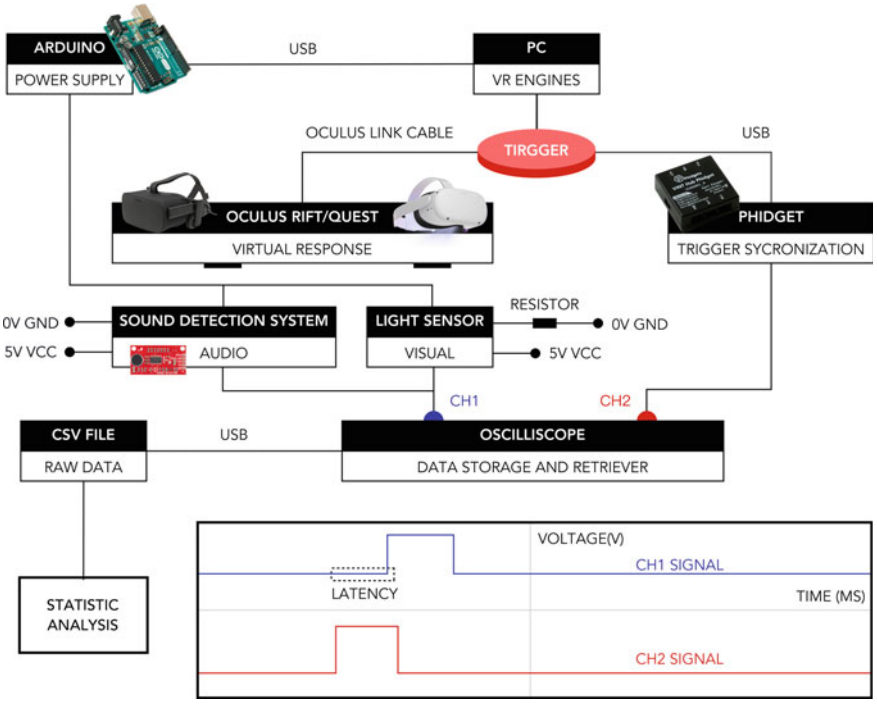



**Fig. 10.5** Schematic diagram of the measurement procedure

Quest 2 had the highest average latency in the Unity Engine with both V-sync modes on/off ≈ 68 ms.

The lowest end-to-end audio latency value was observed using Oculus Rift and Unreal Engine with V-sync on ≈ 145 ms. On the other hand, the highest latency value was observed using Oculus Quest 2 and Unity Engine with V-sync off ≈ 190 ms.

**Table 10.1** Results from visual delay measurements

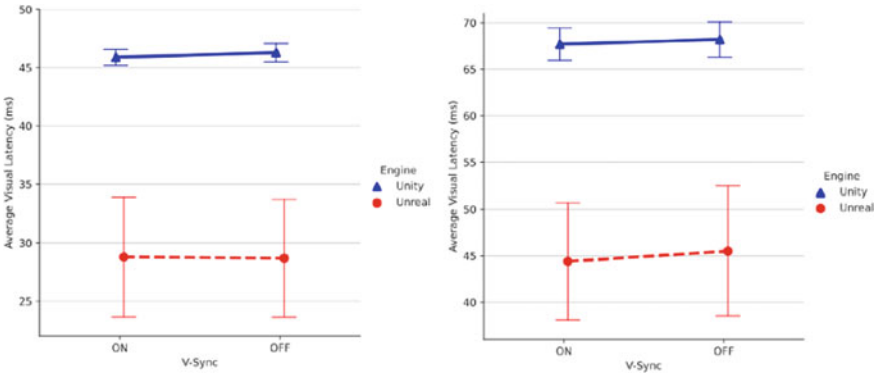
| VR display | Game Engine | V-sync | Min. (ms) | Max. (ms) | Avg. (ms) |
|---|---|---|---|---|---|
| Oculus Rift | Unity | On | 45 | 47 | 46 |
| Oculus Rift | Unity | Off | 45 | 47 | 46 |
| Oculus Quest 2 | Unity | On | 66 | 70 | 68 |
| Oculus Quest 2 | Unity | Off | 66 | 72 | 68 |
| Oculus Rift | Unreal | On | 21 | 32 | 29 |
| Oculus Rift | Unreal | Off | 21 | 32 | 29 |
| Oculus Quest 2 | Unreal | On | 40 | 54 | 44 |
| Oculus Quest 2 | Unreal | Off | 39 | 54 | 46 |

**Table 10.2** Results from audio delay measurements

| VR display | Game Engine | V-sync | Min. (ms) | Max. (ms) | Avg. (ms) |
|---|---|---|---|---|---|
| Oculus Rift | Unity | On | 142 | 158 | 150 |
| Oculus Rift | Unity | Off | 139 | 157 | 150 |
| Oculus Quest 2 | Unity | On | 174 | 198 | 188 |
| Oculus Quest 2 | Unity | Off | 175 | 201 | 190 |
| Oculus Rift | Unreal | On | 133 | 153 | 145 |
| Oculus Rift | Unreal | Off | 137 | 155 | 145 |
| Oculus Quest 2 | Unreal | On | 161 | 186 | 176 |
| Oculus Quest 2 | Unreal | Off | 169 | 192 | 180 |

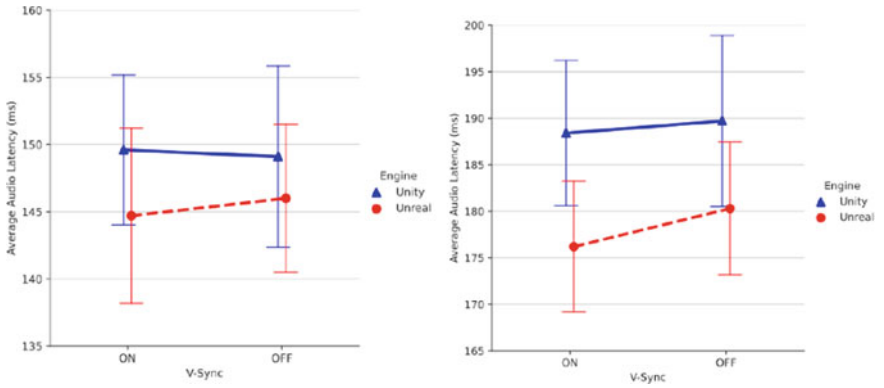
We performed a cross-contrast analysis of the average visual end-to-end latencies for different configurations (Fig. 6a, b). The results of an independent $t$-test with a 95% confidence interval (CI) for different cross configurations have shown a strong statistically significant difference between Oculus Rift and Oculus Quest 2 for Unreal Engine ($t(18) = 5.645$, $p \leq 0.0001$), a strong significant difference between Oculus Rift and Oculus Quest 2 for Unity Engine ($t(18) = 33.483$, $p \leq 0.001$), an overall highly significant difference between Oculus Rift and Oculus Quest 2 ($t(18) = 11.14$, $p \leq 0.0001$) and overall highly significant difference between Unreal and Unity Engines ($t(18) = 10.104$, $p \leq 0.0001$). However, no differences were found between V-sync modes on/off ($t(18) = 0.283$, $p = 0.781$).

The cross-contrast analysis of the average audio end-to-end latencies for different configurations is shown in Fig. 7a, b. The results of an independent $t$-test with a 95% confidence interval (CI) for different cross configurations show strong statistically




(a) Oculus Rift (b) Oculus Quest 2

**Fig. 10.6** Cross-contrast for visual latency measurements. The solid blue line with blue triangles (V-sync on/off) represents visual latency values for Unity engine. The red dashed line with red triangles (V-sync on/off) represents visual latency values for Unreal Engine. The error bars indicate 95% confidence intervals (CI)




(a) Oculus Rift (b) Oculus Quest 2

**Fig. 10.7** Cross-contrast for audio latency measurements. The solid blue line with blue triangles (V-sync on/off) represents audio latency values for Unity Engine. The red dashed line with red triangles (V-sync on/off) represents audio latency values for Unreal Engine. The error bars indicate 95% confidence intervals (CI)

significant difference between Oculus Rift and Oculus Quest 2 for Unreal Engine ($t(18) = 10.624$, $p \leq 0.0001$), a strong significant difference between Oculus Rift and Oculus Quest 2 for Unity Engine ($t(18) = 11.361$, $p \leq 0.0001$), an overall highly significant difference between Oculus Rift and Oculus Quest 2 ($t(18) = 11.02$, $p \leq 0.0001$) and overall significant difference between Unreal and Unity Engines ($t(18) = 2.26$, $p \leq 0.05$). Similarly to visual latency measurements, no differences were found between V-sync modes on/off ($t(18) = 0.474$, $p = 0.641$).

## 10.5 Discussion

This study investigated virtual reality systems' audio and visual latency measurement methods. In contrast to the previous research by Raaen and Kjellmo [12], we found no significant difference between V-sync modes on/off for audio or visual delay measurements which might be because our system configuration and computer parameters differ significantly from their study [12] conducted in 2015. However, it is clear from Tables 10.1 and 10.2 that visual end-to-end latency is considerably lower than audio end-to-end latency.

We found significant differences between HMDs and game engines configurations for audio or visual delay measurements. On average, Oculus Rift's audio and visual latency performance had significantly lower delay values than Oculus Quest 2. Therefore, Oculus Rift is more advisable in VR research where lower latencies are essential (e.g. audio-visual timing), even if the Oculus Quest 2 is more powerful as a newer VR HMD. Since part of VR system latency is inherited through hardware

performance such as signal transferring and graphic display, a possible difference can exist between different generations of VR products.

In addition, the audio and visual latency performance of Unreal Engine had significantly lower delay values in comparison to Unity Engine. Most likely, the reason can be caused by a difference in the programming languages. While Unity's modelling process relies entirely on C#-based programming, Unreal Engine applies C++ as its underlying programming language, which is faster than C#. Unreal Engine also provides an alternative, a blueprint, for those with limited programming experience. In the blueprint, the developer can quickly call wrapped-up user packages through command boxes and intuitively connect them to realise the same function as in the programming working context.

Our system could detect visual and sound delays compared to the previous studies on end-to-end latency measurements. Most earlier solutions could only measure visual or auditory latency alone. However, in the VE, the richer and more realistic the senses are, the more likely the user is to be immersed in the virtual environment. Therefore, measuring multimodal latencies aligns with the actual user experience in VR environments.

### *10.5.1 Limitations*

Despite the contributions, there were also limitations in measurement setup and analysis. For example, although the initial plan was to measure visual and audio end-to-end latency at the same time by retrieving the data from channel 1 (CH1), channel 2 (CH2) and the external trigger channel, it was found that only data from two channels out of three (CH1 and CH2) could be saved at the same time. Thus, we separated the measurements for visual and audio end-to-end latency, with CH1 corresponding to the visual or sound signal and CH2 corresponding to the Phidget signal. We also employed different scaling for light and sound signals to make the first peak observable in the waveform ($X$-axis unit: 100 ms, $Y$-axis unit: 200 mV for light, 2 V for sound and trigger). A more advanced oscilloscope with more channels would allow simultaneous measurement of audio and visual latencies in the future.

Also, due to the configuration differences between HMDs, the light sensor and the sound detector positions were very close but not completely identical. Another factor is the difference between Game Engines in illumination rendering (Unity has a slightly higher brightness of the space than Unreal). However, these differences did not affect the signal detection of the light sensor or sound detector at any time.

## 10.6 Conclusion

This study provided a convenient way to measure (simultaneously or separately) audio and visual end-to-end latency in VR without a strong engineering background using affordable and easy-to-obtain measuring equipment. We introduced this measuring method to simplify quantification and raise awareness of the importance of hardware latency reduction. The latency reduction can improve the efficiency of information exchange between the user and the system, thus enhancing QoE. In the future, we expect to see more elaborated test setups for VR end-to-end latency measurement emerging among other modalities. Our study brings a more comprehensive understanding of hardware latency's role in VR scenarios and contributes to the effective end-to-end latency quantification.

**Acknowledgements** We want to thank the late Armin Kohlrausch for his supervision and considerable inspiration that made this work possible. We also thank the Lab Support Team of the Human-Technology Interaction group at the Eindhoven University of Technology for their continuous assistance in this project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 812719.

## References

1. Basu, A.: A Brief Chronology of Virtual Reality. arXiv:1911.09605 (2019)
2. Becher, A., Angerer, J., Grauschopf, T.: Novel approach to measure motion-to-photon and mouth-to-ear latency in distributed virtual reality systems. arXiv:1809.06320 (2018)
3. van Waveren, J.M.P.: The asynchronous time warp for virtual reality on consumer hardware. In: Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology, pp. 37–46. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2993369.2993375
4. Mine, M., Bishop, G.: Just-in-Time Pixels. Technical Report TR93-005. University of North Carolina, Chapel Hill, NC (1993)
5. Gunn, C., Hutchins, M., Adcock, M.: Combating latency in haptic collaborative virtual environments. Presence Teleoper. Virtual Environ. **14**, 313–328 (2005). https://doi.org/10.1162/105474605323384663
6. Allison, R.S., Harris, L.R., Jenkin, M., Jasiobedzka, U., Zacher, J.E.: Tolerance of temporal delay in virtual environments. In: Proceedings IEEE Virtual Reality 2001, pp. 247–254. IEEE (2001). https://doi.org/10.1109/VR.2001.913793
7. Caserman, P., Martinussen, M., Göbel, S.: Effects of end-to-end latency on user experience and performance in immersive virtual reality applications. In: 1st Joint International Conference on Entertainment Computing and Serious Games (ICEC-JCSG), vol. LNCS-11863, pp. 57–69. Springer International Publishing, Arequipa, Peru (2019). https://inria.hal.science/hal-03652010
8. Feldstein, I.T., Ellis, S.R.: A simple video-based technique for measuring latency in virtual reality or teleoperation. IEEE Trans. Visual. Comput. Graph. **27**, 3611–3625 (2021). https://doi.org/10.1109/TVCG.2020.2980527
9. Stauffert, J.-P., Niebling, F., Latoschik, M.E.: Latency and cybersickness: impact, causes, and measures. A review. Front. Virtual Real. Front. (2020). https://doi.org/10.3389/frvir.2020.582204

10. Gruen, R., Ofek, E., Steed, A., Gal, R., Sinclair, M., Gonzalez-Franco, M.: Measuring system visual latency through cognitive latency on video see-through AR devices. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE (2020). https://doi.org/10.1109/VR46266.2020.00103
11. Kadowaki, T., Maruyama, M., Hayakawa, T., Matsuzawa, N., Iwasaki, K., Ishikawa, M.: Effects of low video latency between visual information and physical sensation in immersive environments. In: Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology. ACM (2018). https://doi.org/10.1145/3281505.3281609
12. Raaen, K., Kjellmo, I.: Measuring latency in virtual reality systems. In: Entertainment Computing—ICEC 2015, pp. 457–462. Springer (2015). https://doi.org/10.1007/978-3-319-24589-8_40
13. Caserman, P., Martinussen, M., Göbel, S.: Effects of end-to-end latency on user experience and performance in immersive virtual reality applications. In: Entertainment Computing and Serious Games, pp. 57–69. Springer (2019). https://doi.org/10.1007/978-3-030-34644-7_5
14. Kijima, R., Miyajima, K.: Measurement of head-mounted display's latency in rotation and side effect caused by lag compensation by simultaneous observation—an example result using oculus rift DK2. In: 2016 IEEE Virtual Reality (VR). IEEE (2016). https://doi.org/10.1109/VR.2016.7504724
15. Pape, S., Kruger, M., Muller, J., Kuhlen, T.W.: Calibratio: a small, low-cost, fully automated motion-to-photon measurement device. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). IEEE (2020). https://doi.org/10.1109/VRW50115.2020.00050
16. Phidgets: Using Phidgets with Unity, Full Examples. Phidgets Project. https://www.phidgets.com/?view=articles&article=UsingPhidgetsWithUnity (2019)
17. Plugins: How to create unreal engine plugins. Unreal Engine 4.27 Documentation. https://docs.unrealengine.com/4.27/en-US/ProductionPipelines/Plugins/
18. Vint Hub Phidget—HUB0000_0 at Phidgets (n.d.). https://www.phidgets.com/?&prodid=643

# Chapter 11
# Development of Virtual CNC Turning Application


**Somkiat Tangjitsitcharoen**

**Abstract** This paper presents a development of virtual Computer Numerical Control (CNC) turning for training and replacing the laboratory classes involving the machining processes in order to enhance the user experience in the field of CNC turning. This research utilizes the advantages of the virtual world for the user to interact with CNC turning machine and to practice the CNC turning processes. A development of virtual reality application contains the key academic concepts taught during CNC machining classes. By wearing a virtual reality headset and using handheld controllers, the content available will allow the user to interact with and inspect the workshop environment, which is modeled after the real world, where a series of turning processes and demonstrations will appear to guide the user. The user can interact with objects in the scene freely with the tracking capability enabled by the two handheld controllers. With these valuable assets available, the application of virtual CNC turning is considered a new generation of the education system. The application of virtual CNC turning can be used in the educational field, training scenario, or manufacturing practice.

## 11.1 Introduction

As the digital twin has been utilized to represent the real manufacturing in the term of virtual manufacturing. Moreover, with the impact of the COVID-19 pandemic, an improvement in virtual education and laboratory is in high demand. Another point of concern is for practices which require specialized equipment and skilled certified instructors to conduct a session. Accessibility in rural institutions is also a concern for practices which consume energy, material, or both to be conducted. Machining processes which deal in the heavy operation can also post physical risks to the trainees and their surroundings. It is beneficial to explore and design a new way that students and trainees can undergo machining training. Generally, Computer

S. Tangjitsitcharoen (✉)
Department of Industrial Engineering, Faculty of Engineering, Chulalongkorn University, Phayathai Road, Patumwan, Bangkok 10330, Thailand
e-mail: somkiat.ta@eng.chula.ac.th



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_11

Numerical Control (CNC) turning [1–3] is one of the most important processes that requires to produce the mechanical parts for automotive parts, injection molds, spindle motors, and aerospace engines. CNC turning was used to perform a wider variety of manufacturing tasks with a large volume and greater accuracy.

However, the surface roughness and straightness models obtained from the previous researches [2, 3] of author will be utilized in the future development to calculate the machined surface roughness and straightness. Moreover, the cross-sectional curvature and shape (inner and outer) of a cylindrical surface have an effect on the roughness parameter. Hence, the surfaces were machined at practically the same cutting speed when performing the final processing step [4]. Therefore, this research involves the development of virtual reality (VR) applications emphasizing the execution of the CNC turning process.

It is widely accepted that virtual reality systems are more suited to educational applications than augmented reality (AR) and mixed reality (MR) systems, particularly if the content involves extensive visual presentation and interaction [5]. However, the majority perception of the public still perceives its usage as limited to the entertainment and video game industries [6]. The virtual reality application has also been implemented into the educational sector, by training for specific and difficult to replicate situations, which allows the users to conduct the training in a safe environment [7].

A common virtual reality system setup consists of the headset, the controllers, and a tracking sensor system. The headset, which is designed to be strapped onto the user head, provides display outputs based on the rendered scenes. The controllers, which usually consist of two handheld controllers, provide the user with means of interacting with the virtual world, capturing and simulating the movement of the user hands and fingers [8]. The tracking sensor system provides the relational positioning of the headset and the controllers, which maps their movements within the real world onto the virtual world.

The basic features of a VR training system for Computer Numerical Control (CNC) are designed and implemented based on the World Tool Kit (WTK) software to support the interactive training for workpiece machining [9]. A simulation platform of automatic CNC loading and unloading production line is developed to assist students in classroom learning. After the simulation, the code can be transferred to the corresponding real device. The results show that the combination of virtual simulation and practical verification can optimize the teaching resources, improve the teaching effect, and improve the teaching quality [10].

Hence, this research aims to develop an application of virtual CNC turning processes. The students can have extremely beneficial experience to practice CNC turning skills including the inability of arranging an in-person workshop class due to social distancing guidelines and availability of the CNC turning machine.

## 11.2 Virtual Reality System for CNC Turning Machine

One of the most popular virtual reality systems is the Oculus Quest product [11, 12], which is a virtual reality headset using the inside-out tracking methodology. The system consists of a headset, which is designed to be mounted on the head of the user, and two handheld controllers as shown in Fig. 11.1, which is adopted to develop the virtual CNC turning application. The Unity program is adopted as the framework with the Oculus Integration SDK [13, 14] as shown in Fig. 11.2.


**Fig. 11.1** Hardware package of Oculus Quest

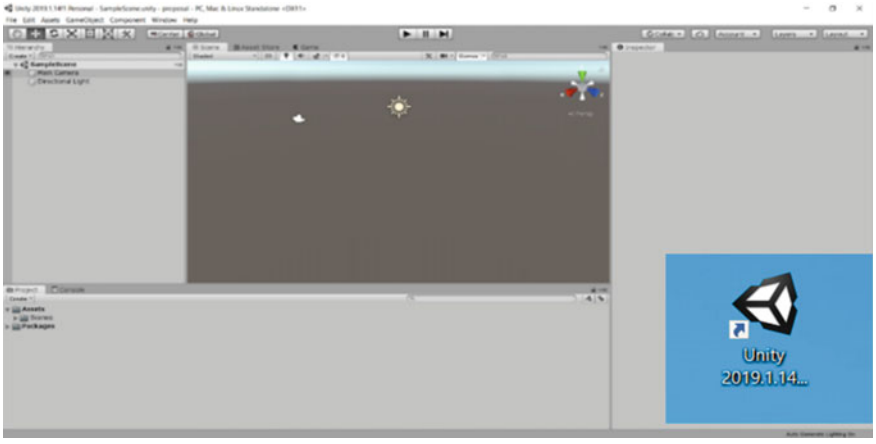



**Fig. 11.2** Illustration of Unity program

However, a computing system consists of three main components, the computing unit, the output device, and the input device, where the input capturing device translates user input in the physical world and communicates the input measurement into the virtual world. The computing unit hence computes the information and the appropriate graphical and auditory output to the headset graphical display and audio system.

The following tools are conducted to develop the VR CNC turning application, which are Unity 3D game engine, 3D modeling software, C# programming languages, video/sound production system, UX/UI, and a feature driven development method. However, the quality of the object might not be as desired. The 3D modeling is obligatory for customized experiences [15, 16].

## 11.3 Development of Virtual CNC Turning Application

In these development procedures, Blender and Unity are the main software used for creating 3D objects and assets for virtual world creation. Firstly, the Unity is employed to compose the scence of application, designs and implement logic, and packaging the software within its extended reality framework. Figure 11.3 shows the 3D model of virtual CNC turning machine and the render of it using Blender as shown in Fig. 11.4.

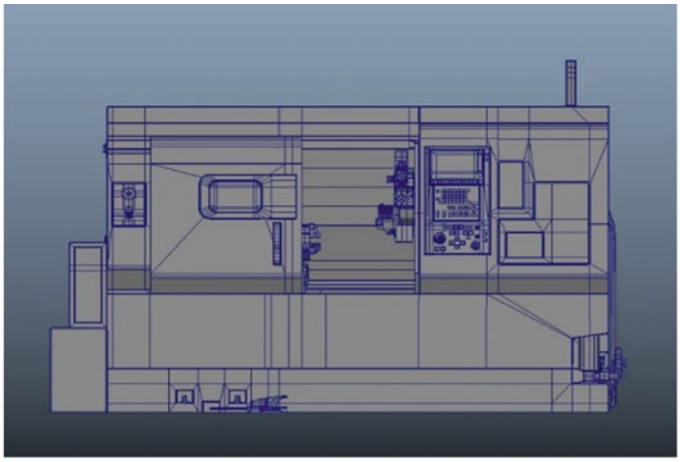
Secondly, the Blender is adopted to color and render the 3D model. Figure 11.5 compares the actual CNC turning machine in the laboratory and the virtual CNC turning machine after rendering.

Thirdly, a development of virtual CNC turning application would be ensured by a constant testing, which would enhance the effectiveness of the feature driven methodology. Once the components of CNC turning machine are setup and implemented, the main function of the application must be created. This will require a design and function which will drive the flow of the turning processes, simulate the necessary surrounding of the processes, and outline the interactions of the applications. The input aspects of the application and the user input available for the user will need to be designed carefully to fulfill the required interactions.

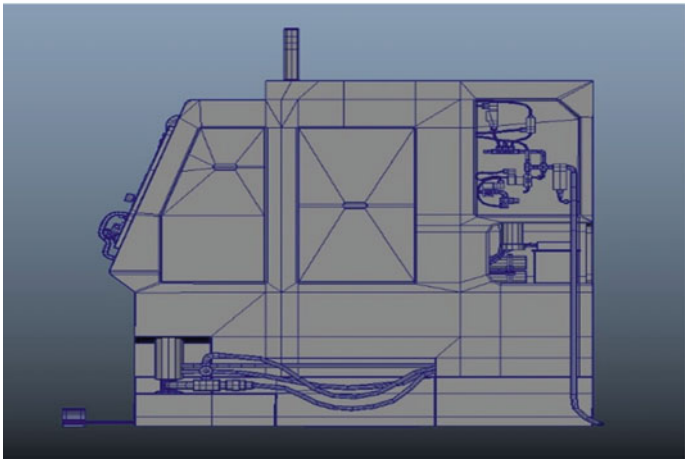
Finally, the validity of virtual CNC turning application referring to CNC turning processes in the real world must be calibrated and executed under a strict testing condition. This will ensure that the virtual CNC turning application is up to standard and robust enough to be handled for actual use in the real environment as shown in Fig. 11.6.

## 11.4 Validation of Virtual CNC Turning Processes

The procedures how to validate and operate the cutting processes on virtual CNC turning are following:



(a) Front view



(b) Side view

**Fig. 11.3** Illustration of 3D model of virtual CNC turning machine

1. Click the menu of 'Operating Mode' and select 'Yes' to start operating the virtual CNC turning as shown in Fig. 11.7.
2. Start the procedures to setup the CNC turning machine on the left controller window as shown in Fig. 11.8. Each procedure needs to be checked after finish on the controller window. Otherwise, the next procedure cannot be executed.
3. Click the menu of 'Coding Mode' and select 'Yes' to key the G-code commands as shown in Fig. 11.9.
4. Select the basic turning processes which have facing, turning, grooving, drilling, and threading, respectively to cut the workpiece as shown in Fig. 11.10.

(a) Front view

(b) Side view

**Fig. 11.4** Render of virtual CNC turning machine

5. Input and run the G-code commands to cut the workpiece using the cutting condition of previous research [17] as shown in Fig. 11.11. Hence, the roundness can be also estimated from the previously obtained model [17] of author in the future application by selecting the cutting condition with plain carbon steel S45C and the cutting force ratio when the cutting tool is still new without flank wear in the developed virtual CNC turning application.

(a) Actual CNC Turning machine

(b) Virtual CNC Turning machine

**Fig. 11.5** Comparison between actual CNC turning machine and virtual CNC turning machine

Once the command block is read and run until end of block, the obtained workpiece will be checked to confirm the G-code commands. The VR CNC turning application is designed to prove out the G-code commands step by step. It means that the users can practice and enhance their skills using the developed VR CNC turning application.

Figure 11.12 illustrates the workpiece after facing, turning, and threading operations, respectively, which can be obtained from the application of virtual CNC turning, and the results are the same as previous work [17]. It is understood that the surface roughness, the straightness, and the roundness can be predicted in advance as functions in the virtual CNC turning application which will be developed in the future work by adopting the models from previous researches [2, 3, 17] of author.

**Fig. 11.6** Illustration of virtual CNC turning application in the real environment

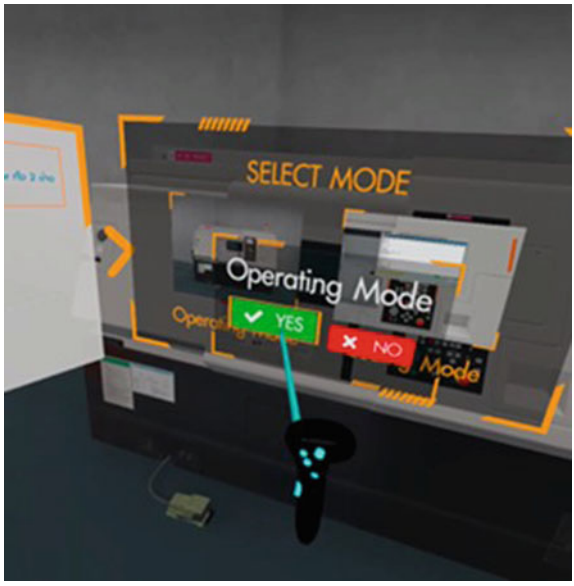



**Fig. 11.7** Illustration of operating mode in virtual CNC turning

It is shown that the virtual CNC turning can be used to practice in order to enhance the user experience and skill as shown in Fig. 11.13. The proposed and developed VR CNC turning application can be used repeatedly without costs and accidents, especially during the COVID-19 period. The VR CNC turning application is tested by users that it runs with the satisfied results.

**Fig. 11.8** Illustration of procedures to setup the CNC machine on the controller windows

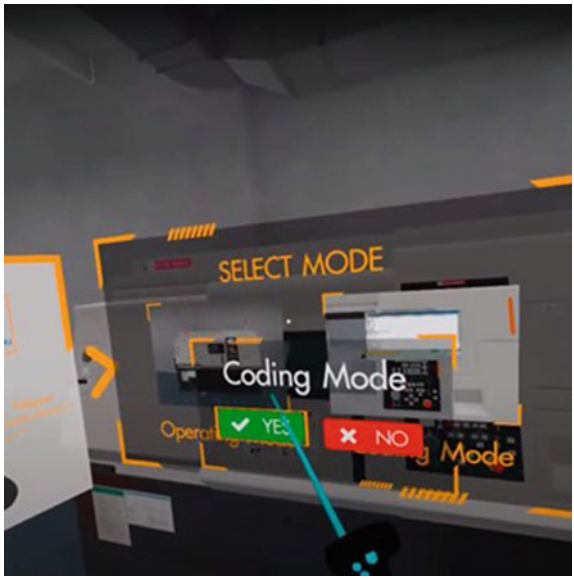



**Fig. 11.9** Illustration of coding mode in virtual CNC turning

## 11.5 Conclusions

This research aims to develop an application of virtual CNC turning utilizing the Unity and the Blender software with the Oculus Integration SDK. It is clear that the VR CNC turning application is beneficial to learn and increase the skill of trainees


**Fig. 11.10** Illustration of basic turning processes to cut the workpiece

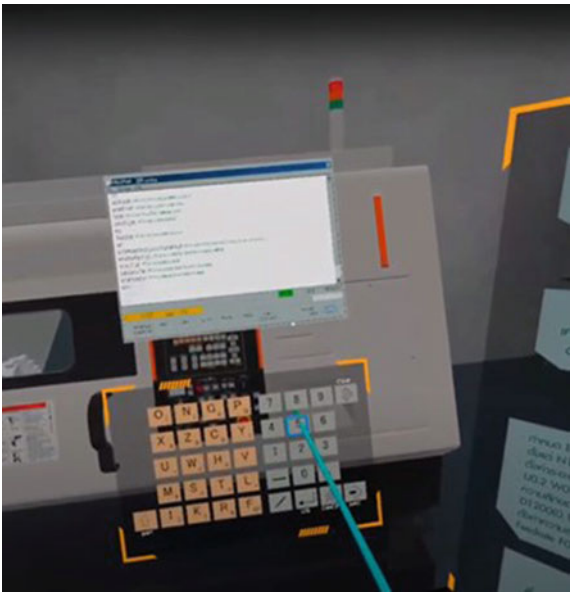


**Fig. 11.11** Illustration of G-code commands to cut the workpiece

or students. The VR CNC turning application can help to avoid the COVID-19 from the participants in the classes or shopfloors. It has been proved that the developed application of VR CNC turning runs satisfactorily.


**Fig. 11.12** Illustration of workpiece obtained from virtual CNC turning application


**Fig. 11.13** Example of practice on the virtual CNC turning application

**Acknowledgements** This work was performed and supported by the partial funding of The National Broadcasting and Telecommunication Commission (NBTC), Thailand, from August 2019 to November 2020.

**Declarations**

**Conflict of Interest** The author declares and confirms that they have no conflict of interest.

**Data Availability** Not Applicable.

**Consent Publications** The author also confirms that the manuscript has not been published elsewhere.

## References


1. Xiao, M., Shen, X., Ma, Y., Yang, F., Gao, N., Wei, W., Wu, D.: Prediction of surface roughness and optimization of cutting parameters of stainless steel turning based on RSM. Math. Probl. Eng. 1–15 (2018)
2. Tangjitsitcharoen, S., Samanmit, K., Ratanakuakangwan, S.: Development of surface roughness prediction by utilizing dynamic cutting force ratio. Proc. Inst. Mech. Eng. **490**, 207–212 (2014)
3. Sassantiwong, M., Tangjitsitcharoen, S.: In-process prediction of straightness in CNC turning by using wavelet transform. In: 2nd International Conference on Green Materials and Environmental Engineering, pp. 199–203 (2015)
4. Kryvyi, P., et al.: Influence of curvature and cross-sectional shape of cylindrical surface formed by turning on its roughness. Arab. J. Sci. Eng. **45**, 5615–5622 (2020)
5. Huang, K.T., et al.: Augmented versus virtual reality in education: an exploratory study examining science knowledge retention when using augmented reality/virtual reality mobile applications. Cyberpsychol. Behav. Soc. Netw. **22**(2), 105–110 (2019)
6. Hagan, J.O., Khamis, M., Williamson, J.R.: Surveying consumer understanding & sentiment of VR. In: Proceedings of the International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE'21) (2021)
7. Baceviciute, S.: Designing Virtual Reality for Learning. University of Copenhagen (2020)
8. Khundam, C., et al.: A comparative study of interaction time and usability of using controllers and hand tracking in virtual reality training. Informatics **8**(3), 60 (2021)
9. Xiaoling, W., et al.: Development an Interactive VR Training for CNC Machining, pp. 131–133. The Association for Computing Machinery, Inc. (2004)
10. Wang, Q., Fu, R., Hu, Y., He, J.: Development of simulation platform for CNC intelligent manufacturing. In: International Conference on Control Science and Electric Power Systems, pp. 184–189 (2021)
11. Boland, M.: Data Point of the Week: 5 Million PSVRs? 10 June 2019. Available from: https://arinsider.co/2019/06/10/data-point-of-the-week-5-million-psvrs/
12. Bellalouna, F.: New approach for industrial training using virtual reality technology. Procedia CIRP **93** (2020)
13. Rowe, S.: Learn VR Development: Guides to Develop VR and AR Applications, 3 Oct 2021. Available from: https://circuitstream.com/blog/programming-development-guides/
14. Parisi, T., Foley, M. (eds.): Learning Virtual Reality Developing Immersive Experiences and Applications for Desktop, Web, and Mobile, 1st edn, p. 127. United States of America: O'Reilly Media, Inc. (2015) (in English)
15. Kristian, M.K., Horvat, M., Oberman, T.: The use of inertial measurement units in virtual reality systems for auralization applications. Presented at the Proceeding of the 23rd International Congress on Acoustics: integrating 4th EAA Euroregio 2019, Aachen, Germany, 9–13 Sept 2019 (2019)
16. L.F. Technologies: Unity Integration, 1.43.0 edn. https://developer.oculus.com/downloads/package/unity-integration/. Last accessed 2020/11/9
17. Tangjitsitcharoen, S., Chanthana, D.: In-process prediction of roundness based on dynamic cutting forces. Int. J. Adv. Manuf. Technol. **94**(5–8), 2229–2238 (2018)

# Chapter 12
# Enhancing Elderly Leisure Experience Through Innovative VTuber Interaction in VR with ChatGPT

**Chi-Hui Chiang and Hsin-Yu Chiang**

**Abstract** This research introduces an innovative approach that integrates a virtual character in a virtual reality setting using ChatGPT as a conversational artificial intelligence system. The program was developed using a programming language and leveraged the Unity game engine's built-in support for virtual reality and natural language processing. In this approach, users engage with the virtual character through natural language conversations, and ChatGPT generates the responses. A study was conducted with individuals living alone to assess the effectiveness of this approach, which yielded promising results in enhancing the user experience of conversational agents within the virtual reality context. The research findings indicate that experiential value significantly influences perceived ease of use and perceived usefulness, and these factors, in turn, impact the intention to use. Therefore, the experience of AI VTuber can provide a sense of leisure and entertainment for the elderly. Designing an innovative VTuber integrated with ChatGPT interaction in virtual reality offers a simple and useful interaction model that can provide a novel experience for the elderly, enhancing their leisure and entertainment experiences while promoting their physical well-being.

## 12.1 Introduction

In recent, ChatGPT is an AI language model developed by OpenAI that has gained widespread attention for its impressive natural language processing capabilities. ChatGPT is a deep learning model that uses a transformer architecture to process and generate natural language. The model is trained on large amounts of text data to learn the statistical patterns and relationships between words and phrases. During the

C.-H. Chiang (✉)
Chia Nan University of Pharmacy & Science, Tainan 71710, Taiwan
e-mail: cscott@mail.cnu.edu.tw

H.-Y. Chiang
National Taiwan University, Taipei 106319, Taiwan
e-mail: b11310038@ntu.edu.tw



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_12

inference stage, the model takes in user input and generates a response by predicting the most likely sequence of words based on the context and input history. Hence, ChatGPT is widely used in various applications, such as chatbots, virtual assistants, and conversational agents. It is also a powerful tool that enables natural language interactions between humans and machines, providing a more personalized and engaging experience for users.

Virtual reality (VR) technology has made remarkable progress, with a significant increase in the use of virtual situations for various applications, such as gaming, education, and therapy [1, 2]. The ability to interact with virtual characters in a realistic and engaging way has been a focus of research in VR [3]. One interesting application of VR is to use the virtual character as a conversational agent to engage users in natural language interactions. This approach has been shown to be effective in various settings, such as education, mental health, and entertainment [4, 5].

In this paper, we propose a novel approach to use the virtual character in the VR situation to interact with users via a conversational AI system called ChatGPT. ChatGPT is an AI language model that can generate human-like responses to user inputs. By integrating ChatGPT with the virtual character in the VR situation, we aim to create a more engaging and immersive conversational experience for users. To the best of our knowledge, there has been limited research in using the virtual character in the VR situation to interact with conversational AI systems. While there have been studies on conversational agents in VR, they mainly focused on text-based interactions or pre-scripted dialogs. Our approach, on the other hand, enables users to engage in natural language conversations with the virtual character that can respond dynamically and adaptively to user inputs.

By enabling natural and engaging interactions with the virtual character, our approach can improve the effectiveness of conversational agents in various applications. For instance, in the context of discussing topics of interest, using a virtual character as a conversational agent in the VR situation can provide a unique and engaging experience for individuals to explore and learn about various topics with an interactive agent [6]. To understand the experience of the silver-haired generation, this research focuses on investigating the relationships among experiential value, perceived usefulness, perceived ease of use, and usage intention. By examining these factors, we aim to gain insights into the usage intention of the silver-haired generation, which can serve as a foundation for future research. Therefore, the objectives of this study are as follows:

(1) To develop a VTuber character and integrate it with ChatGPT to facilitate interactions with the silver-haired generation.
(2) To explore the causal relationships among experiential value, perceived usefulness, perceived ease of use, and usage intention for the silver-haired generation.

## 12.2 Literature

### *12.2.1 ChatGPT Applied in VR*

ChatGPT is based on a deep learning architecture that utilizes a large corpus of text data to generate responses. It employs a transformer-based model that uses attention mechanisms to focus on the relevant parts of the input sequence and generate the corresponding output sequence. The model is trained on a massive amount of data and can generate coherent and contextually appropriate responses to user inputs [7]. While ChatGPT has demonstrated impressive performance in generating human-like responses, it is not without its limitations. In contrast, traditional chatbots may struggle more with handling open-ended questions or more complex dialogs and could potentially provide more rigid or mechanistic responses [8]. One major concern is the potential for bias in the training data, which can lead to biased or discriminatory responses. Another issue is the lack of control over the generated responses, which can lead to inappropriate or offensive content. Despite these limitations, ChatGPT remains a powerful tool for natural language processing and has been widely adopted in various applications. Ongoing research is focused on addressing its limitations and improving its performance in generating human-like responses.

The use of virtual reality (VR) technology in education has become an emerging trend in recent years. Kraus et al. [9] examined the evolution of technological forecasting and social change, tracing the trajectory from the moon landing to the emergence of the metaverse. The past scholars highlighted the potential of VR technology to transform the way we learn, and how it can facilitate collaboration, engagement, and experiential learning. They further noted the importance of understanding the societal impacts of VR, including issues of accessibility and privacy. In a similar vein, Lim et al. [10] explored the implications of generative AI on the future of education. Liaw et al. [11] argued that the use of generative AI could provide users with personalized, interactive, and engaging learning experiences. These studies demonstrate the growing interest in the use of VR and AI in education and highlight the potential benefits and challenges associated with their implementation [12].

The use of the virtual character in the VR situation as a conversational agent has also been explored in various settings, such as education learning, and entertainment [4, 5]. The virtual character can provide a more personalized and immersive experience for users and has been shown to be effective in engaging users in natural language interactions [3]. Studies have also shown that virtual characters can improve user engagement, retention, and satisfaction in various applications [5]. Conversational AI systems, such as ChatGPT, have also been gaining attention as a means of improving interactions with the virtual character in the VR situation. By integrating ChatGPT with the virtual character in the VR situation, users can engage in natural language conversations with the characters that respond dynamically and adaptively to user inputs. The use of ChatGPT in the VR situation has the potential to enhance the user experience of conversational AI systems by creating more engaging and personalized interactions.

However, there has been limited research on using ChatGPT as an API to connect the virtual character in the VR situation. Most studies on conversational agents in VR have focused on text-based interactions or pre-scripted dialogs [3]. According to our study, the utilization of ChatGPT as an API to facilitate natural language interactions with virtual characters in VR settings has been sparsely explored. As research by Barrot [13] indicates, ChatGPT stands as a unique tool that captivates users with its ability to conduct interactive experiences reminiscent of natural and human-like conversations. Thus, the potential for using ChatGPT in VR situations to enhance the user experience of conversational agents remains largely unexplored.

### *12.2.2 Relationship Between Experiential Value, Perceived Usefulness, Perceived Ease of Use, and Usage Intention*

The Technology Acceptance Model (TAM), proposed by Davis [14], is based on the structure and relationships derived from the Theory of Reasoned Action (TRA). In the TAM, perceived usefulness and perceived ease of use are believed to directly influence an individual's attitude. Perceived usefulness (PU) is defined as "the prospective user's subjective probability that using a specific application system will increase his or her job performance within an organizational context". Previous studies have found that perceived usefulness is a major determinant of usage behavior and intention. For instance, Venkatesh and Davis [15] validated the relationship between perceived usefulness (PU) and usage behavior using different technologies. Through structural equation modeling (SEM), it was confirmed that perceived usefulness has a direct effect on usage behavior in the virtual reality [16–18].

In this study, we define perceived usefulness and perceived ease of use as the extent to which individuals believe that engaging with VTuber in virtual reality will enhance their engagement in leisure activities. By drawing upon relevant research findings and applying the Technology Acceptance Model [19, 20], we aim to explore the impact of immersive and interactive experiences provided by VTuber on the perceived usefulness and perceived ease of use among the elderly population. Therefore, the following hypotheses are reasonably proposed:

H1 Experiential value influences the perceived usefulness of engaging with VTuber for the elderly in VR situations.
H2 Experiential value influences the perceived ease of use of engaging with VTuber for the elderly in the VR situation.
H3 Perceived ease of use influences the perceived usefulness of engaging with VTuber for the elderly in the VR situation.

Based on previous research findings, we can observe that perceived ease of use and perceived usefulness are interrelated in their impact on usage intention [18]. In the Technology Acceptance Model, perceived ease of use is considered a direct

determinant of usage intention, while perceived usefulness indirectly affects usage intention through its influence on individual attitudes. Previous studies have demonstrated the significant role of perceived usefulness in usage behavior and intention [14, 16]. Additionally, research has found that perceived ease of use also has a direct impact on usage behavior and intention. Hence, we can infer that in the VR situation, the perceived ease of use and perceived usefulness of engaging in conversations with VTuber will directly influence the usage intention of elderly individuals. Through this study, we aim to further investigate the relationships among these factors to provide a deeper understanding and insights and to offer valuable guidance and recommendations for the use of VTuber by the elderly in virtual reality contexts. Therefore, the following hypotheses are reasonably proposed:

H4 Perceived usefulness influences the usage intention of engaging with VTuber for the elderly in the VR situation.
H5 Perceived ease of use influences the usage intention of engaging with VTuber for the elderly in the VR situation.

## 12.3 Research Design

To achieve our proposed approach of integrating the virtual character in the VR situation with ChatGPT as a conversational AI system, we will use a client–server architecture. This architecture allows for efficient communication and sharing of resources between different machines over a network. The communication protocol between the client and server must be chosen carefully to ensure efficient and secure data transfer. Commonly used protocols include HTTP and TCP/IP [21]. Thus, client–server architecture is a popular design pattern in computer networks [22]. In our research, the client side will be the VR situation that the user interacts with, while the server side will be ChatGPT, which generates responses to the user's inputs, as Fig. 12.1.

The program was written in C# and utilized the Unity game engine's built-in support for VR and natural language processing [23]. These code shows that the VR situation is preparing to connect to the interface of ChatGPT through API. This API Key comes from the registration application on the OpenAI website. After obtaining the API Key code, set it as a parameter in the code. Next, the VR situation will consist of a virtual character (as a VTuber) that can interact with the user via natural language conversations and buttons that the user can click to send their inputs to ChatGPT. Our proposed approach is based on prior research on integrating conversational AI systems in VR situations [3, 4].

The process of interaction between the user and the VTuber will start with the user speaking into a microphone. The user's spoken input will be converted into text using a speech-to-text library. The text will then be sent to the VR situation's input field, where the user can see the text and make any necessary edits. The user will then click a button to send the text to ChatGPT via the API interface. ChatGPT will receive the text input and generate a response based on its natural language

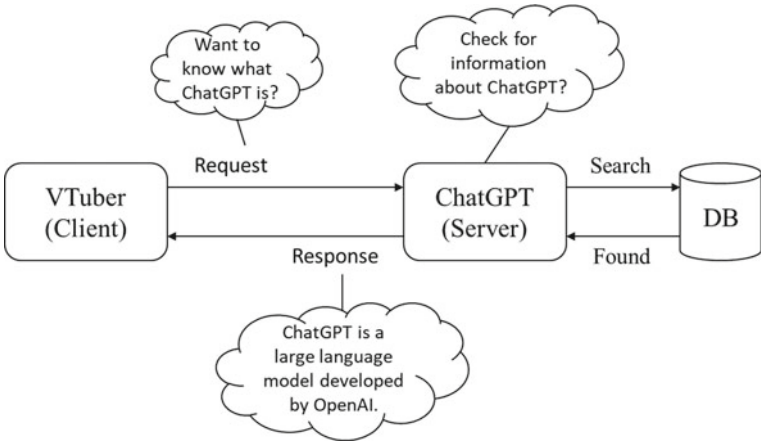



**Fig. 12.1** Client–server architecture for VTuber and ChatGPT

processing capabilities. The response will then be sent back to the VR situation via the API interface. The VR situation will display the response as text on the screen, and the VTuber will use text-to-speech technology to vocalize the response, as Fig. 12.2. Additionally, the VTuber's facial expressions will change based on the text response, adding an extra layer of immersion and naturalness to the conversation.

To test the effectiveness of our proposed approach, we will conduct users study with participants who live alone. The participants will be asked to interact with the VTuber in the VR situation and provide feedback on the naturalness, immersion, and overall experience of the conversation, as Fig. 12.3. The study will also measure the accuracy and coherence of the responses generated by ChatGPT. At the beginning, users need to wear a helmet (such as HTC VIVE) and hold a joystick. The eyes look at the VTuber in the VR situation ahead. Use the joystick to point the microphone through the green wire, then press the button of the joystick, and speak through the

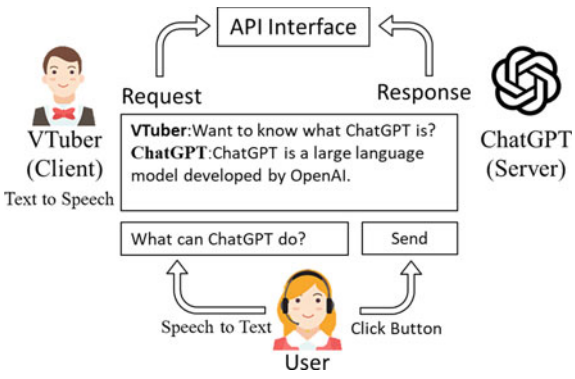



**Fig. 12.2** VTuber to ChatGPT via API


**Fig. 12.3** VTuber in the VR situation with ChatGPT dialog box

microphone. The text field will display what was said. Click the send button, and this VR situation will send the text data of the field to ChatGPT through the API interface. If there is something unclear, ChatGPT will display a message that cannot be explained or explained clearly. The user can say it again, and ChatGPT can give a detailed answer based on the content of the second question. If the VR situation does not work smoothly, the user can use the joystick to click the restart button to restart the interactive operation of the VTuber.

Next, ChatGPT will reply the generated text data in a few seconds and display the screen in the VR situation. VTuber will read the text displayed on the screen in the context. And it is presented to users through body movements and facial expressions. These body movements and facial expressions must be designed before they can be presented in the VR situation.

## 12.4 Research Method

In accordance with a study by Anderson and Gerbing [24], the study had to test hypothesized model using structural equation modeling (SEM). Firstly, the adequacy of the measurement model was assessed and estimated separately before estimating the structural equation model. Next, the study performed path analysis to test main effect hypothesized model (H1 through H5), as Fig. 12.4. Finally, the study conducted path analysis on the elderly samples to see whether there are significant differences in the hypothesized effects and also recorded the feedback from the participants.

A questionnaire was developed from related literature. The items were modified slightly to suit VR situation. Scale items included experiential value (customer

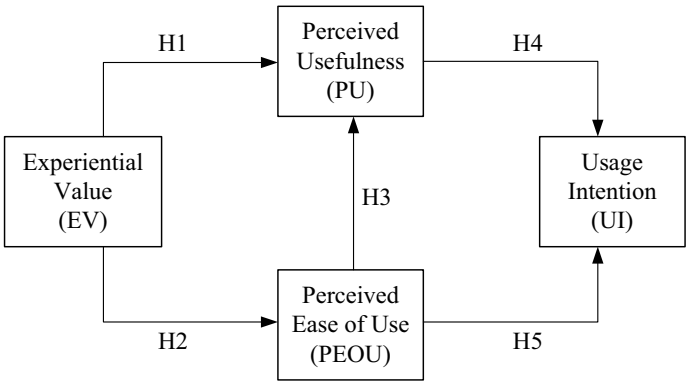



**Fig. 12.4** Research model

return on investment, service excellence, aesthetics, playfulness), perceived usefulness, perceived ease of use, and usage intention as manifest variables. Experiential value was measured by using items adapted from studies by Mathwick et al. [25] and Varshneya and Das [26]. Perceived usefulness, perceived ease of use, and usage intention were measured by using items adapted from studies by Davis [14] and Davis et al. [27], with modifications to suit usage intention for VR situation. Questionnaire items, which included questions modified from previous studies, were rated on a seven-point Likert scale from 1 (strongly disagree) to 7 (strongly agree) [28]. According to Nunnally and Bernstein [29], acceptable values of Cronbach's alpha range from 0.838 to 0.923. Table 12.1 shows that the values of alpha exceeded 0.7, indicating that the scales had good reliability.

## 12.5 Data Analysis

The participants in this study were residents of rural areas where younger family members often left home for work or education, leading to an independent lifestyle. Despite living alone, they maintained good health, adhered to early rising and exercise habits, and actively participated in community activities such as health fitness tests. In order to enhance their understanding of new technologies like virtual reality and artificial intelligence, it was crucial to provide them with opportunities to experience and learn about these advanced techniques. The study yielded 48 valid responses, including 22 male (45.83%) and 26 female participants (54.17%). In terms of age distribution, the largest group was those aged 56–60 years, with 22 participants, accounting for 45.83% of the total. This was followed by the 61–65 age group with 12 people, making up 25.00% of the total. Participants aged 66–70 years totaled 7, representing 14.58% of the total. Those aged 71–75 years accounted for 8.33%

**Table 12.1** Constructs, questionnaire items, and reliabilities

| Constructs | | Items | Questions | Cronbach's alpha | References |
|---|---|---|---|---|---|
| Experiential value | Customer return on investment | CROI1 | Experiencing AI VTuber provides high levels of affirmation and recognition | 0.923 | [25, 26] |
| | | CROI2 | Experiencing AI VTuber situations brings positive benefits | | |
| | | CROI3 | Experiencing AI VTuber situations' interactive feedback provides a sense of fulfillment | | |
| | Service excellence | SE1 | AI VTuber's explanations related to daily life knowledge make me feel a sense of familiarity | 0.842 | [25, 26] |
| | | SE2 | The information provided by AI VTuber is accurate | | |
| | | SE3 | AI VTuber meets my leisure and entertainment needs | | |
| | Aesthetics | AE1 | Experiencing AI VTuber situations is engaging | 0.883 | [25, 26] |
| | | AE2 | Experiencing AI VTuber situations rarely capture my visual attention | | |
| | | AE3 | Experiencing AI VTuber situations are designed with aesthetic value | | |
| | Playfulness | PL1 | Experiencing AI VTuber situations is enjoyable and entertaining | 0.862 | [25, 26] |
| | | PL2 | Experiencing AI VTuber situations helps me temporarily forget about worries | | |
| | | PL3 | The design of AI VTuber situations enhances my enjoyment | | |
| Perceived usefulness | | PU1 | Experiencing AI VTuber situations is practical for me | 0.838 | [14, 27] |
| | | PU2 | Experiencing AI VTuber situations is beneficial for leisure health activities | | |

(continued)

**Table 12.1** (continued)

| Constructs | Items | Questions | Cronbach's alpha | References |
|---|---|---|---|---|
| | PU3 | Experiencing AI VTuber situations is useful for learning new technologies | | |
| Perceived ease of use | PEOU1 | Experiencing AI VTuber situations is easy to learn | 0.847 | [14, 27] |
| | PEOU2 | Experiencing AI VTuber situations is moderately easy for me | | |
| | PEOU3 | Experiencing AI VTuber situations has a user-friendly interface | | |
| Usage intention | UI1 | I am confident that I can become proficient in using AI VTuber situations | 0.840 | [14, 27] |
| | UI2 | I believe that with effort and learning, I can master the use of AI VTuber situations | | |
| | UI3 | I am satisfied with the knowledge and skills acquired through experiencing AI VTuber situations | | |

of the total with four participants. Lastly, the smallest group was those aged 76–80 years, with three individuals, making up 6.25% of the total. Further analysis revealed that the majority of the participants were elderly. About 52.12% of them held a high school diploma or higher level of education. Approximately 50.27% of the elderly individuals had heard of or seen virtual reality, while less than 20% had actual firsthand experience (Table 12.2).

Discriminant validity was tested by comparing the square root of the AVE for each factor with its correlation coefficients with other factors. Table 12.3 shows that the AVE values for all variables were higher than that of the off-diagonal squared correlations, suggesting satisfactorily discriminant validity of the variables [30]. Hence, discriminant validity was also met.

The results obtained using the SEM are illustrated in Fig. 12.5. The corresponding figures show the explanatory powers of these constructs according to the squared multiple correlation ($R$) results. Perceived usefulness had a 80.1% explanatory power ($R^2 = 0.801$) and was most affected by perceived ease of use (PEOU; standardized coefficient = 0.668), followed by experiential value (EV; standardized coefficient = 0.250). Thus, PEOU was the most influential factor for perceived usefulness.

**Table 12.2** Item loadings and *t*-values of related factors ($N = 48$)

| Constructs | | Items | Factor loading | Standard deviation | *t*-value |
|---|---|---|---|---|---|
| Experiential value | Customer return on investment | CROI1 | 0.861 | 0.049 | 17.661 |
| | | CROI2 | 0.862 | 0.041 | 21.228 |
| | | CROI3 | 0.888 | 0.035 | 25.416 |
| | Service excellence | SE1 | 0.851 | 0.044 | 19.225 |
| | | SE2 | 0.829 | 0.048 | 17.101 |
| | | SE3 | 0.784 | 0.061 | 12.866 |
| | Aesthetics | AE1 | 0.890 | 0.028 | 31.691 |
| | | AE2 | 0.847 | 0.046 | 18.383 |
| | | AE3 | 0.819 | 0.064 | 12.766 |
| | Playfulness | PL1 | 0.870 | 0.063 | 13.901 |
| | | PL2 | 0.738 | 0.094 | 7.865 |
| | | PL3 | 0.856 | 0.040 | 21.398 |
| Perceived usefulness | | PU1 | 0.845 | 0.061 | 13.865 |
| | | PU2 | 0.863 | 0.036 | 23.881 |
| | | PU3 | 0.903 | 0.024 | 37.005 |
| Perceived ease of use | | PEOU1 | 0.884 | 0.034 | 26.199 |
| | | PEOU2 | 0.917 | 0.020 | 46.496 |
| | | PEOU3 | 0.822 | 0.061 | 13.534 |
| Usage intention | | UI1 | 0.795 | 0.059 | 13.565 |
| | | UI2 | 0.898 | 0.025 | 36.097 |
| | | UI3 | 0.923 | 0.022 | 42.417 |

**Table 12.3** CR, AVE, and correlation coefficient matrix

| Constructs | CR | AVE | EV | PU | PEOU | UI |
|---|---|---|---|---|---|---|
| EV | 0.967 | 0.710 | **0.843** | | | |
| PU | 0.904 | 0.758 | 0.668 | **0.871** | | |
| PEOU | 0.907 | 0.766 | 0.740 | 0.757 | **0.875** | |
| UI | 0.906 | 0.764 | 0.717 | 0.744 | 0.766 | **0.874** |

This finding suggests that the presence of an easily controllable AI VTuber in a VR situation can influence the intention of the elderly, particularly the silver-haired generation, in using VR experiences. Usage intention had a 83.4% explanatory power ($R^2 = 0.834$) and was most affected by perceived ease of use (PEOU; standardized coefficient = 0.559), followed by perceived usefulness (PU; standardized coefficient = 0.381), meaning that perceived ease of use had a greater influence than perceived usefulness on usage intention.

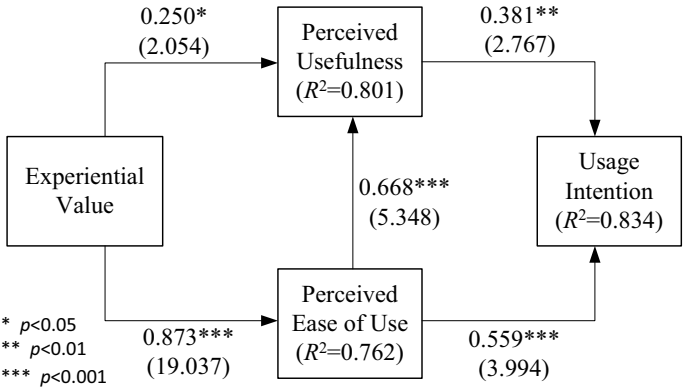



**Fig. 12.5** Empirical results

The result showed that the overall model fit was assessed by using multiple fit criteria, as suggested in the literature. The structural model exhibited an adequate fit ($\chi^2/df = 2.815$, Standardized Root Mean Square Residual (SRMR) = 0.084 and NFI = 0.617). As recommended by Jöreskog and Sörbom [31], $\chi^2/df$ less than 3 indicates an acceptable goodness-of-fit between the hypothesized model and observed data. In summary, the structural model tests, including convergent and discriminant validity measures, were satisfactory [24, 30].

## 12.6 Results and Discussions

The proposed approach of integrating VTuber in the VR situation with ChatGPT as a conversational AI system was successfully implemented and tested. In the user study, participants have interacted with the virtual character as a VTuber in the VR situation and provided feedback on the naturalness, immersion, and overall experience of the conversation. The results showed that the approach was effective in creating engaging and personalized interactions between the users and the VTuber. To this end, the study aimed to explore the relationships between observed variables that influence the usage intentions of the elderly in the VR situation. Specifically, the impact of external variables, including experiential value, perceived usefulness, and perceived ease of use, on the usage intentions of the elderly was examined.

First, the findings of the study revealed a positive correlation between experiential value and perceived usefulness, which aligns with prior research and our initial hypotheses. This relationship was found to be significant and influential. Furthermore, previous studies have highlighted the significant role of consumer return on investment (CRI) in shaping the usage behavior of users [20]. Our research has shown that the perception of investment derived from engaging in leisure activities such as virtual reality, particularly in the emerging technology domain, contributes to the

perception of high perceived benefits and interactive feedback. This is primarily due to the ability of VTubers to engage in conversations with the elderly and provide them with high levels of affirmation. The elderly perceives the VTuber to be highly beneficial and find it easy to interact with them through the interface controls. The VTuber serves as a source of leisure entertainment and social interaction. Therefore, return on investment has been identified as a relevant factor in the entertainment-related investments of the elderly.

> In the VR situation, I experienced the new technology of virtual reality, including the use of a headset and joystick for control. I could ask questions through a microphone and get responses. It made me realize the progress of technology. (Participant #12)

Secondly, based on service excellence, VTuber should prioritize providing convenient service functionalities. In terms of the presentation of the VR situation, the elderly also perceive interacting with AI VTuber as an opportunity to learn new knowledge, such as life-related information. Hence, the study revealed a significant impact with a favorable beta value. This finding aligns with the perspectives of Verhagen et al. [20] and Ros et al. [32], indicating that VTuber can attract potential elderly users through intelligent service functionalities, enhancing their perceived usefulness of the VR situation. The ease of controlling these service functionalities to meet the leisure and entertainment needs of the elderly is also emphasized. Our research findings further highlight the elderly's need for new information in their daily lives.

> Having a conversation with a virtual influencer is a lot of fun. They can respond to everything I ask. It's like an AI service oracle, or you can call it a know-it-all lady. (Participant #35)

Third, this finding suggests that VTuber can fulfill the visual experience and needs of the elderly. Our study revealed that designing an attractive and lively VTuber can make the elderly perceive it as a friendly and conversational character. The designer should carefully plan the appearance of VTubers, creating innovative and appealing characters to attract the elderly for interaction. Specifically, this result indicates that a simple and user-friendly interface can attract the elderly to experience virtual reality situations, enhance their satisfaction, and stimulate their usage behavior.

> Virtual reality is very easy to operate. Interacting with AI characters in VR and chatting is done very well. The virtual field of vision is also excellent, making it a good experience. (Participant #2)

Fourth, the study revealed that playfulness had a significant impact on perceived usefulness and perceived ease of use. The relationship between playfulness and perceived usefulness as well as perceived ease of use was found to be significant. These findings indicate that engaging in playful interactions with VTuber in the VR situation can influence users' perceived usefulness and ease of use of VTuber. Our study emphasizes the importance of pleasant experiences in shaping individuals' perceptions. Therefore, it is important for operators to provide rich content about the VR situation, highlighting its functional advantages. This will contribute to enhancing the perceived usefulness and ease of use of engaging with VTuber in the VR situation, making it an appealing leisure activity for the elderly.

> When operating VR, it's often necessary to know your virtual position in the situation. Interacting with VR situation and VTuber feels very novel, especially being able to respond to questions. It's very interesting. (Participant #9)

Fifth, the results of the study demonstrate that perceived usefulness and perceived ease of use significantly influence usage intention in the interaction between VTuber and the elderly in the VR situation. The study indicates that when the elderly perceives VTuber in the VR situation as valuable and easy to use, they are more likely to engage with them and participate in interactive activities. The perception of usefulness signifies that VTuber provides meaningful knowledge and beneficial leisure experiences, while the perception of ease of use suggests that the interaction between the elderly and VTuber in the VR situation is user-friendly and easily manageable. These positive perceptions play a crucial role in shaping the usage behavior of the elderly as they become more motivated to actively participate in and enjoy the interactive experiences provided by VTuber in the VR situation.

> Integrating AI into daily life is a future trend. Virtual reality is a kind of new technology, and interacting with a virtual influencer is a great experience. It's nice to use it as leisure entertainment when free. (Participant #42)

## 12.7 Conclusions and Future Directions

Based on our research findings, the integration of VTuber in the VR situation with considerations of experiential value, perceived usefulness, and perceived ease of use offers several benefits for the usage intentions of the elderly. Firstly, enhancing experiential value can increase the elderly's interest and engagement in interactive experiences, making them more inclined to use VTuber for conversations and interactions. Secondly, the elderly perceives VTuber in the VR situation as highly useful, which is a significant factor for their usage intentions. They recognize the value and benefits of engaging with VTuber, such as learning new knowledge, gaining entertainment, and fulfilling social needs. Additionally, improving perceived ease of use makes it easier for the elderly to use VTuber for conversations and interactions, reducing barriers and difficulties associated with usage. These research findings demonstrate that integrating VTuber in the VR situation with considerations of experiential value, perceived usefulness, and perceived ease of use can enhance the usage intentions of the elderly. They become more willing to engage in conversations and interactions with VTuber, thereby facilitating their learning, entertainment, and social activities. This technology has practical applications for improving the quality of life and promoting the overall well-being of the elderly.

Future research in this field provides a solid foundation for improving the user experience with conversational AI agents in the VR situation, further understanding the experiential value of VTuber in VR situations, including emotional satisfaction and enjoyment. Researchers can explore how VTuber can enhance the immersive and engaging experiences in virtual situations, leading to greater user satisfaction and enjoyment. Furthermore, this would be worthwhile to compare these findings

with those from the use of chatbots, to analyze the differences in user satisfaction following their experiences. The comparative study could offer valuable insights into the optimal design of user experiences with conversational AI agents in the VR situations.

## References


1. Lu, X., Huang, S., Li, J.: A review of virtual reality and its application in medical field. Ann. Transl. Med. **7**(14), 327 (2019)
2. Kim, J.-H., et al.: Immersive interactive technologies and virtual shopping experiences: differences in consumer perceptions between augmented reality (AR) and virtual reality (VR). Telematics Inform. **77**, 101936 (2023)
3. Gorovoy, V., & Chen, C.: Conversational agents in virtual reality: a study on implications for user experience. In: 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 748–755. ACM (2017)
4. Bickmore, T.W., & Picard, R.W.: Establishing and maintaining long-term human-computer relationships. ACM Trans. Comput Hum. Interact. **12**(2), 293–327 (2005)
5. Huang, H.Y., Li, C.H.: A study of avatar and virtual agents in education. Educ. Technol. Soc. **23**(1), 170–182 (2020)
6. Trunfio, M., Jung, T., Campana, S.: Mixed reality experiences in museums: exploring the impact of functional elements of the devices on visitors' immersive experiences and post-experience behaviours. Inf. Manag. **59**(8), 103698 (2022)
7. Radford, A., et al.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
8. Pandey, S., Sharma, S.: A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning. Healthc. Anal. **3**, 100198 (2023)
9. Kraus, S., et al.: From moon landing to metaverse: tracing the evolution of technological forecasting and social change. Technol. Forecast. Soc. Chang. **189**, 122381 (2023)
10. Lim, W.M., et al.: Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. Int. J. Manag. Educ. **21**(2), 100790 (2023)
11. Liaw, S.Y., et al.: Artificial intelligence in virtual reality simulation for interprofessional communication training: mixed method study. Nurse Educ. Today **122**, 105718 (2023)
12. Peres, R., et al.: Editorial: on ChatGPT and beyond: how generative artificial intelligence may affect research, teaching, and practice. Int. J. Res. Mark. **40**(2), 269–275 (2023)
13. Barrot, J.S.: Using ChatGPT for second language writing: pitfalls and potentials. Assess. Writ. **57**, 100745 (2023)
14. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. 319–339 (1989)
15. Venkatesh, V., Davis, F.D.: A theoretical extension of the technology acceptance model: four longitudinal field studies. Manage. Sci. **46**(2), 186–204 (2000)
16. Huang, Y.-C., et al.: Surfing in virtual reality: an application of extended technology acceptance model with flow theory. Comput. Hum. Behav. Rep. **9**, 100252 (2023)
17. Chen, T., et al.: Path analysis of the roles of age, self-efficacy, and TAM constructs in the acceptance of performing upper limb exercises through immersive virtual reality games. Int. J. Ind. Ergon. **91**, 103360 (2022)
18. Manis, K.T., Choi, D.: The virtual reality hardware acceptance model (VR-HAM): extending and individuating the technology acceptance model (TAM) for virtual reality hardware. J. Bus. Res. **100**, 503–513 (2019)
19. Matsika, C., Zhou, M.: Factors affecting the adoption and use of AVR technology in higher and tertiary education. Technol. Soc. **67**, 101694 (2021)

20. Verhagen, T., et al.: Satisfaction with virtual worlds: an integrated model of experiential value. Inf. Manag. **48**(6), 201–207 (2011)
21. Coulouris, G.F., Dollimore, J., & Kindberg, T.: Distributed Systems: Concepts and Design. Pearson Education (2011)
22. Tanenbaum, A.S., Steen, M.V.: Distributed Systems: Principles and Paradigms. Pearson Education (2007)
23. Unity_Technologies: Unity—Manual: Virtual Reality Overview (2021). Available from: https://docs.unity3d.com/Manual/VROverview.html
24. Anderson, J., Gerbing, D.: Structural equation modeling in practice: a review and recommended two-step approach. Psychol. Bull. **103**(3), 411–423 (1988)
25. Mathwick, C., Malhotra, N., Rigdon, E.: Experiential value: conceptualization, measurement and application in the catalog and Internet shopping environment **77**, 39–56 (2001)
26. Varshneya, G., Das, G.: Experiential value: multi-item scale development and validation. J. Retail. Consum. Serv. **34**(January), 48–57 (2017)
27. Davis, F.D., Bagozzi, R.P., Warshaw, P.R.: Extrinsic and intrinsic motivation to use computers in the workplace. J. Appl. Soc. Psychol. **22**(14), 1111–1132 (1992)
28. Likert, R.: A Technique for the Measurement of Attitudes. The Science Press, New York (1932)
29. Nunnally, J.C., Bernstein, I.H.: Psychometric Theory, 3rd edn. McGraw-Hill, New York (1994)
30. Fornell, C., Larcker, D.F.: Evaluating structural equation models with unobservable variables and measurement error. J. Mark. Res. **18**(1), 39–50 (1981)
31. Jöreskog, K.G., Sörbom, D.: LISREL 8: A Guide to the Program and Applications. SPSS Inc., Chicago (1993)
32. Ros, M., et al.: Applying an immersive tutorial in virtual reality to learning a new technique. Neurochirurgie **66**(4), 212–218 (2020)

# Chapter 13
# Assessing the Utility of GAN-Generated 3*D* Virtual Desert Terrain: A User-Centric Evaluation of Immersion and Realism

**Rahul K. Rai, Reshu Bansal, Shashi Shekhar Jha, and Rahul Narava**

**Abstract** Terrain modeling is increasingly becoming an essential part of robotics and virtual reality (VR). This technology has the potential to transform how robots interact with the world and how we experience virtual reality. In virtual reality, terrain models are used to create immersive and realistic landscapes for users to explore and interact with. However, creating varied, demanding, and realistic terrains in simulation is challenging and time-consuming. This paper presents an automated system capable of generating synthetic terrain and investigates how well the generated virtual terrain suits VR systems. To automate, we leverage the generative adversarial networks (GAN) with post-enhancement to generate a diverse and immersive desert. We used an actual digital elevation model (DEM) (1 m resolution) of the Mojave Desert in California, USA, obtained from the USGS agency to train the generative model. We investigated by employing a head-mounted display (HMD) for a robust quality assessment, and the results indicate our system successfully generated virtual terrain with acceptable realism.

## 13.1 Introduction

Terrain modeling has emerged as a critical aspect of robotics and virtual reality by enabling to interact with realistic environments and providing users with an immersive experience [1]. This has the potential to revolutionize how we experience virtual reality and how robots interact with their surroundings. In recent years, advancements in terrain modeling have opened up new possibilities for these fields, including the ability to simulate a wide range of terrain types, such as hills, valleys, and rugged landscapes. By incorporating terrain modeling into virtual reality simulations, users

R. K. Rai (✉) · S. S. Jha · R. Narava
Indian Institute of Technology Ropar, Rupnagar, PB 140001, India
e-mail: 2018csz0004@iitrpr.ac.in
URL: https://www.iitrpr.ac.in/

R. Bansal
Indian Institute of Technology Mandi, Kamand, HP 175005, India
URL: https://www.iitmandi.ac.in/



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_13

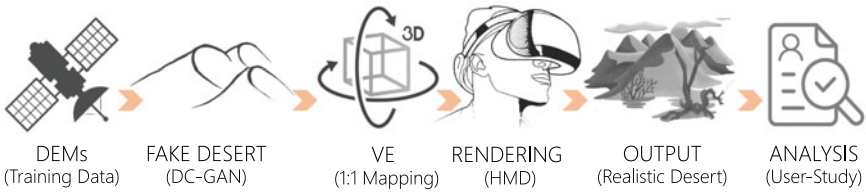



**Fig. 13.1** Overview of system pipeline

can explore and interact with digital landscapes that are virtually indistinguishable from the real world. This has significant implications for a wide range of fields, from entertainment and gaming to education and training (Fig. 13.1).

Traditionally, terrain generation has been a time-consuming and laborious process, often relying on methods such as Perlin noise [2] and diamond-square [3]. However, recent advances in deep generative models, including generative adversarial networks (GANs) [4], variational autoencoders (VAEs) [5], and Pixel-CNN [6] have resulted in the creation of more versatile and complex terrain generation models. Out of these, deep convolutional generative adversarial network (DCGAN) [7]—a specific architecture of GAN that incorporates convolutional neural networks (CNNs) is a powerful and widely used architecture for heightmap generation. For our system, we use the DCGAN architecture with some hyper-parameter tuning.

Since DEMs are limited in resolution, e.g., to around 90 m or 30 m, they cannot capture low-level details on the earth's surface ($\sim$ 1 to $\sim$ 10 m) [8]. In this paper, by using a DEM of around 1-m resolution, we generate realistic synthetic terrain by capturing low-level features.

Regarding the evaluation of GAN models, quantitative measures, such as the structural similarity index (SSI) [9], mean squared error (MSE), inception score (IS) [10], Frechet inception distance (FID) [11], are being used. There is no agreement on which score is the best for evaluating the performance of GAN models [12]. This is because different scores assess different aspects of the image generation process, and a single score is unlikely to provide a comprehensive assessment. Despite the availability of quantitative measures, visual examination of samples by humans remains a common and intuitive way to evaluate the performance of GAN models [12]. Virtual reality (VR), augmented reality (AR), and mixed reality (MR) are a state of the art for visually examining 3D models [13]. The immersive nature of VR enables users to experience a highly realistic and interactive environment, which provides a sense of presence and allows them to inspect 3D models from different angles and perspectives [14]. Compared to traditional 2D screens, VR offers a more intuitive and natural way of exploring complex 3D models. Our study uses the heightmap generated by DCGAN as a seed for rendering 3D terrain, which allows us to take advantage of VR technology for visual examination.

In addition to evaluating GAN models for terrain generation, there is a need to investigate the usefulness of generated terrain for virtual reality applications. While GANs have been shown to be effective in generating high-quality terrain [15], exam-

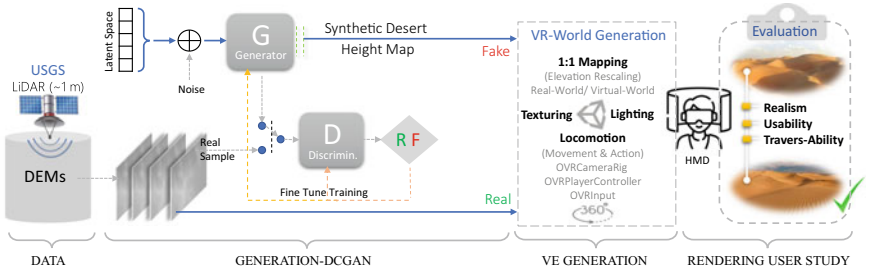



**Fig. 13.2 Illustration of an automated pipeline for generating virtual desert terrain and assessing it in VR**: The input to the system is a real-world desert DEMs. A DCGAN is trained on heightmap data derived from real DEMs. A VE is then modeled using the VR modeling objects and synthetic heightmap generated by DCGAN. The resulting VE is rendered through an HMD and analyzed on usability, traverse-ability, and realism factors

ining how these generated terrains can be utilized in virtual reality settings is essential. Figure 13.2 depicts the pipeline of our system.

This study aims to achieve an automated generation of realistic and diverse desert terrain and analyze its effectiveness for VR systems. Our contributions include the following:

- The implementation of the end-to-end pipeline to demonstrate the concept for automatic generation of realistic and diverse desert virtual environments (VE).
- A virtual reality-based qualitative evaluation method for GAN-generated terrain.

## 13.2 Background

### *13.2.1 HeightMaps*

In video games, simulation software, and other 3D modeling applications, heightmap representation is a common technique for generating 3D terrain. Heightmap is a two-dimensional grayscale image depicting each point's elevation or height on a terrain [16]. Each pixel's grayscale value in the heightmap corresponds to the terrain height at that point. In other words, the lower the height, the darker the pixel, and the higher the height, the brighter the pixel, as shown in Fig. 13.4. DEM derived from satellite data, Li-DAR scans, or other terrain measurement techniques are typically used to generate heightmaps.

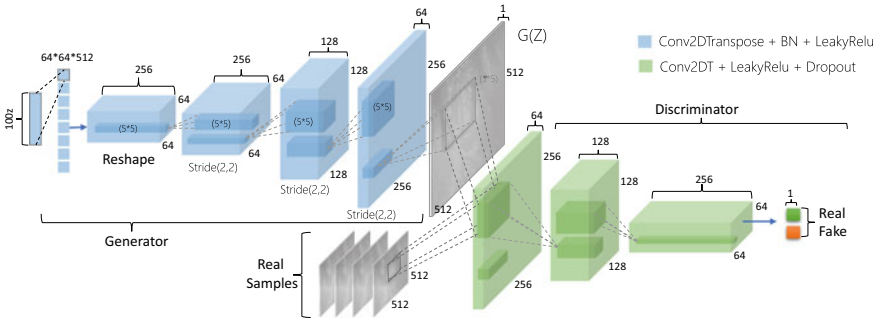



**Fig. 13.3** Architecture of the generator and discriminator network

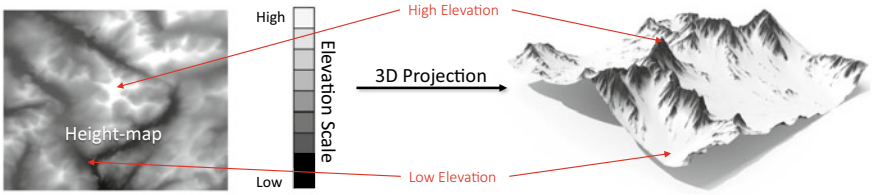



**Fig. 13.4** Heightmap: white pixel represents the highest elevation and black lowest elevation

### *13.2.2 Deep Generative Methods*

Deep generative techniques belong to a category of artificial intelligence that focuses on producing fresh and distinctive data by leveraging learned patterns from existing data. Popular examples of deep generative models are generative adversarial networks (GANs), variational autoencoders (VAEs), and Pixel-CNN. GAN comprises two neural networks: a generator network and a discriminator network. The generator network accepts random noise as input and creates new data samples that aim to resemble the initial data set. On the other hand, the discriminator network takes actual data from the primary data set and generated data from the generator network as input and attempts to differentiate between them. In 2015, Radford's deep convolutional generative adversarial network (DCGAN) improved generated image quality by using CNNs instead of multi-layered perceptrons. We use DCGAN to create pseudo-heightmaps.

### *13.2.3 Literature Review*

In the realm of evaluating 3D models in VR, EAL Lee et al. [17] discuss recent advancements and persisting challenges in virtual reality-based learning environ-

ments. Additionally, Hsinfu Huang et al. conducts a study focusing on improving the learning outcomes of virtual reality 3D modeling. Furthermore, it examines the factors that impact the usability of 3D model learning within a virtual reality environment [18, 19].

In 2018, Christopher et al. [20] first introduced using GANs for procedural terrain generation with a spatial resolution of 1 square km per pixel. Ryan et al. [21] 2019 used satellite RGB data along with DEMs to generate textures much like the real world. Dimitri et al. [15] in 2021 compared different generative models, including GAN, VAE, and Pixel-CNN, for the generation of heightmaps and showed that the GAN architecture provided a more realistic output than others. Some papers have proposed procedural-GAN and spatial-GAN for generating heightmaps [22, 23]. Ramos et al. [24] proposed a system, utilizing the dual critic conditional Wasserstein GAN that transforms low-fidelity sketches into realistic heightmaps, allowing for high visual quality while preserving user control.

To evaluate the effectiveness of the GAN model, quantitative parameters such as SSI and MSE [9] have been used in most of the paper. Despite the availability of quantitative measures, visual examination of samples by humans remains a familiar and intuitive way to evaluate the performance of GAN models [12]. Although terrain generated using GAN models may appear realistic on a $2D$ screen, performing a $3D$ visualization is crucial to ensure that it accurately represents the real-world terrain.

Additionally, there is a need to verify the usefulness of GAN-generated heightmap for Virtual Environment (VE) generation. This study used real-world desert DEMs ($\sim$ 1 m) to generate pseudo-heightmaps. We applied GAN post-processing, rendering them in a VR headset through the Unity Engine for visual analysis.

## 13.3 Methodology

### *13.3.1 Overview*

Capturing low-level features on the desert surface using commonly available Li-DAR satellite sources is challenging [8]. We utilized real-world DEMs of a desert with 1-m resolution. To generate the VE, we used the Unity Engine[1] platform and the Oculus Quest 2[2] as an HMD for visualization. Our methodology (Fig. 13.2) pipeline consists of the following steps:

1. We collected 1-m resolution DEMs of the Mojave Desert in California, USA, using USGS National Map $3DEP$ [25].
2. A 500-pixel window was slid across the DEM, and color compositions less than 80% black were retained to filter out trivial data.

[1] https://unity.com/.

[2] https://www.oculus.com/quest-2.

3. Data augmentation was performed through 4 rotations and 4 flipping, resulting in 2,680 training images.
4. The DCGAN model was trained using the training data.
5. A 1:1 mapping, texturing, and lighting were done in the Unity Engine to resemble real-world terrain.
6. The resulting terrain was rendered in an HMD [26].
7. Qualitative analysis was performed to analyze the realism, traversability, and quality of the generated virtual desert terrain.

### *13.3.2 Data Collection and Preprocessing*

**Area of Investigation** We obtained the data for our study from the USGS National Map 3DEP website, which provides geographical data for various locations in the USA at different resolutions, including 2 arc-second (Alaska—60 m) DEM, 1 arc-second (30 m) DEM, 1/3 arc-second (10 m) DEM, 5-m DEM, and 1/9 arc-second (3 m) DEM, 1-m DEM. The Mojave Desert in California, USA, was selected for our study, and we obtained a DEM of size 10 km × 10 km with a resolution of 1 m. Figure 13.5 shows the geographical location of the study area.

**Training Data Preparation** The heightmap was segmented into 500 × 500-pixel windows, creating a total of 400 images. Only images with a color composition of less than 80% black were kept to filter out data that was trivial to generate by GAN. To expand our data set, we employed data augmentation techniques such as rotating and flipping the heightmap in four different orientations, resulting in 2680

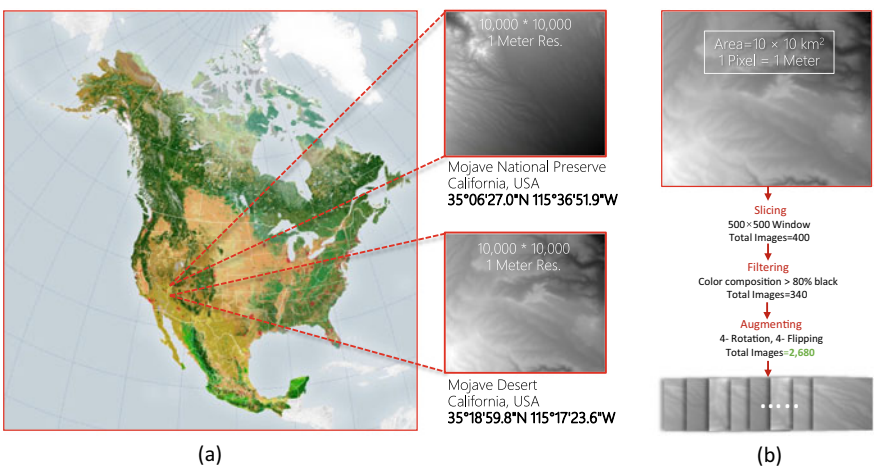



**Fig. 13.5** DATA: **a** sample of DEM selected for this study. **b** Training data preprocessing

unique terrain DEMs for training our DCGAN model. Figure 13.5b illustrates the data preprocessing process for training.

### 13.3.3 DCGAN Training and Parameters

GANs consist of two key components: the Generator G and the Discriminator D. The primary objective of the Generator is to create a new set of data distribution $P_{\text{data}}(x)$ that closely resembles the actual data distribution. At the same time, the discriminator is responsible for determining whether the generated data is real or fake by comparing it to the actual data. To accomplish the desired outcome, a trade-off is required between the discriminator's loss and the generator's loss, as minimizing one leads to an increase in the other.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(x)}[\log(1 - D(G(z)))] \quad (13.1)$$

Here $V(D, G)$ is the objective function, $x$ is a real image sampled from a real distribution $P_{\text{data}}(x)$, and $z$ is the input to the generator network sampled from $p_z(x)$ probability distribution of the random noise vector. The GAN uses the DCGAN (Deep Convolutional GAN) architecture, which typically consists of a generator and discriminator networks. The architecture can be visualized in Fig. 13.3. The training data is first normalized between $[-1, 1]$ and then rescaled to a resolution $512 \times 12$. This is a common practice to ensure that the data is within a reasonable range and has a consistent resolution across samples. The optimizer used for training the GAN is ADAM, a popular choice for training deep neural networks. The learning rate is set to 0.002, which controls how much the networks' weights are updated during each iteration of training. The batch size used for training is 32, which controls how many samples are processed in each training iteration. The depth of the DCGAN architecture is set to 4, kernel size used for both the generator and the discriminator is set to $5\times$. The output size of the generator network is $512 \times 12$, which is the same as the resolution of the rescaled training data (see Table 13.1).

These hyper-parameters and network architecture are typically chosen based on empirical observations and trial-and-error experimentation. Figure 13.6 shows the four images at a time generated at different epochs.

### 13.3.4 VR World Generation

To ensure a seamless and visually impressive virtual reality experience, several adjustments to parameters related to heightmap and the addition of various modules are required before rendering the environment to the HMD.

**Table 13.1** GAN hyper-parameters

| Parameter | Value |
|---|---|
| Optimizer | ADAM |
| Learning rate | 0.002 |
| Batch size | 32 |
| Depth ($N$) | 4 |
| Discriminator Kernel | (5.5) |
| Generator Kernel | (5.5) |
| Dropout | 0.3 |

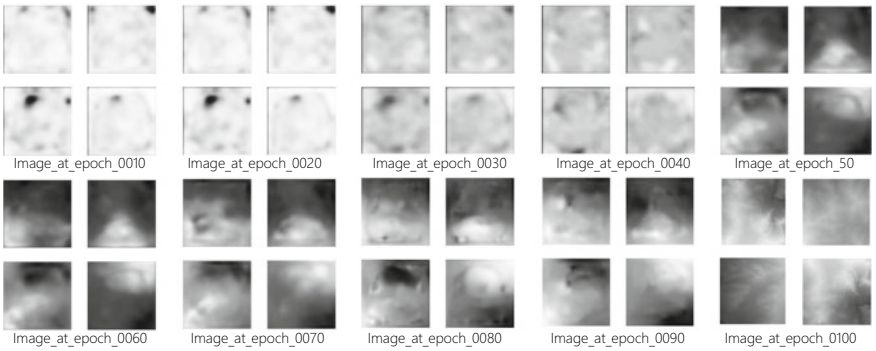



**Fig. 13.6** Images generated at different epochs. Batch size 32, showing 4 at a time

**Environmental Design**: To generate virtual desert terrain from a heightmap in Unity3D, a new GameObject Terrain is created, and a corresponding heightmap image file is imported.

**1:1 Mapping**: Elevation scaling is crucial in accurately representing real-world terrain in VEs. To achieve this, it's essential to ensure that the elevation points of the terrain match with actual world terrain data. Using OpenStreetMap, we obtain the highest and lowest elevation points for our study area and accordingly adjust the parameter of generated desert terrain.

**Texture and lighting design**: After elevation scaling is done, the next step is to add atmospheric lighting and textures to the virtual environment. This includes adding sand texture to the terrain and adding appropriate lighting to closely resemble the real-world desert environment.

**Locomotion**: To enable the user to control their movements and actions within the virtual environment using an Oculus Quest 2, specific GameObjects related to the headset are added to the Unity engine including OVRCameraRig, OVRPlayerController, OVRInput.


**Fig. 13.7** Sample of VR scenes rendered through unity engine

**HMD Rendering**: Once the virtual environment is complete, it is rendered in an HMD Oculus Quest 2, which allows users to experience the environment in virtual reality. Figure 13.7 shows the sample of VR scenes rendered through Unity Engine.

## 13.4 Evaluation and Results

A user study (12 participants) was conducted to evaluate the quality of the virtual desert terrain generated by the DCGAN and to validate whether the GAN-generated heightmaps resemble the real world, thus determining its usefulness for VR applications.

### *13.4.1 Method*

12 volunteers (Ages 23–36, Mdn = 28, 4 Female) were included in the study. The study used Oculus Quest 2 in a 3.0 $\times$2.6 m track space. The experimental design consisted of two VEs, one using real DEM and the other using pseudo-DEM.

### *13.4.2 Procedure*

Before starting, participants were given instructions on using the VR devices. Participants then began to explore and navigate the generated virtual world using the controller for an average of 5 min per participant. First, a VR scene corresponding to a real heightmap is presented, followed by a VR scene corresponding to an

**Table 13.2** Analysis of survey questions

| Question | Factor | Size | Mean | SD | Results |
|---|---|---|---|---|---|
| How confident do you feel when navigating through new areas? | Traversability | 12 | 5.25 | 1.47 | $Z = 0.60$ $P = 0.74$ |
| How easily are you able to judge distances between objects in the VE? | | | | | |
| How does this VE compare to other video games you have played? | Usability | 12 | 6.8 | 1.03 | $Z = 0.32$ $P = 0.86$ |
| How much of the generated content is useful for various applications? | | | | | |
| How realistic do you find the VE? | Immersion | 12 | 4.01 | 1.09 | $Z = 0.61$ $P = 0.73$ |
| How much diversity does the VE have? | | | | | |

enhanced fake heightmap. Finally, each participant was given a set of questionnaires (see Table 13.2) and asked to rate their experience.

### *13.4.3 Data and Analysis*

Questionnaires consist of six different questions that can be further grouped into three major factors: immersion, traversability, and usability. Traverse-ability describes how well the user could move through the landscape without feeling frustrated or like they were exploiting the physics. Immersion is related to the consistency of information in the VE with the real world [1]. Usability describes how well the content suits game design or any VR application. For each question, there is a rating on a scale of

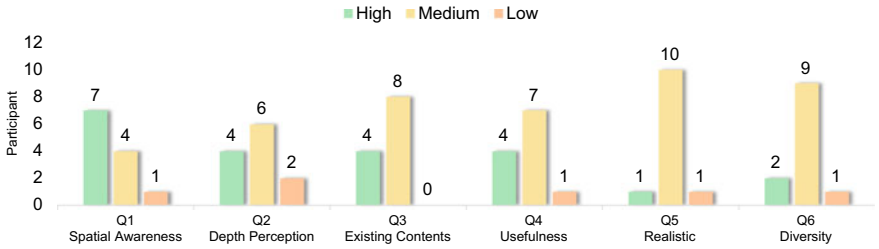



**Fig. 13.8** Data illustration of survey question analysis

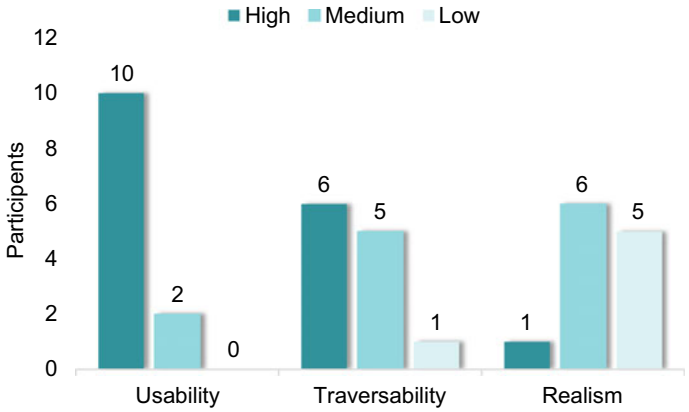



**Fig. 13.9** Distribution of participants across three factors

1–7 (see Fig. 13.8 analysis of survey questions), which is further mapped to a 3-point scale: high (5–7), neutral (4), and low (1–3). Chi-square (2) test analysis [27] was done on this data. The overall rating based on all questions was mean $M = 5.12/7$ and a standard deviation of SD = 0.85. The Average rating were high for usability ($M = 6.8$, SD = 1.03) and traversability ($M = 5.25$, SD = 1.47), and average for immersion ($M = 4.01$, SD = 1.09). Figure 13.9 illustrates the distribution of participants across three factors. The average immersion score in our user study could be because we did not focus extensively on the texturing part. Therefore, applying dynamic sand textures to the environment could prove to be beneficial in improving the overall immersion of the generated desert terrain.

In summary, the study found that a virtual desert generated from real DEM provides satisfying VR experiences with high usability, traversability, and average immersion.

## 13.5 Conclusion and Future Work

In conclusion, the use of GAN models for terrain generation is a powerful technique that enables the creation of realistic and diverse virtual terrain content. Through 1-meter resolution DEM we could generate virtual terrain that more closely mimics the real-world terrain and provide a more realistic and engaging VR experience. Our user study showed high scores for traverse-ability and usability of the generated virtual desert terrain but average scores for immersion.

Future studies may involve a larger sample size of individuals to generalize the results. As a part of future work, to improve immersion, applying dynamics and textures to the environment along with considering sand dune morphology may prove beneficial for generating high-definition immersive desert terrain. Further, future research by comparative analysis of the proposed pipeline with alternative terrain generative techniques or algorithms like Perlin noise, diamond-square algorithm, simplex noise, midpoint displacement, and fractal Brownian motion may provide a broader understanding.

## References

1. Slater, M., Sanchez-Vives, M.V.: Enhancing our lives with immersive virtual reality. Front. Robot. AI **3**, 74 (2016)
2. Perlin, K.: An image synthesizer. **19**(3), 287–296 (1985)
3. Fournier, A., Fussell, D., Carpenter, L.: Computer rendering of stochastic models. Commun. ACM **25**(6), 371–384 (1982)
4. Ian, J., Goodfellow, J.P.-A., Mirza, M., Xu, B., Warde-Farley, D., Courville, A., Bengio, Y., Ozair, S.: Generative adversarial networks (2014)
5. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv:1312.6114 (2013)
6. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with PixelCNN decoders. In: Advances in Neural Information Processing Systems, 29 (2016)
7. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 (2015)
8. Sharp. R.P.: Wind ripples. J. Geol. **71**(5), 617–636 (1963)
9. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
10. Salimans, T., Goodfellow, I., Zaremba., W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs (2016)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems, 30 (2017)
12. Borji, A.: Pros and cons of GAN evaluation measures (2018)
13. Saggio, G., Ferrari, M.: New trends in virtual reality visualization of 3D scenarios. In: Tang, X.-X. (ed.) Virtual Reality, chapter 1. IntechOpen, Rijeka (2012)
14. Radianti, J., Majchrzak, T.A., Fromm, J., Wohlgenannt, I.: A systematic review of immersive virtual reality applications for higher education: design elements, lessons learned, and research agenda. Comput. Educ. **147**, 103778 (2020)

15. Demergis, D.: Comparative analysis of machine learning techniques for island heightmap generation. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2021)
16. Heightmap. UNITY3D. [Online; accessed 01 Jan 2023]
17. Ip, H.S.S., Li, C.: Virtual reality-based learning environments: recent developments and ongoing challenges. In: Hybrid Learning: Innovation in Educational Practices: 8th International Conference, ICHL 2015, Wuhan, China, July 27–29, 2015, Proceedings 8, pp. 3–14. Springer (2015)
18. Huang, H., Lin, C., Cai, D.: Enhancing the learning effect of virtual reality 3D modeling: a new model of learner's design collaboration and a comparison of its field system usability. Universal Access Inf. Soc. **20**, 429–440 (2021)
19. Huang, H., Lee, C.-F.: Factors affecting usability of 3D model learning in a virtual reality environment. Interactive Learn. Environ. **30**(5), 848–861 (2022)
20. Beckham, C., Pal, C.: A step towards procedural terrain generation with GANs (2017)
21. Spick, R.R., Walker, J.: Realistic and textured terrain generation using GANs. CVMP'19. Association for Computing Machinery, New York, NY, USA, 2019
22. Panagiotou, E., Charou, E.: Procedural 3D terrain generation using generative adversarial networks (2020)
23. Wulff-Jensen, A., Rant, N.N., Møller, T.N., Billeskov, J.A.: Deep convolutional generative adversarial network for procedural 3d landscape generation based on dem. In: Brooks, A.L., Brooks, E., Vidakis, N. (eds.) Interactivity, Game Creation, Design, Learning, and Innovation, pp. 85–94. Springer International Publishing, Cham, 2018
24. Ramos, N., Santos, P., Dias, J.: Dual critic conditional Wasserstein GAN for height-map generation. In: Proceedings of the 18th International Conference on the Foundations of Digital Games, FDG'23. Association for Computing Machinery, New York, NY, USA, 2023
25. Shah, C.H.: USGS Science Data Catalog. https://data.usgs.gov/datacatalog/data/USGS:77ae0551-c61e-4979-aedd-d797abdcde0e
26. Sutherland, I.V., et al.: The ultimate display. In: Proceedings of the IFIP Congress, vol. 2, pp. 506–508. New York (1965)
27. Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosoph. Mag. **50**(302), 157–175 (1900)

# Chapter 14
# Application of Lightweight Image Super-Resolution Technology in Smart Grid Management System


**Weixi Feng, Mengqiu Yan, and Haiyuan Xu**

**Abstract** To improve the efficiency of smart grid management system and meet the requirements of real-time operation of power grid lines, we propose a lightweight adaptive weighted attention network (AWAN) for smart grid system. Specifically, a novel lightweight adaptive weighted attention block (AWAB) is carefully designed as the fundamental block. AWAB is actually a multi-branch block, which is made up of linear feature extraction block (LFEB), lightweight channel attention block (LCAB), and nonlinear block (LNB). With these three branches, AWAB can effectively extract discriminative feature and assign more weight to the high-frequency information. The qualitative and quantitative experiments have validated that AWAN can rebuilt better HR images for image super-resolution compared with lightweight method, i.e., SRCNN, FSRCNN, VDSR, and DRCN.

## 14.1 Introduction

Recently, the whole process automation for patrol inspection is on the basis of deep learning technology in the smart grid management system which can improve the management efficiency of the power grid and accelerate the response speed of the system. At present, we need to build an algorithm model for dynamic adjustment of operation and maintenance strategies, apply video terminals to realize automatic generation of plans, control of traffic lights at nodes, automatic identification of defects, automatic generation of patrol reports, and automation of the whole process from planned production to closed-loop control. At the same time, based on the artificial intelligence platform of China Southern Power Grid, the company's self-developed green film, pollution flashover, mountain fire, safety supervision, and other algorithms are integrated to achieve image intelligent identification and early warning. In combination with the special work requirements of special patrol and special maintenance, a chart showing alarms of different algorithm types is formed

W. Feng (✉) · M. Yan · H. Xu
Shenzhen Power Supply Bureau Co., Ltd., Shenzhen 518000, Guangdong, China
e-mail: duanzhiwei69240130@163.com



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_14

to assist the team in formulating operation and maintenance work and support the continuous optimization of the algorithm. In recent years, image super-resolution (SR) algorithms can provide low-cost preprocessing operations for intelligent image recognition.

As a low-level vision technology, image SR is capable of reconstructing low-resolution images as high-resolution ones. Recently, due to the rapid development of deep learning in the field of computer vision, learning the end-to-end mapping relationship between LR and HR images from a large number of training data using convolutional neural network, i.e., CNN, has already become popular. Dong et al. [1] first proposed SRCNN, an image super-resolution reconstruction algorithm based on depth learning. This algorithm first samples the image to be reconstructed to the size of the target image and then trains the mapping relationship between the two, achieving good results. On the basis of SRCNN, Dong et al. [2] continue to propose FSRCNN. Unlike SRCNN, FSRCNN places SRCNN's upsampling module at the end and uses deconvolution for upsampling to reduce the computational complexity and improve the execution speed. Lately, Kim et al. [3] developed a novel VDSR that increased the number of convolutional neural network depth up to 20 and extracted the depth features of the image by increasing the receptive field of the convolutional neural network. The reconstruction quality is higher than FSRCNN. Subsequently, there other methods are devoted to improve the performance of SISR, such as DRCN [4], CRAN [5], RCAN [6], and so forth [7–10].

Although the above models continuously improve the effect of super-resolution reconstruction, the algorithm is getting more complex, and the parameter is large as well as the computational burden, making it hard to apply in actual scenarios.

To cope with these problems, we propose a novel lightweight adaptive weighted attention network (AWAN network). Multi-branch AWAB is adopted for feature extraction and introduces channel attention mechanism to assign more weights to high-frequency information. Finally, different types of features extracted from the three branches are aggregated.

## 14.2 Related Work

### *14.2.1 CNN-Based SR*

Recently, enormous deep learning-based methods have been proposed to address the SR problems and have achieved great success. Different types of networks have been exploited in the task of SR. As the pioneering CNN-based work for single image SR, SRCNN [1] is mainly constructed with three convolutional layers and its performance has exceeded those of traditional models such as bicubic interpolation, K-SVD [11], and ANR [12]. Later, Shi et al. proposed an efficient SR model called ESPCNN [13] that directly extracts feature maps in LR space rather than HR space. Specifically, Shi specially designed a so-called sub-pixel convolution layer or pixel-shuffle layer

to replace the deconvolutional layer or bicubic interpolation to reconstruct the super-resolved HR images. Inspired by the benefits from utilizing deep convolutional neural network, Kim et al. successively proposed more deeper models, i.e., VDSR [3] and DRCN [2] using more than 15 convolutional layers, and unexpectedly improved the state-of-the-art performance for single image SR.

Lan et al. designed a dense lightweight network, named MADNet [14], to extract more multi-scale features and information. Moreover, EDSR [7] developed by Lim et al. has pushed the CNN layers utilized in model more than 100 and therefore attained remarkable performance improvement via removing unnecessary modules (batch normalization) in residual networks. Li et al. [15] built SRFBN via feedback manner based on using hidden states and a feedback block to extract refined powerful high-level representation for image SR task. Zhang et al. [13] introduced dense connection into the network to further enhance the representative ability of the model. However, the aforementioned methods greatly improved the performance for single image SR at the cost of higher computational overhead and a large amount of model parameters.

### *14.2.2 Lightweight Network for SR*

Currently, many SR models based on CNNs often need a large number of computational resources while they improve the performance for SR task, restricting the deployment to mobile devices. Therefore, many researchers are prompted to develop more lightweight and effective algorithms for image SR [2, 16–21]. For example, Tai et al. [16] utilized a serial recursive unit to construct a deep recursive neural network (DRRN) with based on residual blocks [17]. Later, Ahn et al. [18] built a cascading residual network (CARN) incorporating the recursive mechanism with different skip connections as well as group convolution for less FLOPs. Furthermore, Hui et al. [19] introduced a lightweight information multi-distillation network (IMDN) to reduce the amount of channel of output features for single image SR and ranked first in the AIM 2019 SR challenge for their excellent performance. Subsequently, Liu et al. [20] re-developed IMDB and enhanced it as residual feature distillation block (RFDB) and also won championship in the AIM 2020. Nevertheless, designing models using less parameters and FLOPs is not necessary for efficient computation and less running time for application in resources-restricted devices.

To address this issue, we carefully design a novel AWAB with three parallel sub-module branches, namely linear feature extraction block (LFEB), lightweight channel attention block (LCAB), and nonlinear block (LNB) for single image SR.

## 14.3 Proposed Method

Next, the overall structure of our proposed method will be elaborated in detail. The proposed network is called as AWAN—adaptive weighted attention network.

### *14.3.1 Network Architecture*

The AWAN proposed in this paper is inspired by the cascade structure of EDSR network [7], and its architecture is shown in Fig. 14.1. The network is mainly composed of three parts, namely the shallow feature extraction module, the deep feature extraction module, and feature upsampling and reconstruction module. Assuming that $I_{\mathrm{LR}}$ and $I_{\mathrm{SR}}$ denote the LR input image and output image super-resolved by the model, respectively. Concretely, we adopt a $3 \times 3$ convolution kernel to capture the preceding shallow features $F_0$ from the given input. The shallow feature extraction process is denoted as below:

$$F_0 = H_{3\times 3\mathrm{conv}}(I_{\mathrm{LR}}), \tag{14.1}$$

where $F_0$ is the extracted shallow feature. Extracting image features with the $3 \times 3$ convolution in low-dimensional space effectively reduces the amount of computation and parameters, which is helpful to construct a lightweight model. Meanwhile, it is not appropriate to utilize a kernel for the first layer of the super-resolution network with a large receptive field. Due to the fact that every pixel within a downsampled LR input is relevant to the tiny region of input, some unrelated information could be introduced in the process of training.

The deep information extraction module is mainly composed of three cascaded AWABs, and the stacking of AWABs is conducive to the deep discriminative feature extraction for the proposed network. To ease the loss of information during deep network training, the global residual skip connection is combined with AWABs, which is beneficial to training the CNN-based network, and promotes the smooth flow for information across different levels. The deep feature extraction process of $k$-th AWAB can be formulated as follows:

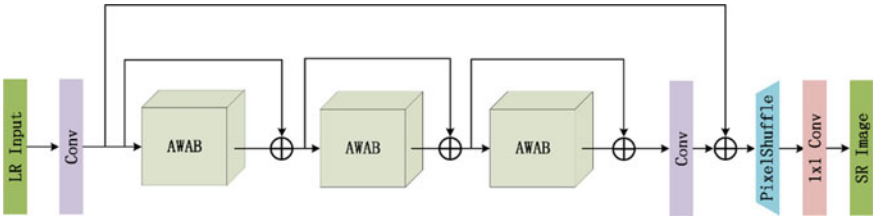



**Fig. 14.1** Fundamental diagram of AWAN

$$F_k = F_{k-1} + H_{\text{AWAB}}(F_{k-1}), \tag{14.2}$$

where $F_k$ is the output feature of $k$-th AWAB.

As shown in Fig. 14.1, the upsampling reconstruction module will expand the extracted feature map to the feature map of desired size. The upsampling method adopted in this module is on the basis of sub-pixel convolution designed by Shi et al. [22]. Specifically, the networks' reconstruction module is mainly composed of a $3 \times 3$ convolutional layer, a $1 \times 1$ convolutional layer, and a sub-pixel convolutional layer. At the end of the network, the sub-pixel convolutional layer is used as the upsampling method to enlarge the input LR image to the HR image of the corresponding target size, which improves the efficiency of model reconstruction. Following the sub-pixel convolutional layer is a $1 \times 1$ convolution incorporated to recover three channels for the super-resolved output. The procedure of reconstruction module is expressed as the below Eq. (14.3):

$$I_{\text{SR}} = H_{\text{up}}(F_0, H_{\text{up}}(H_{3\times 3\text{conv}}(F_3) + I_{\text{LR}})), \tag{14.3}$$

where $H_{\text{up}}(\cdot)$ denotes the sub-pixel convolution operation and $F_3$ denotes the third AWAB's output.

### *14.3.2 Adaptive Weighted Attention Block (AWAB)*

In order to improve the ability of feature extraction and expression of networks, this paper carefully designs an adaptive weighted attention block (AWAB). As depicted in Fig. 14.2, AWAB is made up of three parallel sub-module branches, namely linear feature extraction block (LFEB), lightweight channel attention block (LCAB), and nonlinear block (LNB). The input of $k$-th AWAB will be passed to a $1 \times 1$ convolution, and the output is $f_{\text{in}}$. Firstly, the $3 \times 3$ convolution without following activation function extracts linear feature for the purpose of introducing diversified features and therefore improves the feature expression ability of the network. The output of LFEB can be expressed:

$$F_{\text{LFEB}} = H_{\text{LFEB}}(f_{\text{in}}). \tag{14.4}$$

Secondly, since LR input contains a great number of low-frequency features as well as less high-frequency features. Noticeably, it is easy to extract the former ones from LR inputs. Meanwhile, to improve the ability of capturing high-frequency features for network, LCAB with only two layers of convolution is used to make the model adaptively focus on more complex high-frequency information. The output of LCAB is calculated as follows:

$$F_{\text{LCAB}} = \delta(H_{1\times 1\text{Conv}}(H_{\text{pool}}(H_{3\times 3\text{Conv}}(f_{\text{in}})))). \tag{14.5}$$

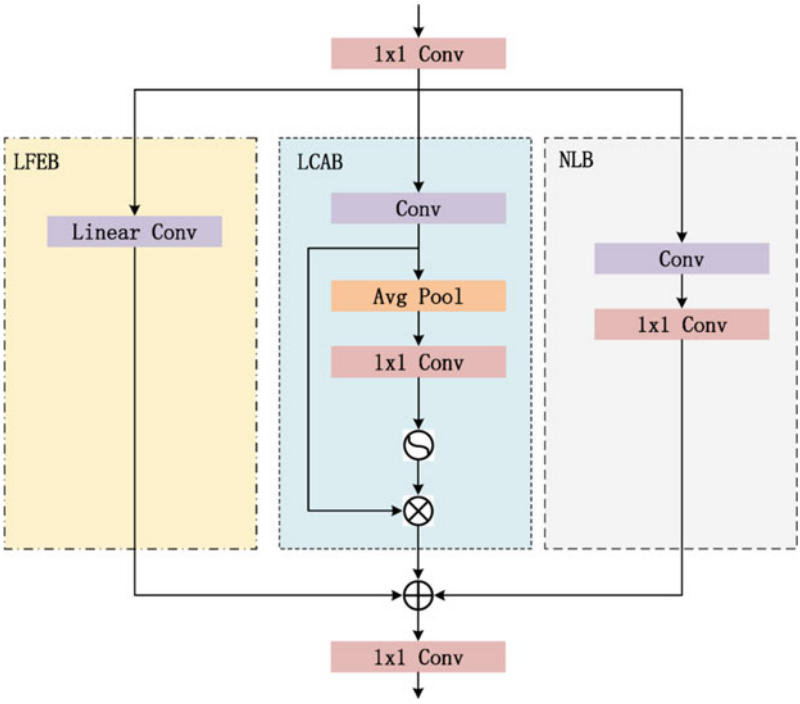



**Fig. 14.2** Adaptive weighted attention block (AWAB)

Then, the adaptive weighted output is point multiplied with the output of $1 \times 1$ convolution, which can be elaborated as:

$$F_{\mathrm{LCAB}} = S_{\mathrm{LCAB}} \cdot H_{3\times 3\mathrm{Conv}}(f_{\mathrm{in}}). \tag{14.6}$$

The output of NLB is:

$$F_{\mathrm{NLB}} = H_{1\times 1\mathrm{Conv}}(H_{3\times 3\mathrm{Conv}}(f_{\mathrm{in}})). \tag{14.7}$$

Finally, the output of AWAB can be computed as:

$$F_{\mathrm{AWAB}} = H_{1\times 1\mathrm{Conv}}(F_{\mathrm{LFEB}} + F_{\mathrm{LCAB}} + F_{\mathrm{NLB}}). \tag{14.8}$$

### 14.3.3 Loss Function

To reduce the reconstruction error, it is necessary to optimize the AWAN utilizing the loss function. At present, the most widely used loss functions are L2 loss function [3] and L1 loss function [6, 7]. Compared with the mean square error (MSE) loss, i.e., L2 loss, L1 loss has less punishment on the relative error and better effect on reconstructing image texture and border. Thus, the L1 loss function is adopted to optimize the network. Given a training dataset $\{(I^j_{\mathrm{LR}}, I^j_{\mathrm{HR}})\}^M_{j=1}$, where $I^j_{\mathrm{LR}}$ and $I^j_{\mathrm{HR}}$ represent LR image and HR image, respectively, $M$ denotes batch size of the training dataset, and $\theta$ denotes the network parameter. Hence, the objectives can be expressed as:

$$L(\Theta) = \frac{1}{M}\sum_{j=1}^{M}\left| f_{\mathrm{AWAN}}(I^j_{\mathrm{SR}}) - I^j_{\mathrm{HR}} \right|, \tag{14.9}$$

where $f_{\mathrm{AWAN}}(\cdot)$ denotes the implicit function of the proposed AWAN.

## 14.4 Experiments

### 14.4.1 Datasets and Metrics

The public dataset DIV2K [23] containing 800 high-quality RGB training images is employed as the training set. The benchmark datasets, i.e., Set5 [24], Set14 [11], BSD100 [25], and Urban100 [26] are utilized as test datasets. Specifically, Set5, Set14, and BSD100 are mainly made up of images from nature scenario, while Urban100 is made up of challenging urban scenario ones. Data augmentation is realized through rotating with degrees of 90, 180, and 270 and rotating horizontally on $64 \times 64$ image patches, which are cropped from LR inputs obtained via bicubic interpolation of DIV2K dataset. To evaluate the model performance, the Peak Signal-to-Noise Ratio (PSNR) and Structure SIMilarity (SSIM) are calculated on the Y channel (i.e., brightness) of YCbCr channel.

### 14.4.2 Implementation Details

In the training phase, Adam optimizer is used to optimize the model convergence speed, where the parameter is set as $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. The initial learning rate is set to $1 \times 10^{-4}$ [27] and halved for every 200 epochs. The batch size is set to 32, and the model trains 800 epochs in total. Each convolution layer is followed by a PReLU activation function, except for the linear convolution layer in

AWAB. Besides, the numbers of filters of each convolutional layer are 64. AWAN is implemented by PyTorch with GeForce 2080ti GPU.

### *14.4.3 Comparison with State-of-the-Arts*

As shown in Table 14.1 is the PSNR and SSIM results on the four test data sets. As you can see, bicubic interpolation yields the smallest values of PSNR and SSIM. As early SR algorithms, i.e., SRCNN [1], FSRCNN [9], VDSR [2], and DRCN [3] have some improvement compared with bicubic interpolation, but the reconstruction effect is still not ideal. AWAN achieves the best results in all test sets for three scale factors. In particular, the single image of the Urban dataset contains the most pixel content, the edge texture and other details in the image are complex, and the size difference of different objects is large, which effectively shows that the method proposed in this paper has a good effect on this kind of image reconstruction.

As depicted in Fig. 14.3, the visual results are also provided to analyze and compare the reconstruction effect of this model with other models. The image reconstructed by bicubic algorithm is blurry and loses more information. Compared with bicubic, SRCNN, FSRCNN, VDSR, and DRCN algorithms, AWAN has significantly improved the reconstruction result in terms of textures and stripes.

## 14.5 Conclusion

This paper has presented a novel adaptive weighted attention network (AWAN) for the smart grid management system. Specifically, a multi-branch block, i.e., adaptive weighted attention block (AWAB) is designed as the basic cascaded block combining the residual skip connection for SISR. AWAB aggregates multiple types of features, which is contributed to improving the representation ability of feature extraction of the model. The experiments suggest the accuracy of AWAN on five public test sets which is superior over the previous lightweight models in terms of objective and subjective evaluation indicators. At present, AWAN has been deployed to the smart grid management system, and the actual operation results show that AWAN can quickly and effectively improve the efficiency of the system.

**Table 14.1** Results of compared methods calculated with PSNR and SSIM

| Scale factor | Method | Params. (K) | Set5 | Set14 | BSD100 | Urban100 |
|---|---|---|---|---|---|---|
| | | | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM |
| ×2 | Bicubic | – | 33.66/ 0.9299 | 30.24/ 0.8688 | 29.56/ 0.8431 | 26.88/ 0.8403 |
| | SRCNN | 8 | 36.66/ 0.9542 | 32.45/ 0.9067 | 31.36/ 0.8879 | 29.50/ 0.8946 |
| | FSRCNN | 13 | 37.00/ 0.9558 | 32.63/ 0.9088 | 31.53/ 0.8920 | 29.88/ 0.9020 |
| | VDSR | 666 | 37.53/ 0.9587 | 33.03/ 0.9124 | 31.90/ 0.8960 | 30.76/ 0.9140 |
| | DRCN | 1774 | 37.63/ 0.9588 | 33.04/ 0.9118 | 31.85/ 0.8942 | 30.75/ 0.9133 |
| | AWAN | 350 | **37.65/ 0.9590** | **33.08/ 0.9121** | **32.02/ 0.8965** | **31.03/ 0.9153** |
| ×3 | Bicubic | – | 30.39/ 0.8682 | 27.55/ 0.7742 | 27.21/ 0.7385 | 24.46/ 0.7349 |
| | SRCNN | 8 | 32.75/ 0.9090 | 29.30/ 0.8215 | 28.41/ 0.7863 | 26.24/ 0.7989 |
| | FSRCNN | 13 | 33.18/ 0.9140 | 29.37/ 0.8240 | 28.53/ 0.7910 | 26.43/ 0.8080 |
| | VDSR | 666 | 33.66/ 0.9213 | 29.77/ 0.8314 | 28.82/ 0.7976 | 27.14/ 0.8279 |
| | DRCN | 1774 | 33.82// 0.9226 | 29.76/ 0.8311 | 28.80/ 0.7963 | 27.15/ 0.8276 |
| | AWAN | 350 | **33.98/ 0.9242** | **30.00/ 0.8345** | **28.96/ 0.7976** | **27.26/ 0.8294** |
| ×4 | Bicubic | – | 28.42/ 0.8104 | 26.00/ 0.7027 | 25.96/ 0.6675 | 23.14/ 0.6577 |
| | SRCNN | 8 | 30.48/ 0.8628 | 27.50/ 0.7513 | 26.90/ 0.7101 | 24.52/ 0.7221 |
| | FSRCNN | 13 | 30.72/ 0.8660 | 27.61/ 0.7550 | 26.98/ 0.7150 | 24.62/ 0.7280 |
| | VDSR | 666 | 31.35/ 0.8838 | 28.01/ 0.7674 | 27.29/ 0.7251 | 25.18/ 0.7524 |
| | DRCN | 1774 | 31.53/ 0.8854 | 28.02/ 0.7670 | 27.23/ 0.7233 | 25.14/ 0.7510 |
| | AWAN | 350 | **31.65/ 0.8876** | **28.27/ 0.7681** | **27.47/ 0.7256** | **25.47/ 0.7542** |

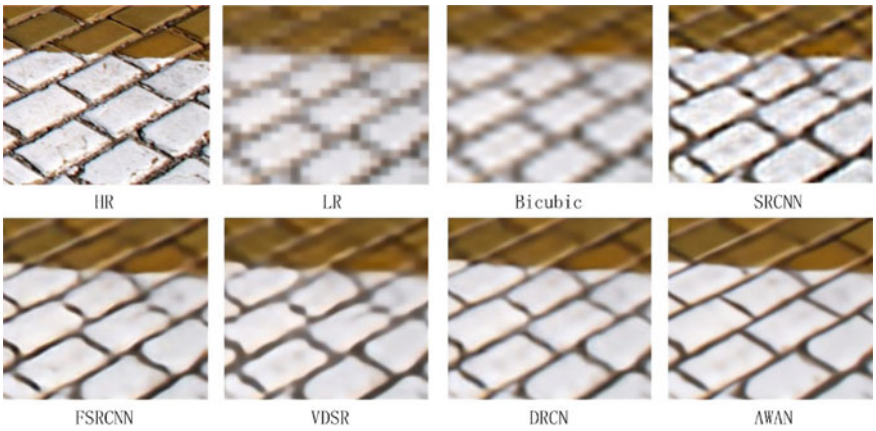



**Fig. 14.3** Visual results of AWAN compared with other models for $\times 4$ scale factor

## References

1. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 184–199. IEEE, Piscataway (2014)
2. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 Oct 2016, Proceedings, Part II 14, pp. 391–407. Springer, Cham (2016)
3. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654. IEEE, Piscataway (2016)
4. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1637–1645. IEEE, Piscataway (2016)
5. Ahn, N., Kang, B., Sohn, K.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 252–268. Springer, Berlin (2018)
6. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301. Springer, Berlin (2018)
7. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144. IEEE, Piscataway (2017)
8. Mo, F., Wu, H., Qu, S., Luo, S., Cheng, L.: Single infrared image super-resolution based on lightweight multi-path feature fusion network. J. IET Image Process. **16**(7), 1880–1896 (2022)
9. Lai, W., Huang, J., Ahuja, N., Yang, M.: Deep Laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 624–632. IEEE, Piscataway (2017)
10. Jin, K., Wei, Z., Yang, A., Guo, S., Gao, M., Zhou, X., Guo, G.: SwiniPASSR: Swin transformer based parallax attention network for stereo image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 920–929. IEEE, Piscataway (2022)

11. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 711–730. Springer, Cham (2012)
12. Timofte, R., De Smet, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1920–1927 (2013)
13. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2472–2481 (2018)
14. Lan, R., Sun, L., Liu, Z., Lu, H., Pang, C., Luo, X.: MADNet: a fast and lightweight network for single-image super resolution. J. IEEE Trans. Cybern. **51**(3), 1443–1453 (2020)
15. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3867–3876
16. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3147–3155. IEEE, Piscataway (2017)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. Ahn, N., Kang, B., Sohn, K.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 252–268 (2018)
19. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2024–2032 (2019)
20. Hui, Z., Wang, X., Gao, X.: Fast and accurate single image super-resolution via information distillation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 723–731 (2018)
21. Song, D., Wang, Y., Chen, H., Xu, C., Xu, C., Tao, D.: Addersr: towards energy efficient image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 15648–15657 (2021)
22. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883. IEEE, Piscataway (2016)
23. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 126–135. IEEE, Piscataway (2017)
24. Bevilacqua, M., Roumy, A., Guillemot, C., Morel, M.A.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: British Machine Vision Conference (BMVC), pp. 1–10. BMVA (2012)
25. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision, ICCV 2001, pp. 416–423. IEEE, Piscataway (2001)
26. Huang, J., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206. IEEE, Piscataway (2015)
27. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)

# Chapter 15
# Research on Portable Intelligent System Based on Lightweight Super-Resolved Image Recognition Algorithm


**Huang Ping, Li Qing, and Ling Letao**

**Abstract** Recently, with the increase of business demand, the whole business field of digital transmission needs to accelerate the construction of a one-stop data collaborative application system, in which the portable intelligent system based on image super-resolution recognition technology is an important part of the system. However, it is too expensive to build and improve the performance of this portable intelligent system in terms of hardware. Remarkably, the image super-resolution technology based on deep convolution neural network has made outstanding progress and dominated the current research on super-resolution technology. However, the improvement of performance is often at the cost of a sharp increase in the number of parameters, which limits the practical application of super-resolution methods. Therefore, we design a lightweight dual-path attention network (LDAN) for single image super-resolution. Specifically, a dual-path attention block (DAB) is carefully designed for assigning more weights to high-frequency information. LDAN is constructed in the manner of stacking several DABs combining global residual skip connection. The experiment demonstrates that compared with the lightweight methods, i.e., FSRCNN, DRRN, LapSRN, and IDN, the proposed LDAN greatly reduces the number of parameters, while the qualitative and quantitative results of the super-resolved images are significantly better.

## 15.1 Introduction

At present, the whole business field of digital transmission needs to build a one-stop data collaborative application system, in which the portable intelligent system is an important part. An integrated hardware system based on image super-resolution recognition technology, namely, portable intelligent magic cube, is portable and integrated, which can process data quickly and automatically on site, realize intelligent upgrading of traditional UAV and camera equipment, and quickly access intelligent

H. Ping (✉) · L. Qing · L. Letao
Shenzhen Power Supply Bureau Co., Ltd., Shenzhen 518000, Guangdong, China
e-mail: quanpwzias94103@163.com



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_15

algorithms. The portable intelligent magic cube has the functions of fast terminal access, video encoding and decoding, image recognition, multiple communication methods, etc. It can integrate different hardware systems through API interfaces, and thus, it has the function of a micro-intelligent data processing workstation, can work in a fully enclosed environment, has a solid metal shell, and is light and hard. However, the cost of improving the performance of these systems through hardware is too expensive. At present, the single image super-resolution (SISR) algorithm based on deep learning (DL) can rapidly improve the image recognition speed of the system, and developing a lightweight network can save a large computational cost.

The goal of the SISR task is to reconstruct high-resolution (HR) results from low-resolution (LR) images. In recent years, most SISR algorithms are based on end-to-end DL technology, that is, directly learning the mapping between LR and HR [1-3]. In 2016, Dong et al. [4] proposed the fast super-resolution convolution neural network (FSRCNN), which uses the original LR image without any preprocessing as the input and uses the deconvolution layer at the end of the network for upsampling, significantly reducing the network computation. With the exploration of convolutional neural network by researchers, the depth and width of the network gradually increase, and the structure of the network is more complex, so the network training becomes more difficult. On the basis of ResNet [5] network, Kim et al. [2] explored the network in a deeper direction and proposed a very deep super-resolution revolutionary network for image SR (VDSR). Lately, Tai et al. proposed a deep recursive residual network (DRRN) [6], which uses recursive learning of multiple residual units to achieve parameter sharing, effectively controlling the network parameters. To quickly breakthrough the bottleneck of SISR, Lim et al. proposed Enhanced Deep Residual Net (EDSR) [7], which greatly reduced model parameters and improved model reconstruction performance by stacking 32 ResNet modules that removed batch normalization layers. Furthermore, Zhang et al. proposed a residual channel attention network (RCAN) [8], which can adaptively learn the importance of different channels and cascade a large number of residual blocks within the residual structure, so that the network can still converge smoothly even when it is very deep and achieve more significant results. Although the deep learning-based SISR method has made great progress, there are still problems in practical application: Better results often rely on deeper networks, which requires longer training and reasoning time, as well as more computing costs. Therefore, its practicability is greatly limited, especially in resource-constrained mobile devices.

To address the above-mentioned problem, a lightweight dual-path attention network (LDAN) is proposed. The basic unit of the network consists of dual-path attention block (DAB). In each DAB, a spatial attention block and a channel attention module are constructed in parallel. DAB contributes to backpropagation and enhances the transmission of characteristic information in the front and back layers of the network and different channels. It can not only solve the problems of network gradient disappearance and gradient explosion, but also enhance the network's weight of high-frequency information in the feature map. LDAN is built by cascading and stacking DABs. Finally, super-resolved high-resolution images with richer texture details are obtained by reconstruction module.

## 15.2 Proposed Method

In this section, we will introduce and elaborate the structure and design of the proposed method in detail. To fully obtain and utilize the LR image feature information, the residual network is improved, and a lightweight dual-path attention network (LDAN) is proposed to achieve image super-resolution reconstruction based on the improved residual attention block. LDAN is consisted of the following indispensable parts: shallow feature extraction module (SFEM), deep feature extraction module (DFEM) stacked by $k$ residual dual-path attention blocks (RDAB), and the image reconstruction module (IRM). The overall network structure is depicted in Fig. 15.1.

### *15.2.1 Network Architecture*

As shown in Fig. 15.1, the feature extraction procedure of this model is divided into two stages, i.e., shallow feature extraction and deep feature extraction. At the initial stage of image super-resolving, the features directly extracted from the shallow network are similar to the input of the network and contain rich HR image information. Therefore, the shallow features extracted from the LR image will be utilized as the network input. The shallow feature extraction module (SFEM) aims to transform LR images into a batch of image features for subsequent processing, which is the basis of image super-resolution reconstruction.

Specifically, the SFEM contains a layer of $3 \times 3$ convolution and a layer of $1 \times 1$ convolution for rich low-frequency features extraction, which is motivated by Lim et al. [7]. Given a LR input image $I_{\mathrm{LR}}$, the process of SFEM can be expressed as follows:

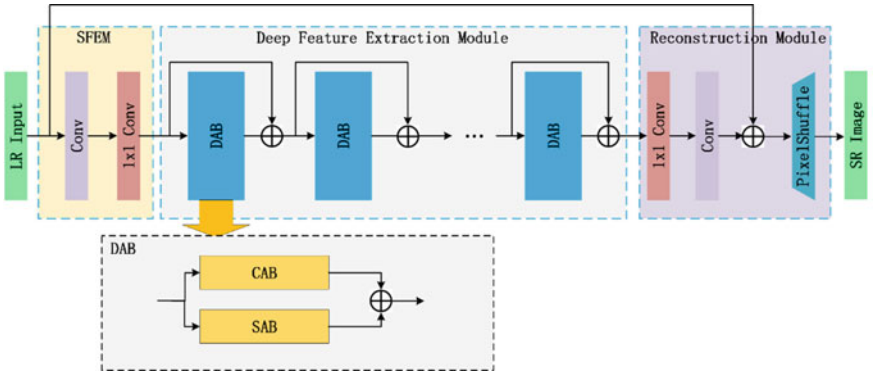



**Fig. 15.1** Overall architecture of LDAN and lightweight dual-path attention block (DAB). SFEM denotes the shallow feature extraction module. CAB and SAB are the channel and spatial attention blocks, respectively

$$F_0 = H_{\text{SFEM}}(I_{\text{LR}}), \tag{15.1}$$

where $H_{\text{SFEM}}$ represents the explicit function of convolutional operation, and $F_0$ is the extracted shallow features.

### *15.2.2 Deep Feature Extraction Based on DAB*

As shown in Fig. 15.1, the deep feature extraction module (DFEM) is composed of $k$ stacked dual-path attention blocks (DABs) and local residual skip connections. The input of each DAB is composed of the output features and the extracted shallow features of all the previous DABs. Residual skip connection mechanism can reduce the number of parameters, compensate for the loss of low-frequency information to a certain extent, and make full use of informative image feature. In addition, the residual skip connection is able to solving the training convergence problem. DEFM is connected by DAB in cascade manner, which is conducive to the flow of information.

The DAB is consisted of two different attention blocks. In order to refabricate the extracted feature and pay more attention to the high-frequency features, we design dual path of attention block included channel attention and spatial attention.

Generally, LR images contain rich low-frequency information and a small amount of high-frequency information. High-frequency information is usually located in the texture, boundary, and other areas of the input LR image, while low-frequency information is usually located in the smooth area of the input LR image. Therefore, jointly using spatial attention mechanism can increase the network's attention to high-frequency information and recover clearer and sharper the textures for the reconstructed images.

Note that the input of the first DAB is $F_0$. Except for the first DAB, the output of other following DAB can be expressed by:

$$\begin{aligned} F_{\text{DAB1}} &= H_{\text{DAB}}(F_0) \\ F_{\text{DAB2}} &= H_{\text{DAB}}(F_0 + F_{\text{DAB1}}) \\ &\cdots \\ F_{\text{DAB}k} &= H_{\text{DAB}}(F_0 + F_{\text{DAB1}} + \cdots + F_{\text{DAB}k-1}), \end{aligned} \tag{15.2}$$

where $F_{\text{DAB}k}$ denotes the output discriminative feature of $k$-th DAB, and $H_{\text{DAB}}(\cdot)$ is the implicit function of DAB.

**Channel Attention Block (CAB)**. Inspired by Zhang et al. [8], we design a lightweight channel attention block (CAB) which is illustrated in Fig. 15.2. Attention mechanism originates from the study of human vision. People usually only focus on a part of the scene or image to obtain important information. The attention mechanism selects the information that is more critical to the current task goal from the global

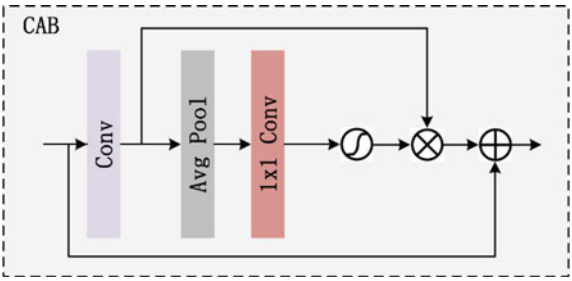



**Fig. 15.2** Structure of lightweight channel attention block (CAB)

information, so as to enhance the representation ability of the network. Different from the convention channel mechanism, we only use two layers of convolution and reintroduce layer normalization (LN) rather batch normalization into CAB. LN makes the distribution of a layer stable by normalizing the dimension of hidden size with respect to different characteristics of the same sample, which has nothing to do with the batch size.

The feature extraction process of CAB can be formulated as:

$$S_{\mathrm{CAB}} = \mathrm{Sigmoid}(H_{\mathrm{Conv}}(H_{\mathrm{pool}}(H_{\mathrm{Conv}}(f_{\mathrm{in}})))), \tag{15.3}$$

where $S_{\mathrm{CAB}}$ is the channel weights obtained by sigmoid function Sigmoid$(\cdot)$, $H_{\mathrm{Conv}}$ represents the convolution-ReLU-LN operation, and $H_{\mathrm{pool}}$ denotes the average pooling operation. $f_{\mathrm{in}}$ is the input feature of DAB.

Then, the obtained channel weights are point multiplied and summed with the feature map of the convolution output and then summed with the original input features, so as to achieve the weight allocation of different channels. The above process can be expressed as follows:

$$F_{\mathrm{CAB}} = S_{\mathrm{CAB}} \cdot H_{\mathrm{conv}}(f_{\mathrm{in}}) + f_{\mathrm{in}}, \tag{15.4}$$

where $F_{\mathrm{CAB}}$ is the feature map output by CAB.

**Spatial Attention Block (SAB)**. To strengthen the weight of high-frequency regions while suppress the weight of other regions, we employ a spatial attention block including convolution layer, average pooling layer, and sigmoid function as shown in Fig. 15.3.

The process of SAB can be formulated as:

$$S_{\mathrm{SAB}} = \mathrm{Sigmoid}(H_{\mathrm{pool}}(H_{\mathrm{Conv}}(f_{\mathrm{in}}))), \tag{15.5}$$

where $S_{\mathrm{SAB}}$ is the spatial weights obtained by sigmoid function Sigmoid$(\cdot)$.

Then, the spatial feature can be obtained by:

$$F_{\mathrm{SAB}} = S_{\mathrm{SAB}} + H_{\mathrm{Conv}}(f_{\mathrm{in}}), \tag{15.6}$$

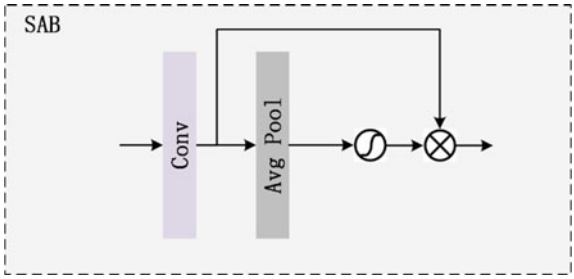



**Fig. 15.3** Structure of lightweight spatial attention block (SAB)

where $F_{\text{SAB}}$ is the feature map output by SAB.

Finally, the output feature of DAB can be the element-wise summation of $F_{\text{CAB}}$ and $F_{\text{SAB}}$.

### *15.2.3 Reconstruction Module (RM)*

At present, there are three common upsampling amplification methods in SISR, namely interpolation, sub-pixel convolution [9], and deconvolution. Follow the practice in [8], we choose sub-pixel convolution to upsample the image for lightweight and fast SR. As shown in Fig. 15.1, the reconstruction module (RM) is consisted by two layers of convolution and sub-pixel convolution. The kernel size of two convolutions is $3 \times 3$ and $1 \times 1$, respectively. The process of RM can be formulated as follows:

$$F_{\text{RM}} = H_{\text{sub - pixel}}(H_{\text{pool}}(H_{\text{Conv}}(g_{\text{in}}))), \tag{15.7}$$

where $F_{\text{RM}}$ is the super-resolved image, and $H_{\text{Conv}}$ represents the convolution-ReLU-LN operation. $g_{\text{in}}$ is the input feature of RM.

### *15.2.4 Loss Function*

At present, the commonly used loss functions are L1 and L2. Although L2 loss function can maximize the PSNR value of the network, L1 loss function makes the network training have better convergence and more stable. To accelerate the convergence of model training, we employ L1 loss to optimize LDAN following the practices in the previous works [7]. Given a LR image training set $\{(f^i_{\text{LR}}, f^i_{\text{HR}})\}^M_{i=1}$, consisting of HR-LR image pairs, the training process of LDAN can be expressed as follows:

$$L(\Theta) = \arg\min \frac{1}{M} \sum_{i=1}^{M} \left\| f_{\mathrm{SR}}^{i} - f_{\mathrm{HR}}^{i} \right\|_1, \tag{15.8}$$

where $M$, $\Theta$, and $\|\cdot\|_1$ are the number of pair-wise images, parameters of network, and the L1-norm, respectively.

## 15.3 Experiments

### *15.3.1 Datasets and Metrics*

The publicly used DIV2K dataset [10] is adopted as the training set, in which 800 HR RGB images are treated as training images. In the test phase, five widely used benchmark datasets Set5 [11], Set14 [12], BSD100 [13], Urban100 [14], and Manga109 [15] are utilized as test sets to evaluate the performance and effectiveness of the image recovery. The performance evaluation metrics are SSIM and PSNR. Note that the larger the value, the better the performance. The network is trained in RGB space and tested in Y channel of YCbCr color space.

### *15.3.2 Implementation Details*

In this paper, the LR image obtained by bicubic interpolation in the training set DIV2K is randomly cropped to $48 \times 48$ image patch and carried out random rotation and horizontal flip of 90°, 180°, and 270° to realize data augmentation. In the training phase, the ADAM optimizer is incorporated to optimize the network parameters, where $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. The learning rate is initialized to 0.0001 in the way of [16] and halved every 200 iterations. Additionally, the batch size of input is set to 16. In this experiment, the PyTorch deep learning framework is selected to build the LDAN model, and LDAN was accelerated in NVIDIA GeForce RTX 2080Ti GPU. Moreover, the kernel size of convolution is set to 64. The number of stacking DAB is $k = 4$ for saving network parameters and computational cost.

### *15.3.3 Ablation Study*

To fully investigate the effect of our DAB, LDAN-SA and LDAN-CA models are obtained by removing channel attention mechanism and spatial attention mechanism from LDAN and compared with the original model. In this paper, LDAN-SA, LDAN-CA, and LDAN are trained for 1000 epochs, respectively, and then tested on five test sets of $\times 2$ SISR. The experimental results are shown in Table 15.1, it can be observed

**Table 15.1** Ablation study's result for $\times 2$ SR on five datasets

| Method | CAB | SAB | Set5 | Set14 | Manga109 | BSD100 | Urban100 |
|---|---|---|---|---|---|---|---|
| | | | PSNR/ SSIM | PSNR/ SSIM | PSNR/SSIM | PSNR/ SSIM | PSNR/ SSIM |
| LDAN-SA | ✗ | ✓ | 37.82/ 0.9601 | 33.28/ 0.9148 | 38.02/ 0.9750 | 32.07/ 0.8984 | 31.26/ 0.9195 |
| LDAN-CA | ✓ | ✗ | 37.83/ 0.9600 | 33.29/ 0.9149 | 38.03/ 0.9751 | 32.08/ 0.8985 | 31.28/ 0.9196 |
| LDAN | ✓ | ✓ | **37.84/ 0.9602** | **33.31/ 0.9151** | **38.04/ 0.9751** | **32.08/ 0.8986** | **31.29/ 0.9198** |

that compared with LDAN, the performance of the other two models has decreased. The experimental results fully demonstrate that the two attention mechanisms can effectively enhance the model's attention to high-frequency features.

### *15.3.4 Comparison with State-of-the-Arts*

For evaluating the efficiency and effectiveness of the proposed model, LDAN is compared with the traditional bicubic algorithm and five other state-of-the-art CNN-based SR algorithms., i.e., SRCNN [1], FSRCNN [4], DRRN [6], LapSRN [17], and IDN [18]. Both qualitative and quantitative experimental results are demonstrated with respect to three scale factors, i.e., $\times 2$, $\times 3$, and $\times 4$.

Table 15.2 shows the PSNR and SSIM obtained by different SR reconstruction algorithms on the five testing datasets of Set5, Set14, Manga109, BSD100, and Urban100. Quantitatively, it can be obviously seen from Table 15.2 that LDAN attains the best results on five testing datasets among all the compared methods for three scale factors. Note that the network parameters of LDAN are less than LapSRN and IDN, while LDAN exceeds them in terms of evaluation metrics. The experimental results have validated that reasonably aggregating the spatial attention and channel attention mechanism is a feasible manner for lightweight image SR.

As shown in Fig. 15.4, the reconstructed image of LDAN proposed in this paper has clearer texture, richer details, and better overall visual effect. Bicubic algorithm, SRCNN, FSRCNN, DRRN, and LapSRN fail to utilize attention mechanism in feature extraction, so their reconstruction results have less texture information. Compared with IDN, LDAN gains more clear and better borders because LDAN additionally employs spatial attention mechanism.

**Table 15.2** Experimental results on five datasets

| Scale factor | Method | Params. (K) | Set5 | Set14 | Manga109 | BSD100 | Urban100 |
|---|---|---|---|---|---|---|---|
| | | | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM |
| ×2 | Bicubic | – | 33.66/ 0.9299 | 30.24/ 0.8688 | 30.80/ 0.9339 | 29.56/ 0.8431 | 26.88/ 0.8403 |
| | SRCNN | 8 | 36.66/ 0.9542 | 32.45/ 0.9067 | 35.60/ 0.9663 | 31.36/ 0.8879 | 29.50/ 0.8946 |
| | FSRCNN | 13 | 37.00/ 0.9558 | 32.63/ 0.9088 | 36.67/ 0.9710 | 31.53/ 0.8920 | 29.88/ 0.9020 |
| | DRRN | 298 | 37.74/ 0.9591 | 33.23/ 0.9136 | 37.88/ 0.9749 | 32.05/ 0.8973 | 31.23/ 0.9188 |
| | LapSRN | 251 | 37.52/ 0.9591 | 32.9/ 0.9124 | 37.27/ 0.9740 | 31.80/ 0.8952 | 30.41/ 0.9103 |
| | IDN | 553 | *37.83/ 0.9600* | *33.30/ 0.9148* | *38.01/ 0.9749* | *32.08/ 0.8985* | *31.27/ 0.9196* |
| | LDAN | 325 | **37.84/ 0.9602** | **33.31/ 0.9151** | **38.04/ 0.9751** | **32.08/ 0.8986** | **31.29/ 0.9198** |
| ×3 | Bicubic | - | 30.39/ 0.8682 | 27.55/ 0.7742 | 26.95/ 0.8556 | 27.21/ 0.7385 | 24.46/ 0.7349 |
| | SRCNN | 8 | 32.75/ 0.9090 | 29.30/ 0.8215 | 30.48/ 0.9117 | 28.41/ 0.7863 | 26.24/ 0.7989 |
| | FSRCNN | 13 | 33.18/ 0.9140 | 29.37/ 0.8240 | 31.10/ 0.9210 | 28.53/ 0.7910 | 26.43/ 0.8080 |
| | DRRN | 298 | 34.03/ 0.9244 | 29.96/ 0.8349 | 32.71/ 0.9379 | 28.95/ 0.8004 | *27.53/ 0.8378* |
| | LapSRN | 251 | 33.81/ 0.92203 | 29.79/ 0.8325 | 32.21/ 0.9350 | 28.82/ 0.7980 | 27.07/ 0.8275 |
| | IDN | 553 | *34.11/ 0.9253* | *29.99/ 0.8354* | *32.71/ 0.9381* | *28.95/ 0.8013* | 27.42/ 0.8359 |
| | LDAN | 325 | **34.13/ 0.9255** | **30.02/ 0.8359** | **32.74/ 0.9390** | **28.96/ 0.8014** | **27.54/ 0.8380** |
| ×4 | Bicubic | - | 28.42/ 0.8104 | 26.00/ 0.7027 | 24.89/ 0.7866 | 25.96/ 0.6675 | 23.14/ 0.6577 |
| | SRCNN | 8 | 30.48/ 0.8628 | 27.50/ 0.7513 | 27.58/ 0.8555 | 26.90/ 0.7101 | 24.52/ 0.7221 |
| | FSRCNN | 13 | 30.72/ 0.8660 | 27.61/ 0.7550 | 27.90/ 0.8610 | 26.98/ 0.7150 | 24.62/ 0.7280 |
| | DRRN | 298 | 31.68/ 0.8888 | 28.21/ 0.7720 | *29.45/ 0.8946* | 27.38/ 0.7284 | *25.44/ 0.7638* |
| | LapSRN | 251 | 31.54/ 0.8852 | 28.09/ 0.7700 | 29.09/ 0.8900 | 27.32/ 0.7275 | 25.21/ 0.7562 |

(continued)

**Table 15.2** (continued)

| Scale factor | Method | Params. (K) | Set5 | Set14 | Manga109 | BSD100 | Urban100 |
|---|---|---|---|---|---|---|---|
| | | | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM |
| | IDN | 553 | *31.82/ 0.8903* | *28.25/ 0.7730* | 29.41/ 0.8942 | *27.41/ 0.7297* | 25.41/ 0.7632 |
| | LDAN | 325 | **31.85/ 0.8908** | **28.27/ 0.7734** | **29.46/ 0.8976** | **27.43/ 0.7304** | **25.43/ 0.7640** |

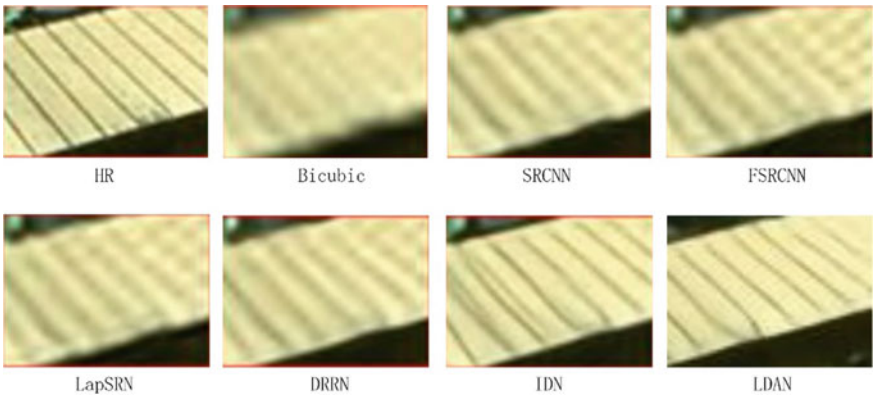



**Fig. 15.4** Visual results of various methods on BSD100 dataset for ×4 scale factor

## 15.4 Conclusion

A lightweight dual-path attention network (LDAN) is properly designed for portable intelligent system. We propose a dual-path attention block (DAB) as the basic block for extracting informative and refabricating the extracted features. Based on DAB, we carefully establish LDAN model for lightweight image SR. Extensive experimental results have proved that LDAN is not only superior over the compared methods but attain a better trade-off between model's performance and computational cost. The proposed LDAN achieves real-time SR on the portable intelligent system designed ourself and obtains good single image reconstruction results.

## References

1. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 184–199. IEEE, Piscataway (2014)

2. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654. IEEE, Piscataway (2016)
3. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1637–1645. IEEE, Piscataway (2016)
4. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 Oct 2016, Proceedings, Part II 14, pp. 391–407. Springer, Cham (2016)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE, Piscataway (2016)
6. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3147–3155. IEEE, Piscataway (2017)
7. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144. IEEE, Piscataway (2017)
8. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301. Springer, Berlin (2018)
9. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883. IEEE, Piscataway (2016)
10. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 126–135. IEEE, Piscataway (2017)
11. Bevilacqua, M., Roumy, A., Guillemot, C., Morel, M.A.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: British Machine Vision Conference (BMVC), pp. 1–10. BMVA (2012)
12. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 711–730. Springer, Cham (2012)
13. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision, ICCV 2001, pp. 416–423. IEEE, Piscataway (2001)
14. Huang, J., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206. IEEE, Piscataway (2015)
15. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. J. Multimed. Tools Appl. **76**(20), 21811–21838 (2017)
16. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
17. Lai, W., Huang, J., Ahuja, N., Yang, M.: Deep Laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 624–632. IEEE, Piscataway (2017)
18. Hui, Z., Wang, X., Gao, X.: Fast and accurate single image super-resolution via information distillation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 723–731. IEEE, Piscataway (2018)

# Chapter 16
# Facial Expression Retargeting from a Single Character


**Ariel Larey, Omri Asraf, Adam Kelder, Itzik Wilf, Ofer Kruzel, and Nati Daniel**

**Abstract** Video retargeting for digital face animation is used in virtual reality, social media, gaming, movies, and video conference, aiming to animate avatars' facial expressions based on videos of human faces. The standard method to represent facial expressions for 3D characters is by blendshapes, a vector of weights representing the avatar's neutral shape and its variations under facial expressions, e.g., smile, puff, blinking. Datasets of paired frames with blendshape vectors are rare, and labeling can be laborious, time-consuming, and subjective. In this work, we developed an approach that handles the lack of appropriate datasets. Instead, we used a synthetic dataset of only one character. To generalize various characters, we re-represented each frame to face landmarks. We developed a unique deep learning architecture that groups landmarks for each facial organ and connects them to relevant blendshape weights. Additionally, we incorporated complementary methods for facial expressions that landmarks did not represent well and gave special attention to eye expressions. We have demonstrated the superiority of our approach to previous research in qualitative and quantitative metrics. Our approach achieved a higher mean opinion score (MOS) of 68% and a lower mean square error (MSE) of 44.2% when tested on videos with various users and expressions.

## 16.1 Introduction

Various applications use video retargeting for digital face animation. These domains include social media, gaming, movies, and video conferences. Video retargeting systems aim to translate human facial expressions into 3D characters, eventually mimicking the real footage expressions.

A common method to represent facial expressions of 3D characters is by blendshapes. In this method, different mesh shapes serve as targets where each mesh

A. Larey, O. Asraf, and N. Daniel: These authors contributed equally to this work.

A. Larey · O. Asraf · A. Kelder · I. Wilf · O. Kruzel · N. Daniel (✉)
Huawei Research, Tel Aviv, Israel
e-mail: snatidaniel@gmail.com



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_16

target has its corresponding blendshape weight that determines its significance in the desired expression. Mathematically, the mesh targets serve as eigenvectors and the blendshape weights as linear combination coefficients, resulting in an interpolated expressed mesh. In the case of video retargeting, the objective is to translate expressions from real footage videos to sequences of blendshape weights to control 3D characters' facial animation. Particularly in this study, we used a 3D character with 62 mesh targets with semantic meanings such as smile, puff, and blink in addition to multiple Visemes.

The large variation of facial expressions and character shapes poses a significant obstacle in training efficient models: creating a large, representative dataset of video frame—blendshape weights pairs. This dataset type can be created manually by labeling each video frame with its blendshape weights.

Yet, such a process is labor-intensive, time-consuming, and highly subjective—when individual observers can interpret the same image differently, causing a nondeterministic labeling process. A reasonable solution is training the machine learning model using a synthetic dataset.

In this approach, realistic 3D characters deform into numerous expressions based on pre-defined blendshape weights, which serve as the dataset's deterministic ground truth. Next, each mesh is rendered into realistic scenes used as input images during the training procedure. However, a significant challenge in this approach is overcoming the domain gap between the synthetic scenes the model encounters during training and the real scenes it encounters during inference. Producing a diverse synthetic dataset using high-poly realistic 3D characters that represent the photorealistic scenes sufficiently can, on the one hand, narrow the domain gap but, on the other hand, can be very expensive. We combat that multiplicative growth of efforts by training the video to blendshape weights conversion on a single character. To narrow the domain gap between the trained single character to the realistic human facial scenes, we use a well-known representation of face structure that captures expressions quite well—facial landmarks, particularly the standard set of 68 landmarks [15]. The conversion of a face image, depicting an actual person or a human-like ('realistic') 3D character, to a set of landmarks performs data reduction into a 'symbolic' representation. We train a blendshape translation network in that space using a single character.

Furthermore, the local nature of both blendshape coefficients and face landmarks allows partitioning the problem into subsets—e.g., eye landmarks versus eye blendshapes. Moreover, working with blendshape provides an additional advantage—the ability to apply the predicted blendshape coefficients to characters of identical topology but different geometry, thus generalizing from our single character to a wide range of avatars.

Section 16.2 reviews related work, highlighting specific challenges. Our method is described in Sect. 16.3, starting with our data preparation. We generate synthetic data according to a real-world distribution of head poses and by modifying blendshape coefficients to generate plausible expressions from a single 3D character that can be matched to video frames of real actors.

We further apply landmark detection to the synthetic images and describe how to regress blendshape weights locally from landmarks. While the landmark networks

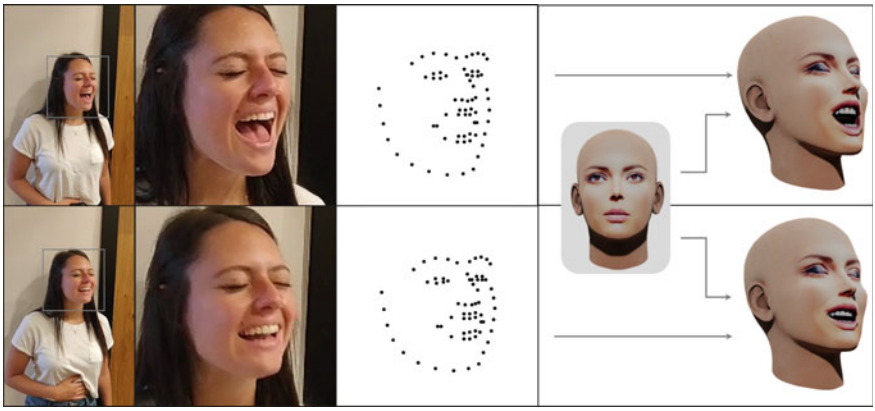


**Fig. 16.1** Results of our method on two video frames. From left to right: input image, cropped input image, extracted facial landmarks, our output expressed 3D character

reliably present many expressions, they fail to replicate eye blinks and eye gaze directions. Other expressions, like puff and sneer, occur in face regions, poorly represented by landmarks. Thus, we added complementary methods for these regions and expressions.

Finally, our results in Sect. 16.4 show that our approach scales well to videos with different users and various expressions, outperforming previous research in qualitative and quantitative metrics (Fig. 16.1).

## 16.2 Related Work

Video-driven facial animation, particularly in the context of Video retargeting/Motion capture, is a challenging task that has gained significant attention in recent years. One of the primary objectives of video-driven facial animation is to automatically transfer facial expressions from videos onto 3D characters.

Hence, it enables a wide variety of product applications in virtual reality and augmented reality eras. In particular, in social media, gaming, movies, music clips, and video conference scenarios, driving a 3D avatar gives the illusion of the real world and improves the viewer and user experience.

Several techniques and approaches have been proposed to realistically address the complexities of capturing and transferring facial expressions. In this section, we present a review of the relevant literature and highlight the key contributions and advancements in the field.

### 16.2.1 Facial Expressions Animation Transfer

The emergence of deep learning has revolutionized many areas of computer vision and computer graphics, including facial animation transferring. Many researchers have leveraged deep neural networks, generative adversarial networks, and 2D/3D facial landmarks for facial animation retargeting, whose goal is typically to capture the facial performances of the source actor and then transfer the expression to a target character.

Cao et al. [6] suggests a system that learns to regress blendshapes values based on 3D facial landmarks. This approach used a paired database of 2D images and user-specific blendshapes for the training process. Siarohin et al. [26] developed a novel deep model based on a generative adversarial network (GAN) [13] that transfers the captured source facial expressions to a different actor, thus allowing for personalized retargeting.

Furthermore, [5] suggests a deep learning solution that regresses a displacement map to predict dynamic expression based on inferring accurate 2D facial landmarks and the geometry displacements from an actor video. Cao et al. [4] derived a system that learns the dynamic rigidity of prior images from 2D facial landmarks and motion-guided correctives [17]. In addition, [2] developed a method for 2D-3D facial expression transfer by estimating the rigid head pose and non-rigid face expression from 2D actor facial landmarks using an energy-based optimization solved by the non-linear least square problem. Peng et al. [24] introduced a method that combines optical-low estimation with a mesh deformation model to establish correspondences between the given actor video to the target 3D character.

Besides, other approaches have recently been using domain transfer methods such as [1, 21] to animate any 3D characters from human videos. Specifically, [1] method learns to predict the facial expressions in a geometrically consistent manner and relies on the 3D rig parameters. This requires a large database of synthetic 2D image characters aligned to 3D facial rig [9, 18, 19, 22]. While the [21] method learns to transfer animations between distinct 3D characters without consistent rig parameters and any engineered geometric priors.

### 16.2.2 Facial Expressions Animation Synthesis

Recently, there has been a shift toward 3D-based approaches for facial expression synthesis. With the availability of depth sensors and advanced 3D modeling techniques, researchers have explored methods to capture and represent facial expressions in three dimensions. This has led to the development of more accurate and detailed facial expressions models, such as 3D Morphable Models (3DMMs) and blendshape-based models, which enable more realistic retargeting of facial expressions onto different actors [7, 8, 11, 25].

## 16.3 Method

### *16.3.1 Preliminaries*

Our method relies on datasets with reasonable facial expressions and with natural head poses, similar to the way achieved in our approach. The uniqueness of this process is that the main training is performed via only one character that is required to be a polygonal mesh of fixed topology that deforms to different blendshape targets linearly.

Moreover, we assume the character is a realistic avatar with human facial geometry to reduce the domain gap between the training character and the inferred real actors. The blendshape coefficients predicted during inference could also be applied to other characters, even to stylized avatars with unnatural geometries. Yet, we assume these characters support the same blendshape targets as the original character used for training. In addition, we assume these blendshape targets are constructed with a semantic meaning.

Furthermore, our methodology does not require any temporal information in both domains. On the one hand, we do not use any information regarding the character's animation or continuously rendered frames for training. On the other hand, our pipeline could work during inference on actors' frames that lack temporal relation.

### *16.3.2 Pipeline Architecture*

Our pipeline comprises several building blocks that process a facial image and predict its corresponding blendshape weights. The main route starts by detecting the face boundaries using a face detection model [3] and cropping the image based on the bounding box. Next, facial landmarks are extracted by a pre-trained model as well [3]. As a final pre-process step, the landmarks are aligned into a frontal position with a resolution of $128 \times 128$ pixels by an sRT transformation.

The aligned landmarks are used as the input to a dedicated deep model that predicts the blendshape weights that serve as the coefficients for the blendshapes linear combination. As a post-process, we fine-tune the predicted weights to verify that they are plausibly visible. To accomplish that, we constructed in advance an array (coined Reasonable-Array) that consists of all blendshape target pairs. The Reasonable-Array provides a binary indication of whether each pair of targets should be enabled together. When a pair of targets is not reasonable, the weight prediction of the smaller target is zeroed. Furthermore, when the input frames are part of a video sequence, they are processed by an Exponential Moving Average operation to smooth the temporal dynamics. Moreover, the head pose of the actor is predicted via Hope-net [10] for rendering orientation knowledge.

Yet, some targets are prone to errors when predicting them using only landmarks. Eyes expressions detection using landmarks is a challenging task where the landmarks' prediction errors propagate to the final predictions. Moreover, for some

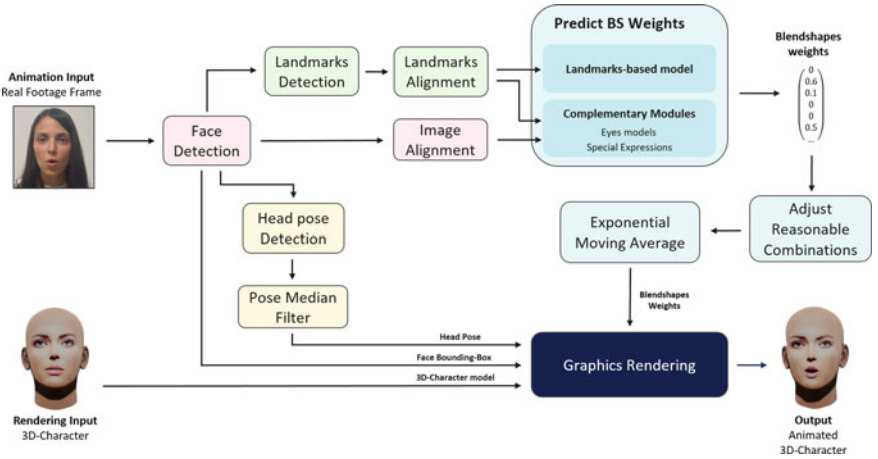



**Fig. 16.2** Overview of facial expression retargeting inference pipeline. The full solution includes sub-modules applied to real footage, such as face detection to the given video frame and extracting its face landmarks as a pre-processing phase. Then, both face images and landmarks go through an alignment process. The given aligned face and landmarks are the input for deep convolutional neural networks (landmarks to blendshapes weights network and complementary networks) that predict the corresponding blendshape weights of the given character model. The predicted weights are post-processed with the corrective expressions method (reasonable combinations) and stabilized based on previous frames' information (exponential moving average). Finally, a character is then rendered from real footage head pose, and the predicted blendshape weights to demonstrate the animated 3D character

blendshape targets, such as Puff and Sneer expressions, only 68 standard landmarks fail to cover relevant facial regions. Subsequently, the model miss-predicts the corresponding facial expressions. Thus, to achieve weights prediction over the entire set of blendshapes, complementary modules are being used in the challenging cases where the landmarks model fails. During these special scenarios, the cropped images are aligned into $128 \times 128$ pixels resolution and serve as an input to the complementary modules.

Finally, the predicted blendshape weights, face bounding box, and head pose are applied to a 3D-mesh rendering procedure. Figure 16.2 illustrates the overall pipeline for inferring a 2D image, predicting its blendshape weights, and retargeting them to a 3D mesh.

### *16.3.3 Data Preparation*

The main deep learning model was trained using a single 3D character. The objective of the data preparation phase is to create 2D landmarks of the character in various head poses and reasonable expressions while maintaining its blendshape weights as the ground truth for the training procedure.

**Synthetic Data Pre-process** First, we learn the distribution of natural head poses by predicting the orientation of real-life video frames that reflect the use-case scenario. The head pose prediction is performed by Hope-net [10], and the pose of each frame is stored in a collection of realistic head poses.

Furthermore, the procedure requires creating manually in advance a Reasonable-Array that is adjusted to the character that defines blendshapes pairs that could be enabled simultaneously.

**Synthetic Data Creation** To create pairs of landmarks and ground truth blendshape weights for the training procedure, we start by rendering the 3D character into 2D images in various poses and expressions.

The head pose is randomly sampled from the realistic head poses constructed in the pre-processing phase. The expression of each frame is generated randomly, with the following limitations: (1) No more than five blendshapes are active, where active weights are considered as weights with values larger than zero. (2) Each active blendshape weight ranges from (0,1]. (3) All active weight combinations are reasonable, based on the Reasonable-Array.

Next, we process the high-resolution rendered images, similar to the process done in the inference pipeline. Still, in this case, it is performed over the rendered character images and not on real actor images. First, we obtain the bounding box of the character's face, extract its facial landmarks, and finally, align the landmarks into a frontal $128 \times 128$ pixels resolution.

**Real Footage Data Creation** Based on the real actor images dataset, specific blendshape targets not accurately reflected by the standard landmarks are predicted differently. Two eyes blinking weights are predicted by a dedicated model. In this case, we capture videos from different actors, where the actors must perform the same eye expression during the entire video, while other facial expressions, head pose, and distance from the camera alter.

This technique enabled a simple and convenient labeling procedure, where all decoded frames from the same video are labeled the same. Each video contains one of the following eye expressions: (1) close both eyes, (2) natural open eyes, (3) wink, (4) close eyes partly. All videos are decoded into frames where each frame is assigned to the label of the entire video. In addition to the blinking, a dedicated model is trained for 'Puff' and 'Sneer' facial expressions using the same data-collecting technique.

### *16.3.4 Landmarks-to-Weights*

This main module aims to translate the knowledge represented by the facial landmarks to the facial expressions, and Visemes are reflected as the blendshape weights. Thus, the inputs to the model are 68 aligned landmarks, each represented by its two horizontal and vertical coordinates ($68 \times 2$ shape). In contrast, the model outputs 62 blendshape weights that indicate the coefficients of the linear combination of the blendshape targets.

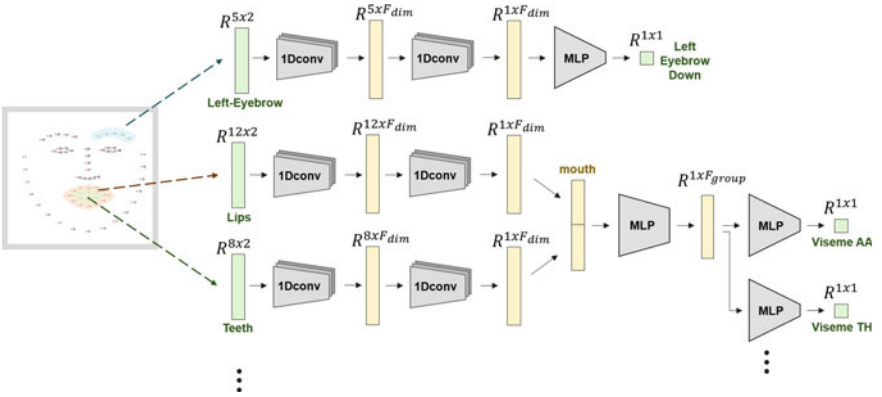



**Fig. 16.3** Landmarks to blendshape weights network architecture. The architecture stages: (1) Separating the landmarks into facial regions. (2) Feature extraction for each region by 1D-convolution layers. (3) Grouping and connecting the regions to the relevant landshape weights

The simple approaches of training regressors between the source landmarks domain to the target blendshape weights domain did not converge well. However, separating the landmarks into facial regions and propagating the information in a hierarchical route showed high performance. We pre-defined eight landmarks regions: eyebrow-left, eyebrow-right, eye-left, eye-right, nose, nostril, teeth, and lips. 1D-convolution layers processed each region of 2-dimensional landmarks into a 1-dimensional hidden layer representing the region's extracted knowledge.

The next step depends on the behavior of the blendshape targets. A grouping layer is required when blendshape targets are influenced by more than one landmarks region. In this case, the relevant regions' hidden layers are concatenated and processed by an additional MLP. Finally, each blendshape target has its dedicated MLP that outputs a scalar value in the range of [0,1] that represents the blendshape weight's value.

For example, as demonstrated in Fig. 16.3, the target representing the 'AA' Viseme depends on the 'Teeth' and 'Lips' regions of landmarks. Thus 'Lips' and 'Teeth' landmarks regions are grouped into the 'Mouth' group in advance. On the other hand, when the blendshape target is affected directly by only one region of landmarks, the region's hidden layer is regressed directly to predict the corresponding blendshape weight. For example, the blendshape target representing the left lowered eyebrow depends only on the left-eyebrow region of landmarks (Fig. 16.3).

### *16.3.5 Complementary Modules*

**Blink Detection** Predicting blinking targets given facial landmarks is challenging due to the high diversity of eye structure and surrounding textures that cause landmarks extractors' failures. These errors propagate into the landmarks-based model when predicting blendshape weights.

Thus, we train a dedicated model to directly predict two blinking blendshape weights from the given image using the real footage dataset. The input images are aligned and cropped around the eyes to the resolution of $16 \times 40$ pixels and applied to a ResNet18 model [14].

Yet, predictions do not account for different eye geometries, i.e., eyes with narrow geometry could be interpreted as partially closed. Therefore, we adjust the prediction range of values to the individual actors' eye geometry. For each frame, the distance between the lower eyelid to the upper eyelid is calculated for both eyes, which is then classified by online $K$-means into one of two classes: 'opened eyes' and 'closed eyes'. $K$-means averaged values of the two classes are updated during the video progress by the distances calculated per frame. The 'opened eyes' class value reflects the actor's natural eyelids distance and is converted by a linear transformation to a threshold between 0 and 1, which serves as the new low edge for the blinking prediction range of values.

**Gaze Detection** Herein, we derived a practical approach for accurately determining and monitoring the direction of an actor's gaze. This involves identifying whether the actor's gaze is directed straight ahead (the Primary position) or in one of the secondary positions, namely up, down, right, or left.

Our method relies on comprehensive facial eye landmarks, including the iris, inner corner keypoints, and outer corner keypoints. Specifically, the calculations are based on the distance between the iris and the inner and outer corner keypoints.

To predict the coefficients for detecting the direction of the eyes' gaze, we outline the following three steps:

1. Horizontal Eye Line Calculation: We begin by calculating the properties of the horizontal eye line using the key points of the eye corners. These properties include the mid-point, line slope, bias, and the L2 distance of the horizontal eye.
2. Intersection Point Determination: Next, we determine the intersection point between the iris projection and horizontal eye lines.
3. Secondary Positions Detection: By analyzing the obtained intersection point, we identify the secondary positions of the gaze:
   (a) Left and Right Gaze: We measure the distance of the intersection point relative to the mid-point. This distance is normalized based on the individual's horizontal eye L2 distance, resulting in a unique value for each actor. The direction of the eye is correlated with the position of the intersection point relative to the mid-point.
   (b) Up and Down Gaze: We measure the distance between the intersection and iris points to detect the upward or downward gaze. The eye's direction is correlated with the position of the iris point relative to the horizontal eye line.

In summary, leveraging comprehensive facial eye landmarks and employing specific calculations can predict the gaze blendshape coefficients for various eye positions while ensuring smooth and reliable results.

**Special Expressions Detection** 'Puff' and 'Sneer' expressions are part of the Facial Action Coding System (FACS) and refer to facial muscle movements. The standard landmarks are not sufficient to capture these expressions. Thus, to detect the corresponding blendshape weights, we train a ResNet18 model [14], using the real footage dataset.

## 16.4 Experiments

We used three datasets encompassing various facial expressions and video sequences to conduct our experiments. In addition, we demonstrated our study's methodological advancement through reproducibility, transparency, and comparability. Finally, we evaluated our approach's performance using qualitative and quantitative measures. Assessing the quality and performance of facial expressions retargeting algorithm is a crucial aspect of research in this domain. Therefore, various evaluation metrics have been proposed to measure the realism, accuracy, and perceptual quality of retargeted facial expressions, poses, and identities.

### *16.4.1 Datasets*

**Synthetic Character Landmarks Dataset** We used a 3D character consisting of 12.8K vertices for each of its 62 blendshape targets to train the landmarks-based model. The blendshapes included various facial expressions and mimics in addition to 14 Visemes (Attribution 4.0 International license [27]). We created manually in advance a Reasonable-Array that is adjusted to this specific set of blendshapes and performed the data preparation as described above.

We used the Blender tool to render 30,000 images of the character, saved their corresponding blendshapes weights as ground truth, and extracted their aligned facial landmarks. The head pose of each frame scene was sampled from a natural head pose distribution. This distribution was obtained by detecting head pose information from relevant Youtube and Denver Intensity of Spontaneous Facial Action (DIFSA) videos [20].

**Real Footage Labeled Dataset** We collected 200 videos from 40 actors targeting blinking, Puff, and Sneer expressions. The videos were decoded at 30 fps yielding 17,101 real footage frames and their corresponding labeled blendshape weights. This dataset was dedicated to supervised training of the Special Expressions models.

**Real Footage Unlabeled Dataset** To evaluate the performance of our pipeline, we captured additional real footage videos from 20 identities. The actors were requested to perform various FACS expressions and Visemes. The videos were decoded at 30 fps resulting in 25,075 real footage frames.

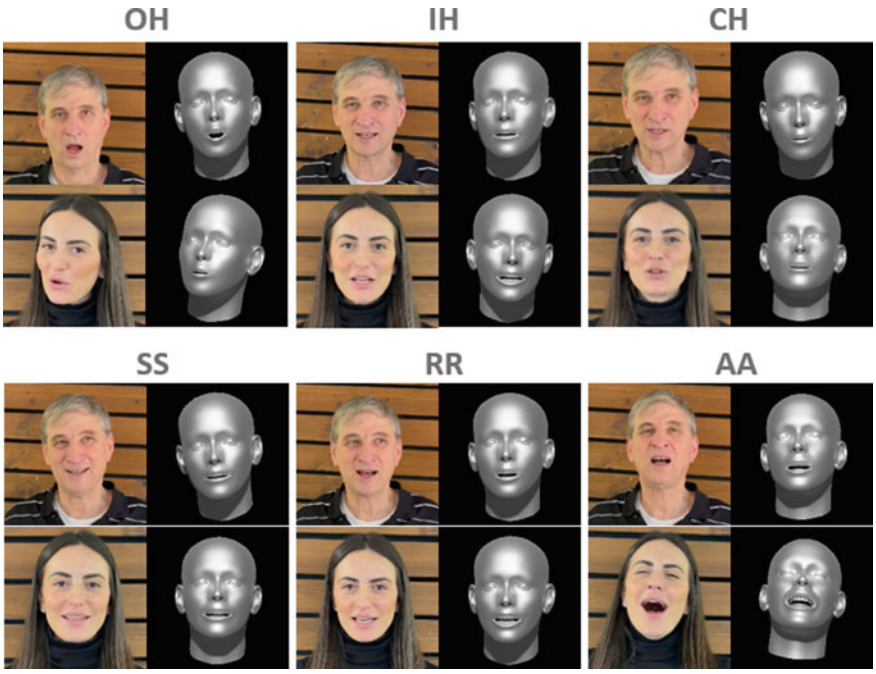



**Fig. 16.4** Examples of prediction performance to a variety of Visemes. For each Viseme from left to right: (1) input image, (2) our output expressed 3D character

### *16.4.2 Training Procedure*

The updated model was trained and optimized using Pytorch [23] framework on a single NVIDIA GeForce GTX 1080 with 24 GB GPU memory.

The hyperparameters of the model that optimize convergence were examined using Adam Solver [16] with beta1 = 0.5 and beta2 = 0.999, eps of 1e−10, weight decay of 1e−7, a minibatch of size 16, a learning rate of 5e−5 for 100 epochs. In contrast, we decay the learning rate to zero by gamma = 0.5 every three epochs. Weights were initialized from a uniform distribution described in [12]. We use Leaky ReLUs with slope 1e−2 for the convolutional layers and a fully connected layer, in addition to Sigmoid activation at the end of the fully connected layer.

The optimization loss function contains two terms. First, a mean square error (MSE) between all ground truth and predicted blendshape weights. Second, we used an MSE between active ground truth and predicted blendshape weights. At the same time, the loss function elements were weighted with values of 1 and 0.1, respectively.


**Fig. 16.5** Examples of prediction performance to a variety of expressions. For each expression from left to right: (1) input image, (2) our output expressed 3D character

### *16.4.3 Results*

**Facial Expressions and Visemes** We examined our method using the Real Footage Unlabeled Dataset. The video frames were subjected to our pipeline, producing Blendshape weights, which serve as coefficients of the linear combination between the mesh geometry targets. Figures 16.4 and 16.5 show examples of real footage frames and their corresponding rendered mesh with the translated facial expression. The former presents different Visemes as part of a video speaking sequence, while the latter presents other common facial expressions frames.

**Multiple Characters** The uniqueness of our method stems from the single-character training procedure where only one 3D character is enough to obtain sufficient results. Yet, we can drive any desired character during inference if it supports the same geometrical space and can be animated using the same semantic blendshapes. Figure 16.6 shows the robustness of our model over multiple actor identities and multiple 3D characters constructed by their unique geometries and textures. These characters were not part of the training but are represented in the same semantic blendshape space as the single mesh used for training [27].

**Competitive Comparison** As a baseline to the single-character video retargeting pipeline, we implemented the algorithm proposed in [21] where a semi-supervised approach that included an image translation technique was introduced. We inferred the baseline model over the Real Footage Unlabeled Dataset as well. Examples of our method's results compared to [21] method are presented in Fig. 16.7.

For qualitative evaluation, we conducted a subjective user study where 80 individuals were required to rate the degree of compatibility between pairs of actors and their corresponding rendered 3D character images in facial expressions aspects. The user study contained 34 pairs, where each actor frame appeared twice, once with a prediction obtained by our method and once with a prediction obtained by [21] method (each time in random order). The subjects rated each pair on the Likert scale (scores between one to five), and we reported their mean opinion score (MOS) separately for Visemes representative frames and other facial expressions representative frames.

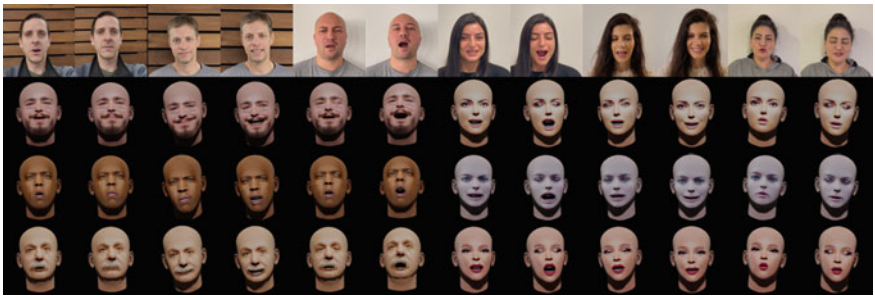


**Fig. 16.6** Examples of performance when retargeting to multiple 3D characters from different actors. For each actor from top-to-bottom: (1) input image, (2) our results expressed on three different characters

**Table 16.1** Quantitative and qualitative performance of ours and Moser et al. approaches

| | FACS | | Visemes | | Overall | |
|---|---|---|---|---|---|---|
| Evaluation-metric\Method | Moser et al. | Ours | Moser et al. | Ours | Moser et al. | Ours |
| Quantitative (MSE) ↓ | 40.43 | 29.95 | 42.73 | 22.59 | 40.92 | 28.37 |
| Qualitative (MOS) ↑ | 2.43 | 4.28 | 2.38 | 3.81 | 2.41 | 4.05 |

The quantitative evaluation represents the MSE of landmarks similarity on 39 videos. The qualitative evaluation represents the subjective results (MOS) of a survey of 80 participants
*MSE* mean square error—a quantitative measure; *MOS* mean opinion score—a qualitative measure
Downarrow symbol indicates lower is better; Uparrow symbol indicates higher is better

For quantitative evaluation, we introduce the landmarks similarity metric. Herein, the source actor frame and the corresponding rendered 3D character image are compared based on facial landmarks. First, both images are cropped around their face. Next, facial landmarks are extracted from both crops and are processed by landmarks alignment procedure to the same template using an sRT transformation. A mean square error (MSE) is calculated between aligned landmarks, representing the distance between the source actor's facial expression and the translated expression over the 3D character. Table 16.1 shows that our method outperforms [21] method in both qualitative and quantitative metrics.

**Ablation Study** In this section, we provide ablation experiments substantiating the need for the different grouping layers of the landmarks to blendshape weights network. The No-Grouping network replaces all layers with a simple MLP model that directly regresses the blendshape weights from the facial landmarks. The Conv-Grouping model only uses our grouping method for the 1D-convolution layers (i.e., each facial region has its convolution weights). The convolution features are propagated to the blendshape weights directly via an MLP network eliminating the last grouping layers. The Full-Grouping model contains all grouping layer components. In Table 16.2, all these effects are reported by calculating the MSE between the facial

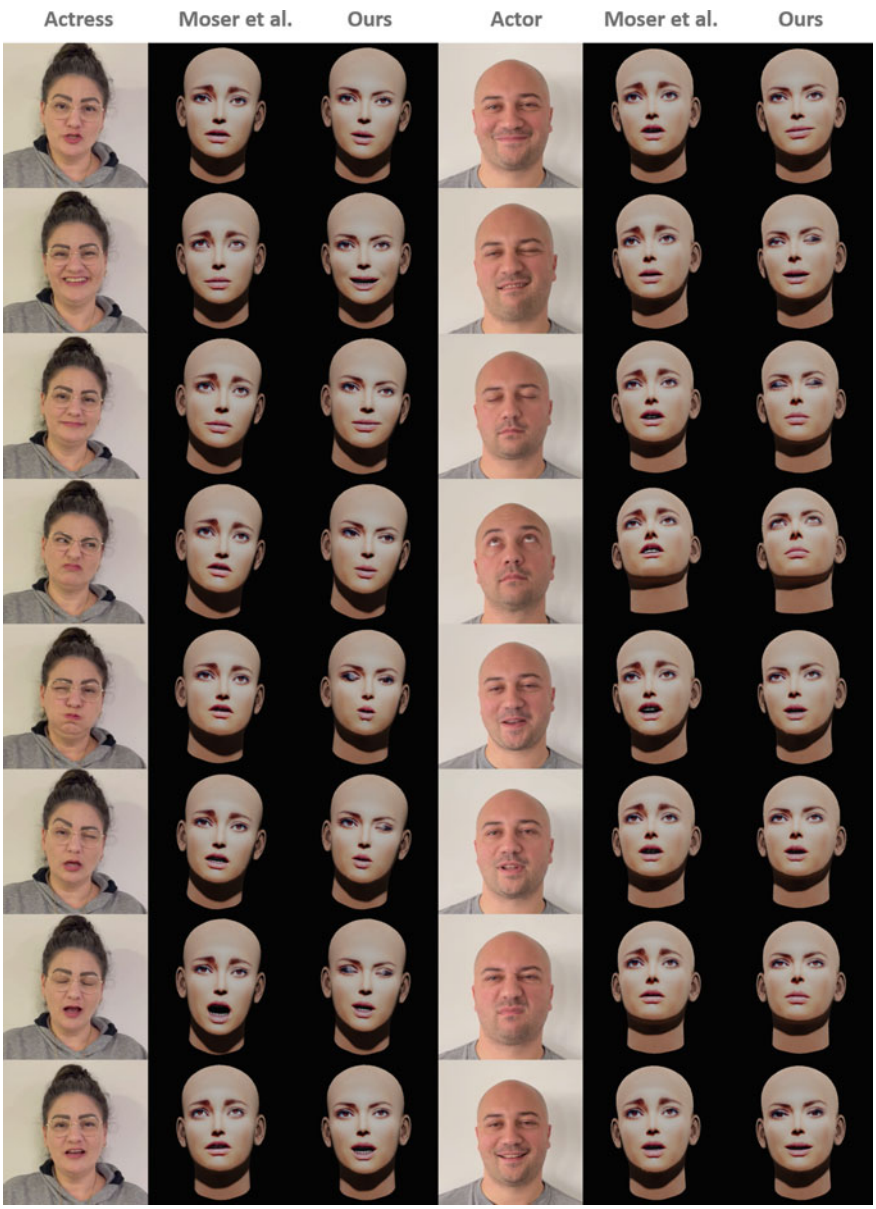



**Fig. 16.7** Video retargeting comparing. Examples of ours versus Moser et al. pipeline results of different facial expressions for two real actors

**Table 16.2** Ablation study for the different grouping layers of the landmarks to blendshape weights network

| | Overall results | | | |
|---|---|---|---|---|
| Evaluation-metric\Method | Moser et al. | Ours no-grouping | Ours conv-grouping | Ours full-grouping |
| MSE ↓ | 40.92 | 31.52 | 28.48 | 28.37 |

*MSE* mean square error—a quantitative measure
Downarrow symbol indicates lower is better

landmarks of the actors to the 3D characters on the test dataset. One can observe an improvement in accuracy at the introduction of every proposed component with a final reduction of 10% in the Error over the baseline.

## 16.5 Conclusion

One of the main challenges of establishing a video retargeting system is acquiring an optimal dataset for the supervised learning approach. Labeling each video frame manually with its corresponding blendshape weights (62 blendshapes in our case) can be laborious, time-consuming, and subjective. A reasonable solution could be training an AI model via a synthetic dataset. Yet, this approach requires an expensive dataset that contains a diverse range of 3D characters rendered in high-quality photorealistic scenes. In this work, we describe a technique that overcomes these challenges to produce a facial animation model trained with only a single 3D character.

Here, we introduce a full pipeline that benefits from facial landmarks to reduce the domain gap between the synthetic 3D character encountered during training to the real footage inference actors. This approach eliminates using a large-scale, expensive dataset and enables us to achieve sufficient performance with only one 3D character.

Through this pipeline, we translate the facial landmarks information into blendshape weights by a unique grouping approach where each spatial region of landmarks is grouped, and the knowledge propagates hierarchically until it reaches the corresponding target shape. We further demonstrate a technique to complete the expressions range by implementing target-shape-specific sub-modules.

We show the effectiveness of our method in various aspects. The proposed pipeline captures the real footage of actors' facial expressions in both Visemes and FACS frames. We demonstrate the robustness of our method by capturing multiple actors and applying their frames to the pipeline over multiple 3D characters that were excluded from the training procedure. We further compared our results to Moser et al. qualitatively and quantitively, achieving a higher mean opinion score (MOS) by 68% and a lower mean squared error (MSE) by 30.7%.

Overall, our work provides a state-of-the-art solution for video retargeting using a single 3D character in a high-level pipeline and low-level deep architecture.

**Acknowledgements** The authors thank all the human actors and actresses who participated in the work experiments.

## References

1. Aneja, D., Chaudhuri, B., Colburn, A., Faigin, G., Shapiro, L., Mones, B.: Learning to generate 3d stylized character expressions from humans. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 160–169. IEEE (2018)
2. Barros, J.M.D., Golyanik, V., Varanasi, K., Stricker, D.: Face it!: a pipeline for real-time performance-driven facial animation. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 2209–2213. IEEE (2019)
3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1021–1030 (2017)
4. Cao, C., Chai, M., Woodford, O., Luo, L.: Stabilized real-time face tracking via a learned dynamic rigidity prior. ACM Trans. Graph. (TOG) **37**(6), 1–11 (2018)
5. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. (TOG) **33**(4), 1–10 (2014)
6. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3d shape regression for real-time facial animation. ACM Trans. Graph. (TOG) **32**(4), 1–10 (2013)
7. Chaudhuri, B., Vesdapunt, N., Shapiro, L., Wang, B.: Personalized face modeling for improved face reconstruction and motion retargeting. In: European Conference on Computer Vision, pp. 142–160. Springer, Berlin (2020)
8. Chaudhuri, B., Vesdapunt, N., Wang, B.: Joint face detection and facial motion retargeting for multiple faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9719–9728 (2019)
9. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2106–2112. IEEE (2011)
10. Doosti, B., Naha, S., Mirbagheri, M., Crandall, D.J.: Hope-net: a graph-based model for hand-object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6608–6617 (2020)
11. Egger, B., Smith, W.A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models-past, present, and future. ACM Trans. Graph. (TOG) **39**(5), 1–38 (2020)
12. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, pp. 249–256 (2010)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR (2015). arXiv preprint arXiv:1512.03385
15. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980
17. Kroeger, T., Timofte, R., Dai, D., Van Gool, L.: Fast optical flow using dense inverse search. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 Oct 2016, Proceedings, Part IV 14, pp. 471–488. Springer, Berlin (2016)

18. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops, pp. 94–101. IEEE (2010)
19. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: DISFA: a spontaneous facial action intensity database. IEEE Trans. Affect. Comput. **4**(2), 151–160 (2013)
20. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: DISFA: a spontaneous facial action intensity database. IEEE Trans. Affect. Comput. **4**(2), 151–160 (2013). https://doi.org/10.1109/T-AFFC.2013.4
21. Moser, L., Chien, C., Williams, M., Serra, J., Hendler, D., Roble, D.: Semi-supervised video-driven facial animation transfer for production. ACM Trans. Graph. (TOG) **40**(6), 1–18 (2021)
22. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: 2005 IEEE International Conference on Multimedia and Expo, p. 5. IEEE (2005)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
24. Peng, Z., Jiang, B., Xu, H., Feng, W., Zhang, J.: Facial optical flow estimation via neural non-rigid registration. Comput. Vis. Media **9**(1), 109–122 (2023)
25. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7763–7772 (2019)
26. Siarohin, A., Lathuiliere, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019
27. Skullvez: Synthetic 3d characters data creator. https://sketchfab.com/skullvez. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/legalcode

# Chapter 17
# Research on Intelligent Fault Identification Method Based on UAV Power Inspection


**Feng Weixi, Li Qing, and Xu Peng**

**Abstract** To build an intelligent fault detection system based on UAV patrol and improve the efficiency of technicians, a novel multilevel and multi-scale feature fusion network (M2F2N) is developed. Specifically, to assign more weights to important features, such as high-frequency information containing image edge details, we use two parallel channel attention and spatial attention branches in dual-attention feature extraction block (DAFEB) to promote information interaction and learn interdependency across different channels. In addition, a multi-scale future fusion block (MSFFB) is designed based on stacking several DAFEBs to extract more effective features. Finally, the features extracted from different hierarchies are fused and aggregated via concatenation. Adequate experiments validate that M2F2N exceeds the compared models, i.e., SRCNN FSRCNN, VDSR, DRRN, and IDN, in terms of quantitative and qualitative results.

## 17.1 Introduction

In the current digital transmission business, the intelligent fault identification technology based on UAV power patrol inspection can effectively improve the management and production efficiency of power companies and can also realize the digital transformation of traditional technical analysis work. It liberates technical workers from offline "cousins" and "cousins" and changes them into online data analysis engineers, effectively improving the efficiency of fault analysis and emergency repair. The intelligent fault identification system based on UAV electric power patrol inspection can access the line fault location device; integrate equipment accounts, hidden dangers, defects, lightning, and other multi-source heterogeneous data; and achieve rapid fault location of transmission lines, automatic export of preliminary analysis reports, and automatic sending of SMS. At the same time, intelligent terminals such as video, micrometeorology, and grounding circulation are used to realize fault linkage

F. Weixi (✉) · L. Qing · X. Peng
Shenzhen Power Supply Bureau Co., Ltd., Shenzhen 518000, Guangdong, China
e-mail: duanzhiwei69240130@163.com



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_17

analysis and video one click capture, support on-site fault finding, and improve the efficiency of fault finding by more than 50%. At present, the popular image super-resolution technique on the basis of deep learning can provide greater help for the intelligent fault identification system based on UAV power inspection. It is capable of not only saving computation resource, but also improving the processing accuracy and efficiency of the system's fault identification.

Single image super-resolution (SISR) reconstruction technique is capable of reconstructing high-resolution (HR) inputs from one or more existing low-resolution (LR) images in the same scene using image processing technology and machine learning technology. Recently, with the rapid development of convolutional neural network (CNN), SR method based on deep learning (DL) has achieved excellent reconstruction results. Dong et al. proposed the first DL-based super-resolution convolution neural network (SRCNN) [1] to achieve end-to-end mapping function learning across LR inputs and HR ones, which greatly improved the image reconstruction effect. However, SRCNN uses the upsampled image of LR image as the input of the model, resulting in many network parameters and low efficiency. Therefore, Dong and Tang et al. proposed an improved network, namely FSRCNN [2], which directly uses the original LR image as the input, and used deconvolution method to reconstruct the image at the end of the network, speeding up the network training process. Kim et al. developed an image SR reconstruction method, i.e., VDSR, utilizing more than 20 layers of convolution based on depth convolution networks, which deepened the network to 20 layers, combined with residual learning and adaptive gradient clipping to accelerate the training and learning of models, and improved the reconstruction effect [3]. Subsequently, various DNN-based methods [4–7] were proposed to enhance the performance of SISR. However, these methods not only ignore to utilize the attention to learn the interdependency across various channel, but also bring too much parameters with deeper network. In other words, the network consumes too much computational resources, and it is not suitable to be applied to actual production scene.

To solve the above-mentioned issues, a well-designed lightweight multilevel and multi-scale feature fusion network (M2F2N) is developed for SISR, which can extract and aggregate multi-scale and multilevel discriminative features.

## 17.2 Related Work

### *17.2.1 CNN-Based SR*

Recently, enormous deep learning-based methods have been proposed to address the SR problems and have achieved great success. Different types of networks have been exploited in the task of SR. Among them, Dong et al. [1] firstly introduced the CNN layers to the field of SR and therefore designed a simple three-layered network called SRCNN for the single image SR. Subsequently, Kim et al. [3] stacked 20 CNN layers

and thus built a well-known network named VDSR. Motivated by the performance improvements of deep neural network, a series of deep neural networks [2, 4, 6-11] is designed for SR tasks, which greatly improved the performance SR performance. Li et al. [12] realized feedback manner based on using hidden states and therefore constructed a feedback block extract refined powerful high-level representation for single image SR.

### *17.2.2 Lightweight Network for SR*

There are many works designed for lightweight image SR. Tai et al. [5] established a deep recursive neural network (DRRN) with recursive unit based on residual blocks [13]. Later, Ahn et al. [14] proposed a cascaded model named CARN combining the recursive mechanism with different residual skip connections. Moreover, Hui et al. [15] carefully designed a lightweight information multi-distillation network (IMDN) for single image SR.

In this paper, we carefully design a novel MSFFB for single image super-resolution. MSFFB is a deep feature extraction block which is made up of a series of multi-scale future fusion blocks.

## 17.3 Proposed Method

Here, we will discuss in detail the network framework of our proposed method. For simplicity, the proposed network is named as M2F2N—multilevel and multi-scale feature fusion network.

M2F2N is mainly divided into three parts: shallow feature extraction (SFE), deep multilevel feature extracting block (DMFE), and reconstruction part. The shallow feature extraction block is composed of one convolution layer, which is used to extract shallow features of the given LR input. The deep feature extraction module is composed of a series of multi-scale future fusion blocks (MSFFBs). The residual skip connections mechanism can be utilized to transfer the output features and shallow features of each MSFFB to the multilevel feature fusion module for feature aggregation. The image reconstruction module consists of a $3 \times 3$ convolution layer and an upsampling reconstruction layer. The sub-pixel convolution proposed by Shi et al. [8] is used to realize the upsampling operation in the image reconstruction stage. The overall structure of our proposed model is depicted in Fig. 17.1.

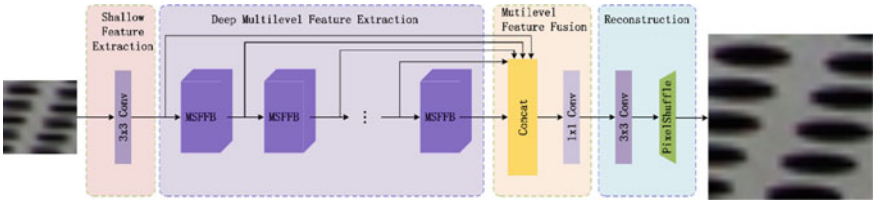



**Fig. 17.1** Framework of our proposed multilevel and multi-scale feature fusion network (M2F2N)

### *17.3.1 Shallow Feature Extraction Layer*

The feature extraction process of M2F2N is divided into two stages: shallow feature extraction and deep feature extraction. As shown in Fig. 17.1, the shallow feature extraction module contains a layer of convolution. The process of extracting shallow features is stated as follows:

$$F_0 = H_{\mathrm{SFE}}(I_{\mathrm{LR}}), \tag{17.1}$$

where $H_{\mathrm{SFE}}(\cdot)$ denotes the implicit function of $3 \times 3$ convolution, $I_{\mathrm{LR}}$ is the input of model, and $F_0$ represents the extracted shallow feature, which is the initial input of deep feature extraction module.

### *17.3.2 Deep Feature Extraction Based on MSFFB*

As shown in Fig. 17.1, the deep feature extraction module is composed of 4 MSFFBs, and a multilevel feature fusion structure is established. The input of multilevel feature fusion module is from all the previous MSFFBs. Multilevel feature fusion structure can reduce the number of parameters to some extent and maximize the usage of image feature information from various hierarchies. The process can be represented as:

$$F_{\mathrm{MSFF}} = H_{1\times 1\mathrm{conv}}([F^1_{\mathrm{MSFF}}, \ldots, F^4_{\mathrm{MSFF}}]), \tag{17.2}$$

where $H_{1\times 1\mathrm{conv}}(\cdot)$ indicates the operation of $1 \times 1 \times$ convolution, $F^k_{\mathrm{MSFF}}$ represents output of the $k$-th MSFFB, while $[\cdot]$ denotes the concatenation. $F_{\mathrm{MSFF}}$ is the aggregated feature.

Figure 17.2 shows the structure of multi-scale future fusion block (MSFFB). The MSFFB is composed of two convolutional branches of various kernel sizes ($3 \times 3$ and $5 \times 5$). Then, the multi-scale features extracted by these two convolutions are concatenated and aggregated by one $1 \times 1$ convolution. Finally, the output is fed into dual-attention feature extraction block (DAFEB). The above procedure can be expressed as follows:

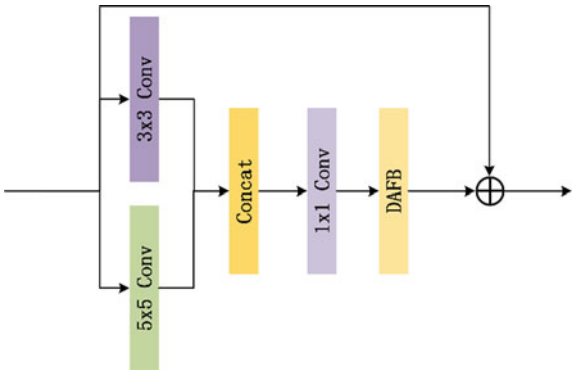



**Fig. 17.2** The multi-scale future fusion block (MSFFB)

$$F_{\mathrm{MSFF}}^{k} = H_{\mathrm{DAFEB}}(H_{1\times1}([H_{3\times3}, H_{5\times5}])), \tag{17.3}$$

where $H_{n\times n}(\cdot)$ and $H_{\mathrm{DAFEB}}(\cdot)$ denote $n \times n$ convolution part and the operation of DAFEB, respectively.

### *17.3.3 Dual-Attention Feature Extraction Block (DAFEB)*

To learn and acquire the importance of different information of each feature channel, a dual-attention feature extraction block (DAFEB) is carefully developed, which is depicted in Fig. 17.3. Generally, channel attention (CA) can suppress the information of low-frequency channel and strengthen the importance of information of high-frequency channel. Besides, not only the significance of each feature map is different, but also the texture details at different locations on each feature map are different. Therefore, the spatial attention (SA) is introduced to capture the textures and border for rebuilding HR images. The calculation process of DAFEB can be expressed as:

$$F_{\mathrm{MSFF}}^{k} = H_{\mathrm{CAB}}(f_{\mathrm{in}}) + H_{\mathrm{SAB}}(f_{\mathrm{in}}), \tag{17.4}$$

where $H_{\mathrm{CAB}}(\cdot)$ and $H_{\mathrm{SAB}}(\cdot)$ indicate the implicit function of channel attention and spatial attention, respectively.

### *17.3.4 Image Reconstruction Module*

As shown in Fig. 17.1, the image reconstruction module consists of a sub-pixel convolution layer and a reconstruction layer. The reconstruction process can be formulated as follows:

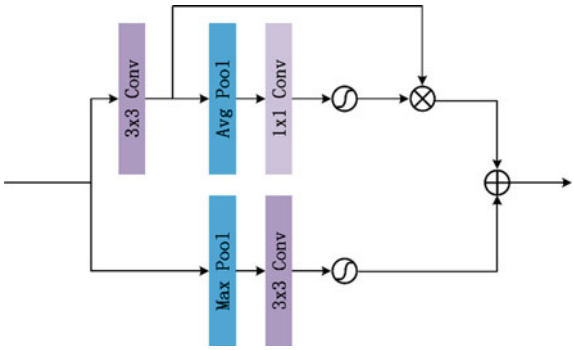



**Fig. 17.3** Dual-attention feature extraction block (DAFEB)

$$I_{\mathrm{SR}} = H_{\text{sub-pixel}}(H_{1\times1}(F_{\mathrm{MSFF}})), \tag{17.5}$$

where $H_{\text{sub-pixel}}(\cdot)$ and $I_{\mathrm{SR}}$ denote the sub-pixel convolution and reconstructed feature, respectively.

### *17.3.5 Loss Function*

We adopt L1 loss to optimize M2F2N following the practices within most of the previous works [16]. We assumed that the LR image training set is $\{(I^j_{\mathrm{LR}}, I^j_{\mathrm{HR}})\}^M_{j=1}$, which is composed of pair-wise HR-LR images. The loss function is represented as:

$$L(\Theta) = \arg\min \frac{1}{M}\sum_{j=1}^{M}\left|I^j_{\mathrm{SR}} - I^j_{\mathrm{HR}}\right|, \tag{17.6}$$

where $M$ and $\Theta$ indicate the amount of pair-wise HR-LR image and network parameters, respectively.

## 17.4 Experiments

### *17.4.1 Datasets and Metrics*

The DIV2K [17] dataset is used to train the model, which consists of 800 images for training, 100 verification images as well as the same number of test ones. Specifically, the experiment uses 800 of these training images for network model training. In addition, in order to prevent overfitting during network training, images in the original training set are randomly rotated by 90°, 180°, and 270° during image preprocessing and then horizontally flipped to obtain 3200 image enhancement datasets. The new

dataset is generated based on the DIV2K dataset, which can narrow the gap between the training set and the verification set and extract more discriminative information. Five widely utilized benchmark datasets are employed as test datasets from Urban100 [18], Manga109 [19], BSD100 [20], Set5 [21] to Set14 [22]. are used as test datasets. The SSIM and PSNR are computed on YCbCr color space's Y channel.

### *17.4.2 Implementation Details*

In the experiment, the training epoch is set to 900, the training batch size is set to 32, the input and output channels are set to 64, and the initial value of all adaptive weights is set to 1. In each training batch, the LR image with the size of $45 \times 45$ is cropped as the input, and then its HR ground truth is used to train. During training process, the Adam optimizer is employed to update the weight parameters of the network. During the update process, the exponential decay rate of Adam optimizer is set to $\beta_1 = 0.9$, $\beta_2 = 0.99999$, and $\varepsilon = 1 \times 10^{-8}$, we initially set learning rate as 0.0002, and it is halved after every $2 \times 10^5$ back-propagation iterations. The Kaiming initialization is adopted [23].

### *17.4.3 Comparison with State of the Arts*

To verify the effectiveness of the proposed algorithm in this paper, M2F2N is compared with six existing image super-resolution methods from Bicubic, SRCNN [1], FSRCNN [2], DRRN [5], VDSR [3], to DRCN [4] in terms of objective evaluation indicators and subjective visual results on three different scale factors (from $\times$ 2, $\times 3$, to $\times 4$).

Table 17.1 shows the PSNR and SSIM obtained by different SISR reconstruction algorithms on five publicly utilized datasets. It can be seen from the data in Table 17.1 that most of the objective evaluation indicators of M2F2N proposed in this paper are at the leading level. Notably, the network parameter of M2F2N is 510K, which is less than that of VDSR (666K) and DRCN (1774K), while higher PSNR values are achieved by M2F2N on all test datasets with respect to three scale factors. These observations effectively demonstrate that the proposed M2F2N achieves a good trade-off in terms of computational resources and network parameters.

As illustrated in Fig. 17.4, the reconstruction of M2F2N is clearer and shaper than that of the compared methods. Note that compared to other methods, M2F2N yields the best reconstruction effect, especially the problem of line direction disorder is significantly reduced, and the image details are clearer.

**Table 17.1** Experimental results of various algorithms under different scale factors

| Scale factor | Method | Params. (K) | Set5 | Set14 | Manga109 | BSD100 | Urban100 |
|---|---|---|---|---|---|---|---|
| | | | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM |
| ×2 | Bicubic | – | 33.66/ 0.9299 | 30.24/ 0.8688 | 30.80/ 0.9339 | 29.56/ 0.8431 | 26.88/ 0.8403 |
| | SRCNN | 8 | 36.66/ 0.9542 | 32.45/ 0.9067 | 35.60/ 0.9663 | 31.36/ 0.8879 | 29.50/ 0.8946 |
| | FSRCNN | 13 | 37.00/ 0.9558 | 32.63/ 0.9088 | 36.67/ 0.9710 | 31.53/ 0.8920 | 29.88/ 0.9020 |
| | DRRN | 298 | *37.74/ 0.9591* | *33.23/ 0.9136* | *37.88/ 0.9749* | *32.05/ 0.8973* | *31.23/ 0.9188* |
| | VDSR | 666 | 37.53/ 0.9587 | 33.03/ 0.9124 | 36.67/ 0.9710 | 31.90/ 0.8960 | 30.76/ 0.9140 |
| | DRCN | 1774 | 37.63/ 0.9588 | 33.04/ 0.9118 | 37.55/ 0.9732 | 31.85/ 0.8942 | 30.75/ 0.9133 |
| | M2F2N | 510 | **37.84/ 0.95995** | **33.33/ 0.9149** | **38.03/ 0.9752** | **32.06/ 0.8976** | **31.26/ 0.9197** |
| ×3 | Bicubic | – | 30.39/ 0.8682 | 27.55/ 0.7742 | 26.95/ 0.8556 | 27.21/ 0.7385 | 24.46/ 0.7349 |
| | SRCNN | 8 | 32.75/ 0.9090 | 29.30/ 0.8215 | 30.48/ 0.9117 | 28.41/ 0.7863 | 26.24/ 0.7989 |
| | FSRCNN | 13 | 33.18/ 0.9140 | 29.37/ 0.8240 | 31.10/ 0.9210 | 28.53/ 0.7910 | 26.43/ 0.8080 |
| | DRRN | 298 | *34.03/ 0.9244* | *29.96/ 0.8349* | *32.71/ 0.9379* | *28.95/ 0.8004* | *27.53/ 0.8378* |
| | VDSR | 666 | 33.66/ 0.9213 | 29.77/ 0.8314 | 32.01/ 0.9340 | 28.82/ 0.7976 | 27.14/ 0.8279 |
| | DRCN | 1774 | 33.82/ 0.9226 | 29.76/ 0.8311 | 32.24/ 0.9343 | 28.80/ 0.7963 | 27.15/ 0.8276 |
| | M2F2N | 510 | **34.06/ 0.9245** | **30.01/ 0.8362** | **32.86/ 0.9397** | **28.99/ 0.8011** | **27.53/ 0.8384** |
| ×4 | Bicubic | – | 28.42/ 0.8104 | 26.00/ 0.7027 | 24.89/ 0.7866 | 25.96/ 0.6675 | 23.14/ 0.6577 |
| | SRCNN | 8 | 30.48/ 0.8628 | 27.50/ 0.7513 | 27.58/ 0.8555 | 26.90/ 0.7101 | 24.52/ 0.7221 |
| | FSRCNN | 13 | 30.72/ 0.8660 | 27.61/ 0.7550 | 27.90/ 0.8610 | 26.98/ 0.7150 | 24.62/ 0.7280 |
| | DRRN | 298 | *31.68/ 0.8888* | *28.21/ 0.7720* | *29.45/ 0.8946* | *27.38/ 0.7284* | *25.44/ 0.7638* |
| | VDSR | 666 | 31.35/ 0.8838 | 28.01/ 0.7674 | 28.83/ 0.8870 | 27.29/ 0.7251 | 25.18/ 0.7524 |

(continued)

**Table 17.1** (continued)

| Scale factor | Method | Params. (K) | Set5 | Set14 | Manga109 | BSD100 | Urban100 |
|---|---|---|---|---|---|---|---|
| | | | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM |
| | DRCN | 1774 | 31.53/ 0.8854 | 28.02/ 0.7670 | 28.93/ 0.8854 | 27.23/ 0.7233 | 25.14/ 0.7510 |
| | M2F2N | 510 | **31.81/ 0.8905** | **28.31/ 0.7746** | **29.48/ 0.8965** | **27.43/ 0.7295** | **25.53/ 0.7643** |

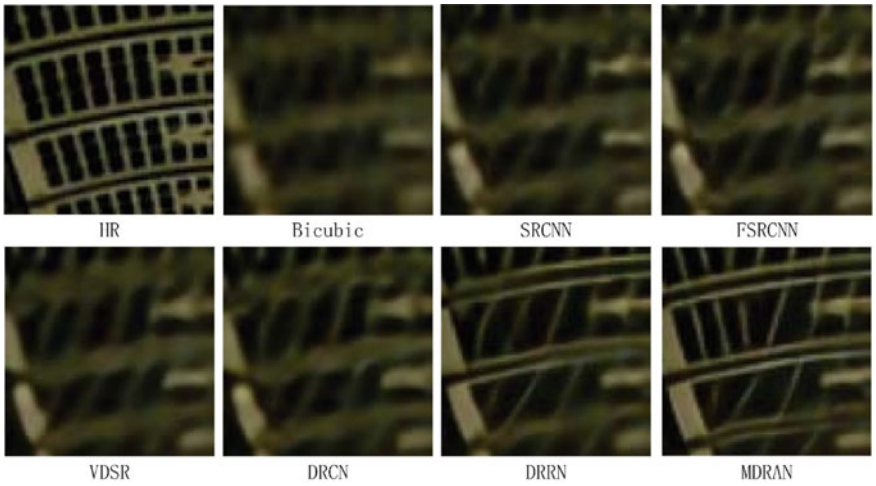



**Fig. 17.4** Reconstruction results of the compared methods for ×4 scale factor

### *17.4.4 Ablation Study*

To fully study the effect of our MSFFB, M2F2N-1 and M2F2N-2 models are built by removing the max-pooling branch channel attention and average-pooling branch channel attention mechanism from M2F2N and compared with the original model. In the experiments, M2F2N-1, M2F2N-2, and M2F2N are trained for 800 epochs, respectively, and then tested on five test sets for ×2 scale SISR. The experimental results are listed in Table 17.2, from which it can be observed that compared with M2F2N, the accuracy of the other two variants decreases. The experimental results fully demonstrated that the two-channel attention branch both contributes to the reconstruction results and can effectively improve the ability of extracting more important discriminative information from the LR input for image SR.

**Table 17.2** Ablation study's result for ×2 SR on five datasets

| Method | Max-pool branch | Average-pool branch | Set5 | Set14 | Manga109 | BSD100 | Urban100 |
|---|---|---|---|---|---|---|---|
| | | | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM |
| M2F2N -1 | ✗ | ✓ | 37.82/ 0.9597 | 33.30/ 0.9147 | 38.03/ 0.9751 | 32.05/ 0.8974 | 31.23/ 0.9194 |
| M2F2N -2 | ✓ | ✗ | 37.83/ 0.9598 | 33.31/ 0.9148 | 38.02/ 0.9750 | 32.05/ 0.8973 | 31.25/ 0.9194 |
| M2F2N | ✓ | ✓ | **37.84/ 0.9599** | **33.33/ 0.9149** | **38.03/ 0.9752** | **32.06/ 0.8976** | **31.26/ 0.9197** |

## 17.5 Conclusion

In this paper, a novel multilevel and multi-scale feature fusion network (M2F2N) is carefully presented to help build an intelligent fault detection system based on UAV patrol, and this paper proposes a novel network. The network is established based on the multi-scale future fusion block (MSFFB), which is capable of extracting multi-scale discriminative feature like high-frequency information. Besides, dual-attention feature extraction block (DAFEB) embedded in MSFFB is developed to assign more weights to the high-frequency information. The extensive experiments suggest that compared with the mainstream algorithm, the performance of M2F2N is better, and the reconstructed image has more details. At present, M2F2N has been applied to the intelligent fault detection system based on UAV patrol, which effectively helps and enhances the working efficiency for our system.

## References

1. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 184–199. IEEE, Piscataway (2014)
2. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 Oct 2016, Proceedings, Part II 14, pp. 391–407. Springer, Cham (2016)
3. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654. IEEE, Piscataway (2016)
4. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1637–1645. IEEE, Piscataway (2016)
5. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3147–3155. IEEE, Piscataway (2017)
6. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144. IEEE, Piscataway (2017)

7. Lai, W., Huang, J., Ahuja, N., Yang, M.: Deep Laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 624–632. IEEE, Piscataway (2017)
8. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883. IEEE, Piscataway (2016)
9. Jin, K., Wei, Z., Yang, A., Guo, S., Gao, M., Zhou, X., Guo, G.: SwiniPASSR: Swin transformer based parallax attention network for stereo image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 920–929. IEEE, Piscataway (2022)
10. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: a persistent memory network for image restoration. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4539–4547 (2017)
11. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4799–4807 (2017)
12. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3867–3876 (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 770–778 (2016)
14. Ahn, N., Kang, B., Sohn, K.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 252–268 (2018)
15. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2024–2032 (2019)
16. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference Computer Vision (ECCV), pp. 286–301. Springer, Berlin (2018)
17. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 126–135. IEEE, Piscataway (2017)
18. Huang, J., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206. IEEE, Piscataway (2015)
19. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. J. Multimed. Tools Appl. **76**(20), 21811–21838 (2017)
20. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision, ICCV 2001, pp. 416–423. IEEE, Piscataway (2001)
21. Bevilacqua, M., Roumy, A., Guillemot, C., Morel, M. .: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: British Machine Vision Conference (BMVC), pp. 1–10. BMVA (2012)
22. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 711–730. Springer, Cham (2012)
23. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)

# Chapter 18
# Surface Defect Detection Using Deep Learning: A Comprehensive Investigation and Emerging Trends


**Fajar Pitarsi Dharma and Moses Laksono Singgih**

**Abstract** Surface defect detection is currently a topic that contributes important things in identifying and assessing defects based on surface appearances, finding widespread applications in diverse manufacturing industries. This approach involves the effective handling and analysis of surface appearances using image processing techniques, coupled with the utilization of deep learning methods for defect detection in several materials such as fabric, steel, aluminum, welding, and others. However, the existing research in this field is confronted with several limitations pertaining to the accuracy, speed, and balance of defect detection outcomes. In response to these challenges, this research paper presents a comprehensive investigation into deep learning techniques for surface defect detection in some applications in industries. With the growing demand for efficient and accurate defect detection in various industries, this study aims to explore the current state of research, identify key research gaps, and shed light on the emerging trends in leveraging deep learning for surface defect detection. Through a meticulous review investigation of relevant literature and an in-depth analysis of existing studies, this research provides valuable insights into the advancements, challenges, and potential future directions in this topic area.

## 18.1 Introduction

The indicator of an advanced civilization is marked by the process of transformation in industries known as industrialization. Industrialization entails a shift in manufacturing processes from human labor to machine power, also referred to as the industrial revolution [1]. The current industrial revolution has shifted toward technological advancements. The development of technology has made manufacturing

F. P. Dharma · M. L. Singgih (✉)
Department of Industrial and System Engineering, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
e-mail: moseslsinggih@ie.its.ac.id

F. P. Dharma
Teknik Pembuatan Benang, AK-Tekstil Solo, Solo 57126, Indonesia



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_18

processes increasingly complex, as technological advancements bring forth demands for process improvement [2].

Since its inception in the eighteenth century, the industrial revolution has undergone remarkable transformations with the objective of enabling companies to maintain their existence and continuously improve in response to evolving market demands and product requirements [3]. The current peak of the industrial revolution is supported by rapidly advancing technologies and their integration into cyber-physical systems [4], as predicted by numerous experts [5] smart technology, artificial intelligence, automation, robotics, and algorithms, collectively referred to as STARAA, encompass the broad categories of technological advancements [5, 6].

The application of technologies such as digital technology has shown positive and significant impacts on economic and environmental performance in manufacturing companies in China [7]. Similarly, in a different region, South Africa, a positive relationship has been found between the adoption of technology, namely knowledge of big data analytics (BDA) and artificial intelligence (AI), and sustainable manufacturing and circular economy capabilities in the automotive component and related product manufacturers [8]. These empirical findings align with the research findings of [9]. Indicating that the economic performance, environmental performance, and operational performance of companies receive positive influence from the implementation of technologies such as Industry 4.0, with the greatest impact observed in the operational performance [10].

Digital disruption and technology, in essence, serve as complements to technological advancement, with the aim of achieving leaner, more flexible, and even more complex production processes [11]. However, the implementation of digital disruption is currently limited to certain industries, and the diffusion of these technologies may not occur in the near future, at least not in smaller industries [12].

Studies and research on the application of innovation technology in the textile industry are predominantly limited to literature review research [13-17], and some studies focus on bibliometric analyses [18, 19]. Furthermore, research applying quality improvement with innovation technology only focuses on defect detection for a single type of defect, such as hairiness detection in fabric [20] or yarn breakage prediction [21, 22]. Moreover, no research has utilized innovation technology for quality improvement involving multiple defect categories, nor has any research integrated it to support decision-making processes related to these defects.

## 18.2 Methodology

This study's methodology employs a comprehensive investigation technique, which includes searching the Scopus database for relevant scholarly papers. The retrieved papers will then be evaluated with VOS Viewer software to gain an overview of the scope of study on the issue and to investigate the chronological distribution of these studies (Fig. 18.1).

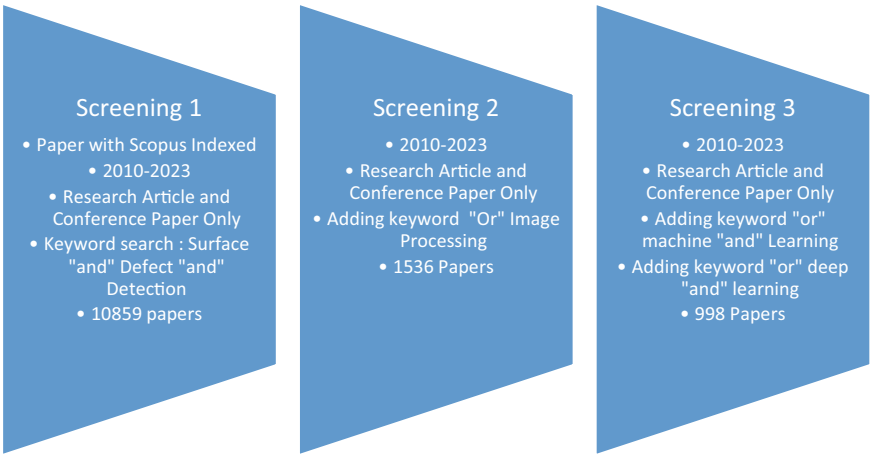



**Fig. 18.1** Screening the literature

The papers were acquired using the Scopus paper database in three steps. The first stage was to limit the filter year to only research articles and conference articles from 2010 to 2023. Following that, conduct a keyword search for surface "and" defect "and" detection, yielding 10,859 papers. Using the same filter as the first, the second stage adds keywords "or" picture "or" processing, yielding 1536 papers. The third stage is the same filter as the first and second steps, with the addition of keyword search "or" machine "or" deep "and" learning, yielding 998 papers.

Following a screening process, 998 papers were determined as being highly related to the research topic. Following that, use VOS Viewer to view the classification and clustering of these publications, as well as to find notable research groups that are constantly increasing. This analysis also sheds light on prospective future research areas that might be pursued within the subject.

## 18.3 Result

The final visualization map is shown below after completing an analysis on 998 papers using VOS Viewer and applying "binary counting" and "occurrence" approaches for ten selected terms. The resulting visualization map is presented in Fig. 18.2.

According to the network visualization, research development is classified into numerous dominant hues. The green color represents the dominance of the topic "surface defect detection," the blue color represents discussions about attention and defect samples, the red color represents discussions about cracks, machine learning, and techniques, and the yellow color represents discussions about crack images and scores (Fig. 18.3).

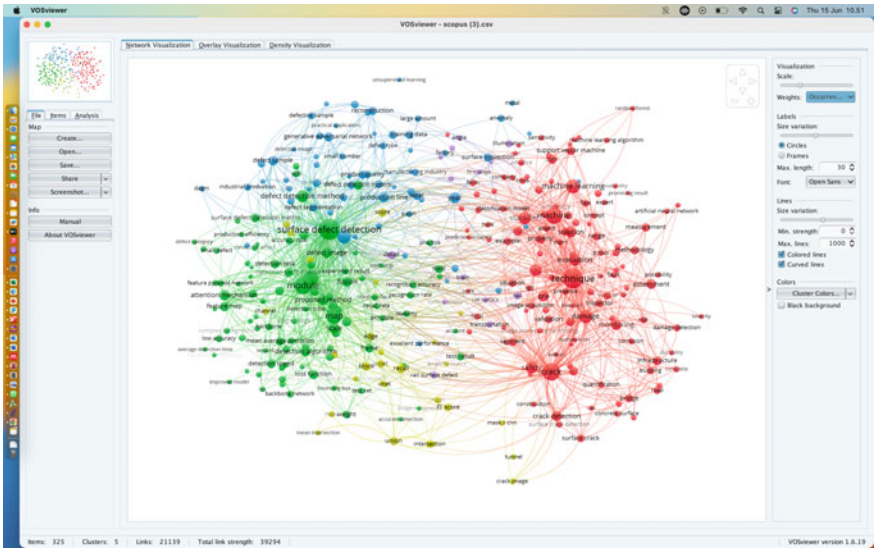



**Fig. 18.2** Network visualization


**Fig. 18.3** Overlay visualization

**Table 18.1** Top ten occurrences based on VOS Viewer

| Rank | Keywords | Number of occurrence | Total link strength |
|---|---|---|---|
| 1 | Deep learning | 414 | 720 |
| 2 | Defect detection | 222 | 394 |
| 3 | Convolutional neural network | 76 | 156 |
| 4 | Machine vision | 73 | 162 |
| 5 | Machine learning | 87 | 133 |
| 6 | Surface defect detection | 68 | 124 |
| 7 | Image processing | 55 | 114 |
| 8 | Computer vision | 52 | 107 |
| 9 | Transfer learning | 48 | 93 |
| 10 | Object detection | 40 | 87 |

Meanwhile, for the overlay visualization, also known as the co-occurrence map, the research is color-coded and organized by year. The darker colors represent the years preceding 2020, while the brighter colors, particularly yellow, depict the more recent years beginning in 2021.

The next step is to look at the top ten most frequently occurring keywords from the 889 papers linked to the previously defined topic.

Deep learning, defect detection, and convolutional neural network (CNN) are the three highest co-occurrences among the top ten co-occurrences in Table 18.1. The next step is to identify the top ten papers that are most relevant to these three co-occurrences by filtering out papers with the highest level of relevance.

## *18.3.1 Deep Learning*

#### 18.3.1.1 A Learning-Based Approach for Surface Defect Detection Using Small Image Datasets (2020)

The problem in this research is to reduce and solve the unbalanced image representation in rare defect detection accuracy, with the aim that it can be applied in the manufacturing industry practically. In addition, the research approach aims to reduce as much as possible the false negative rate (FNR), because FNR can interfere with and reduce the accuracy of surface defect detection [23].

The research methodology of this paper is based on learning based on small-size image datasets and ensures automatic defect detection can work. By using Wasserstein generative adversarial nets (WGANs) which is a transfer learning technique used on feature extraction and a multi-model ensemble framework [23].

#### 18.3.1.2 Automated Visual Inspection of Fabric Image Using Deep Learning Approach for Defect Detection (2021)

The main challenge in this research is to automatically detect fabric damages in complex scenarios, involving the complexities of textile textures, defects, and intraclass differences [24].

This study introduces a dual-phase method that merges innovative and conventional algorithms to improve image and defect detection. The initial phase employs a unique fusion of domain-centered local and global image improvement algorithms, utilizing block-based alpha-rooting. The subsequent phase involves creating a neural network using modern frameworks for the precise identification of fabric defects. This approach permits more precise defect localization compared to conventional machine learning and state-of-the-art deep learning techniques [24].

#### 18.3.1.3 An Automatic Welding Defect Location Algorithm Based on Deep Learning (2021)

The study focuses on assessing defects in welded joints on various items, leveraging deep learning's strong feature representation abilities. An automated technique for locating defects is introduced, utilizing an improved Unet network and digital X-ray images. This approach incorporates data augmentation and strategies for defect localization [25].

For improved localization accuracy, data augmentation is utilized to expand the dataset of defects welds for network training. Using this, an enhanced defect localization approach is suggested, employing a Unet network, to attain automated and highly precise defect detection [25].

### *18.3.2 Defect Detection*

#### 18.3.2.1 A Public Fabric Database for Defect Detection Methods and Results (2019)

The study aims to mitigate errors in textile industry inspections by creating a publicly annotated database containing plain fabrics with and without defects, characterized by uniform fabric textures. This facilitates precise comparisons among existing methods and potential future investigations. Thus, the benefits of each approach can be thoroughly understood through this database [26].

The applied analysis techniques encompass tasks related to texture, such as classification, segmentation, synthesis, shape analysis, and image restoration within the database. The defect detection approach employed for testing images from the presented database in this study involves the utilization of Gabor filters. Gabor filters, which are spectral methods rooted in texture analysis, are widely used for defect

detection. Among non-feature extraction detection methods, Gabor filters are recognized as highly effective for identifying fabric defects. The objective of this research is not to establish the superiority or appropriateness of specific methods, nor is it focused on method comparison. Instead, it aims to illustrate instances using the proposed database in this investigation [26].

#### 18.3.2.2 EDDs: A Series of Efficient Defect Detectors for Fabric Quality Inspection (2021)

The research is addressing fabric defect detection through the utilization of a streamlined deep convolutional neural network (DCNN) architecture, an evolved iteration of the convolutional neural network (CNN) [27].

The strategy adopted in this study to enhance fabric defect detection efficiency is known as efficient defect detection (EDD). In detail, the approach consists of these key elements: opting for a lightweight backbone from Efficient-Nets, incorporating L-FPN for effective multi-scale feature integration, and employing a structure reminiscent of Retina-Net for tasks involving classification and bounding box regression. This section can be adjusted using the recommended R compound scaling approach to create a variety of detectors suitable for various resource limitations [27].

#### 18.3.2.3 Unified Detection Method of Aluminum Profile Surface Defects: Common and Rare Defect Categories (2020)

Automating the visual identification of defects on aluminum profile surfaces (APSD) is a complex task owing to the varied classes, irregular forms, haphazard arrangement, and skewed sample distribution. By harnessing attention mechanisms, a comprehensive approach for defect detection is put forth, aimed at overcoming these difficulties for both prevalent and rare defects [28].

In this study, the approach is structured as a derivation of multiple learning algorithms, designed to identify both prevalent and uncommon defect classes. Initially, a category representation network is utilized to extract common category maps (CCMs). Following this, a subject module is introduced to create proposal maps (PMs) for individually infrequent classes. Lastly, the transformation of rare category maps (RCMs) from CCMs is guided by the information present in PMs [28].

#### 18.3.2.4 Multistage GAN for Fabric Defect Detection (2020)

Fabric (textile product) defect detection presents an intriguing and demanding subject. Numerous approaches have been suggested to address this issue, but they remain less than ideal due to the intricate variety of fabric textures and defects [29].

This study introduces a framework for fabric defect detection based on generative adversarial networks (GANs). Addressing real-world complexities, the proposed

system learns from available fabric defect examples and flexibly adjusts to diverse fabric textures across distinct application scenarios. The core of this approach involves the adaptation of a deep semantic segmentation network, enabling the identification of diverse defect types. Furthermore, our efforts include training a hierarchical GAN to artificially create convincing defects within new defect-free samples [29].

### *18.3.3 Convolutional Neural Network (CNN)*

#### 18.3.3.1 Mobile-Unet: An Efficient Convolutional Neural Network for Fabric Defect Detection (2020)

The existing fabric manufacturing conditions demand methods with enhanced real-time capabilities. Furthermore, fabric defects, when used as samples, are significantly less common compared to normal samples, leading to imbalanced data. This, in turn, poses a challenge for training deep learning-based models [30].

To accomplish end-to-end defect segmentation, a notably well-organized convolutional neural network named Mobile-Unet is put forth. The irregular distribution of defect instances is leveraged to tackle the concern of defect sample representation. Moreover, Mobile-Unet integrates depth-wise independent convolutions, significantly diminishing computational complication and network size. The architecture comprises two segments: an encoder and a decoder. The MobileNetV2 has a feature extractor functions as the encoder, followed by the inclusion of five deconvolutional layers to serve as the decoder [30].

#### 18.3.3.2 Fabric Defect Detection System Using Stacked Convolutional Denoising Auto-encoders Trained with Synthetic Defect Data (2020)

With the growing diversity and advancement in machine vision-based defect detection, the utilization of deep learning methods is becoming more prevalent. Lately, various investigations have been conducted regarding defect identification and categorization through image segmentation, detection, and classification. These techniques yield positive results; nevertheless, they necessitate a substantial volume of authentic defect data. Yet, procuring an ample amount of genuine defect data within industrial environments presents a considerable challenge [31].

The study introduces an approach for identifying defects through the application of stacked convolutional autoencoders. The devised autoencoder is trained solely on defect-free data and artificially generated flawed data, which is created based on expert knowledge-driven defect attributes [31].

#### 18.3.3.3 Detecting Textile Micro-defects: A Novel and Efficient Method Based on the Visual Gain Mechanism (2020)

Considered a crucial technique in machine learning, the faster region-based convolutional neural network (Faster RCNN) has surfaced as a hopeful framework, exhibiting commendable efficiency in object detection. Nonetheless, the detection of diminutive entities like micro-defects within textiles continues to be a complex endeavor for the Faster RCNN [32].

Creating an innovative detection model to enhance the capacity for detecting small-sized entities. Initially, through an examination of the interplay between reading and visual mechanisms enhancement process, it's discerned that mechanisms connected to attention-driven visual enhancement can modify response amplitudes while retaining selectivity, consequently ameliorating visual perception acumen. Subsequently, these pertinent mechanisms are integrated into the Faster RCNN framework, culminating in the formation of a novel model termed Faster VG-RCNN. To assess the suggested class of detection, a distinctive micro-textile defect database is established as a reference point for micro-defect detection. Additionally, spacious experimental validation is carried out, encompassing diverse design alternatives [32].

Furthermore, Table 18.2 illustrates the positions and comparisons of the ten selected papers, highlighting their relative strengths and weaknesses. The table provides insights into the existing gaps in the research and identifies potential areas for future research and development. This analysis aids in understanding the current landscape of the field and guides further investigations to advance the research in this area.

Table 18.3 illustrates that within the context of defect detection, the topic of automatic defect detection remains a substantial and ongoing concern. Scholars continue to advance their understanding in this domain by delving into relevant literature through the examination of citations. This underscores the evolving nature of research endeavors in this particular area.

## 18.4 Discussion

This paper presents a comprehensive review of research studies indexed in Scopus that are closely related to surface defect detection. The reviewed papers focus on various aspects of surface defect detection analysis, including subjects, methods, and reference outcomes. Some studies aim to enhance defect quality and refine detection methods, while others explore cost–benefit analyses to uncover differences in defect reading approaches. The findings from these papers have significantly influenced the topic area of surface defect detection, particularly in terms of tested samples.

However, despite the progress made, there remain several aspects that require further improvement and development to achieve an optimal defect detection model. These aspects include the selection of derivative methods for artificial intelligence, types of defect readings, accuracy, error reduction, and addressing the issue of false

**Table 18.2** Paper method comparison and future research

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | √ | √ | √ | √ | | | | | | | |
| 2 | √ | | | | √ | √ | | | | | |
| 3 | √ | | | √ | | | √ | | | | |
| 4 | √ | | | | √ | | | √ | | | |
| 5 | √ | | | | | | | | √ | √ | |
| 6 | √ | | | √ | | | | | √ | | |
| 7 | √ | | | √ | √ | | | | | | |
| 8 | √ | | | √ | √ | √ | | | | | |
| 9 | √ | | | √ | √ | √ | | | | | √ |
| 10 | √ | | | √ | √ | | | √ | √ | | |
| 11 | √ | | | √ | √ | √ | | √ | | | |
| 12 | √ | | | √ | √ | | √ | | √ | √ | |
| 13 | √ | | | | | | | | | | |
| 14 | √ | | √ | √ | √ | √ | √ | √ | √ | | √ |

Table description:
A = Surface defect detection
B = Wasserstein generative adversarial nets (WGANs)
C = Multi-model ensemble framework
D = Fabric/textile industry
E = Convolutional neural network (CNN)
F = Computer processing/machine learning
G = Image processing/computer vision
H = Automatic defect recognition
I = Deep learning
J = UNet framework efficient
K = Defect detectors (EDDs)
1 = Le et al. (2020)
2 = Fu et al. [33]
3 = Silvestre-Blanes et al. [26]
4 = Gao et al. (2020)
5 = Yang et al. [25]
6 = Jin and Niu (2021)
7 = Li et al. (2019)
8 = Wei et al. [32]
9 = Zhou et al. [27]
10 = Han and Yu [31]
11 = Liu et al. [35]
12 = Jing et al. [30]
13 = Zhang et al. [28]
14 = Future research

**Table 18.3** Paper method comparison based on citation

| Article | Google Scholar citation | Scopus citation |
|---|---|---|
| Le et al. (2020) | 57 | 47 |
| Fu et al. [33] | 4502 | 3100 |
| Silvestre-Blanes et al. [26] | 74 | 48 |
| Gao et al. [34] | 26 | 23 |
| Yang et al. [25] | 57 | 41 |
| Jin and Niu (2021) | 126 | 58 |
| Wei et al. [32] | 26 | 23 |
| Zhou et al. [27] | 13 | 11 |
| Han and Yu [31] | 18 | 13 |
| Liu et al. [35] | 1653 | 4173 |
| Jing et al. [30] | 71 | 123 |
| Zhang et al. [28] | 29 | 26 |

negative ratio. Collectively, these factors highlight the need for continued attention and research in the field of surface defect detection.

To address these research gaps, this study aims to elaborate on the existing literature and introduce innovative technology in surface defect detection within the manufacturing industry. Specifically, the use of convolutional neural networks (CNNs) for fabric defect detection will be explored, aiming to refine and reduce errors encountered in previous CNN research [36].

Processing data grids, images, and videos is the main function and special design of CNN which is a neural network architecture. It effectively recognizes patterns and features present in spatial data. Training CNN involves the use of deep learning techniques, specifically backpropagation, which iteratively adjusts network weights and parameters to minimize prediction errors. This process allows CNN [36] to automatically learn relevant structures from input data deprived of the need for manual feature engineering.

CNN has emerged as a highly successful architecture in image processing and computer vision, surpassing traditional methods in several tasks including image classification, object detection, segmentation, and face recognition. Moreover, CNN [36] has found applications in other domains such as natural language processing, speech recognition, and bioinformatics.

Given these reasons, the field of artificial intelligence focused on image detection, particularly CNN, is highly relevant for future research and aligns with the identified research gap. Additionally, CNN is particularly well-suited for texture-based readings, making it an optimal choice for surface defect detection in materials like fabric.

## 18.5 Conclusion

This paper discusses potential directions for further research in the field of defect detection analysis, focusing on the incorporation of multiple methods to enhance the detail and accuracy of defect identification. Furthermore, the introduction of AI-based decision-making techniques is proposed to support stakeholders in making informed decisions concerning defective products. By integrating these aspects, this research aims to contribute to the existing body of knowledge, which can be further developed from various perspectives, utilizing different tools and samples. Additionally, the measurement and comparison of research outcomes are emphasized to showcase how emerging technologies can effectively assist industries in improving product quality.

Defect detection, particularly in the domain of fabric or textile product defect detection, has generated a considerable body of related literature. These connections can be categorized into various sub-connections, including object relevance, employed methods, obtained results, branches of science employed, and functional relevance. To gain a comprehensive understanding of these relationships, it is necessary to explore specific research gap topics that identify areas requiring further investigation to ensure continued relevance and advancement in the field.

**Acknowledgements** The authors wish to acknowledge the support of Lembaga Pengelola Dana Pendidikan (LPDP) in facilitating and supporting the publication of this article.

## References

1. Chehri, A., Zimmermann, A., Schmidt, R., Masuda, Y.: Theory and practice of implementing a successful enterprise IoT strategy in the industry 4.0 era. Procedia Comput. Sci. 4609–4618 (2021). https://doi.org/10.1016/j.procs.2021.09.239
2. Reiman, A., Kaivo-oja, J., Parviainen, E., Takala, E.P., Lauraeus, T.: Human factors and ergonomics in manufacturing in the industry 4.0 context—a scoping review. Technol. Soc. **65** (2021). https://doi.org/10.1016/j.techsoc.2021.101572
3. Badri, A., Boudreau-Trudel, B., Souissi, A.S.: Occupational health and safety in the industry 4.0 era: a cause for major concern? Saf. Sci. **109**, 403–411 (2018). https://doi.org/10.1016/j.ssci.2018.06.012
4. Schwab, K.: The Fourth Industrial Revolution (2016). [Online]. Available: www.weforum.org
5. Brougham, D., Haar, J.: Employee assessment of their technological redundancy. Labour Ind. J. Soc. Econ. Relat. Work **27**(3), 213–231 (2017). https://doi.org/10.1080/10301763.2017.1369718
6. Brougham, D., Haar, J.: Smart technology, artificial intelligence, robotics, and algorithms (STARA): employees' perceptions of our future workplace. J. Manag. Organ. **24**(2), 239–257 (2018). https://doi.org/10.1017/jmo.2016.55
7. Li, M., et al.: A decision support system using hybrid AI based on multi-image quality model and its application in color design. Futur. Gener. Comput. Syst. **113**, 70–77 (2020). https://doi.org/10.1016/j.future.2020.06.034
8. Bag, S., Pretorius, J.H.C., Gupta, S., Dwivedi, Y.K.: Role of institutional pressures and resources in the adoption of big data analytics powered artificial intelligence, sustainable manufacturing

practices and circular economy capabilities. Technol. Forecast. Soc. Chang. **163** (2021). https://doi.org/10.1016/j.techfore.2020.120420.
9. Lopes de Sousa Jabbour, A.B., Chiappetta Jabbour, C.J., Choi, T.M., Latan, H.: 'Better together': evidence on the joint adoption of circular economy and industry 4.0 technologies. Int. J. Prod. Econ. **252** (2022). https://doi.org/10.1016/j.ijpe.2022.108581
10. Singh, L.B., Srivastava, S.: Linking workplace ostracism to turnover intention: a moderated mediation approach. J. Hosp. Tour. Manag. **46**, 244–256 (2021). https://doi.org/10.1016/j.jhtm.2020.12.012
11. Seçkin, M., Seçkin, A.Ç., Coşkun, A.: Production fault simulation and forecasting from time series data with machine learning in glove textile industry. J. Eng. Fiber Fabr. **14** (2019). https://doi.org/10.1177/1558925019883462
12. Pardi, T.: Fourth industrial revolution concepts in the automotive sector: performativity, work and employment. J. Ind. Bus. Econ. **46**(3), 379–389 (2019). https://doi.org/10.1007/s40812-019-00119-9
13. de Oliveira, C.R.S., da Silva Júnior, A.H., Mulinari, J., Immich, A.P.S.: Textile re-engineering: eco-responsible solutions for a more sustainable industry. Sustain. Prod. Consumption **28**, 1232–1248 (2021). https://doi.org/10.1016/j.spc.2021.08.001
14. Liu, Y., Bao, R., Tao, J., Li, J., Dong, M., Pan, C.: Recent progress in tactile sensors and their applications in intelligent systems. Sci. Bull. **65**(1), 70–88 (2020). https://doi.org/10.1016/j.scib.2019.10.021
15. Manavalan, R.: Towards an intelligent approaches for cotton diseases detection: a review. Comput. Electron. Agric. **200** (2022). https://doi.org/10.1016/j.compag.2022.107255
16. Wang, J., Xu, C., Zhang, J., Zhong, R.: Big data analytics for intelligent manufacturing systems: a review. J. Manuf. Syst. **62**, 738–752 (2022). https://doi.org/10.1016/j.jmsy.2021.03.005
17. Xia, Z., Xu, W.: A review of ring staple yarn spinning method development and its trend prediction. J. Nat. Fibers **10**(1), 62–81 (2013). https://doi.org/10.1080/15440478.2012.763218
18. Arora, S., Majumdar, A.: Machine learning and soft computing applications in textile and clothing supply chain: bibliometric and network analyses to delineate future research agenda. Expert Syst. Appl. **200** (2022). https://doi.org/10.1016/j.eswa.2022.117000
19. Pitt, C., Park, A., McCarthy, I.P.: A bibliographic analysis of 20 years of research on innovation and new product development in technology and innovation management (TIM) journals. J. Eng. Technol. Manag. JET-M **61** (2021). https://doi.org/10.1016/j.jengtecman.2021.101632
20. Khanal, S.R., Silva, J., Gonzalez, D.G., Castella, C., Perez, J.R.P., Ferreira, M.J.: Fabric hairiness analysis for quality inspection of pile fabric products using computer vision technology. Procedia Comput. Sci. **204**, 591–598 (2022). https://doi.org/10.1016/j.procs.2022.08.072
21. Azevedo, J., et al.: Predicting yarn breaks in textile fabrics: a machine learning approach. Procedia Comput. Sci. **207**, 2301–2310 (2022). https://doi.org/10.1016/j.procs.2022.09.289
22. Cioară, I., Cioară, L., Onofrei, E.: Forecast of yarn breakages during the weaving process. Res. J. Text. Appar. **8**(1), 20–24 (2004). https://doi.org/10.1108/RJTA-08-01-2004-B003
23. Le, X., Mei, J., Zhang, H., Zhou, B., Xi, J.: A learning-based approach for surface defect detection using small image datasets. Neurocomputing **408**, 112–120 (2020). https://doi.org/10.1016/j.neucom.2019.09.107
24. Voronin, V.V., Sizyakin, R., Zhdanova, M., Semenishchev, E.A., Bezuglov, D., Zelemskii, A.A.: Automated visual inspection of fabric image using deep learning approach for defect detection. SPIE Int. Soc. Opt. Eng. 23 (2021). https://doi.org/10.1117/12.2592872
25. Yang, L., Wang, H., Huo, B., Li, F., Liu, Y.: An automatic welding defect location algorithm based on deep learning. NDT E Int. 120 (2021). https://doi.org/10.1016/j.ndteint.2021.102435
26. Silvestre-Blanes, J., Albero-Albero, T., Miralles, I., Pérez-Llorens, R., Moreno, J.: A public fabric database for defect detection methods and results. Autex Res. J. (2019). https://doi.org/10.2478/aut-2019-0035
27. Zhou, T., Zhang, J., Su, H., Zou, W., Zhang, B.: EDDs: a series of Efficient Defect Detectors for fabric quality inspection. Measurement (Lond.) **172** (2021). https://doi.org/10.1016/j.measurement.2020.108885

28. Zhang, D., Song, K., Xu, J., He, Y., Yan, Y.: Unified detection method of aluminium profile surface defects: common and rare defect categories. Opt. Lasers Eng. **126** (2020). https://doi.org/10.1016/j.optlaseng.2019.105936
29. Liu, J., et al.: Multistage GAN for fabric defect detection. IEEE Trans. Image Process. **29**(202), 3388–3400 (2020)
30. Jing, J., Wang, Z., Rätsch, M., Zhang, H.: Mobile-Unet: an efficient convolutional neural network for fabric defect detection. Text. Res. J. **92**(1–2), 30–42 (2022). https://doi.org/10.1177/0040517520928604
31. Han, Y.J., Yu, H.J.: Fabric defect detection system using stacked convolutional denoising auto-encoders trained with synthetic defect data. Appl. Sci. (Switz.) **10**(7) (2020). https://doi.org/10.3390/app10072511
32. Wei, B., Hao, K., Gao, L., Song Tang, X.: Detecting textile micro-defects: a novel and efficient method based on visual gain mechanism. Inf. Sci. (N.Y.) **541**, 60–74 (2020). https://doi.org/10.1016/j.ins.2020.06.035
33. Fu, J., et al.: Dual attention network for scene segmentation. In: CVPR (2019) [Online]. Available: https://github.com/junfu1115/DANet/
34. Chang, J., et al.: Sequential Recommendation with graph neural networks. In: SIGIR 2021—Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 378–387. Association for Computing Machinery, Inc., July 2021. https://doi.org/10.1145/3404835.3462968
35. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph CNN for learning on point clouds. ACM Trans. Graph **38**(5) (2019). https://doi.org/10.1145/3326362
36. Li, F., Li, F.: Bag of tricks for fabric defect detection based on Cascade R-CNN. Text. Res. J. **91**(5–6), 599–612 (2021). https://doi.org/10.1177/0040517520955229

# Chapter 19
# Lightweight Real-Time Intelligent Inspection System for Digital Transmission Security


**Feng Weixi, Huang Ping, and Yan Mengqiu**

**Abstract** The whole process automation of patrol inspection can effectively improve the management and production efficiency of current digital transmission business. The rapidly developed image super-resolution (SR) technology is helpful to achieve this goal. Therefore, we specially design a lightweight effective multi-level dual residual attention network (MDRAN) for single image super-resolution network (SISR). Firstly, to fully extract and utilize the dependency information between different channels, a dual residual attention block (DRAB) is proposed. This module can adaptively adjust the weight proportion between features across various channels, so as to accurately extract high-frequency information containing rich details and texture information. Meanwhile, to maximize the usage of the characteristics from various levels, the DRAB is cascaded combining with residual skip connection, which can not only reduce the loss of information as the network depth increases, but also ease the difficulty of network training. The experimental results on multiple benchmark datasets show that the PSNR and SSIM indicators of MDRAN are the highest among the comparison models so as to the greatly reduced amount of model parameters. Furthermore, the visual effect of MDRAN is richer in texture and border.

## 19.1 Introduction

In the current digital transmission business, the whole process automation of patrol inspection is able to effectively improve the management and production efficiency for current power companies. The construction of dynamic adjustment algorithm model of operation and maintenance strategy and the application of video terminals are conducive to the automatic generation of plans. In addition, the implementation of node traffic light control, automatic defect identification, and automatic generation of patrol inspection reports is conducive to the automation of the whole process from planned production to closed-loop control. At the same time, the patrol plan

F. Weixi (✉) · H. Ping · Y. Mengqiu
Shenzhen Power Supply Bureau Co., Ltd., Shenzhen 518000, Guangdong, China
e-mail: duanzhiwei69240130@163.com



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_19

is dynamically adjusted according to the equipment risks, status, and changes in special sections, greatly reducing the repetitive work of the team in preparing the plan, and effectively controlling the scientific implementation of the plan. At the same time, based on the artificial intelligence platform of China Southern Power Grid, the company's self-developed green film, pollution flashover, mountain fire, safety supervision and other algorithms are integrated to achieve image intelligent identification and early warning. In combination with the special work requirements of special patrol and special maintenance, a chart showing alarms of different algorithm types is formed to assist the team in formulating operation and maintenance work and support the continuous optimization of the algorithm. Recently, image technologies based on deep learning (DL) have made incredible achievements. Among them, the image super-resolution (SR) algorithm on the basis of convolutional network can provide a better preliminary assistance for the intelligent patrol system.

Single image super-resolution (SISR) is capable of reconstructing a high-resolution (HR) image from a given low-resolution (LR) one, which is classified as a type of ill-posed process in that one single LR input may correspond to multiple HR results. In recent years, the performance of SISR based on convolutional neural network (CNN) is extremely superior over that of the traditional methods, so it has been developed rapidly. In 2016, Dong et al. proposed the Super-Resolution Using Convolutional Neural Networks (SRCNN) [1] based on convolutional neural networks, which includes three steps: feature extraction, nonlinear mapping, and reconstruction. Subsequently, Dong et al. [2] proposed the accelerated FSRCNN, using a single upsampling method to improve image resolution and obtain HR image with higher accuracy. Due to the limited network capacity, it is impossible to learn complex mapping. Literature [3] constructed a type of very deep convolutional super-resolution networks (VDSR) utilizing 20 layers of convolution and gained the significantly improved results. Lately, many other algorithms are devoted to elevate the accuracy for SISR, such as DRCN [4], DRRN [5], IDN [6], and so on [7, 8].

Though these models gradually lift the performance of super-resolution reconstruction, the architecture of network becomes more complex and fails to make a good trade-off between computational cost and network parameters, which hinders the actual application of image SR models. Moreover, they ignore to utilize the attention mechanism to extract high-frequency information.

To address these problems, a lightweight effective multi-level dual residual attention network (MDRAN) for single image super-resolution network (SISR). The building block is the dual residual attention block (DRAB), which can adaptively adjust the weight proportion of features across various channels.

## 19.2 Related Work

### *19.2.1 DL-Based SR*

With the remarkable development of deep learning technology, convolutional neural network (CNN) has been widely applied in the domain of super-resolution. The earliest model SRCNN designed by Dong et al. [1] is a three-layered CNN model. However, due to the shallow layer network of SRCNN, many researchers began to exploit more CNN layers to build more complex SR networks to improve the performance of super-resolution. Motivated by the design principles of VGG-net [9, 10] built a type of 20-layered deep CNN-based model named as VDSR, which straightly stacks 20 convolutional layers. Later, Hui et al. [6] proposed EDSR, namely a wider neural network and a very deeper network MDSR, which greatly increased the SR accuracy. Furthermore, Liang et al. [11] applied the transformer architecture to the domain of image restoration on the basis of Swin Transformer [12], which yields an obvious improvement and elevates the accuracy for SR. However, most of the aforementioned methods gain a lot in accuracy for SR in the cost of consuming a huge computational burden, prompting SR community to design more effective and lightweight models for SR. In terms of lightweight networks, Ahn et al. [13] proposed a cascading network called CARN which adopted both the recursive mechanism and residual connections. Moreover, Hui et al. [14] improved their previously designed IDB to a lightweight one, i.e., IMDB, for image SR.

## 19.3 Proposed Method

In this section, we will detailly discuss the network framework of the proposed method. For convenience, the proposed network is named as MDRAN—multi-level dual residual attention network.

### *19.3.1 Network Architecture*

The detailed architecture of MDRAN is depicted in Fig. 19.1. MDRAN reconstruction algorithm uses stacked residual network structure, including shallow information extracting layer (SIE), deeper information extracting (DIE) part, as well as image reconstruction module. Assumed that the LR input image is $I_{\mathrm{LR}}$ and the output result is super-resolved HR image $I_{\mathrm{SR}}$, the overall expression is as follows:

$$I_{\mathrm{SR}} = H_{\mathrm{MDRAN}}(I_{\mathrm{LR}}). \tag{19.1}$$

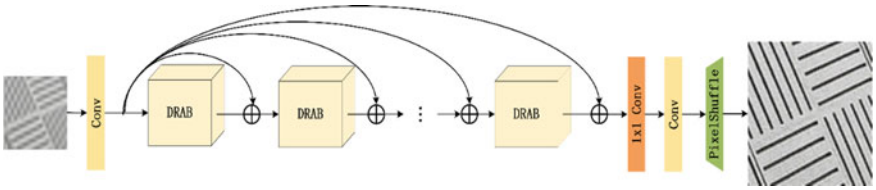



**Fig. 19.1** Overall architecture of the MDRAN

Firstly, we use one $3 \times 3$ convolution to extract the shallow feature, i.e., low-frequency information, which is motivated by [6]. The process of shallow feature extraction can be expressed as:

$$F_0 = H_{\mathrm{SIE}}(I_{\mathrm{LR}}), \tag{19.2}$$

where $H_{\mathrm{SIE}}(\cdot)$ is the operation of $3 \times 3$ convolution, and $F_0$ is the shallow feature.

The DIE is consisted of four cascaded dual residual attention blocks (DRABs). To alleviate the loss of low-frequency information, the global residual skip connection is combined into DIE. The high-frequency feature extraction process of $k$-th DRAB can be formulated as:

$$F_k = H_{\mathrm{DRAB}}(F_{k-1} + F_0), \tag{19.3}$$

where $H_{\mathrm{DRAB}}(\cdot)$ indicates the implicit function of the proposed DRAB, and $k = 2, 3, 4$. When $k$ equals to 1, $F_1 = H_{\mathrm{DRAB}}(F_0)$.

Finally, the image reconstruction part uses one $1 \times 1$ convolution, one $3 \times 3$ convolution, and sub-pixel convolution layer [15] to reconstruct the LR input. The rebuilt HR output can be expressed as the below equation:

$$I_{\mathrm{SR}} = H_{\mathrm{recon}}(H_{3\times3\ \mathrm{conv}}(H_{1\times1\ \mathrm{conv}}(F_0 + F_4))), \tag{19.4}$$

where $H_{\mathrm{recon}}(\cdot)$, $H_{3\times3\ \mathrm{conv}}(\cdot)$, and $H_{1\times1\ \mathrm{conv}}(\cdot)$ represents the sub-pixel convolution, $3 \times 3$ convolution, and $1 \times 1$ convolutional layer, respectively.

### *19.3.2 Dual Residual Attention Block (DRAB)*

In this part, more details of dual residual attention block (DRAB) will be introduced. As shown in Fig. 19.2, the DRAB is composed of dual way of channel attention branches combining with global residual skip connection. It has been proven that the attention mechanism has a good effect on capturing high-frequency information such as texture and border details [16]. Channel attention mechanism means that the network adaptively acquires the significance of each channel and assigns a weight

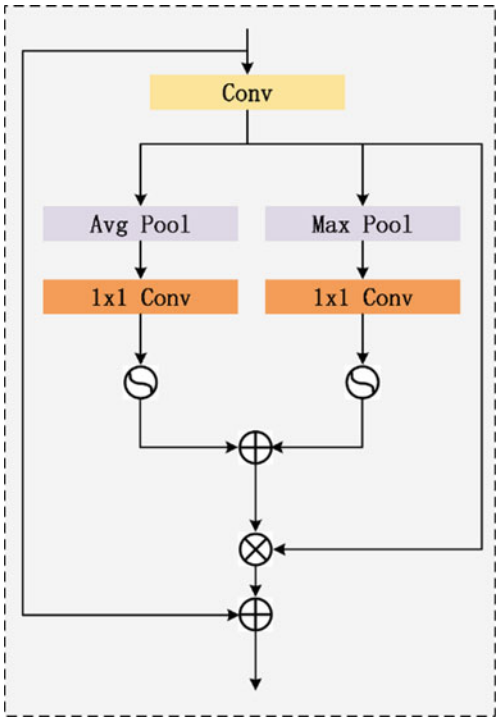



**Fig. 19.2** Dual residual attention block (DRAB)

value to each channel accordingly, so that the network will focus on the features with more high-frequency information and recalibrate the previously obtained features.

Specifically, we utilize two different pooling (average pooling and max pooling) to build DRAB. The equation of two different branches is:

$$F_{\text{sig1}} = \sigma(H_{1\times1\ \text{conv}}(H_{\text{Avg}}(H_{3\times3\ \text{conv}}(f_{\text{in}})))), \tag{19.5}$$

$$F_{\text{sig2}} = \sigma(H_{1\times1\ \text{conv}}(H_{\text{max}}(H_{3\times3\ \text{conv}}(f_{\text{in}})))), \tag{19.6}$$

where $H_{\text{max}}$ and $H_{\text{Avg}}$ are the implicit function of max pooling and average pooling, respectively. $\sigma(\cdot)$ represents the sigmoid function. Eventually, the output feature of $k$-th ($k = 1, 2, 3, 4$) DRAB can be represented as follows:

$$F_k = H_{1\times1\ \text{conv}} \cdot ((F_{\text{sig1}} + F_{\text{sig2}})) + f_{\text{in}}. \tag{19.7}$$

### 19.3.3 Loss Function

To achieve better reconstruction effect, the experiment uses L1 function as the loss function, which is basically adopted to compute the average absolute value of the difference between every single pixel of the input image as well as that of ground-truth image. The L1 loss function is popularly utilized in the domain of SISR, and its equation is:

$$L(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \left\| H_{\mathrm{DRAN}}(I_{\mathrm{LR}}^{i}) - I_{\mathrm{HR}}^{i} \right\|_{1}. \tag{19.8}$$

## 19.4 Experimental Results

### 19.4.1 Metrics and Datasets

To obtain higher-quality reconstructed images, the DIV2K [17] dataset is used to train the network model. The dataset contains 1000 RGB three channel images with a resolution of about 2000 pixels in total, and 800 of them are selected for training. At the same time, in order to compare with the classical network model with respect to peak signal-to-noise ratio (PSNR) and structure similarity (SSIM) evaluation indicators, experiment mainly uses several commonly used open datasets for testing, namely: Set5 [18], Manga109 [19], BSD100 [20], Urban100 [10] to Set14 [21]. These test datasets mainly include images of structures, people, animals, and natural landscapes in different scenes. For example, Urban100 test set mainly focuses on buildings, including image datasets of building facilities with similar structures. It is noted that we only calculate PSNR and SSIM in terms of the luminance (Y) channel following the practice of previous works [14, 22–24].

### 19.4.2 Implementation Details

LR images are achieved through downsampling the HR images using Bicubic algorithm for different scale factors training. In order to prevent the occurrence of overfitting to a certain extent, 800 training images are augmented during image preprocessing. Data augmentation can extract more information from the original dataset, thereby narrowing the gap between the trainset and the verification one. The specific operation is to rotate these images randomly by 90°, 180°, and 270°, and then turn them horizontally to obtain enhanced datasets. Then the image patch with the size of $20 \times 20$ is randomly cut from the LR image, and then it is input into the network

for training. The batch size during training is set to 32. Besides, L1 loss function and Adam optimizer are incorporated for training, where $\beta_1 = 0.9$, $\beta_2 = 0.9999$ as well as $\varepsilon = 10^{-8}$. Notably, the training rate is adjusted by cosine annealing strategy. We set the learning rate [25] as 0.0004 halved after every $2 \times 10^5$ iterations. MDRAN is implemented with PyTorch and trained on GeForce 2080ti GPU.

### *19.4.3 Comparison with State of the Arts*

To better validate the effectiveness of MDRAN, MDRAN is utilized to compare with the following methods: Bicubic interpolation, SRCNN [1], FSRCNN [2], VDSR [3], DRRN [5], and IDN [6] Besides, we list the quantitative combined with quantitative comparison into Table 19.1 and Fig. 19.3 from $\times$ 2, $\times$ 3, to $\times$ 4 scale factors.

It can be seen from Table 19.1 that the proposed MDRAN achieves the best performance in all three scale factors (i.e., $\times$ 2, $\times$ 3, and $\times$ 4) of the five benchmark datasets compared with all previous advanced methods with respect to quantitative evaluation indicators. The quantitative results show that the theory of pioneering SRCNN algorithm is groundbreaking, but the reconstruction effect is not ideal. The subsequent deep learning SR reconstruction has improved the depth for network to a very level. Though the amount of network parameters and inference time of our proposed MDRAN is more than SRCNN, FSRCNN, and DRRN, the values of PSNR/SSIM are far exceed those of the aforementioned models. As for those networks with fewer parameters than that of MDRAN, such as DRRN and FSRCNN, their PSNR/SSIM results are lower than those of MDRAN. For example, VDSR and IDN have improved their evaluation indicators by optimizing the network structure and deepening the number of network layers. In contrast, MDRAN makes reasonable use of different levels of features and attention mechanisms and effectively improves the performance of the algorithm, attaining a great trade-off across total amount of parameters as well as performance.

The proposed MDRAN in this paper is compared with the selected Bicubic, FSRCNN, VDSR, and IDN algorithms in terms of the visual results of the reconstructed image. The randomly selected images from the reference dataset are input into each network model. As depicted in Fig. 19.3, compared with other state-of-the-art methods, the HR outputs rebuilt by MDRAN are clearer and sharper with respect to edge and texture details, and much more approximated to the visual effect of the HR ground-truth image. It demonstrates that the proposed MDRAN can greatly reduce the computational burden and cost while yielding satisfactory image reconstruction performance.

**Table 19.1** PSNR and SSIM test results of different methods for scale factor from ×2, ×3, to ×4

| Scale factor | Methods | Params(K) | Set5 | Set14 | Manga109 | BSD100 | Urban100 |
|---|---|---|---|---|---|---|---|
| | | | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM |
| × 2 | Bicubic | – | 33.66/ 0.9299 | 30.24/ 0.8688 | 30.80/ 0.9339 | 29.56/ 0.8431 | 26.88/ 0.8403 |
| | SRCNN | 8 | 36.66/ 0.9542 | 32.45/ 0.9067 | 35.60/ 0.9663 | 31.36/ 0.8879 | 29.50/ 0.8946 |
| | FSRCNN | 13 | 37.00/ 0.9558 | 32.63/ 0.9088 | 36.67/ 0.9710 | 31.53/ 0.8920 | 29.88/ 0.9020 |
| | DRRN | 298 | 37.74/ 0.9591 | 33.23/ 0.9136 | 37.88/ 0.9749 | 32.05/ 0.8973 | 31.23/ 0.9188 |
| | VDSR | 666 | 37.53/ 0.9587 | 33.03/ 0.9124 | 36.67/ 0.9710 | 31.90/ 0.8960 | 30.76/ 0.9140 |
| | IDN | 553 | 37.83/ 0.9600 | 33.30/ 0.9148 | 38.01/ 0.9749 | 32.08/ 0.8985 | 31.27/ 0.9196 |
| | MDRAN | 450 | **37.86/ 0.9604** | **33.33/ 0.9151** | **38.03/ 0.9752** | **32.10/ 0.8985** | **31.29/ 0.9199** |
| × 3 | Bicubic | – | 30.39/ 0.8682 | 27.55/ 0.7742 | 26.95/ 0.8556 | 27.21/ 0.7385 | 24.46/ 0.7349 |
| | SRCNN | 8 | 32.75/ 0.9090 | 29.30/ 0.8215 | 30.48/ 0.9117 | 28.41/ 0.7863 | 26.24/ 0.7989 |
| | FSRCNN | 13 | 33.18/ 0.9140 | 29.37/ 0.8240 | 31.10/ 0.9210 | 28.53/ 0.7910 | 26.43/ 0.8080 |
| | DRRN | 298 | 34.03/ 0.9244 | 29.96/ 0.8349 | 32.71/ 0.9379 | 28.95/ 0.8004 | 27.53/ 0.8378 |
| | VDSR | 666 | 33.66/ 0.9213 | 29.77/ 0.8314 | 32.01/ 0.9340 | 28.82/ 0.7976 | 27.14/ 0.8279 |
| | IDN | 553 | 34.11/ 0.9253 | 29.99/ 0.8354 | 32.71/ 0.9381 | 28.95/ 0.8013 | 27.42/ 0.8359 |
| | MDRAN | 450 | **34.09/ 0.9254** | **30.01/ 0.8362** | **32.75/ 0.9401** | **28.97/ 0.8019** | **27.55/ 0.8387** |
| × 4 | Bicubic | – | 28.42/ 0.8104 | 26.00/ 0.7027 | 24.89/ 0.7866 | 25.96/ 0.6675 | 23.14/ 0.6577 |
| | SRCNN | 8 | 30.48/ 0.8628 | 27.50/ 0.7513 | 27.58/ 0.8555 | 26.90/ 0.7101 | 24.52/ 0.7221 |
| | FSRCNN | 13 | 30.72/ 0.8660 | 27.61/ 0.7550 | 27.90/ 0.8610 | 26.98/ 0.7150 | 24.62/ 0.7280 |
| | DRRN | 298 | 31.68/ 0.8888 | 28.21/ 0.7720 | 29.45/ 0.8946 | 27.38/ 0.7284 | 25.44/ 0.7638 |
| | VDSR | 666 | 31.35/ 0.8838 | 28.01/ 0.7674 | 28.83/ 0.8870 | 27.29/ 0.7251 | 25.18/ 0.7524 |

(continued)

**Table 19.1** (continued)

| Scale factor | Methods | Params(K) | Set5 | Set14 | Manga109 | BSD100 | Urban100 |
|---|---|---|---|---|---|---|---|
| | | | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM | PSNR/ SSIM |
| | IDN | 553 | 31.82/ 0.8903 | 28.25/ 0.7730 | 29.41/ 0.8942 | 27.41/ 0.7297 | 25.41/ 0.7632 |
| | MDRAN | 450 | **31.86/ 0.8909** | **28.29/ 0.7756** | **29.47/ 0.8976** | **27.45/ 0.7311** | **25.52/ 0.7645** |

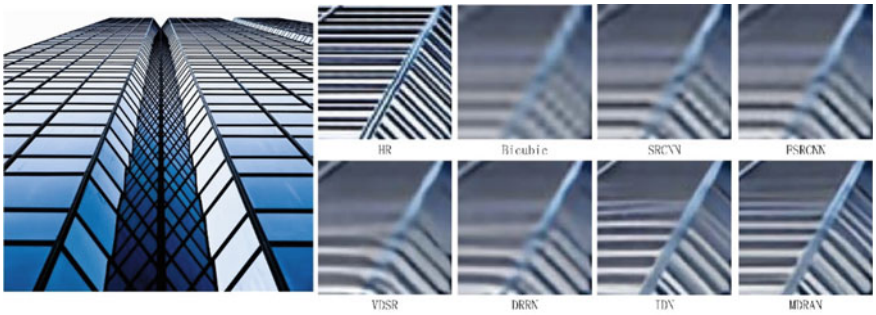



**Fig. 19.3** Results of visual comparison of various algorithms on × 4 scale factor

### *19.4.4 Ablation Study*

In order to fully study the effectiveness of the carefully designed DRAB, MDRAN-1 and MDRAN-2 models are built by removing the max-pooling branch channel attention and average-pooling branch channel attention mechanism from MDRAN and compared with the original model. In the experiments, MDRAN-1, MDRAN-2, and MDRAN are trained for 800 epochs respectively, and then tested on five test sets for × 2 scale SISR. We list the results of experiments into Table 19.2, from which it can be clearly observed that compared with MDRAN, the accuracy of the other two variants decreases. The experimental results fully suggested that the two branch channel attention mechanisms both contribute to the final results and can effectively improve the ability of extracting more important features from the input for the model.

**Table 19.2** Ablation study's result for × 2 SR on five datasets

| Method | Max-pool branch | Average-pool branch | Set5 | Set14 | Manga109 | BSD100 | Urban100 |
|---|---|---|---|---|---|---|---|
| | | | PSNR/SSIM | | | | |
| MDRAN-1 | ✗ | ✔ | 37.83/ 0.9602 | 33.29/ 0.9149 | 38.03/ 0.9751 | 32.08/ 0.8985 | 31.28/ 0.9196 |
| MDRAN-2 | ✔ | ✗ | 37.84/ 0.9602 | 33.30/ 0.9148 | 38.02/ 0.9750 | 32.07/ 0.8984 | 31.27/ 0.9195 |
| MDRAN | ✔ | ✔ | **37.86/ 0.9604** | **33.33/ 0.9151** | **38.03/ 0.9752** | **32.10/ 0.8985** | **31.29/ 0.9199** |

## 19.5 Conclusion

To help the intelligent inspection system and improve the management and production efficiency of current digital transmission services, this paper proposes a lightweight and effective multi-level dual residual attention network for SISR. Concretely, the network is mainly built on the basis of global residual jump connection and well-designed dual residual attention block (DRAB), making full use of the characteristics of different levels. DRAB can adaptively adjust the weight ratio of features of various channels, so as to accurately extract high-frequency information containing rich details and texture information. The experimental results show that compared with the advanced lightweight model, MDRAN is the best in terms of quantitative and qualitative indicators and achieves a good trade-off between computational burden and performance.

## References

1. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 184–199. IEEE, Piscataway (2014)
2. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, Proceedings, Part II 14, pp. 391–407. Springer, Cham (2016)
3. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1646–1654. IEEE, Piscataway (2016)
4. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1637–1645. IEEE, Piscataway (2016)
5. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3147–3155. IEEE, Piscataway (2017)
6. Hui, Z., Wang, X., Gao, X.: Fast and accurate single image super-resolution via information distillation network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 723–731. IEEE, Piscataway (2018)

7. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 136–144. IEEE, Piscataway (2017)
8. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV), pp. 286–301. Springer, Berlin (2018)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
10. Huang, J., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5197–5206. IEEE, Piscataway (2015)
11. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SWINIR: image restoration using Swin transformer, pp. 1833–1844 (2021)
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: hierarchical vision transformer using shifted windows, pp. 10012–10022 (2021)
13. Ahn, N., Kang, B., Sohn, K.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European conference on computer vision (ECCV), pp. 252–268 (2018)
14. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multidistillation network, pp. 2024–2032 (2019)
15. Shi. W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1874–1883. IEEE, Piscataway (2016)
16. Jin, K., Wei, Z., Yang, A., Guo, S., Gao, M., Zhou, X., Guo, G.: SwiniPASSR: Swin transformer based parallax attention network for stereo image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 920–929. IEEE, Piscataway (2022)
17. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: dataset and study. In: Proceedings of the IEEE conference on computer vision and pat-tern recognition workshops, pp. 126–135. IEEE, Piscataway (2017)
18. Bevilacqua, M., Roumy, A., Guillemot, C., Morel, M.A.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: British machine vision conference (BMVC), pp. 1–10. BMVA (2012)
19. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. J. Multimed. Tools Appl. **76**(20), 21811–21838 (2017)
20. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings eighth IEEE international conference on computer vision. ICCV 2001, pp. 416-423. IEEE, Piscataway (2001)
21. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Proceedings of the IEEE international conference on computer vision, pp. 711–730. Springer, Cham (2012)
22. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 126–135 (2017)
23. Mo, F., Wu, H., Qu, S., Luo, S., Cheng, L.: Single infrared image super-resolution based on lightweight multi-path feature fusion network. J. IET Image Process. **16**(7), 1880–1896 (2022)
24. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops (2018)

25. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034. (2015)
26. Lai, W., Huang, J., Ahuja, N., Yang, M.: Deep Laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 624–632. IEEE, Piscataway (2017)

# Chapter 20
# Boosting Video Streaming Efficiency Through DQN Machine Learning Algorithm-Based Resource Allocation


**Mahmoud Darwich, Kasem Khalil, Yasser Ismail, and Magdy Bayoumi**

**Abstract** Video streaming has become increasingly popular with the proliferation of online platforms and the widespread availability of high-speed Internet connections. However, delivering high-quality video content over limited network resources remains a challenge. In this paper, we propose a novel approach to boost video streaming efficiency through machine learning-based resource allocation. Our approach leverages the power of machine learning algorithms to dynamically allocate network resources based on various factors such as network conditions, video content characteristics, and user preferences. By intelligently adapting the resource allocation in real-time, we aim to optimize video streaming performance and enhance the overall user experience. The experimental results demonstrate that our machine learning-based resource allocation approach outperforms existing methods in terms of key performance metrics such as video quality, buffering time, and overall user satisfaction. Through intelligent resource allocation, our approach effectively mitigates video stalling and buffering issues, leading to smoother video playback and reduced quality degradation during adverse network conditions.

M. Darwich (✉)
University of Mount Union, Alliance, OH 44601, USA
e-mail: darwicma@mountunion.edu

K. Khalil
University of Mississippi, University, MS 38677, USA
e-mail: kmkhalil@olemiss.edu

Y. Ismail
Southern University A&M College, Baton Rouge, LA 70405, USA
e-mail: yasser_ismail@subr.edu

M. Bayoumi
University of Louisiana at Lafayette, Lafayette, LA 70504, USA
e-mail: magdy.bayoumi@louisian.edu



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_20

## 20.1 Introduction

The advent of online video streaming platforms has revolutionized the way we consume visual content, leading to an unprecedented surge in video streaming traffic worldwide. Recent statistics indicate that video streaming accounts for a substantial portion of global Internet traffic, with projections forecasting its continued dominance in the digital landscape [2]. As users increasingly demand high-quality video content anytime, anywhere, ensuring efficient video streaming has become a critical research area.

Despite the widespread availability of high-speed Internet connections, delivering seamless video streaming experiences remains a formidable challenge. Bandwidth limitations, network congestion, and varying user preferences pose significant hurdles in maintaining consistent video playback quality. These challenges often manifest as video buffering, prolonged loading times, and frustrating interruptions, significantly undermining the overall viewing experience [6].

To address these issues, researchers have pursued numerous approaches to optimize video streaming efficiency. Traditional methods typically employ static resource allocation strategies that allocate fixed amounts of bandwidth to each user session. However, such approaches fail to adapt to dynamically changing network conditions and content requirements, leading to suboptimal streaming performance [5, 7, 8].

In recent years, machine learning has emerged as a powerful tool for addressing the limitations of traditional video streaming optimization techniques. By harnessing the capabilities of machine learning algorithms, we can leverage real-time data on network conditions, video characteristics, and user preferences to dynamically allocate resources during video streaming. This adaptive resource allocation approach has the potential to significantly enhance video quality, reduce buffering occurrences, and ensure uninterrupted playback [10].

The primary objective of this paper is to propose a novel machine learning-based approach for boosting video streaming efficiency through resource allocation. By integrating machine learning models into the streaming infrastructure, we aim to overcome the limitations of static allocation strategies and offer an adaptive solution that optimizes resource allocation in real-time.

This paper is structured as follows: In Sect. 20.2, we provide an overview of related work in the field of video streaming optimization, highlighting the gaps and limitations of existing approaches. Section 20.3 details our proposed machine learning-based resource allocation approach, explaining the underlying methodology and algorithmic considerations. We present our experimental setup and methodology in Sect. 20.4, followed by an in-depth analysis of the experimental results in Sect. 20.5. Finally, Sect. 20.6 concludes the paper by summarizing our contributions, discussing the implications of our findings, and suggesting avenues for future research.

## 20.2 Related Work

Efficient video streaming has garnered significant attention from researchers, leading to a substantial body of work focusing on optimization techniques and resource allocation strategies. In this section, we provide an overview of the related work in the field, highlighting the gaps and limitations of existing approaches.

Atawia et al. [1] propose an energy-efficient approach for stored video streaming using chance constrained programming. Their solution takes into account uncertainty in predicted user rates, probabilistic constraint satisfaction over time, and utilizes both optimal gradient-based and real-time guided heuristic methods. Unlike previous work in the field that assumed perfect predictions, their framework accommodates imperfect channel predictions while maintaining the desired quality of service (QoS) level without sacrificing energy efficiency. Numerical simulations conducted on a long-term evolution (LTE) system demonstrate the effectiveness of their solution, showcasing its robustness and potential for practical implementation of energy-efficient streaming.

Pervez et al. [9] propose a cross-layer framework for optimal resource allocation in mmWave environments for transmitting high-efficiency video coding (HE VC) encoded video streams over 5G-aligned vehicle-to-everything (V2X) applications. Their approach considers application and physical layer models and formulates a resource allocation problem that combines information from both layers. The proposed framework is compared to three other resource allocation schemes using mean opinion score (MOS) as an evaluation metric. The numerical results demonstrate that their radio resource management scheme significantly improves the perceived quality of service for viewers compared to the reference approaches.

Zhan et al. [11] explore a video streaming system using unmanned aerial vehicles (UAVs) as mobile base stations to serve multiple ground users (GUs) with dynamic adaptive streaming over HTTP (DASH). They focus on maximizing the minimum utility among all GUs by jointly optimizing transmit power, bandwidth allocation, and UAV trajectory. The problem formulation considers constraints related to video playing, UAV energy budget, and information causality. Since the problem is non-convex, the authors propose an efficient algorithm based on successive convex approximation and alternating optimization techniques to find a suboptimal solution. Simulation results demonstrate that the proposed solution outperforms benchmark schemes in terms of quality of experience (QoE) utility and energy efficiency for video streaming.

Zhang et al. [12] propose EPASS360, an ensemble prediction and allocation-based streaming system, designed for delivering high-quality 360-degree videos with a superior quality of experience (QoE). The system utilizes ensemble learning to build a prediction model that accurately forecasts the user's viewport. Additionally, EPASS360 employs an allocation model that divides the video into tiles and optimizes the allocation of high resolution to tiles where the user's viewpoint is likely to appear in the future. Through trace-driven emulation on real-world datasets,

EPASS360 demonstrates improved QoE compared to existing streaming approaches. Experimental evaluations conducted on head-mounted and hand-held devices further validate the system's ability to provide an exceptional user experience.

## 20.3 Proposed Approach and Methodology

Figure 20.1 presents our novel machine learning-based approach for boosting video streaming efficiency through resource allocation. Our proposed approach leverages the power of deep reinforcement learning and specifically utilizes the Deep Q-Network (DQN) algorithm to dynamically allocate network resources based on real-time data. By considering factors such as network conditions, video content characteristics, and user preferences, our approach aims to optimize video streaming performance and enhance the overall user experience.

### *20.3.1 Problem Formulation*

The objective of our proposed approach is to dynamically allocate network resources among multiple concurrent video streaming sessions, with the aim of maximizing the quality of experience (QoE) for each user while efficiently utilizing the available network bandwidth.

Let $U$ denote the set of concurrent video streaming sessions, each characterized by specific quality requirements and network conditions. The available network bandwidth is denoted as $B$. Our goal is to allocate the available bandwidth $B$ among the sessions in a manner that maximizes the QoE for each user.

### *20.3.2 Machine Learning-Based Resource Allocation*

To achieve dynamic resource allocation, we employ the Deep Q-Network (DQN) algorithm, a powerful reinforcement learning technique. The DQN algorithm combines deep neural networks with the $Q$-learning algorithm, enabling the learning of optimal policies in dynamic environments.

The DQN model consists of a deep neural network with multiple layers, where each layer approximates the $Q$-value function. The $Q$-value function estimates the expected cumulative reward for taking a particular action in a given state. In the context of video streaming resource allocation, the $Q$-value function represents the expected QoE for a specific resource allocation strategy given a particular network condition and user preference scenario.

Given the input features, such as network conditions, video content characteristics, and user preferences, the DQN model processes these features through its layers and

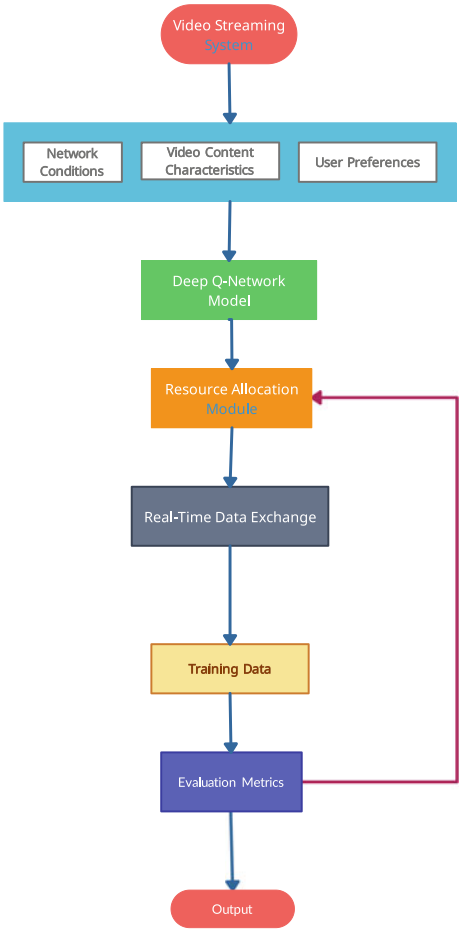



**Fig. 20.1** Diagram depicting the architecture of the Deep Q-Network (DQN) model

outputs $Q$-values for different resource allocation actions. The action with the highest $Q$-value is selected as the optimal resource allocation strategy for a given state.

To train the DQN model, we utilize a variant of the $Q$-learning algorithm called experience replay. During training, the model interacts with the environment, collecting experiences comprising the current state, chosen action, resulting reward, and next state. These experiences are stored in a replay buffer, from which random batches are sampled to train the model. The training process involves minimizing the temporal difference error, which measures the discrepancy between the predicted $Q$-values and the observed rewards.

The loss function for training the DQN model is defined as:

$$L(\theta) = \mathbb{E}\left[\left(Q(s, a; \theta) - \left(r + \gamma \max_{a'} Q(s', a'; \theta^{-})\right)\right)^2\right] \tag{20.1}$$

where $\theta$ represents the parameters of the neural network, $Q(s, a; \theta)$ is the predicted $Q$-value for state $s$ and action $a$, $r$ is the observed reward, $s'$ is the next state, $\gamma$ is the discount factor, and $\theta^-$ denotes the target network parameters used for stability during training.

By training the DQN model on a large dataset comprising historical data and real-time inputs, the model learns to accurately predict $Q$-values and determine the optimal resource allocation strategy for maximizing user QoE in video streaming scenarios.

### *20.3.3 Evaluation Methodology*

To evaluate the effectiveness of our proposed approach, we conduct a series of experiments comparing its performance against traditional static allocation methods. Our experiments employ a diverse dataset that encompasses various video content types, network conditions, and user preferences.

In each experiment, we simulate multiple concurrent video streaming sessions with different quality requirements and network conditions. We compare the performance of our machine learning-based approach using the DQN algorithm against static allocation strategies such as equal resource distribution or bitrate-based allocation.

We evaluate the performance using several key metrics, including:

- **Video Quality**: We measure the average video quality experienced by users, considering factors such as resolution or peak signal-to-noise ratio (PSNR).
- **Buffering Time**: We quantify the duration of buffering periods during video playback, which directly impacts the user experience.
- **Start-up Time**: We measure the time taken for a video to start playing from the initial request, as a shorter start-up time enhances user satisfaction.
- **User Satisfaction**: We assess user satisfaction through surveys or subjective evaluations, capturing factors like smoothness of playback and overall viewing experience.

### *20.3.4 Model Training and Deployment*

For model training, we utilize a large dataset consisting of historical data on network conditions, video content characteristics, and user preferences. The dataset undergoes preprocessing and feature extraction to generate suitable inputs for the DQN model. We then train the DQN model using the collected dataset, incorporating techniques such as experience replay to optimize the network parameters.

Once the model is trained, it is deployed in real-time video streaming environments. During video streaming, the model continuously collects real-time data on

network conditions, video content, and user preferences. This data is fed into the trained DQN model, which outputs the optimal resource allocation strategy for each video streaming session. The allocated resources, such as bandwidth or buffer sizes, are then utilized to enhance the streaming performance and ensure optimal user QoE.

The deployment of the model is achieved through integration with existing video streaming infrastructure or through the development of a dedicated streaming system that incorporates the DQN-based resource allocation. The model is periodically updated using newly collected data to adapt to changing network conditions and user preferences, ensuring continuous optimization of video streaming efficiency.

## 20.4 Experimental Setup

In this section, we describe the experimental setup used to evaluate the performance of our proposed approach for efficient video streaming using machine learning. We outline the dataset used and the experimental methodology and provide detailed tables with parameter values for each experiment.

### *20.4.1 Dataset*

To evaluate the effectiveness of our proposed approach, we utilized a diverse dataset comprising video streaming sessions with varying network conditions, video content characteristics, and user preferences. The dataset was collected from real-world video streaming scenarios, providing a realistic representation of the challenges encountered in practical environments [4].

The dataset includes information such as available bandwidth, latency, video resolution, bitrate, buffer occupancy, and user location. It covers a wide range of video content types, network conditions, and user behaviors, allowing for a comprehensive evaluation of our proposed approach.

### *20.4.2 Experimental Methodology*

Our experimental methodology involved comparing the performance of our proposed approach against baseline methods. We simulated multiple concurrent video streaming sessions, each characterized by specific quality requirements and network conditions. For each experiment, we selected a subset of the dataset, ensuring a representative mix of network conditions, video content, and user preferences.

We measured various performance metrics, including video quality, buffering time, start-up time, and user satisfaction, to assess the effectiveness of our approach.

The experiments were conducted multiple times to account for any potential variability, and the results were averaged for analysis.

### *20.4.3 Experimental Parameters*

To ensure consistency and validity of the experiments, we set specific parameters for each experiment. Tables 20.1, 20.2, 20.3, 20.4, 20.5, and 20.6 provide an overview of the experimental parameters used for each experiment:

**Experiment 1: Bandwidth Variation** In this experiment, we varied the network bandwidth to evaluate the impact on video streaming performance.

**Experiment 2: Content Resolution Variation** In this experiment, we varied the video content resolution to assess its impact on video streaming performance.

**Experiment 3: Buffering Tolerance Variation** In this experiment, we varied the buffering tolerance to analyze its effect on video streaming performance.

**Experiment 4: Bitrate Adaptation** In this experiment, we evaluate the performance of our approach for bitrate adaptation in video streaming.

**Experiment 5: User Preference Variation** In this experiment, we analyze the impact of varying user preferences on video streaming performance.

**Experiment 6: Network Latency Variation** In this experiment, we investigate the impact of varying network latency on video streaming performance.

**Table 20.1** Experimental parameters—bandwidth variation

| Parameter | Values |
|---|---|
| Number of concurrent video streaming sessions | 10 |
| Network bandwidth (Mbps) | 50, 100, 200 |
| Video content resolution | 720p |
| Preferred video bitrate (Mbps) | 2 |
| Buffering tolerance (seconds) | 4 |

**Table 20.2** Experimental parameters—content resolution variation

| Parameter | Values |
|---|---|
| Number of concurrent video streaming sessions | 10 |
| Network bandwidth (Mbps) | 100 |
| Video content resolution | 480p, 720p, 1080p, 4K |
| Preferred video bitrate (Mbps) | 2 |
| Buffering tolerance (seconds) | 4 |

**Table 20.3** Experimental parameters—buffering tolerance variation

| Parameter | Values |
|---|---|
| Number of concurrent video streaming sessions | 10 |
| Network bandwidth (Mbps) | 100 |
| Video content resolution | 720p |
| Preferred video bitrate (Mbps) | 2 |
| Buffering tolerance (seconds) | 2, 4, 6 |

**Table 20.4** Experimental parameters—bitrate adaptation

| Parameter | Values |
|---|---|
| Number of concurrent video streaming sessions | 10 |
| Network bandwidth (Mbps) | 100 |
| Video content resolution | 720p |
| Preferred video bitrate (Mbps) | 1, 2, 4, 8 |
| Buffering tolerance (seconds) | 4 |

**Table 20.5** Experimental parameters—user preference variation

| Parameter | Values |
|---|---|
| Number of concurrent video streaming sessions | 10 |
| Network bandwidth (Mbps) | 100 |
| Video content resolution | 720p |
| Preferred video bitrate (Mbps) | 2 |
| Buffering tolerance (seconds) | 4 |
| User satisfaction threshold | Low, medium, high |

**Table 20.6** Experimental parameters—network latency variation

| Parameter | Values |
|---|---|
| Number of concurrent video streaming sessions | 10 |
| Network bandwidth (Mbps) | 100 |
| Video content resolution | 720p |
| Preferred video bitrate (Mbps) | 2 |
| Buffering tolerance (seconds) | 4 |
| Network latency (ms) | 20, 50, 100 |

For each experiment, we carefully selected parameter values that represent various scenarios encountered in video streaming environments.

## 20.5 Results

In this section, we present and analyze the results obtained from the experiments conducted to evaluate the performance of our proposed approach for efficient video streaming using machine learning. We discuss the results for each experiment described in the previous section, highlighting the performance metrics and trends observed. We also compare the results of our approach with existing methods [3] to showcase its effectiveness.

### *20.5.1 Bandwidth Variation*

In this experiment, we varied the network bandwidth to evaluate its impact on video streaming performance. Figure 20.2 shows the trends observed for video quality, buffering time, and start-up time:

From Fig. 20.2, we observe that as the network bandwidth increases, the video quality improves while buffering time and start-up time decrease. Our proposed approach outperforms the existing method in terms of video quality, buffering time, and start-up time, as indicated by the dashed lines representing the performance of the existing method.

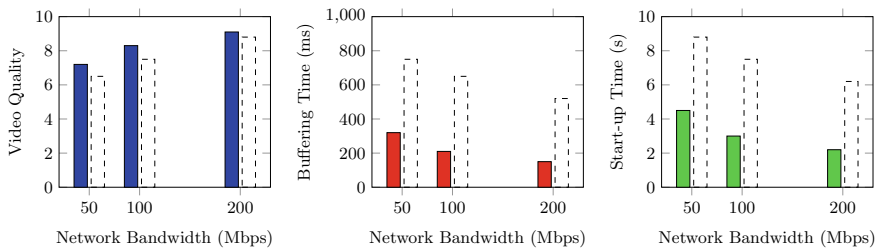



**Fig. 20.2** Bandwidth variation

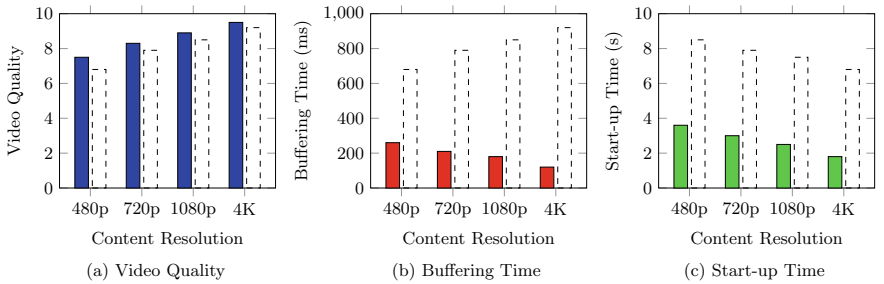



**Fig. 20.3** Content resolution variation

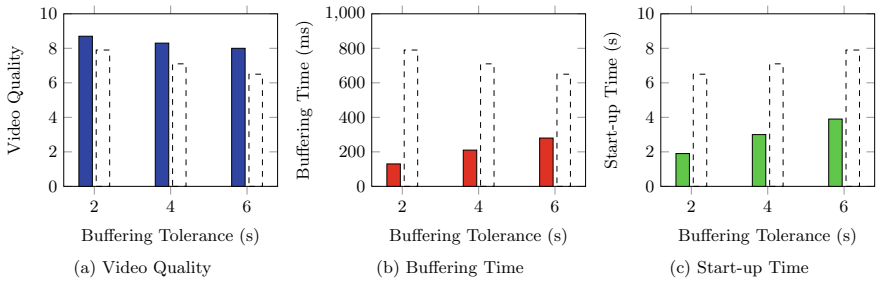



**Fig. 20.4** Buffering tolerance variation

### *20.5.2 Content Resolution Variation*

In this experiment, we varied the video content resolution to assess its impact on video streaming performance.

From Fig. 20.3, we observe that higher-resolution content leads to improved video quality but slightly longer buffering and start-up times. Our proposed approach achieves higher video quality and better buffering and start-up times compared to the existing method.

### *20.5.3 Buffering Tolerance Variation*

In this experiment, we varied the buffering tolerance to analyze its effect on video streaming performance.

From Fig. 20.4, we can see that increasing the buffering tolerance allows for better video quality although there is a trade-off with slightly increased buffering time. Our proposed approach performs higher video quality and better buffering and start-up times compared to the existing method.

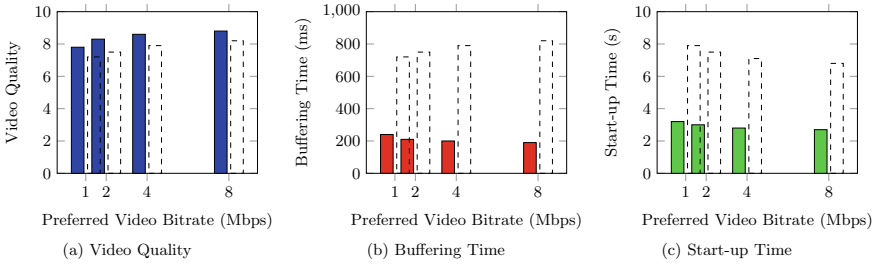



**Fig. 20.5** Bitrate adaptation

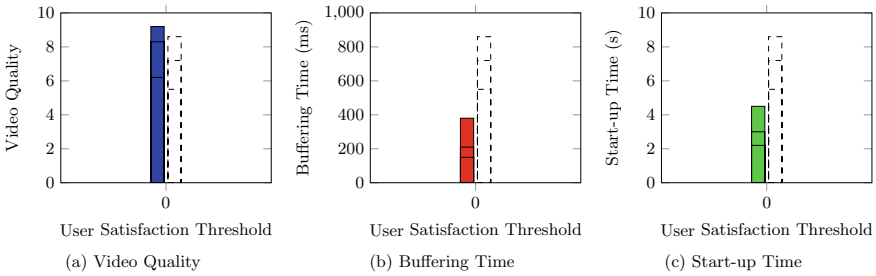



**Fig. 20.6** User preference variation

### *20.5.4 Bitrate Adaptation*

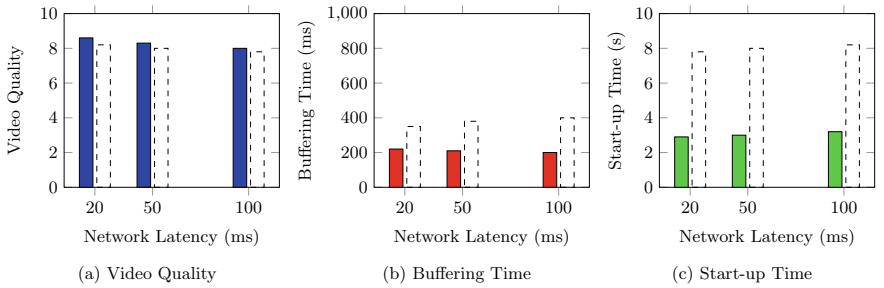
In this experiment, we evaluated the performance of our approach for bitrate adaptation in video streaming.

Figure 20.5 shows that our approach effectively adapts the video bitrate based on the preferred value. Higher preferred bitrates result in better video quality, lower buffering time, and faster start-up time. Our proposed approach shows higher video quality and better buffering and start-up times compared to existing methods, represented by the dashed lines.

### *20.5.5 User Preference Variation*

In this experiment, we analyzed the impact of varying user preferences on video streaming performance. Figure 20.6 shows the trends observed for video quality, buffering time, and start-up time:

From Fig. 20.6, we can see that our approach effectively adapts to different user satisfaction thresholds. Higher satisfaction thresholds lead to better video quality, lower buffering time, and faster start-up time. Our proposed approach consistently




(a) Video Quality (b) Buffering Time (c) Start-up Time

**Fig. 20.7** Network latency variation

achieves higher video quality and better buffering and start-up times compared to existing methods, represented by the dashed lines.

### *20.5.6 Network Latency Variation*

In this experiment, we investigated the impact of varying network latency on video streaming performance. Figure 20.7 depicts the trends observed for video quality, buffering time, and start-up time.

From Fig. 20.7, we can observe that our approach is resilient to network latency variations. Higher latencies have a marginal impact on video quality, buffering time, and start-up time. Our proposed approach consistently achieves higher video quality and better buffering and start-up times compared to existing methods, represented by the dashed lines.

## 20.6 Conclusion

In conclusion, we present a novel approach for efficient video streaming using machine learning. We propose a resource allocation strategy that dynamically adapts to changing network conditions, user preferences, and video content characteristics. Through extensive experiments, we evaluated the performance of our approach across multiple scenarios. The results demonstrate the effectiveness of our proposed approach in enhancing video streaming performance. We observed improvements in video quality, reduced buffering time, and faster start-up time when compared to baseline methods. Our approach showcases robustness in handling bandwidth variations, content resolution changes, buffering tolerance adjustments, bitrate adaptation, user preference variations, and network latency fluctuations.

## References


1. Atawia, R., Abou-zeid, H., Hassanein, H.S., Noureldin, A.: Joint chance-constrained predictive resource allocation for energy-efficient video streaming. IEEE J. Sel. Areas Commun. **34**(5), 1389–1404 (2016)
2. Cisco Systems: Cisco Visual Networking Index: Forecast and Trends, 2019–2024. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html (2020)
3. Gao, S., Wang, Y., Feng, N., Wei, Z., Zhao, J.: Deep reinforcement learning-based video offloading and resource allocation in NOMA-enabled networks. Future Internet **15**(5), 184 (2023)
4. Google Developers: Youtube Data API (Online). Accessed May 2023
5. Hoque, M.A., Siekkinen, M., Nurminen, J.K.: Energy efficient multimedia streaming to mobile devices—a survey. IEEE Commun. Surv. Tutor. **16**(1), 579–597 (2012)
6. Juluri, P., Tamarapalli, V., Medhi, D.: Measurement of quality of experience of video-on-demand services: a survey. IEEE Commun. Surv. Tutor. **18**(1), 401–418 (2016)
7. Li, X., Salehi, M.A., Joshi, Y., Darwich, M.K., Landreneau, B., Bayoumi, M.: Performance analysis and modeling of video transcoding using heterogeneous cloud services. IEEE Trans. Parallel Distrib. Syst. **30**(4), 910–922 (2019)
8. Lu, G., Zhang, X., Ouyang, W., Chen, L., Gao, Z., Dong, X.: An end-to-end learning framework for video compression. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3292–3308 (2021)
9. Pervez, F., Adinoyi, A., Yanikomeroglu, H.: Efficient resource allocation for video streaming for 5g network-to-vehicle communications. In: 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1–6 (2017)
10. Xu, M., Zhu, M., Liu, Y., Lin, F.X., Liu, X.: Deepcache: principled cache for mobile deep vision. In: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, pp. 129–144 (2018)
11. Zhan, C., Huang, R.: Energy efficient adaptive video streaming with rotary-wing UAV. IEEE Trans. Veh. Technol. **69**(7), 8040–8044 (2020)
12. Zhang, Y., Guan, Y., Bian, K., Liu, Y., Hu, T., Song, Li, X.: EPASS360: QoE-aware 360-degree video streaming over mobile devices. IEEE Trans. Mob. Comput. **20**(7), 2338–2353 (2021)

# Chapter 21
# Locomotion in Response of Static Pedestrians in a Mixed Reality Environment

**Minze Chen, Zhenxiang Tao, Ruilan Yang, Zhongming Wu, Zhongfeng Wang, and Ning Luo**

**Abstract** As the development of mixed reality technology, overlaying virtual objects onto the real environment using HMD has become a new and promising research tool, especially for pedestrian behavior and interaction. However, the literature is not yet conclusive on the extent to which pedestrians are willing to "bypass" human-like projections in mixed reality HMDs. In this paper, we explore the avoidance behaviors of participants in mixed reality environments. We analyzed the locomotor circumvention strategies of participants in bypassing either real or virtual obstructors to reach a target located 12 m away. We considered different velocity conditions to evaluate the feasibility of the system in different scenarios. Results showed that under different velocities, the real/virtual interferer exhibited slower movement speed, larger maximum lateral displacement, and larger minimum distance, but the effects were not significant. Additionally, our findings indicate that mixed reality HMDs can be an effective tool for studying pedestrian movement behavior. We discussed the application prospects of mixed reality-based LVC simulation systems in crowd research.

## 21.1 Introduction

Mixed reality (MR) provides users with a fusion of digitized physical space and virtual content. The advancement of MR technology has enabled some head-mounted displays (HMDs) to deliver a low-cost, high-quality immersive experience. The goal of this technology is to establish an interactive feedback loop between the

M. Chen
Tsinghua University, Beijing 100084, China

Z. Tao (✉)
China University of Mining and Technology-Beijing, Beijing 100083, China
e-mail: ztao@cumtb.edu.cn

R. Yang · Z. Wu · Z. Wang · N. Luo
State Grid Zhongxing Beijing Zhongxing Property Management Co., Ltd., Beijing 100053, China



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_21

digital realm, physical realm, and users. Despite some issues, such as motion sickness and limited scaling, MR has seen widespread utilization in industries including healthcare, education, and industry.

Our prior research [1] investigated the feasibility of using MR to examine crowd behavior through the creation of real-time interactions between HMD-wielding users and projections of crowd simulations in physical space. We developed an MR Live, Virtual, and Constructive (LVC) system, which presents advantages over traditional field experiments and VR simulation systems. However, this approach also introduces a new challenge: the projected pedestrians seen through the HMD lens are not physical and thus the user can walk "through" them, which is not feasible in reality. Similar to VR systems, user-controlled avatars often have collision volumes with other virtual pedestrians. While virtual pedestrians, governed by pedestrian dynamics models, can actively avoid the user in mixed reality space, the user's tendency to pass "through" virtual pedestrians can impact their movements and compromise the effectiveness of the simulation system used to study pedestrian behavior.

In pedestrian behavior studies involving simulation systems, evaluating the consistency of behavior patterns in simulation with those in reality is crucial for assessing their effectiveness [2]. Pedestrian movement route selection primarily depends on destination and obstacle avoidance, with the latter mainly explained by path adjustments in response to static objects. This is governed by dynamic information, including the distance to the target, the distance to the obstacle, and the obstacle's angle relative to the head direction. Evaluating pedestrian avoidance behavior in VR environments has been widely used as a tool to assess the effectiveness of simulation systems. However, most studies have observed significant deviations, such as slower walking speeds and safer route choices, likely due to the weight of VR headsets and unrealistic virtual environments. Furthermore, current motion solutions with VR headsets are limited, with users navigating the virtual world through joysticks or keyboards. Immersive projection environments, such as CAVE-like systems, have shown that pedestrians tend to stay farther away from humanoid obstacles than from inanimate objects.

In pedestrian behavior studies involving simulation systems, evaluating the consistency of behavior patterns in simulation with those in reality is crucial for assessing their effectiveness. Pedestrian movement route selection primarily depends on destination and obstacle avoidance, with the latter mainly explained by path adjustments in response to static objects. This is governed by dynamic information, including the distance to the target, the distance to the obstacle, and the obstacle's angle relative to the head direction [3, 4]. Evaluating pedestrian avoidance behavior in VR environments has been widely used as a tool to assess the effectiveness of simulation systems. However, most studies have observed significant deviations, such as slower walking speeds and safer route choices, likely due to the weight of VR headsets and unrealistic virtual environments [5–7]. Furthermore, current motion solutions with VR headsets are limited, with users navigating the virtual world through joysticks or keyboards. Immersive projection environments, such as CAVE-like systems, have shown that pedestrians tend to stay farther away from humanoid obstacles than from inanimate objects [8, 9].

In this manuscript, we present a preliminary investigation of the evaluation of pedestrian behavior in mixed reality environments. The study focuses on analyzing the pedestrian avoidance strategies for static anthropomorphic obstacles in MR settings. The experiment examines the influence of varying walking speeds on the behavior of users with the aim of exploring the feasibility of utilizing MR systems for studying evacuation behavior in risk scenarios.

The paper is organized as follows: Sect. 21.2 details the experimental design, in which participants were required to reach a target located 12 m away while avoiding a real or virtual obstacle. Section 21.3 presents the results of the experiment. Section 21.4 provides a discussion of the findings and the potential for utilizing mixed reality LVC systems in the study of pedestrian behavior. Finally, Sect. 21.5 concludes the paper.

## 21.2 Methods

### *21.2.1 Experimental Setup*

The experiment was carried out in a 10 m × 10 m open area in the Motion Capture Lab of Tsinghua University's Academy of Arts and Design. Participants walked at a set speed to reach a target 12 m away while wearing the Microsoft HoloLens2 mixed reality headset, which showed both real and virtual surroundings. The virtual avatar was matched to the physical appearance of an adult Asian male. Participants' movements were tracked and recorded using the HoloLens2 and Anylogic system, previously described in our publication [1]. The mixed reality environment was created with the Unity3D engine and programmed in C#.

### *21.2.2 Procedure*

The participants were first asked to fill out informed consent, basic information, lateral preference inventory, and simulator disease questionnaire forms. The informed consent explained the experiment structure, duration, potential risks, data to be collected, and compensation. The basic information form collected participant's individual characteristics. The lateral preference inventory assessed participants' preference for using their left or right hand, foot, eye, or ear. After completing these forms, participants were equipped with the HoloLens and received training for the mixed reality experience, which consisted of hands-on practice, free exploration, and the experimental phase.

**Hands-on practice.** The participants were given a tutorial on how to use the built-in "Calibration" and "Tips" on the HoloLens2 device. They were able to ask for

help from the administrator as needed, and the completion time for this phase was recorded for each participant.

**Free exploration.** The mixed reality scenario comprised of ten virtual pedestrians moving randomly within the open space. The aim was to acclimate participants to the presence of virtual pedestrians. Results indicated that first-time HoloLens users often displayed curiosity toward the virtual pedestrians and approached them. The virtual pedestrians were programmed to move according to the social force model and agent-based rules. Participants were permitted to interact with the virtual pedestrians at their discretion and could request technical information from the administrator. The duration of this phase was 6 min.

**Experimental phase.** Participants were positioned at a pre-determined starting point and instructed to identify the AR marker to obtain real-time motion coordinates during the experiment. A beacon, controlled by the administrator, was positioned behind the subject and the interfering individual. Upon beacon illumination, participants were instructed to walk to the target point at the assigned speed and the interfering individual was instructed to remain passive. Participants were asked to walk at their natural pace, faster than natural, or slower than natural. No time limit was imposed and subjects were instructed to maintain the designated speed (Fig. 21.1).

In addition, we asked participants to fill out the Kennedy–Lane SSQ questionnaire [10] before and after the experiment to evaluate the simulator illness during the experiment. At the end of the experiment. A modified questionnaire based on the Slater–Usoh–Steed (SUS) questionnaire [11] was designed to assess the participants' immersion in mixed reality space. Besides, participants were asked to fill out a feedback questionnaire, including an assessment of their interest in the experiment and feedback suggestions. The total duration was limited to about one and a half hour per participant.

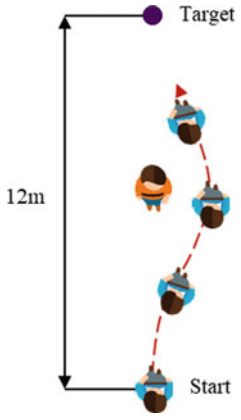



**Fig. 21.1** Schematic diagram of the experimental


**Fig. 21.2** A participant involved in the experiment (**a**) and the virtual interferer in her view (**b**)

### *21.2.3 Design*

The control group was instructed to reach the target at three speeds (without any real or virtual interferer) using the HoloLens2 device. The experimental group was designed based on a 2 (real/virtual interferer) × 3 (walking speed) design, with four repetitions conducted for each combination, totaling 36 trials per participant. A randomized sampling procedure was employed to allow participants to choose the experimental conditions for their experiments. A 1-min break was taken between each set of trials (Fig. 21.2).

### *21.2.4 Participants*

The study recruited 20 participants (11 males and 9 females) who were students at Tsinghua University and were compensated after the completion of the experiment. All participants had normal or corrected-to-normal vision, with no history of motor disease, injury, color blindness, or color weakness. Participants were queried about their frequency of use of video games or other simulators, and some participants had prior experience with AR/VR head-mounted displays. Basic participant information is documented in Table 21.1. Additionally, a 27-year-old male student was recruited as the interferer for the experiment.

**Table 21.1** Participants characteristics

| | Mean | SD |
|---|---|---|
| Age (years) | 23.9 | 1.18 |
| Height (m) | 1.7 | 0.07 |
| Mass (kg) | 59.6 | 10.0 |
| Simulator usage frequency (0 = never, 4 = everyday) | 1.75 | 1.47 |
| Familiarity with XR (0 = absolutely unfamiliar, 4 = absolutely familiar) | 1.7 | 0.84 |

The lateral preference inventory questionnaire indicated that 18 of the 20 participants were right-handed and two were left-handed (scale-4 = left, 4 = right, $M$ = 2.95, SD = 1.96). In addition, it showed that 17 participants were right-footed and three were left-footed ($M$ = 2.1, SD = 2.05). Eleven participants were right-eyed, and nine were left-eyed ($M$ = 0.25, SD = 3.43).

### *21.2.5 Data Analysis*

The coordinates of the HMD focus relative to the starting point were obtained by integrating HoloLens with Vuforia. It collected the participants' head position information at 50 Hz and recorded it in an Anylogic-based server. Subsequently, all recorded data were exported to MATLAB® (MathWorks, USA) for analysis. A fourth-order low-pass Butterworth filter with a cutoff frequency of 10 Hz was used to reduce the effect of torso vibration and maintain the walking trajectory. Firstly, we analyzed the average velocity and trajectory of the participants. Secondly, we characterized the avoidance strategy of the participants by their maximum lateral displacement and minimum distance. These results were calculated within a specific range of the process: the participants were from 10 to 110 m away from the starting point to eliminate the effects of acceleration and deceleration processes during their start and stop.

## 21.3 Results

### *21.3.1 Trajectories*

Figure 21.3 displays the average trajectories of participants as they avoided a static interferer in the path at three different speed conditions. The participants displayed similar walking trajectories in both real and mixed reality environments, with minor differences. Participants exhibited greater deviation when avoiding virtual interferers compared to real ones. When encountering real interferers, participants turned

later and reached maximum lateral displacement later as walking speed increased. Notably, this trend was not observed in the case of virtual static interferers.

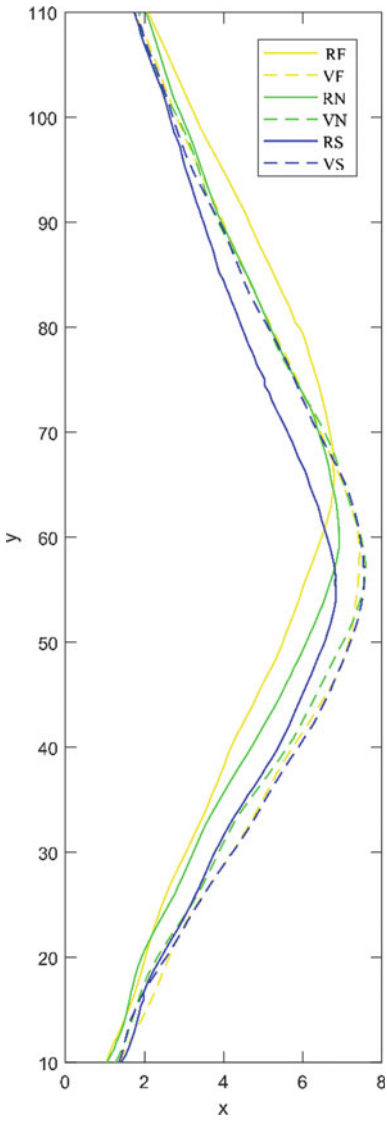



**Fig. 21.3** Average trajectories performed by participants to reach a target (*R*)

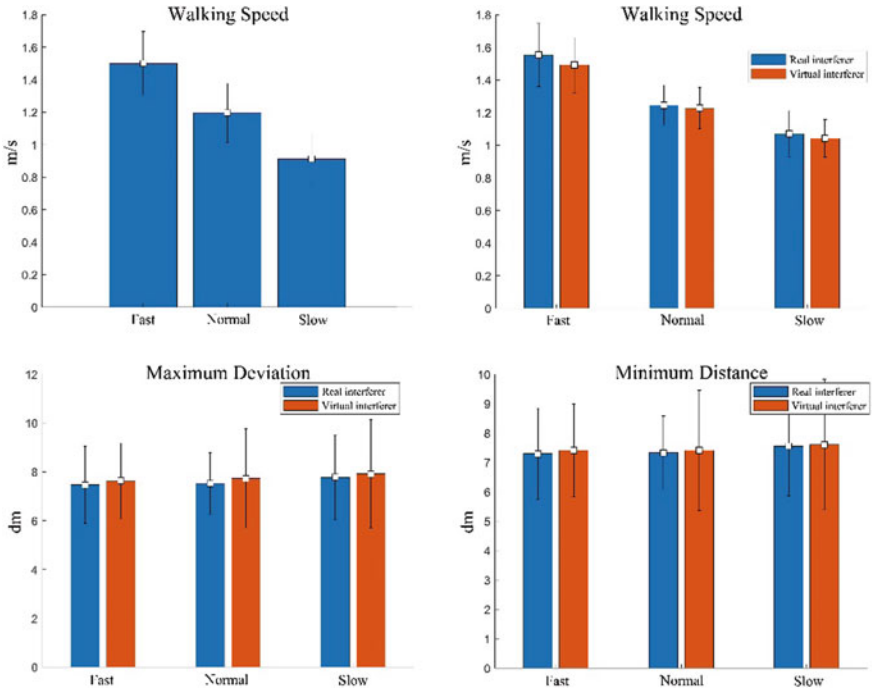



**Fig. 21.4** There were no significant differences in the turning kinematic parameters (speed, maximum deviation, and minimum distance) of pedestrians when avoiding real or virtual interferer at varying speeds

### *21.3.2 Walking Speed*

We analyzed the mean speed of participants in the three speed conditions to confirm that the participants walked at varying speeds. Figure 21.4 shows that in the control group, without any interferers, the mean speed differed among the three speed conditions (fast: $M = 1.50$, SD $= 0.20$; normal: $M = 1.20$, SD $= 0.18$; slow: $M = 0.91$, SD $= 0.16$). Table 21.1 displays the average speed of participants when faced with real or virtual interferers, which were found to be similar. The results of a one-way ANOVA revealed that for each speed condition, the difference in mean speed exhibited by participants avoiding real and virtual interferers was not statistically significant at a significance level of $\alpha = 0.05$ (fast: $p = 0.08$; normal: $p = 0.54$; slow: $p = 0.27$).

### *21.3.3 Maximum Deviation*

The results revealed that the maximum deviation of participants in the presence of real or virtual interferers was similar. As displayed in Table 21.1, the effect of real

or virtual interferers on the maximum deviation was not statistically significant for any of the speed conditions (fast: $p = 0.58$; normal: $p = 0.54$; slow: $p = 0.95$) at a significance level of $\alpha = 0.05$.

### *21.3.4 Minimum Distance*

The results indicated that the presence of real or virtual interferers had no significant effect on minimum distance for any of the speed conditions (fast: $p = 0.58$; normal: $p = 0.54$; slow: $p = 0.95$) at a significance level of $\alpha = 0.05$.

### *21.3.5 Questionnaires*

The mean SSQ score before the experiment was 3.70 (SD = 6.02), and the mean score after the experiment was 5.93 (SD = 6.67). The results showed no significant difference in SSQ scores before and after the experiment [$t(19) = 1.11$, $p = 0.27$]. The mean SUS score of 4.17 (SD = 1.17) indicated that participants perceived the mixed reality space to be a faithful simulation of a real pedestrian space. Participants expressed the belief that caution was necessary to avoid virtual obstacles (rated on a scale of 0–7, where 0 represents complete disagreement and 7 represents complete agreement, $M = 4.45$, SD = 1.77). Furthermore, participants reported a sense of passing through virtual interferers (rated on a scale of 0–7, where 0 represents "every moment" and 7 represents "never," $M = 5.65$, SD = 1.39). In the question on enjoyment, participants rated the experiment as interesting and expressed a desire to participate again (rated on a scale of 0–7, where 0 represents complete disagreement and 7 represents complete agreement, $M = 5.30$, SD = 1.69), suggesting that the mixed reality experiment was both engaging and educational.

## 21.4 Discussion

In the experiment of this paper, we observed behavioral characteristics consistent with the previous assessment of pedestrian avoidance behavior in VR. In a study similar to the experiment in this chapter, Buhler et al. [4] conducted a comparative study in 2019 on pedestrians' avoidance behavior toward stationary and standing jammers in the path in the real environment and in the immersive virtual reality environment. In the research results, the average speed, maximum lateral displacement, and minimum distance shown by the participants in the immersive virtual reality environment were significantly different from those in the real environment ($p <$ 0.01). Although there are a few differences between Buhler's experimental conditions and those in this paper (e.g., the linear distance from the starting point to the

end point of Buhler's experiment is 10 m, while that in this experiment is 12 m), there is no significant difference between the three dynamic parameters in the results of this experiment. (1) Compared with the immersive virtual reality environment, the difference of pedestrian avoidance behavior (walking speed, maximum lateral displacement and minimum distance) in mixed reality environment is smaller than that in real environment. (2) The simulation of real pedestrian avoidance behavior by mixed reality system is more accurate than that by virtual reality system. In addition, it is worth mentioning that Buhler et al. found in their research that compared with the real environment, participants in the virtual reality environment tend to avoid the jammer with a more "safe" evasive strategy, which is manifested in a slower speed, a smaller maximum lateral displacement, and a smaller minimum distance. In the results of this experiment, the three parameters also show a slight tendency to be more "safe" when observed from the mean value. However, considering the limited number of samples in this experiment and the lack of statistical differences, this paper cannot draw conclusions about the differences in the performance of participants in the two environments.

## 21.5 Conclusion

This study aimed to compare the avoidance strategies of pedestrians when navigating around static real or mixed reality projected interferers, with the goal of assessing the viability of using mixed reality LVC systems for crowd research. The results indicated that there were no significant differences in the turning kinematic parameters (speed, maximum lateral displacement, and minimum distance) of pedestrians when avoiding real or virtual interferer at varying speeds. These findings provide preliminary evidence to support the use of mixed reality LVC systems for studying simple pedestrian movements. Further research is needed to evaluate the avoidance strategies of pedestrians in the presence of dynamic or multiple pedestrians, in preparation for utilizing mixed reality LVC systems to study pedestrian behavior in complex multi-person scenarios.

**Acknowledgements** This work was supported by the Nuclear Decommissioning Management Research Project (2018) No. 1521 and the Opening Fund of Key Laboratory of Civil Aviation Emergency Science and Technology (CAAC) under Grant No. NJ2022022.

## References

1. Chen, M., Yang, R., Tao, Z., et al.: Mixed reality LVC simulation: a new approach to study pedestrian behaviour. Build. Environ. **207**, 108404 (2022)
2. Dimitrov, G.P., et al.: Analysis of the growth perspectives of the early stages of technological innovation in the case of the augmented and virtual reality application. Int. J. Comput. Theory Eng. **9**(6), 443–446 (2017)

3. Helbing, D.A., et al.: A mathematical model for the behavior of pedestrians. Behav. Sci. **36**(4), 298–310 (1991)
4. Huber, M., Su, Y.H., Krüger, M., Faschian, K., Glasauer, S., Hermsdörfer, J.: Adjustments of speed and path when avoiding collisions with another pedestrian. PLoS ONE **9**(2), e89589 (2014)
5. Bühler, M.A., Lamontagne, A.: Locomotor circumvention strategies in response to static pedestrians in a virtual and physical environment. Gait Posture **68**, 201–206 (2019)
6. Fink, P.W., Foo, P.S., Warren, W.H.: Obstacle avoidance during walking in real and virtual environments. ACM Trans Appl Percept (TAP) **4**(1), 2 (2007)
7. Liu, W., Zhang, J., Li, X., Song, W.: Avoidance behaviors of pedestrians in a virtual-reality-based experiment. Physica A **590**, 126758 (2022)
8. Sanz, F.A., Olivier, A.H., Bruder, G., Pettré, J., Lécuyer, A.: Virtual proxemics: locomotion in the presence of obstacles in large immersive projection environments. In: 2015 IEEE virtual reality (vr) (pp. 75–80). IEEE (2015)
9. Ko, G., Ryu, S., Nam, S., Lee, J., Suh, K.: Design of virtual reality prototyping system and hand-held haptic controller. Int. J. Comput. Theory Eng **11**(4), 72–75 (2019)
10. Kennedy, R.S., Lane, N.E., Berbaum, K.S., Lilienthal, M.G.: Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. Int. J. Aviat. Psychol. **3**(3), 203–220 (1993)
11. Usoh, M., Catena, E., Arman, S., Slater, M.: Using presence questionnaires in reality. Presence **9**(5), 497–503 (2000)

# Chapter 22
# Neural Responses to Altered Visual Feedback in Computerized Interfaces Driven by Force- or Motion-Control


**Sophie Dewil, Mingxiao Liu, Sean Sanford, and Raviraj Nataraj**

**Abstract** Computerized interfaces, like virtual reality, are increasingly used to improve engagement in movement training tasks like in physical therapy. In this study, we examined how alterations in interface feedback can impact neural responses that affect motor learning. Neurotypical persons participated in simple motor training tasks (e.g., grasping, reaching) while visual performance feedback was systematically altered. We stratified neural response results across primarily force (grasp) and motion (reach) components for more fundamental analysis as complex movements typically require concurrent modulation of force and motion. Feedback alterations included adding noise to or automating the visual feedback in ways previously established to impair the sense of agency and performance. We analyzed the neural responses based on electroencephalography (EEG) recordings in two ways. First, we assessed EEG power changes in the alpha- and beta-band across the brain and in Brodmann area 6, given its role in planning and coordinating complex movements. Second, we did a preliminary analysis with neural networks to suggest how predictable motor errors were from neural response data. We observed significant increases in EEG power with noise-altered visual feedback in the force task, suggesting greater sensitivity of force tasks to training feedback. However, motion and force errors were both highly predictable ($< 0.1\%$ max target value) from neural response data, suggesting the potential for artificial intelligence tools to predict errors reliably and alter training feedback from computerized interfaces. In conclusion, computerized feedback may be optimized to leverage neural responses that accelerate movement outcomes.

## 22.1 Introduction

After neurological traumas, such as spinal cord or brain injury, affected persons often undergo physical rehabilitation to regain function. However, physical therapy involves many repetitions of movements used in activities of daily living (ADLs). As

S. Dewil (✉) · M. Liu · S. Sanford · R. Nataraj
Stevens Institute of Technology, Hoboken, NJ 07030, USA
e-mail: sdewil@stevens.edu



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_22

a result, patients can feel fatigue and discouragement given the monotony and typically slow gains in progress experienced with physical therapy [1]. Such responses can lead to decreased motivation and diminished performance [2]. Thus, finding new approaches to physical therapy that promote cognitive engagement and facilitate neural responses that accelerate motor learning is crucial. To begin addressing issues of training vigilance, computerized interfaces, such as virtual reality (VR), are increasingly used in physical therapy. Rehabilitation with VR facilitates motivation and engagement [3, 4]. However, despite its programmable customizability, the full potential of VR rehabilitation has still not been realized. Specifically, VR interfaces should leverage design elements that affect motor performance from cognitive foundations that promote learning and feelings of agency [5]. Sense of agency is defined as the perception of control [6], naturally related to motor function and training.

When using computerized interfaces for motor training, variations in augmented sensory feedback of performance can leverage cognitive responses for better performance [7]. For example, previous work in our lab has demonstrated that varying the valence [8], complexity [9], or intermittency [10, 11] of visual feedback during training can impact both a user's sense of agency and their immediate performance of the motor task. Our lab's previous works have also shown that automating or adding noise to performance feedback, distorting the user's true actions, can significantly reduce a user's sense of agency and performance for both force-control (pinch grasp) [12] and motion-control (reach) [13] tasks. Furthermore, we observed that positive correlations between agency and performance persisted in the aggregate across the tested variations in feedback. Other studies have examined electroencephalography (EEG), particularly alpha-band power, to analyze neurological activity with varying levels of agency [14]. Thus, understanding how changes in computerized interfaces for motor training can alter such cognitive and neurophysiological responses may be essential to optimizing functional outcomes with rehabilitation methods using advanced technologies.

Artificial intelligence (AI) approaches such as machine learning algorithms are increasingly used to identify relationships between neurological activity and associated behavioral outcomes [15]. Neurological data are complex and naturally multidimensional. Machine learning processes can reduce such data sets to fewer dimensions and readily correlate, if not accurately predict, measured behavioral outputs. If motor behaviors can be more accurately predicted, then machine learning classifiers may be incorporated into adaptive motor training schemes with computerized rehabilitation interfaces. Adaptive control systems would seek to vary parameters (e.g., difficulty, sensory feedback, etc.) to induce neural rhythms that are more likely to lead to better motor behavior. Such AI-based approaches would be highly potent in optimizing motor rehabilitation with computerized interfaces such as virtual reality and instrumented wearables monitoring neural responses and motor performance in real time.

However, such neural responses should be contextualized in terms of the nature of the motor task used for rehabilitation training. In our previous works implicating relationships between agency and performance [8, 12, 13], we examined motor tasks that were primarily force-driven (e.g., grasp) or motion-driven (e.g., reach). Complex

functional movements typically have distinct motion and force-control components. For example, the reach-and-grasp task is generally viewed as a singular action. Still, it includes force (grasp) and motion (reach) sub-tasks that are controlled by distinct neurological modules [16, 17] that function independently [16, 18].

This study examines neural responses characterized by EEG data across various feedback modes for motor tasks, primarily force- or motion-control. The force-control task involved precision pinch (index finger and thumb) to apply loads upon a force-sensitive apparatus to track a dynamic ramp target; i.e., applied force must linearly increase with time, displayed on a computer monitor. The performance objective for this task was to minimize deviations (errors) between a displayed performance trace, sensitive to applied forces by the participant, and the target ramp. The motion-control task involved participants reaching to drive the motion of a virtual prosthetic hand displayed to the user in a VR environment. This task's performance objective is to minimize the virtual hand's reaching pathlength from its initial position to a highlighted spherical target.

These tasks were appropriate for fundamental analysis of how altered feedback of performance uniquely affects neural responses for force or motion modulation of a motor task. Both tasks were relatively simple in that the objective was readily evident, and effort for each task was perceived as minimal (e.g., low force magnitudes, comfortable reaching volume) for neurotypical participants. Furthermore, each task required either force or motion, i.e., not both, to drive their respective computerized interfaces. We hypothesized that variations in feedback modes known to alter agency would also induce significant changes in neural activity. Changes in the alpha-band and beta-band would respectively suggest the allocation and usage of computational resources for movement [19]. We further hypothesize that these responses are generally characterizable by artificial intelligence (AI) based on neural network prediction of force or motion errors from respective neural responses. Confirming these hypotheses would demonstrate the potency of computerized interfaces to provide feedback during movement training. Specifically, such findings will suggest how computerized interfaces can uniquely impact neural responses associated with motor learning based on the nature of the motor task, i.e., primarily force- or motion-control.

## 22.2 Methods

### *22.2.1 Participants and Equipment*

The same neurotypical participants ($n = 11$) participated in protocols for the force-control and motion-control motor tasks. All participants signed an informed consent form approved by the Stevens Institutional Review Board. Participants wore a 32-channel scalp-surface cap (g.USBamp, g.tec), sampling EEG data at 256 Hz in all trials. All participants were right-hand dominant, using that hand for both motor tasks. **Force-control (grasp) task**: A custom pinch apparatus with two 6-DOF load cells

(*Mini40, ATI Industrial Automation*) was used to record forces from precision pinch (index finger, thumb) sampled at 100 Hz. Participants controlled a dynamic (moving rightward in time) force trace that was displayed against a target trace to be tracked (matched) as performance feedback (*SIMULINK, Mathworks*). The height of the force trace under participant control was computed as the sum of the magnitudes of the 3D force vectors recorded for the index finger and thumb. **Motion-control (reach) task**: Marker-based motion capture using nine infrared cameras sampling at 120 Hz (*Prime17W, Optitrack*) was employed to track retroreflective markers worn on the participant's reaching hand. Independent marker clusters defining coordinate systems were placed on the back of the hand and on the nails of the distal segments of the index finger and thumb. Position and orientation changes of these coordinate systems drove the motion of respective segments of a virtual prosthetic hand represented in a customized environment (*MuJoCo*) displayed to the participant in a VR headset (*HTC Vive*).

### *22.2.2 EEG Signal Processing and Analysis*

Recorded EEG data was bandpass filtered for frequencies of 0.1–60 Hz. Subsequent data processing was done using functions in EEGLAB (MATLAB, Mathworks). The mean EEG power in the alpha (8–12 Hz) and beta (13–30 Hz) frequency bands were computed for each channel, participant, task, and feedback mode. For each trial, channels were examined and assessed for having aberrantly large power outputs indicating they were outliers from other channels. Trial data for these channels were effectively rejected by specifying their outputs to be zero when doing subsequent power analyses. Per trial, the average number of channels (out of 32) rejected were $3.3 \pm 3.1$ and $0.8 \pm 0.7$ for the force-control and motion-control tasks, respectively. Thus, neural response data was largely preserved on a trial-by-trial basis.

### *22.2.3 Experimental Protocols*

**Force-control Task**: We replicated the protocol described in [12] but now additionally recorded EEG, as seen in Fig. 22.1. Participants applied grasp force onto the pinch apparatus to command vertical displacement (i.e., height) of a displayed force trace. The force trace moved rightward with time across the screen at 2.35 inches per second. Participants were instructed to match the height of their trace (green) to a target ramp trace (red) to the best of their ability. The target ramp trace began at a force of zero and rose at a slope of ~ 2 inches/s. This slope was equivalent to a grasp force rate of 1.25 s/N as participants applied force from 0 to 5 N over a 4-s ramp period. Thus, the participant was tasked to steadily (linearly) increase their grasp

force at the same rate Performance was assessed by the average *error* (difference) between the force and target traces.

Participants repeated this force-tracking task in three 20-trial blocks. Performance was displayed within each block under a different (altered) mode of visual feedback, intended to represent possibly perceived distortions in device control as previously done in [12]. The first feedback mode was the "*Default,*" whereby the displayed force trace reflected the force applied without alteration. The second mode was "*Auto,*" whereby the force trace progressively (automatically) adhered to the target trace with time. This adherence was based on a weighted average between the participant's force output and the ramp, whereby the weighting toward the ramp increased (linearly) with time. Finally, the third mode was "*Noise,*" whereby a random low-level force (< 0.5 N) was superimposed to the displayed force trace based on the participant's force output.

**Motion-control Task:** We replicated the protocol described in [13] while additionally recording EEG, as seen in Fig. 22.2. Groups of three non-collinear markers, serving as clusters for defining local 3D coordinate systems, were placed on the reaching hand. As mentioned, a cluster was placed on the dorsal (back) of the hand and on each nail of the index finger and thumb. Position and orientation for these segments were tracked by the motion capture system to animate these respective parts of the virtual prosthetic hand observed through the VR headset. Cluster position and orientation data were streamed in real time (MATLAB, Mathworks), and inverse kinematic solutions (MuJoCo) were used to mitigate any discrepancies in kinematic constraints across the animated independent hand segments.

This task's procedures (i.e., number of trials, feedback modes, etc.) largely mirror the force-control tasks. Points unique to the motion protocol are as follows: (1) in each trial, participants would reach toward a highlighted spherical target, and performance was positively assessed as minimizing the pathlength of the reaching motion; (2) all 6-DOFs were controllable by the participant, and represented to the participant via hand motions; (3) there was no explicitly displayed target path to follow (e.g., straight

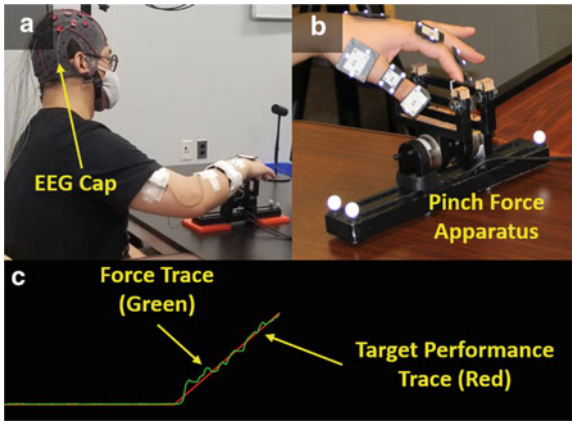



**Fig. 22.1** **a** Experimental setup for the force-control task, **b** device recording pinch forces, **c** participant view of performance feedback

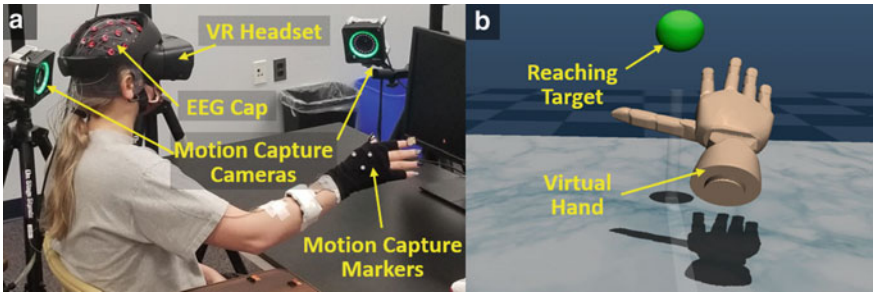



**Fig. 22.2** **a** Experimental setup for the **motion-control task**, **b** participant view in VR

line to track); (4) in the "*Auto*" mode, the virtual hand progressively adhered to a straight line (i.e., "optimal path") from the hand's initial starting position to the target sphere; (5) in the "*Noise*" mode, a random position displacement was applied to the virtual hand position in any direction. Typically, these displacements were < 1 cm, but the magnitude of displacement was proportional to hand velocity (1 cm displacement per 10 cm/s hand velocity); (6) the *error* in this task was computed as the average distance between the hand position (defined by cluster on back of hand) and the optimal path; the trial ceases upon virtual hand contact with the target.

### *22.2.4 Statistical Analyses*

The mean EEG power (alpha, beta) value per trial was tabulated across participants and reported per feedback mode for each motor task. A 1-way repeated-measures ANOVA with post hoc analysis was applied to determine significant differences in neural responses between feedback modes within each task. The ANOVA was applied independently for each motor task to determine whether simple effects in neural responses were discernible from altered feedback. As such, we could readily verify the dependence of neural responses on the task's fundamental type (i.e., force-control or motion-control). A two-sample t-test was then applied to determine whether neural responses generally varied between motor task types.

### *22.2.5 Creating Neural Network to Predict Motor Errors*

As a preliminary exploration into the feasibility of developing AI applications that could adapt computerized interfaces for motor training based on neural responses, we evaluated neural networks predicting motor errors based on EEG inputs. First, we assessed the EEG and error data for every "active" time point. The "active" time points are when the force ramp rises or the virtual hand moves. Thus, for each time point,

we have 32 inputs of EEG against a single output of error. In pooling data across all participants, modes, and trials, we created data sets of approximately 500,000 (500 k) points for each motor task. We created a 2-layer feedforward network with 33 hidden neurons, using a rule-of-thumb whereby the hidden layer size is based on summing the number of inputs and outputs. We trained the network with a random sampling of 100, 200, or 300 k points for each network evaluation with each task. Thus, we trained a total of six neural networks. Each network was trained using backpropagation (Levenberg–Marquardt algorithm) with toolkits in MATLAB (Mathworks). Each randomly sampled data set was divided as follows before training: 70% training, 15% testing, and 15% validation. A maximum of 1000 iterations was allotted for training, with default tolerance on validation used for the early-stopping criterion. By demonstrating basic efficacy in predicting motor errors from EEG activity with a simple feedforward network, there would be high likelihood of success to employ AI predictors with real-time control systems when employing more sophisticated network structures and learning algorithms (e.g., recurrent inputs, deep learning).

## 22.3 Results

There were significant ($p < 0.001$) differences in alpha-band power between feedback modes for the force-control task but not the motion-control task (Fig. 22.3). Significant post hoc differences were observed when pairing the default and noise modes and when pairing the auto and noise modes. The noise mode exhibited the highest alpha power overall for the force task.

Similarly, there were significant ($p < 0.05$) differences in beta-band power between feedback modes for the force-driven task but not the motion-control task (Fig. 22.4). However, no significant post hoc differences were observed. Again, the noise mode exhibited the highest beta power overall for the force task.

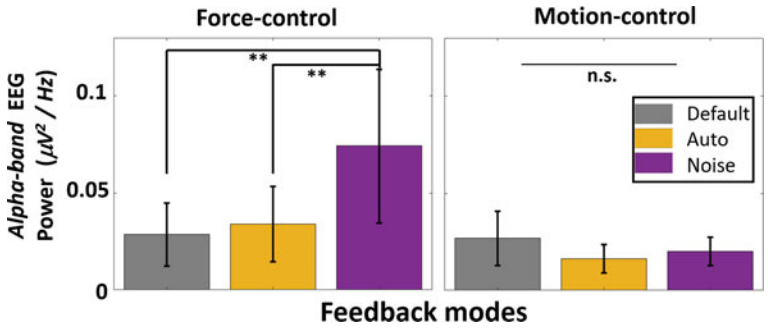



**Fig. 22.3** Neural response (EEG alpha-band power) across feedback modes in the force-control and motion-control tasks. ** $p < 0.001$

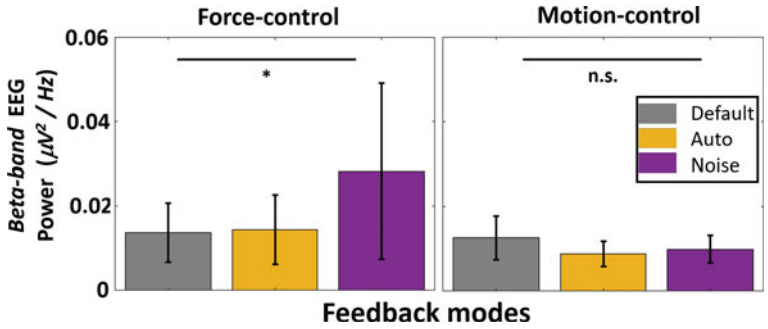



**Fig. 22.4** Neural response (EEG beta-band power) across feedback modes in the force-control and motion-control tasks. * $p < 0.05$

When pooling data across all participants and modes within each motor task, a two-sample t-test demonstrated a significant difference in alpha-band activity and beta-band activity between the force-control and motion-control tasks (Fig. 22.5). The force-control task displayed greater overall neural response than the motion-control task within both the alpha-band and beta-band.

The overall (across the entire brain) EEG activation plots are shown for the alpha-band and beta-band in Fig. 22.6. As supported by the previous results, there are generally evident increases in brain activations, both regionally and overall, with force-control over motion-control tasks and with the noise mode over other modes. As such, the greatest activation overall is observed when coupling the noise mode with the force task. In this case, there also appears to be increased bursting in regions associated with sensorimotor processing. When examining regional changes in brain activity due to alterations in feedback modes, we observed significant differences in alpha-band and beta-band activity for Brodmann area 6 ("ba06"), which has a central role in movement planning. Related to the motor tasks in this study, this

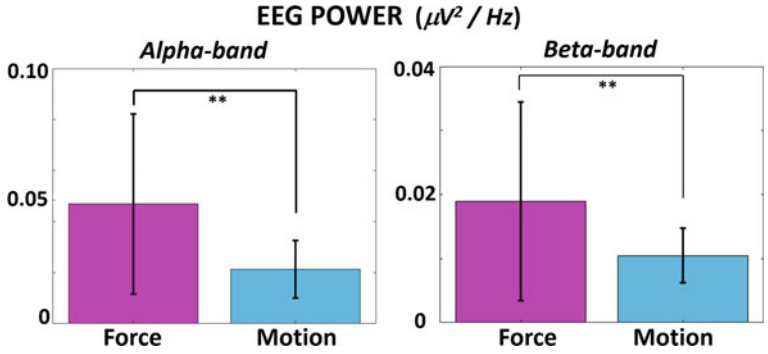



**Fig. 22.5** Neural response for EEG alpha-band and beta-band power overall (across all modes) in force-control versus motion-control tasks. *$p < 0.001$

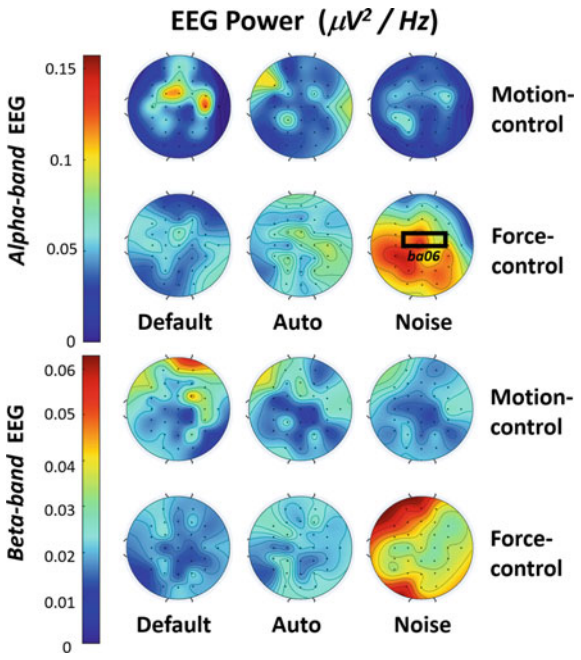



**Fig. 22.6** EEG power shown as brain activation maps for alpha- (top) and beta- (bottom) frequency bands across motor tasks (force-control, motion-control) and feedback modes (default, auto, noise). Brodmann area 6 denoted as "ba06"

area also includes several subareas associated with visual guidance of reaching, eye movements, or control of grasping [20].

When re-examining alpha-band and beta-band activation patterns within ba06, we again observe that only the force-control task distinguishes significant differences between feedback nodes (Fig. 22.7). Furthermore, the Noise mode demonstrated the highest alpha-band and beta-band power. However, these activation patterns are more persistent, as indicated by the relatively smaller standard deviations and the noise mode being significantly greater in pairwise comparison against the other two modes for both the alpha-band and beta-band.

Given the clear variation in alpha-band neural responses with changes in feedback modes in the force task, we further examined this case for an existing relationship between performance and alpha power as a function of altered feedback. First, we plotted the individual participant mean values for performance (error) versus neural response (alpha activity) for the modes in which feedback was altered (i.e., noise, auto) as a "change" from baseline with no feedback alteration (i.e., default). A linear regression indicated a significant correlation between performance and neural response when represented as a function of altered visual feedback. The performance errors increased with higher alpha power for the force-control task (Fig. 22.8).

Each of the previously described six neural networks (i.e., training across a subset of 100, 200, or 300 k data points for each task—motion and force) was trained once. With each training, weights and biases were randomly initialized; thus, there was no assurance of the number of iterations to convergence. All six networks converged,

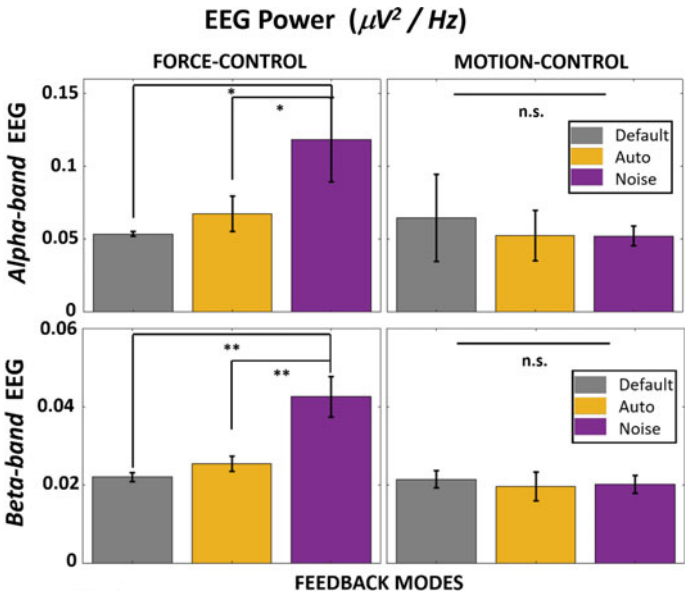



**Fig. 22.7** Neural response (EEG beta-band power) across feedback modes in the force-control and motion-control tasks. * $p < 0.05$, ** $p < 0.001$

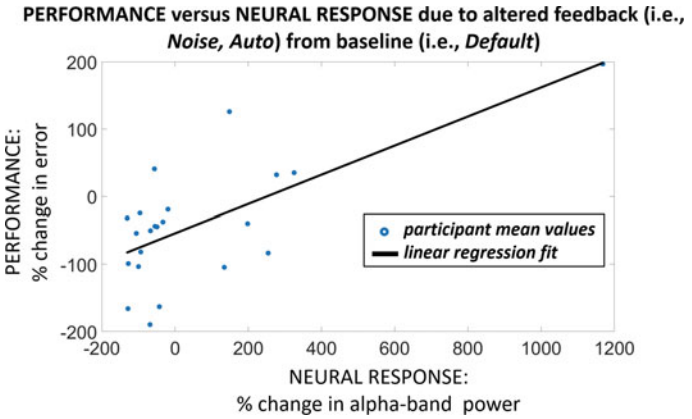



**Fig. 22.8** Linear regression for performance versus neural response as a function of altered visual feedback. Slope = 0.22 ($p = 0004$), $y$-int = −54.23

based on the early-stopping criterion on the validation data set, within 200–400 iterations. Prediction accuracy was not significantly different across the three data sets tested within each motor task type. However, prediction accuracy was excellent for all networks. The prediction error for motor errors in the force task was on the order of $1 \times 10^{-4}$ N. The prediction error for motor errors in the motion task was

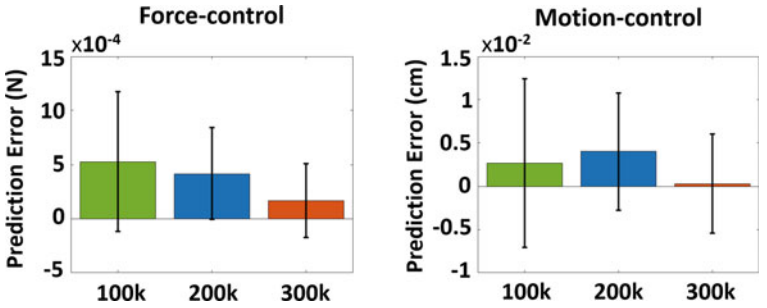



**Fig. 22.9** Average prediction error of motor error by neural networks trained on random sampling of 100 k, 200 k, or 300 k points for either force or motion tasks

on the order of $1 \times 10^{-2}$ cm. When normalized by their maximum respective target force (5 N) or motion excursion (67 cm), the accuracy against these maximum values is < 0.01% and < 0.02% for force and motion tasks (Fig. 22.9).

## 22.4 Discussion

This study demonstrated that altering the presentation of performance feedback can elicit significant changes in neural response, represented by alpha- and beta-band power, in force-control (grasp). However, neural response results were inconclusive regarding motion-control (reach) tasks. This may be explained, in part, by a limitation in this study in which we re-created the protocols from our previous works [12, 13], now with the addition of EEG measures. In those studies, we initially examined and reported the performance and agency results with alterations in visual feedback that represented potential distortions perceived in the control of an assistive device. However, in directly comparing the two protocols, although each is predicated on primarily driving the computerized interface with either force-control or motion-control, they differ in complexity and dimensionality. Although 3D forces were recorded at each digit (thumb, index finger), the force magnitude was computed for each and then summed to present the participant with a single degree-of-freedom of control, i.e., total grasp force. The motion protocol allowed the participant to control 3D position and 3D orientation of the virtual hand, although they had a singular goal to move the hand toward the highlighted target. Despite the simple task objective, the motion task provided an additional stream of feedback, which may have lowered the sensitivity to alterations. Thus, the complexity of both the task and feedback may also play a role in how alterations in feedback can modulate neural responses, but these considerations are beyond the scope of the current study.

Despite this limitation, the results observed for the force-control task still represent a crucial first-level observation in validating the use of virtual reality in a more systematic way to modulate neural responses for improved motor learning. While alpha and

beta activities fundamentally indicate distinct brain functions associated with mindfulness and arousal, respectively, both have established roles in motor control. Alpha activity is central to motor preparation [21], while beta activity supports sustained motor engagement [22]. While alpha and beta waves have been targeted with neuromodulation techniques involving stimulation to facilitate improved motor learning [23], virtual reality offers a more organic way, i.e., via representations of performance, to positively modulate neural responses for better motor function. Furthermore, our results suggest VR-based modulation still produces localized changes in activation (e.g., Brodmann area 6), whereby neuroplasticity would desirably occur at brain regions primarily associated with sensorimotor activity. Thus, basic alterations in feedback from computerized interfaces generating notable alpha and beta activation changes suggest the potential to personalize VR motor rehabilitation more effectively with targeted results using intelligent design of select interface parameters.

Furthermore, inducing larger neural response with modes altering feedback from the default mode, most notably the noise mode, may affect alpha activity through the cognitive mechanism of agency. Agency, or perception of control, is crucial not only for normative movement function [24] but is readily modulated within computerized environments such as virtual reality [25]. In our previous works [12, 13], the default mode, which presumably represents user actions most accurately, produced the highest agency and performance outcomes for both force and motion tasks. This finding was expected since all non-default modes should be perceived as systematic distortions of intended actions by neurotypical participants. Such disruptions between intended action and observed outcome directly oppose the binding of voluntary actions to expected outcomes in participants, which can indicate a sense of agency [6]. Therefore, our finding of increased alpha-band activity in a low agency mode (i.e., noise) can be seen as a logical extension of previous work showing an inverse relationship between agency and alpha-band activity [26]. Thus, computerized environments are a potentially powerful tool to systematically modulate neural responses for better motor function via cognitive mechanisms such as agency.

However, the question becomes and is inducing higher or lower alpha-band power during computerized motor rehabilitation desirable, and if so, when? Increases in alpha-band power have been correlated with differential learning strategies and can suggest early consolidation of motor learning features from a pre-learning state [27]. Consequently, systematic distortions such as noise may accelerate learning with higher alpha activity for force-driven tasks despite reducing agency [14]. Our finding of a negative correlation between performance and alpha-band power follows as well, as lower alpha-band power is related to higher agency, and higher agency is related to higher performance. Thus, the ideal training regime may seek to induce higher alpha activity during or after training for earlier sessions in a longitudinal protocol. However, we may expect to see reductions in alpha activity with follow-up sessions such that a higher sense of agency is being adopted, which in fact supports greater motor function. In any case, pursuing such advanced neural response profiles to maximize learning trajectories may be supportable through intelligent alterations in computerized feedback.

Our final analysis in this study involved observing how well a neural network model could predict force or motion errors in these protocols based on EEG inputs across a spectrum of data entailing alterations in feedback. Both force and motion errors could be predicted with high fidelity. Despite this study's limitation regarding the difference in feedback dimensionality for the motion and force tasks, error prediction remained high in both cases. This finding suggests the universal potential for neural network models to be developed for real-time applications. These applications would entail interface parameters, such as feedback alterations, being varied within individual sessions to modulate neural responses to induce smaller performance errors during training. Such considerations should be balanced against broader considerations, such as following neural response profiles across multiple sessions that support long-term motor learning, as mentioned previously. In any case, our study should inspire further investigation into how computerized interfaces, like virtual reality, for motor rehabilitation can be deployed more intelligently such that it can leverage cognitive mechanisms for neural engagement that accelerates functional outcomes.

## References

1. Burke, J.W., McNeill, M.D.J., Charles, D.K., Morrow, P.J., Crosbie, J.H., McDonough, S.M.: Optimising engagement for stroke rehabilitation using serious games. Vis. Comput. **25**(12), 1085–1099 (2009). https://doi.org/10.1007/s00371-009-0387-4
2. Maclean, N., Pound, P.: A critical review of the concept of patient motivation in the literature on physical rehabilitation. Soc. Sci. Med. **1982**(50), 495–506 (2000). https://doi.org/10.1016/S0277-9536(99)00334-2
3. Howard, M.C.: A meta-analysis and systematic literature review of virtual reality rehabilitation programs. Comput. Hum. Behav. **70**, 317–327 (2017)
4. Lim, D.Y., Hwang, D.M., Cho, K.H., Moon, C.W., Ahn, S.Y.: A fully immersive virtual reality method for upper limb rehabilitation in spinal cord injury. Ann. Rehabil. Med. **44**(4), 4 (2020). https://doi.org/10.5535/arm.19181
5. Nataraj, R., Sanford, S., Liu, M., Harel, N.Y.: Hand dominance in the performance and perceptions of virtual reach control. Acta Physiol. (Oxf) **223**, 103494 (2022). https://doi.org/10.1016/j.actpsy.2022.103494
6. Moore, J.W., Obhi, S.S.: Intentional binding and the sense of agency: a review. Conscious. Cogn. **21**(1), 546–561 (2012). https://doi.org/10.1016/j.concog.2011.12.002
7. Liu, M., Wilder, S., Sanford, S., Saleh, S., Harel, N.Y., Nataraj, R.: Training with agency-inspired feedback from an instrumented glove to improve functional grasp performance. Sensors **21**(4), 1173 (2021)
8. Nataraj, R., Hollinger, D., Liu, M., Shah, A.: Disproportionate positive feedback facilitates sense of agency and performance for a reaching movement task with a virtual hand. PLoS ONE **15**(5), e0233175 (2020)
9. Sanford, S., Liu, M., Selvaggi, T., Nataraj, R.: Effects of visual feedback complexity on the performance of a movement task for rehabilitation. J. Motor Behav. **53**(2), 243–257 (2020). https://doi.org/10.1080/00222895.2020.1770670
10. Sanford, S., Collins, B., Liu, M., Dewil, S., Nataraj, R.: Investigating features in augmented visual feedback for virtual reality rehabilitation of upper-extremity function through isometric muscle control. Front. Virtual Real. **3**, 943693 (2022). https://doi.org/10.3389/frvir.2022.943693

11. Sanford, S., Liu, M., Nataraj, R.: Concurrent continuous versus bandwidth visual feedback with varying body representation for the 2-legged squat exercise. J. Sport Rehabil. **30**(5), 794–803 (2021). https://doi.org/10.1123/jsr.2020-0234
12. Nataraj, R., Sanford, S.: Control modification of grasp force covaries agency and performance on rigid and compliant surfaces. Front. Bioeng. Biotechnol. **8**, 1544 (2021)
13. Nataraj, R., Sanford, S., Shah, A., Liu, M.: Agency and performance of reach-to-grasp with modified control of a virtual hand: implications for rehabilitation. Front. Hum. Neurosci. **14**, 126 (2020)
14. Bu-Omer, H.M., Gofuku, A., Sato, K., Miyakoshi, M.: Parieto-occipital alpha and low-beta EEG power reflect sense of agency. Brain Sci. **11**(6), 6 (2021). https://doi.org/10.3390/brainsci11060743
15. Rasheed, M.A., et al.: Use of artificial intelligence on electroencephalogram (EEG) waveforms to predict failure in early school grades in children from a rural cohort in Pakistan. PLoS ONE **16**(2), e0246236 (2021). https://doi.org/10.1371/journal.pone.0246236
16. Casadio, M., Pressman, A., Mussa-Ivaldi, F.A.: Learning to push and learning to move: the adaptive control of contact forces. Front. Comput. Neurosci. (2022). https://doi.org/10.3389/fncom.2015.00118
17. Chib, V.S., Krutky, M.A., Lynch, K.M., Mussa-Ivaldi, F.A.: The separate neural control of hand movements and contact forces. J. Neurosci. **29**(12), 3939–3947 (2009). https://doi.org/10.1523/JNEUROSCI.5856-08.2009
18. Shadmehr, R., Mussa-Ivaldi, F.A.: Adaptive representation of dynamics during learning of a motor task. J. Neurosci. **14**(5), 3208–3224 (1994). https://doi.org/10.1523/JNEUROSCI.14-05-03208.1994
19. Brinkman, L., Stolk, A., Dijkerman, H.C., de Lange, F.P., Toni, I.: Distinct roles for alpha- and beta-band oscillations during mental simulation of goal-directed actions. J. Neurosci. **34**(44), 14783–14792 (2014). https://doi.org/10.1523/JNEUROSCI.2039-14.2014
20. Stein, J.: Sensorimotor Control. In: Reference Module in Neuroscience and Biobehavioral Psychology. Elsevier, New York (2017). https://doi.org/10.1016/B978-0-12-809324-5.06855-3
21. Deiber, M.-P., Sallard, E., Ludwig, C., Ghezzi, C., Barral, J., Ibañez, V.: EEG alpha activity reflects motor preparation rather than the mode of action selection. Front. Integr. Neurosci. **6**, 59 (2012). https://doi.org/10.3389/fnint.2012.00059
22. Kristeva-Feige, R., Fritsch, C., Timmer, J., Lücking, C.-H.: Effects of attention and precision of exerted force on beta range EEG-EMG synchronization during a maintained motor contraction task. Clin. Neurophysiol. **113**(1), 124–131 (2002). https://doi.org/10.1016/s1388-2457(01)00722-2
23. Pollok, B., Boysen, A.-C., Krause, V.: The effect of transcranial alternating current stimulation (tACS) at alpha and beta frequency on motor learning. Behav. Brain Res. **293**, 234–240 (2015). https://doi.org/10.1016/j.bbr.2015.07.049
24. Moore, J.W., Fletcher, P.C.: Sense of agency in health and disease: a review of cue integration approaches. Conscious. Cogn. **21**(1), 59–68 (2012). https://doi.org/10.1016/j.concog.2011.08.010
25. Kong, G., He, K., Wei, K.: Sensorimotor experience in virtual reality enhances sense of agency associated with an avatar. Conscious. Cogn. **52**, 115–124 (2017). https://doi.org/10.1016/j.concog.2017.04.018
26. Kang, S.Y., et al.: Brain networks responsible for sense of agency: an EEG study. PLoS ONE **10**(8), e0135261 (2015). https://doi.org/10.1371/journal.pone.0135261
27. Henz, D., Schöllhorn, W.I.: Differential training facilitates early consolidation in motor learning. Front. Behav. Neurosci. **10**, 199 (2016). https://doi.org/10.3389/fnbeh.2016.00199

# Chapter 23
# Towards Enhancing Extended Reality for Healthcare Applications with Machine Learning


**Pranav Parekh and Richard O. Oyeleke**

**Abstract** Extended Reality is expanding in healthcare with regard to the quality of vision the operator experiences. On the other hand, machine learning has provided numerous capabilities to enhance user interaction. This paper focuses on using the visualizations of extended reality (XR) methods alongside functionalities provided by intelligent systems. We propose three hypothetical methods for integrating the two fields to provide healthcare solutions. The three use cases of treatment visualization, therapeutic strategies and decision-making are proposed to enhance healthcare. The treatment visualization case uses image segmentation and virtual reality (VR). We segment the medical image that the medical worker visualizes through the VR headset. We propose an adaptive mixed reality system that can change according to the user's mood for appropriate therapy. Neural models recognize the user's emotions and adapt the system accordingly. An augmented reality chatbot has been proposed to influence timely decision-making during medical situations. Each case study uses different tools from the other and focuses on a specific healthcare-related solution.

## 23.1 Introduction

Significant technological advances in these futuristic times have allowed us the visualization of virtual worlds or the enhancements of real-world objects and scenarios through multiple sensory nodes [1]. Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR), collectively known as Extended Reality (XR) have altered the way users perceive entertainment and have changed the gaming industry significantly. More importantly, they are used extensively in healthcare and education since visualization enhancement can help improve the efficiency of both sectors. We establish how AR, VR, and MR differ as follows [2]:

P. Parekh (✉) · R. O. Oyeleke
Stevens Institute of Technology, Hoboken, NJ 07030, USA
e-mail: pparekh5@stevens.edu

R. O. Oyeleke
e-mail: ooyeleke@stevens.edu



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_23

- VR: Overwrites the physical environment so the subject is placed entirely within the virtual world. Requires a head-mounted display as its interface.
- AR: Provides a view from which digital objects can be visualized within the physical world. The interface is usually a handheld device.
- MR: Digital objects are added to the physical world, however, the key difference is that interactions are enabled with such imaginary objects. Requires a head-mounted display as its interface.

Advancements in machine learning and artificial intelligence have produced significantly countless users in the field of healthcare [3]. Machine learning can be integrated with XR to develop therapeutic strategies, effective decision-making, medical visualizations, and physiological and exercise-based training. Rogers[4] extracted features of the operation performed by a surgical student and trained a machine learning model over the extracted features to give us insight into how he has performed. A machine learning model is used as a secondary entity integrated with a virtual reality system for several scenarios. For treatment visualization, we primarily use computer vision on a medical diagram which is then visualized through a headset [5]. On the other hand, we have neural rendering models comprising both computer graphics and machine learning algorithms as primary entities [6]. Since we have introduced ML and XR, we would like to establish the primary objectives of the manuscript as follows:

- To focus on three broad areas of healthcare: visualization during surgery, efficient therapy and quick decision-making.
- To propose using XR to provide visual aids in all three cases.
- To provide functionalities to the system through machine learning: Image segmentation, mood-adaptive systems, and personal chatbots.
- To propose three end-to-end hypothetical frameworks integrating the two fields.

3D imaging systems such as X-rays, magnet-resource (MR), have revolutionized how medical workers visualize body fragments [7]. Recent applications in the field include immersive real-time simulations of surgery and medical operations. Through the first case study, we propose a hypothetical framework that employs computer vision methods in the simulations. Strategy in therapy is a concept that many physiologists and psychologists follow for maximizing improvements in the outcome of their treatment. Sound results take repeated exercise and conversation [8]. The second case study proposes a theoretical technological solution for achieving the same. Making quick decisions can be essential when dealing with a medical problem. This is why our third case study, proposes a AR chatbot that guides us as soon as we communicate with it.

## 23.2 Related Work

In this section, we introduce three important use cases of extended reality and machine learning in healthcare; treatment visualization, therapeutic strategies, and effective decision-making. We establish the role that each field has to play for the adequate execution of the application. Furthermore, we discuss the surrounding technologies used in our case studies.

### *23.2.1 Extended Reality for the Case Study Applications*

Precise visualization is a crucial element when dealing with body structures. Extended reality helps enhance how we view the human anatomy by improving the quality of vision and precision. Augmented reality does so by superimposing necessary markers over impacted areas, illustrating which portion requires remedy [9]. The principal advantage of such a scenario is that the physician can conduct the procedure with a higher degree of accuracy. AR displays also present the advantage of not obscuring the medic's vision [10]. An example is the portrayal of a three-dimensional scar for an animal ablation procedure generated by Jang et al. [11]. Cardiologists have used 2D AR displayed on Google Glass to assist in the restoration of blood flow in a patient's coronary artery [12].

During clinical surgeries, surgeons require multiple visualization instruments such as cardiac monitors, endoscopes, laparoscopes. Highly delicate surgeries may even require microscopes [13]. However, the microscope has a small operable interface and can be bulky. Smart glasses that have adjustable focal lengths have recently been shown to be helpful [14]. VR and MR are used to impart surgical education to budding surgeons. A virtual world is created that resembles an actual operation theater. The student visualizes how a surgery is performed through which he can get accustomed to the intricacies and pressure of an operation theater. Examples of such simulations include: orthopedics [15], spine surgery [16], cataract surgery [17], plastic surgery [18], and many more.

Immersive simulations are also a therapeutic agency to relieve patients from psychological or physical disorders. The best way to alleviate pain is through distractions. When immersed in a computer-generated world, their mind is diverted from their discomforts. Carlin et al. [19] created a computer-generated immersive environment called Spider World, which treats spider phobia. When immersed in Spider World, the pain decreased in two teenagers suffering from excruciating burns was noted. It was observed that Spider World managed to alleviate pain to a greater degree than a Nintendo game like Mario Kart. It was concluded that VR reduces the amount of pain-related brain activity rather than just changing how patients interpret incoming pain signals [20].

Virtual avatars represent the embodiment of self in the physical or virtual space. Such avatars are an adequate representation of the characteristics and mannerisms of

the user. The study by [21] uses virtual agents to reduce anxiety and facilitate effective therapy through an interface. It was observed that the setup positively affected the user's perceived enjoyment and sense of calm. The study by [22] performs effective avatarization by introducing a taxonomy for the virtual embodiment experiences. Complete virtual embodiment is highly prominent in VR systems, but this study has effectively been extended to AR applications, which can also enhance our standpoint on such applications.

Augmented reality allows for the presence of virtual objects in space. These virtual objects are called markers, and these can be viewed through a handheld device [23]. These markers are not interactive, as seen in mixed reality, but are still used widely in numerous applications. Cardiopulmonary Resuscitation (CPR) is a lifesaving technique used when a person's heartbeat or breathing has stopped. The study by Boonbrahm et al. [24] uses AR markers for CPR training among children. The advantage provided by AR is that it is more specific and systematic than traditional training. Each marker is programmed to be interactive to denote the trainee's performance. Marker-based AR is also prominently used for posture correction. Track markers are visualized in a smartphone, representing what the correct posture should be [25]. AR can also be location-based or markerless; however, the method suggested in this study required a review of marker-based AR.

### *23.2.2 What Role Can Machine Learning Play?*

For ease of communication, we use machine learning as a collective term for computer vision, deep learning, and artificial intelligence algorithms. Machine learning is omnipresent in the present climate since the techniques of prediction and classification are used as insightful tools in every domain [26]. Accurate machine learning methodologies that process images and text as input data are critical for healthcare problems. Extended reality enables us to visualize effectively; however, we rely on machine learning techniques to provide for the functionalities within the scenarios.

Medical ultrasound dates back to the 1950s, with the first experiment done using radar and sonar equipment by Wild [27] and Edler [28]. Henceforth, it has become an integral tool for medical imaging systems. It is used in diagnosing and studying internal body structures and is therefore essential in disease detection. It has been observed that the images obtained are contaminated with speckle noise, thereby affecting diagnosis [29]. Traditional methods include filters, while recent studies have utilized anomaly detection concepts to improve the quality of output images. The paper by Chen et al. [30] presented robust PCA for the reconstruction of the MRI, which had better results in removing noise than the traditional methods. Noisy data in images can hinder how we view body structures through a headset. Using machine learning methods to provide spotless images for visualization through XR would be ingenious.

The use of XR for therapy could be enhanced further if the system could perceive the emotions that the user is undergoing within the headset. Emotion recognition

has been done extensively using speech, facial expressions, or both [31]. Our facial expressions are the sum of our facial landmarks and alignments. The neural model extracts features from the same and infers the emotion as an output label. Tautkute et al. [32] uses the Deep Alignment Network architecture(DAN and EmotionalDAN) while [33] uses a neurofuzzy network. The study by Tarnowski et al. [34] uses Multilayer Perceptron (MLP) to predict six labels: sadness, joy, neutral, fear, surprise, anger, and disgust. A CNN architecture extended that further to enable real-time emotion recognition, achieving an accuracy of 96.43% [35].

Chatbots try to mimic conversations that take place between two individuals [36]. Chatbots that do not use natural language processing (NLP) can answer general healthcare questions. These questions have databases to store their corresponding answers. Examples of such chatbots include Casper for users that have insomnia [37], Endurance for people suffering from dementia [38], and MedWhat for general healthcare-related FAQs [39], among others. Such chatbots are monotonous and need to establish smart communication with the user. Chatbots that use NLP have the ability to imitate the tone and expressions that are often precisely spoken by an individual. They also provide more specific answers to the end user's questions.

Intelligent chatbots use a Natural Language Understanding (NLU) engine that performs four primary tasks: word segmentation, parts-of-speech (POS) tagging, dependency parsing, and pattern recognition, as depicted in Fig. 23.1. Segmentation refers to breaking the text into smaller, more meaningful units known as tokens. Nagabhushan and Javed [40] used character spaces while [41] performed word segmentation using Natural Language Toolkit (NLTK). POS tagging refers to assigning grammatical annotations to a particular token; for example, nouns, adjectives, prepositions, and many more. The NLTK toolkit uses an inbuilt tagger; however, these are not the most accurate [42]. Dependency parsing establishes a relationship between words and is extremely important for correct results in a chatbot. Traditional methods include using dependency and parsing trees [43]. Pattern recognition is responsible for matching the intent between two sentences. In this study, we shall not be relying on traditional methods. Instead, we discuss neural language models for performing natural language processing tasks.

### *23.2.3 Connecting Technologies*

It would be unjust to define extended reality in terms of its devices since these headsets and software may fall out of favor in a year or two [44]. Continuous improvements in headsets and software are taking place. Game engines like Unity and Unreal add new assets and require regular updates to improve our capabilities in creating better visualizations. Game engines are beneficial in speedy visualization but they have their limitations. They are usually generic and do not perform for specific problems. A few examples where game engines fail are generating photo-realistic representations of images like X-rays or CT scans, intricate depictions of internal body structures, and producing segmented images. Headsets allow us to import external images not

created by a game engine. Therefore, thinking of XR beyond game engines and headsets is essential. It relies on foundational computer graphics and we will discuss these methods since they are relevant to the case studies proposed in the following section.

In medicine, the relative presence and volume of anatomical structures are indispensable. CT scans provide a 3D view of the internal body structures. Visualization of these medical images in a virtual environment offers six degrees of freedom for a deeper understanding. Kratz et al. [45] have used a virtual environment called Studierstube [46] and performed volumetric rendering through ray-casting. The ray-casting algorithm renders an image by casting rays from the viewpoint through each pixel. The ray is resampled and all successful sample contributions are composited [47]. The sampling rate decides the interval between each sample. Another method of volumetric rendering is through the visualization of image stacks iteratively. A transfer function is used as the conversion factor for generating the 3D image from the stack. The transfer function assigns optical properties to each voxel of the visualized scene [48].

Machine learning research typically focuses on modeling, optimization, and testing, but deployment in a public setting requires more. Deployment issues get little attention, but effective deployment is necessary to have a real-world impact, effective deployment is necessary [49]. Machine learning research is often done on static data, but deployment must focus on real-time input. Netflix had offered one million$ for a model that reduces the error rate of its recommendation system by ten %. Although more accurate, the model selected was too complex for production. Therefore, apart from the models used and the virtual simulations discussed, we examine a deployment pipeline for connecting models to mobile applications.

The model is trained on the dataset. Once the training is done, the parameters and hyperparameters are tuned, and the model is ready for deployment. The learning rate, regularization parameters, presence of normalization, and amount of training data are important hyperparameters to consider while developing an ML model [50]. If these judgments are not made carefully, there is a high chance that the model could overfit or underfit. After tuning, the model is pickled. Pickle is a useful Python tool that saves the ML model from avoiding lengthy retraining. We can now share, commit, and re-load the pre-trained machine learning models [51]. Once we have the saved model, we use the Flask API to convert the file into JSON. The JSON file is connected to our mobile or web application. Hence, the model is deployed and ready for the input real-time data stream. The entire deployment pipeline is depicted in Fig. 23.3.

## 23.3 Proposed Case Studies

The following section discusses three case studies on how the machine learning methods discussed above are integrated with extended reality to enhance and benefit user experience. The three case studies intend to solve a problem related to health-

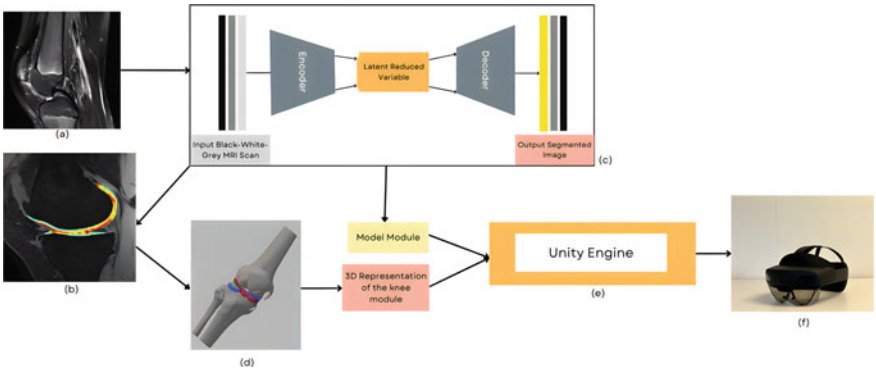



**Fig. 23.1** Framework for Case Study 1. **a** Unsegmented initial image, **b** segmented images received as output from the model, **c** encoder-decoder model for image segmentation, **d** 3D model created from output of the model, **e** 3D model rendered using unity, **f** HoloLens 2 for visualization of the hologram

care. We rely on XR for adequate visualization and ML to provide the underlying functionalities.

### *23.3.1 Case Study 1: Treatment Visualization*

As discussed in the section above, extended reality can visualize body structures intricately. Machine learning, through image segmentation, can accurately mark the affected regions. The case study combines these abilities to help the physician visualize and diagnose the treated area. It intends to enhance visualization of the affected portions of the human anatomy to give the medical specialist a clearer picture. X-rays and MRIs are widely used 2D representations. Mixed reality holograms go one step further by providing a 3-dimensional view and enabling interactions with the body structures. The basic outline is as follows:

- Step 1: Use Image segmentation for identifying the affected portion.
- Step 2: Volumetric rendering to convert the image into a 3D structure.
- Step 3: Use a mixed reality headset to visualize the area in the 3D space.

Numerous studies have focused on one step out of the three rather than considering the entire scenario. However, the recent manuscript by Oulefki et al. [52] merged the idea of segmentation with XR. They used geometrical methods to perform segmentation and discuss the concept of visualization through a virtual reality headset. They perform the image segmentation using computer vision methods but do not extend to rendering the segmented image over an MR headset. More insight into how the segmented image is to be rendered is required.

Before deployment for practical uses, ensuring satisfactory results from our segmentation procedure is wise. Datasets like DICOM have CT-scan and X-ray images of many body structures. They can be used to tune the hyperparameters of the model. The model is to be pickled and saved to be loaded into the testing pipeline directly. Image segmentation can be implemented using encoder-decoder architectures like Seg-Net and U-Net ML models. They are specialized in the segmentation of X-ray and CT images. Experience in Python and libraries like Numpy, Pandas, OpenCV, TensorFlow, and Keras is essential for implementing the model. A powerful GPU-based computer system is recommended while handling machine learning problems of such high scale and precision.

Volumetric rendering is done using the ray-casting algorithm discussed in the previous section. An NVIDIA graphic card would be required since volumetric rendering requires high computation. A Google Hololens 2 mixed reality headset can be used for the 3D visualization of the segmented image. The hologram is implemented using the mixed reality toolkit (MRTK), which comprises rendering and projection functions.

The testing pipeline takes real-time images as input and gives us the 3D hologram of the segmented image as output. We propose a generalized framework for a test input applicable to any body structure. The neural model that is saved through pickling is a significant component of the pipeline. We have included a basic encoder-decoder architecture within our current framework where the encoder converts the input image into latent vectors while the decoder decodes it into the segmented output image. They require heavy computation but significantly improve accuracy.

Once we have the segmented image, we render this onto our mixed reality display as a hologram. We do so using the ray-casting algorithm and image-based meshing. Image-based meshing creates computer-generated models from flat 3D images like X-rays or CT-scans. Meshing is done for two characteristics of the image: volume and surface mesh generation. After meshing, we apply smoothing over the generated 3D model and imbibe fleshy textures for muscular organs and bony textures for bones and skeletons. We use tools like CAD to create the 3D model on the computer. Once the 3D model is generated, we render it on the mixed reality headset by importing the model file into the Unity engine. Using MRTK, the hologram is generated and viewed through the Hololens 2. The entire framework is discussed in Fig. 23.2.

### *23.3.2 Case Study 2: Therapeutic Strategies*

In the preceding section, we discuss how VR can be used to reduce pain and anxiety through appropriate simulations. Using machine learning methods, we want the system to adapt to the user's moods and emotions. An adaptive system can facilitate more adequate therapy for a particular user. This makes the system more user-to-user oriented rather than generic for all users. The paper by Bălan et al. [53] focused on how therapy modules can be adaptive through machine learning, but they did not extend it to VR simulations.

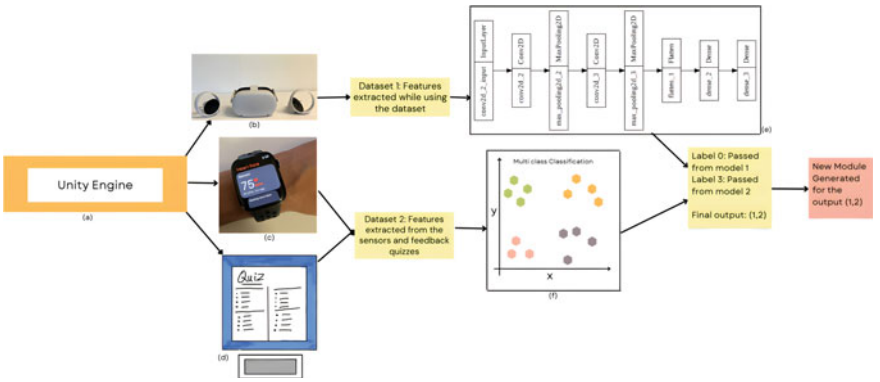



**Fig. 23.2** Framework for Case Study 2. **a** Unity engine for creating healthcare-related scenarios **b** Oculus quest 2 for visualization, **c** calculation of heart rate, **d** quiz taken by the patient, **e** basic CNN for emotion recognition from facial features, **f** multiclass classification with 4 different classes as our output label

The modules are created using the Unity game engine and visualized through the Oculus Rift headset. We rely on a camera, heart rate, breathing sensor, and quiz modules for data collection. Health experts must formulate the quiz modules. Smartwatches are effective sensors for measuring $O_2$ level, heart rate and breathing. A CNN model extracts features to recognize the user's emotions. The model is pickled and saved and then added to the testing pipeline. Since our training dataset is enormous, we would require a powerful GPU computing system. A regression model is used for multiclass classification on the numeric dataset since the number of features is significantly less. Knowledge of Python and its libraries is crucial for implementing neural models. Experience in the Unity game engine is essential since each module is generated through the game engine. Table 23.1 consists of neural models that can be effectively used for the case study.

The user fills up an introductory quiz to initiate the testing pipeline. Based on his answers, the system projects a VR module visualized through the headset. The camera captures images of the user while the sensor records the heartbeat and breathing rate. The facial images acquired while wearing the headset serve as one form of a dataset. The other dataset is obtained by heart rate, breathing sensors, and quiz modules. The quiz module consists of multiple-choice questions that ask the user about their experience and if any improvement has been felt. The average heart rate and oxygen levels are calculated over the period of use. The collective sample dataset is represented in Fig. 5. If each module is of 60 min, the camera is to capture three images: one at the start, one at 30 min, and one just before the module is completed (Fig. 23.3).

Once the two datasets are acquired, we input them parallelly into the pre-trained machine learning models. We have included a CNN within the framework for emotion recognition since examples like LeNet have given high accuracies while solving the problem statement. The second dataset uses the regression model since it is numeric

**Fig. 23.3** The dummy of the dataset described

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| Usage Per | Avg. Hear | Avg. O2 | Q1 | Q2 | Q3 |
| 23 mins | 67 bpm | 96.2 | 2 | 4 | 1 |
| 36 mins | 75 bpm | 95.5 | 2 | 3 | 1 |
| 61 mins | 74.4 bpm | 93.4 | 2 | 2 | 2 |
| 12 mins | 78 bpm | 97.8 | 1 | 4 | 4 |
| 18 mins | 69.5 bpm | 98 | 2 | 4 | 3 |
| 77 mins | 76 bpm | 96 | 3 | 2 | 4 |

and requires lower computation. Logistic regression has been added to the framework since it is practical for multiclass classification. Since the neural models are pre-trained and saved, they are added to the testing pipeline. They are expected to take in the input and generate the output label. The output labels are as follows:

- Emotion Recognition problem: Output labels are 0, 1, 2, 3, 4 for sad, angry, neutral, calm, and happy: Requires neural networks
- Body features problem: Output labels are 0, 1, 2, 3 where each label denotes the level of performance from bad to good: can be done using logistic regression.

The two labels form an output set. A new module with the correct content is uploaded based on the set. For example, if the output from the CNN model is 2 and the output from the regression model problem is 1, our output set will be (2,1). According to (2,1), a new module shall be uploaded onto the headset, making the system adaptive. This loop continues until the patient completes all required modules and has undergone appropriate therapy. Figure 23.4 depicts the framework for the same.

### *23.3.3 Case Study 3: Decision Making*

The literature review section discusses the creation of AR avatars and markers as visual aids within a handheld application. In this case study, we would like to propose an intelligent chatbot application that uses AR visual aids. An introduction and review of intelligent chatbots have been done in the prior section. The use of visual aids along with the chatbot functionality is done to personalize the application according to the user. The chatbot application can impart speedy and convenient healthcare services to a patient in need. The paper by Wu et al. [54] uses augmented reality with traditional chatbots to impart nutritional education to students. We want to expand on this idea by making the chatbots intelligent through NLP and using it for general healthcare questions.

The end goal of the application is its presence on a handheld device like an iPad or a Smartphone. The AR visual aids are generated by Vuforia, which is downloaded from the App Store or Google Play. An NLP unit that performs four primary tasks has been discussed in the section above. The four tasks require a neural language model;

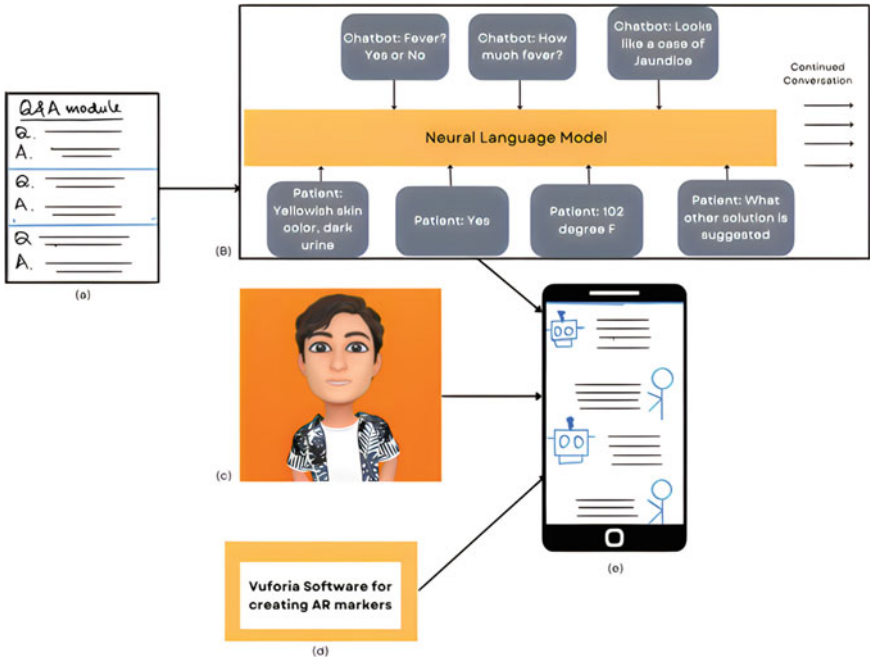



**Fig. 23.4** Framework for Case Study 3. **a** Question and answer modules input as training data, **b** sample conversation through a natural language processing unit, **c** sample augmented reality avatar, **d** Vuforia for creating AR markers, **e** the 3 modules input into the mobile application to create the chatbot

the model used is usually common for all four tasks, i.e., once pickled, it is reused for performing the next task. Many healthcare-related question-and-answer modules are required to train the model accurately. A powerful GPU-based computer system is suggested since the dataset is huge. Flask API is used to deploy the model on the smartphone application. Again, python and its libraries are essential in implementing the model.

The testing pipeline consists of three modules imported into the smartphone application through Flask APIs.

- NLP module: It consists of a pre-trained language model that gives a text output for each particular text input. For example, a person using the application is asked what he/she is suffering from. The input is processed by an NLP model, which predicts the next question connected to the previous answer. The user answers each question while the model predicts the next question. We want the model to establish long-term dependencies. Therefore Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs), or Gated Recurrent Units (GRUs) (with or without attention) would be appropriate choices. Transformer models like BERT can also be remarkably effective here.

- The avatar module: This module comprises patient characteristics and provides a personalized environment to the user. It is a life-like representation of the patient and a doctor-like representation of the bot interacting with the patient. Avatars can be created using augmented reality software such as Vuforia.
- Visual Aids module: Using the AR software, we can enhance the user experience by giving them exercise suggestions and projecting markers to help them perform their exercises. This is convenient since they require a small open space where these visual aids can be projected. For example, someone requiring physiotherapy can discuss the issues they are going through with the chatbot. Based on the different exercise suggestions, the application displays projections on the surrounding area, following which the exercises can be performed.

The three modules together function to make an intelligent AR chatbot application. Figure 6 depicts the framework for the same. Table 23.1 summarizes all of the mentioned ML models, and their suitability for each case study.

## 23.4 Limitations and Challenges

Implementing the three methods will come with specific challenges. For the case study of treatment visualization, most of our issues are created due to technical limitations. Precision is indispensable for a medical image since a millimeter deviation can affect the physician's judgment. Precision is improved significantly by better models and more training data, but this immensely increases the amount of computation [55]. Another issue commonly associated with medical images is anomalies in the image produced by the machine. Medical images can have speckled noise and must be filtered before we apply this procedure. We have trained our model on a trusted dataset; however, we must be wary of the test input in the testing pipeline [56].

Once we have the segmented image with the required accuracy, rendering the image into a 3D model using ray-casting is again computationally expensive. Both steps require a powerful GPU system. Finally, once we have the model, we can visualize it through the Hololens 2. The visualization of the model will have specialized applications and cannot be generalized across healthcare solutions. For example, they cannot be used in surgery because the headset's weight can cause discomfort and loss of focus for the surgeon. Surgeries can go on for periods longer than the device's battery life [57]. The output visualization through the device can be used for diagnosis or educational purposes rather than as a critical healthcare solution. It would take several revisions before such a framework was used practically for such solutions.

For case study 2, maintaining data privacy and security is paramount. Healthcare-related data like mental health issues and psychology can be susceptible. Maintaining that degree of trust and enabling effective security measures to prevent data theft is essential. Security enhancement can be done using cryptography, while differential privacy increases data privacy [58]. Differential privacy protects the individual's

**Table 23.1** Summary of suggested ML models pertaining to each case study

| Model name | Description | Advantages | Limitations | Case study usage |
|---|---|---|---|---|
| U-Net | Stacked encoder with one block consisting of convolution, ReLU, and maxpooling layer. Similarly, we have a stacked decoder for the decoding operations | It requires fewer data for efficient training and accurate results and is effective in recognizing finer structures compared to other models | Training is very time-consuming, even on smaller datasets | Case Study 1: A U-Net can be highly effective for treatment visualization due to the presence of finer details in the output image segmentation |
| SegNet-Basic | A 4 encoder-4 decoder architecture: similar to U-Net but instead of using pool indices, the entire feature map transfers from the encoder to the decoder | Greatly improves boundary lineation, and the concept of upsampling and downsampling can be employed to other encoder-decoder architectures | It makes the model larger and requires a lot more memory | Case Study 1: Can be used for a rough segmentation of the input image but U-Net provides finer details |
| N-layer ConvNet | N ranges from 4 to 8 stacked (convolution + maxpool) layers. The architecture is similar to a VGG-16, however, they are not as deep as the architecture above | Lower computation than the architectures above, and also quite accurate | Does not include real-time prediction of features and not apt for complex computer vision problems | Case Study 2: Appropriate for emotion recognition, since the computational complexity of the model matches the level of complexity of the problem |
| RNN-based Language model | Uses an RNN for performing prediction of the next work/sentence. An RNN model works on sequential or time-dependent data and is therefore highly useful for NLP problems | These are much more effective than statistical models since they avoid data sparsity making them crucial for larger datasets | They are not able to establish long-term dependencies due to the vanishing gradients | Case Study 3: For the implementation of the chatbot, we require an efficient language model. Using an RNN model can be effective |
| LSTM-based Language model | LSTM is an RNN variant with the major difference being that it serves as a memory device that can store significant data content. It does so using gates: the input, output, and forget gate | They can establish long-term dependencies therefore making them more accurate | They take a lot of time to compute and are much more complex | Case Study 3: This would be a better option for the same as compared to an RNN-based model since it improves accuracy |
| GRU-based Language model | GRU is another RNN variant that uses an update and reset gate instead of three gates, like the LSTM | They are quicker to compute than LSTMs | LSTM is a better choice when there is a lot of training data | Case Study 3: Since there would be a lot of training data, it would be better to use an LSTM-based neural language model as compared to this |

data by adding random noise while cryptography encrypts it. It is also easy to get addicted while being immersed in virtual environments. Even while treating phobias and pain, the good feelings induced by the environment can provide a false sense of high. Therefore, setting time limits while using the framework is necessary. Each module should be designed for brief intervals, after which the user should pause from using the environment [59].

In case study 3, we rely on NLP for the functioning of the chatbot. Again, for the high accuracy of the model, training over a large dataset is required, which increases the need for computation. Apart from the computation, NLP models are usually more complex to understand than other ML models, requiring more studying and engineering. The conversation between the user and the application must be encrypted for data security. Due to the presence of AR, there are high risks of the application being addictive. A very famous example of an addictive AR application is Pokemon GO. Users worldwide were hooked to the location-based AR game [60]. Considering these challenges, a practical framework must be developed for this scenario.

## 23.5 Opportunities and Future Direction

This study proposes three frameworks integrating XR and ML for effective healthcare solutions. An accurate implementation of each framework could help physicians and medics; however, several revisions and alterations are necessary before they can be used practically. The next step for us as researchers would be to validate each of the theoretical frameworks proposed as implementations. We can also explore integrating the two fields for competitive sports, posture rehabilitation, etc. Plenty can be done when the two mighty fields complement each other. However, through this paper, we hoped to provide an initial idea of how this can be done.

Case study one sees excellent prospects for the review of internal body structures for medical examination and education. The segmentation of body portions helps recognize the exact part and separates the affected area from the rest of the structure. Case study two offers personal and adaptive therapy sessions through VR simulations. They can lessen our reliance on psychologists and help develop confidence. Psychologists do not focus on specific phobias, but a user can overcome his/her fears through simulations. Case study three can offer speedy recovery remedies to the user. The user does not have to wait for anyone since this makes it highly convenient to take appropriate measures. This would also help the physicians in picking more significant medical cases rather than ones that could be solved easily by the AR chatbot.

Continuous improvements in ML models help improve the accuracy of the output, while new innovative XR devices can add more features to a virtual environment. However, each case study in the paper uses ML and XR as two distinct entities. Research on neural rendering is booming, in which new landscapes and images are generated based on neural networks and computer graphics techniques. When visualized through a headset, these landscapes can enhance user experience through the headset [61, 62].

## 23.6 Conclusion

In the following paper, we discuss how a tool like ML provides underlying functionality to the content created by XR for healthcare. We have recognized three applications in healthcare where we want to review and propose the use of ML and XR. The related works section studies how the two entities are employed individually in the three applications. Once they have been established individually for the specific use case, we want to find a way of integrating the two technologies. The first case study is treatment visualization. We use image segmentation to classify the affected areas from the rest, while volumetric rendering and mixed reality render their interactive holograms onto the headset. The second case discusses therapeutic strategies in which we perform emotion recognition based on the body features and use virtual reality to create therapy-based simulations. The third case study focuses on effective decision-making. We employ NLP for predicting the next question. On the other hand, augmented reality creates the avatar and provides visual aids for exercises. After the three case studies are proposed, we discuss their limitations and the opportunities they provide.

## References

1. Parekh, P., Patel, S., Patel, N., Shah, M.: Systematic review and meta-analysis of augmented reality in medicine, retail, and games. Vis. Comput. Ind. Biomed. Art **3**(1) (2020). https://doi.org/10.1186/s42492-020-00057-7
2. Chavan, S.: Augmented reality vs. virtual reality: differences and similarities. Semantic Scholar (1970). Retrieved 7 Oct 2022, from https://www.semanticscholar.org/paper/Augmented-Reality-vs.-Virtual-Reality-Differences-Chavan/7dda32ae482e926941c872990840d654f9e761ba
3. Moawad, G.N., Elkhalil, J., Klebanoff, J.S., Rahman, S., Habib, N., Alkatout, I.: Augmented realities, artificial intelligence, and machine learning: clinical implications and how technology is shaping the future of medicine. J. Clin. Med. **9**(12), 3811 (2020). https://doi.org/10.3390/jcm9123811
4. Rogers, M.P., DeSantis, A.J., Janjua, H., Barry, T.M., Kuo, P.C.: The future surgical training paradigm: virtual reality and machine learning in surgical education. Surgery, **169**(5), 1250–1252 (2021). https://doi.org/10.1016/j.surg.2020.09.040
5. Oulefki, A., Agaian, S., Trongtirakul, T., Benbelkacem, S., Aouam, D., Zenati-Henda, N., Abdelli, M.-L.: Virtual reality visualization for computerized COVID-19 lesion segmentation and interpretation. Biomed. Signal Process. Control **73**, 103371 (2022). https://doi.org/10.1016/j.bspc.2021.103371
6. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., et al.: State of the art on neural rendering. In: Computer Graphics Forum, vol. 39, no. 2, pp. 701–727 (2020)
7. Gross, M.H.: Computer graphics in medicine: from visualization to surgery simulation. ACM Siggraph Comput. Graph. **32**(1), 53–56 (1998)
8. Haley, J., Richeport-Haley, M.: The Art of Strategic Therapy. Routledge (2004)
9. Vávra, P., Roman, J., Zonča, P., Ihnát, P., Němec, M., Kumar, J., Habib, N., El-Gendi, A.: Recent development of augmented reality in surgery: a review. J. Healthc. Eng. 1–9 (2017). https://doi.org/10.1155/2017/4574172
10. Andrews, C., Southworth, M.K., Silva, J.N., Silva, J.R.: Extended reality in medical practice. Curr. Treat. Options Cardiovasc. Med. **21**, 1–12 (2019)

11. Jang, J., Tschabrunn, C.M., Barkagan, M., Anter, E., Menze, B., Nezafat, R.: Three-dimensional holographic visualization of high-resolution myocardial scar on HoloLens. PloS One **13**(10), e0205188 (2018)
12. Opolski, M.P., Debski, A., Borucki, B.A., Szpak, M., Staruch, A.D., Kepka, C., Witkowski, A.: First-in-man computed tomography-guided percutaneous revascularization of coronary chronic total occlusion using a wearable computer: proof of concept. Can. J. Cardiol. **32**(6), 829–e11 (2016)
13. Gong, X., JosephNg, P.S.: Technology behavior model-beyond your sight with extended reality in surgery. Appl. Syst. Innov. **5**(2), 35 (2022)
14. Rauschnabel, P.A.: Augmented reality is eating the real-world! The substitution of physical products by holograms. Int. J. Inf. Manag. **57**, 102279 (2021)
15. Tsai, M.D., Hsieh, M.S., Jou, S.B.: Virtual reality orthopedic surgery simulator. Comput. Biol. Med. **31**(5), 333–351 (2001)
16. Thomsen, A.S.S., Bach-Holm, D., Kjærbo, H., Højgaard-Olsen, K., Subhi, Y., Saleh, G.M., Konge, L.: Operating room performance improves after proficiency-based virtual reality cataract surgery training. Ophthalmology **124**(4), 524–531 (2017)
17. Gasco, J., Patel, A., Ortega-Barnett, J., Branch, D., Desai, S., Kuo, Y.F., Roitberg, B.Z.: Virtual reality spine surgery simulation: an empirical study of its usefulness. Neurol. Res. **36**(11), 968–973 (2014)
18. Kim, Y., Kim, H., Kim, Y.O.: Virtual reality and augmented reality in plastic surgery: a review. Arch. Plast. Surg. **44**(03), 179–187 (2017)
19. Carlin, A.S., Hoffman, H.G., Weghorst, S.: Virtual reality and tactile augmentation in the treatment of spider phobia: a case report. Behav. Res. Ther. **35**(2), 153–158 (1997)
20. Hoffman, H.G.: Virtual-reality therapy. Sci. Am. **291**(2), 58–65 (2004)
21. Jeong, H., Yoo, J. H., Song, H.: Virtual Agents with Augmented Reality in Digital Healthcare
22. Genay, A., Lécuyer, A., Hachet, M.: Being an avatar "for real": a survey on virtual embodiment in augmented reality. IEEE Trans. Vis. Comput. Graph. **28**(12), 5071–5090 (2021)
23. Masmuzidin, M.Z., Aziz, N.A.A.: The current trends of augmented reality in early childhood education. Int. J. Multimedia Appl. (IJMA) **10**(6), 47 (2018)
24. Boonbrahm, P., Kaewrat, C., Boonbrahm, S.: Interactive marker-based augmented reality for CPR training. Int. J. Technol. **10**(7), 1326–1334 (2019)
25. Basiratzadeh, S., Lemaire, E.D., Baddour, N.: Augmented reality approach for marker-based posture measurement on smartphones. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp 4612–4615. IEEE (2020)
26. Shailaja, K., Seetharamulu, B., Jabbar, M.A.: Machine learning in healthcare: a review. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (2018). https://doi.org/10.1109/iceca.2018.8474918
27. Wild, J.J.: The use of ultrasonic pulses for the measurement of biologic tissues and the detection of tissue density changes. Surgery **27**(2), 183–188 (1950)
28. Edler, I.: The use of ultrasonic reflectoscope for the continuous recording of the movements of the heart walls. Kungliga Fysiografiska Sallskapetsi I Lund Forhandlingar **24**, 1 (1954)
29. Jensen, J.A.: Medical ultrasound imaging. Prog. Biophys. Mol. Biol. **93**(1–3), 153–165 (2007)
30. Chen, J., Liu, S., Huang, M.: Low-rank and sparse decomposition model for accelerating dynamic MRI reconstruction. J. Healthc. Eng. (2017)
31. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., et al.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th International Conference on Multimodal Interfaces, pp. 205–211 (2004)
32. Tautkute, I., Trzcinski, T., Bielski, A.: I know how you feel: emotion recognition with facial landmarks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1878–1880
33. Ioannou, S.V., Raouzaiou, A.T., Tzouvaras, V.A., Mailis, T.P., Karpouzis, K.C., Kollias, S.D.: Emotion recognition through facial expression analysis based on a neurofuzzy network. Neural Networks **18**(4), 423–435 (2005)

34. Tarnowski, P., Kołodziej, M., Majkowski, A., Rak, R.J.: Emotion recognition using facial expressions. Procedia Comput. Sci. **108**, 1175–1184 (2017)
35. Ozdemir, M.A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., Akan, A.: Real time emotion recognition from facial expressions using CNN architecture. In: 2019 Medical Technologies Congress (Tiptekno), pp. 1–4. IEEE (2019)
36. Bhirud, N., Tataale, S., Randive, S., Nahar, S.: A literature review on chatbots in healthcare domain. Int. J. Sci. Technol. Res. **8**(7), 225–231 (2019)
37. Ayanouz, S., Abdelhakim, B.A., Benhmed, M.: A smart chatbot architecture based NLP and machine learning for health care assistance. In: Proceedings of the 3rd International Conference on Networking, Information Systems & Security, pp. 1–6 (2020)
38. Bulla, C., Parushetti, C., Teli, A., Aski, S., Koppad, S.: A review of AI based medical assistant chatbot. Res. Appl. Web Dev. Des. **3**, 1–14 (2020)
39. Shaikh, A., More, D., Puttoo, R., Shrivastav, S., Shinde, S.: A survey paper on chatbots. Int. Res. J. Eng. Technol. (IRJET) **6**(04), 2395–0072 (2019)
40. Nagabhushan, P., Javed, M.: Word and character segmentation directly in run-length compressed handwritten document images (2019). arXiv preprint arXiv:1909.05146
41. Lee, N., Kim, K., Yoon, T.: Implementation of robot journalism by programming custombot using tokenization and custom tagging. In: 2017 19th International Conference on Advanced Communication Technology (ICACT), pp. 566–570. IEEE (2017)
42. Yang, L., Zhang, M., Liu, Y., Sun, M., Yu, N., Fu, G.: Joint POS tagging and dependence parsing with transition-based neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(8), 1352–1358 (2017)
43. Chen, B., Ji, D.: Chinese semantic parsing based on dependency graph and feature structure. In: Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, vol. 4, pp. 1731–1734. IEEE (2011)
44. http://lavalle.pl/vr/vrch1.pdf
45. Kratz, A., Hadwiger, M., Fuhrmann, A., Splechtna, R., Bühler, K.: GPU-based high-quality volume rendering for virtual environments. In: International Workshop on Augmented Environments for Medical Imaging and Computer Aided Surgery (AMI-ARCS), vol. 2006, pp. 33–38 (2006)
46. Fuhrmann, A.L., Purgathofer, W.: Studierstube: an application environment for multi-user games in virtual reality. In: GI Jahrestagung (2), pp. 1185–1190 (2001)
47. Levoy, M.: Display of surfaces from volume data. IEEE Comput. Graph. Appl. **8**(3), 29–37 (1988)
48. El Beheiry, M., Doutreligne, S., Caporal, C., Ostertag, C., Dahan, M., Masson, J.B.: Virtual reality: beyond visualization. J. Mol. Biol. **431**(7), 1315–1321 (2019)
49. Ackermann, K., Walsh, J., De Unánue, A., Naveed, H., Navarrete Rivera, A., Lee, S.J., et al.: Deploying machine learning models for public policy: a framework. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 15–22 (2018)
50. Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A.: Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol. Model. **406**, 109–120 (2019)
51. Yaganteeswarudu, A.: Multi disease prediction model by using machine learning and Flask API. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 1242–1246. IEEE (2020)
52. Oulefki, A., Agaian, S., Trongtirakul, T., Benbelkacem, S., Aouam, D., Zenati-Henda, N., Abdelli, M.-L.: Virtual reality visualization for computerized COVID-19 lesion segmentation and interpretation. Biomed. Sig. Process. Control **73**, 103371 (2022). https://doi.org/10.1016/j.bspc.2021.103371
53. Bălan, O., Moldoveanu, A., Leordeanu, M.: A machine learning approach to automatic phobia therapy with virtual reality. In: Modern Approaches to Augmentation of Brain Function, pp. 607–636. Springer, Cham (2021)

54. Wu, W.C., Yu, Y.H.: Combination of augmented reality with chatbots for visual aids in nutrition education. In: 2022 IEEE International Conference on Consumer Electronics-Taiwan. IEEE, pp. 203–204 (2022)
55. Thompson, N.C., Greenewald, K., Lee, K., Manso, G.F.: The computational limits of deep learning (2020). arXiv preprint arXiv:2007.05558
56. Goyal, B., Agrawal, S., Sohi, B.S.: Noise issues prevailing in various types of medical images. Biomed. Pharmacol. J. **11**(3), 1227 (2018)
57. Khor, W.S., Baker, B., Amin, K., Chan, A., Patel, K., Wong, J.: Augmented and virtual reality in surgery-the digital surgical environment: applications, limitations and legal pitfalls. Ann. Transl. Med. **4**(23) (2016)
58. De Cristofaro, E.: An overview of privacy in machine learning (2020). arXiv preprint arXiv:2005.08679
59. Luciana, R.P.: One minute more: adolescent addiction for virtual world. Procedia Soc. Behav. Sci. **2**(2), 3706–3710 (2010)
60. Abd Wahab, S.A., Jamalludin, N.H., Wok, S.: Factors determining Pokémon Go Addiction in Malaysia. J. Manag. Mark. Rev. **2**(2), 73–78 (2017)
61. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., et al.: State of the art on neural rendering. In: Computer Graphics Forum, vol. 39, no. 2, pp. 701–727 (2020)
62. Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., et al.: Advances in neural rendering. In: Computer Graphics Forum, vol. 41, no. 2, pp. 703–735 (2022)

# Chapter 24
# KP-RNN: A Deep Learning Pipeline for Human Motion Prediction and Synthesis of Performance Art


**Patrick Perrine and Trevor Kirkby**

**Abstract** Digitally synthesizing human motion is an inherently complex process, which can create obstacles in application areas such as virtual reality. We offer a new approach for predicting human motion, KP-RNN, a neural network which can integrate easily with existing image processing and generation pipelines. We utilize a recently introduced human motion dataset of performance art, Take The Lead, as well as the existing motion generation pipeline, the Everybody Dance Now system, to demonstrate the effectiveness of KP-RNN's motion predictions. We have found that our neural network can predict human dance movements effectively, which serves as a baseline result for future works using the Take The Lead dataset. Since KP-RNN can work alongside a system such as Everybody Dance Now, we argue that our approach could inspire new methods, such as those for rendering human avatar animation. This work also serves to benefit the visualization of performance art in digital platforms by utilizing accessible neural networks.

## 24.1 Introduction

Human motion synthesis involves digitally extracting the geometry of the human body for analysis or use in software applications. Some current research themes in human motion synthesis involve using deep neural networks for predicting human motion [1]. Such methods can have lent themselves to generating graphics for mixed reality [2]. An overarching question to our work is: how effectively can deep neural networks synthesize human motion? This leads us to the more specific question: how well can deep neural networks learn from human performance art?

P. Perrine (✉) · T. Kirkby
California Polytechnic State University, San Luis Obispo, CA 93407, USA
e-mail: paperrin@calpoly.edu



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_24

In this research we intended to accomplish the following:

- Implement an instance of OpenPose [3] for processing human skeletal poses from video and the Everybody Dance Now system [4] for video generation.
- Design a custom-tuned deep learning model for human motion prediction, specifically for performance art.
- Apply a new dataset to all these models.

This model could also have applications in mixed reality, computer animation, and graphics. Being able to generate new human motion and a subsequent video of predicted motion could ease the processing of human avatars. This model could also be used by dance choreographers to analyze their dance pieces. Artists sometimes are concerned with subverting or reinforcing expectations, so being able to measure the predictability of their performance could allow them greater control to work to be more predictable or less predictable in their movements.

## 24.2 Background

To understand human motion synthesis, one must understand the principles of computer vision. This field is concerned with utilizing digital cameras as "eyes" for computer programs to "see." If one has some intuition of computer graphics, one can conceptualize computer vision as reverse computer graphics. Whereas work in computer graphics is concerned with generating visuals based on data, computer vision is about extracting data from visuals. There has been much cross-fertilization in research for vision and graphics over decades. This has resulted in various sub-disciplines such as human motion synthesis, which has conceptual roots as early as the mid-1970's [5]. Studies in human motion saw a significant surge in the late 1990's/early 2000's [6] and have been consistently present in computer vision literature since.

Being familiar with various machine learning models, such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Generative Adversarial Networks (GANs) are often required to understand the architecture of neural network models used for human motion synthesis.

Generative Adversarial Networks are systems in which two distinct neural networks are trained in competition. One of them optimizes toward distinguishing genuine and synthesized examples from a dataset, while the other optimizes toward synthesizing convincing examples to fool the first network. Generative adversarial networks have been used successfully to create or transform images, and more specifically for generating human motion [4].

## 24.3 Related Work

There have been various approaches to understanding human motion in digital form. A common approach is to intake video of humans performing actions, generate representations of their bodies, and then perform motion prediction [7, 8]. Deep neural networks have commonly been applied with motion prediction, sometimes in conjunction with the popular OpenPose framework [3]. Other approaches that have used graph neural networks have yielded strong motion prediction results [8]. However, by not utilizing a sequence model, such as an RNN, such methods would not lead to the capability to generate new sequences of motion, which could benefit the previously mentioned areas of virtual reality and animation.

There exists a related line of research in human motion forecasting, which focuses on predicting human movement trajectories from a distant, top-down perspective [1]. However, such methods do not account for the intricacies of human poses, rather the general direction that a given human moves within a geographic space. Motion forecasting would not lend itself as well to the visualization of full human avatars, as motion synthesis often can.

Interest in prediction methods have also lent themselves to head position prediction for virtual reality [2]. While such methods can be valuable when understanding the visual perspective of a given user in virtual reality environments, we are more concerned with entire scenes for which users could experience. Such scenes could involve the motion of various humanoid figures, rather than simply the head movements of human figures.

## 24.4 Data Acquisition

Our data was obtained primarily from the Take The Lead (TTL) Dataset [9]. This data was originally collated from publicly available videos of human dance posted on YouTube. Videos were downloaded and labeled by hand to reflect several different genres of dance. We acknowledge that there may exist some heterogeneity within the image data, due to inconsistent video capture procedures. This inconsistency could affect results drawn from models trained on such data. Sample data from the Everybody Dance Now system was also used as a reference point. This includes footage of collaborators in the Everybody Dance Now project performing basic human motions optimized for testing the Everybody Dance Now system. The sample data used in the previous experiments is linked publicly in the associated repository here: https://github.com/carolineec/EverybodyDanceNow. There were no human subjects involved within the course of this work.

As with virtually all machine learning systems, the products of our system may inherit biases present in the data the system is trained upon. This issue has continued to be addressed within the academic communities of machine learning [10]. The potential of bias is difficult to eliminate from machine learning, and it is important

to consider the potential biases introduced through datasets. A significant mitigating factor of bias in this work is that all video of humans is pre-processed down to simple "stick figures," which only convey information about a subject's physical pose. Since our system is only aware of humans as stick figures, we believe it is difficult for our system to inherit biases from information such as race, gender, et cetera. The datasets we used were also acquired from two existing works that were vetted prior to publication. We suggest the investigation of ethical frameworks when creating new datasets to be released publicly [10].

## 24.5 System Design

Our system (see Fig. 24.1) learns from the contents of YouTube videos of dancing, obtained from Take The Lead. These are stored as sequences of PNG images for each video frame. From there, the OpenPose system is used to estimate the human poses in each frame, creating a skeletal representation of the human in a given frame. The skeletal representation is translated into a JSON format that gives two-dimensional coordinates for various keypoints (head, chest, shoulders, et cetera). At this point, the data may be used in two separate models: the Everybody Dance Now system and our Long Short-Term Memory (LSTM) neural network.

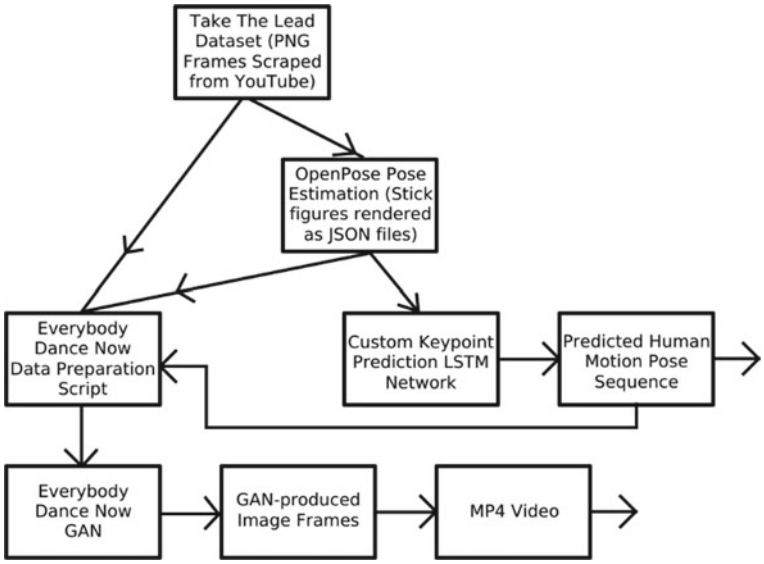



**Fig. 24.1** Our video-to-video system. We can intake 2D videos of human motion and output predicted sequences of motion, as well as output synthesized video of said predicted sequences. See Fig. 24.2 for a visual of the Everybody Dance Now GAN procedures and see Fig. 24.3 for the Keypoint Prediction LSTM Architecture

To make the Everybody Dance Now (EDN) system work for videos scraped from YouTube, which often contain multiple people, the OpenPose JSON files are altered to only keep track of the poses for a single person in each video. Then a data preparation script provided in the EDN repository is used to convert the JSON files to images of stick figures, which are provided as inputs to EDN's conditional GAN. EDN trains from footage of a specific person performing movements and can then create footage of that same person performing new movements. The footage is provided as sequential image frames, which are resized to a correct aspect ratio and then combined into a single video.

The custom keypoint predicting LSTM network, (KP-RNN, seen in Fig. 24.3) accepts a sequence of poses as input. This input is represented as 50 numbers, which are the $(x, y)$ coordinates of the 25 keypoints created by OpenPose to model the position of a single person. It has several recurrent layers and two densely connected layers at the end. We use conventional stochastic gradient descent to optimize mean squared error. The output of KP-RNN is the predicted pose in the next frame of the video for the person, again represented as 50 numbers. It should be noted that the predictions of this network can be used to generate a sequence of poses, which can then be provided as an input to Everybody Dance Now to generate a new video. There was some fine-tuning of the architectural design of KP-RNN, but not so much as to overfit our data.

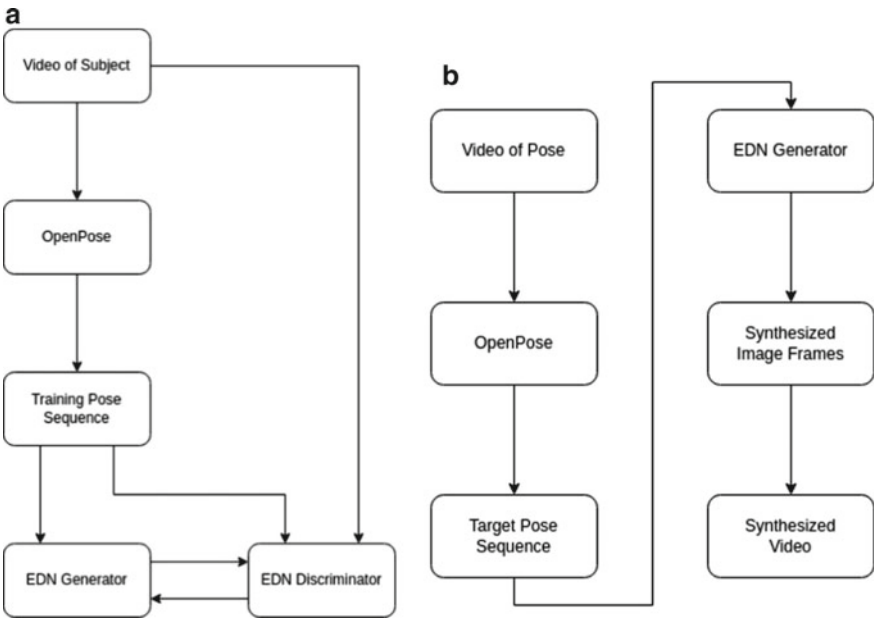



**Fig. 24.2** **a** The training procedure for Everybody Dance Now. **b** The generation procedure for Everybody Dance Now

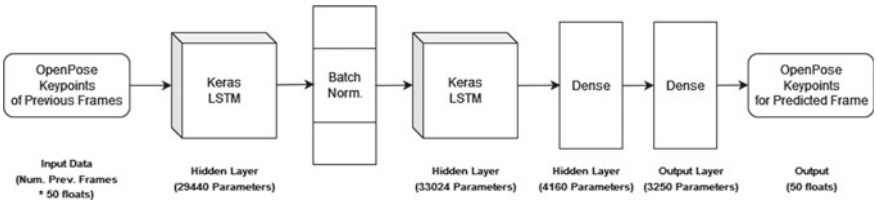



**Fig. 24.3** Our keypoint recurrent neural network architecture (KP-RNN)

### *24.5.1 Implementation*

This system was set up on an instance of Google Colab Pro running with a Tesla P100 GPU as a development environment. The system is implemented using primarily Python and Bash, and both Everybody Dance Now and the custom architecture KP-RNN are implemented using Tensorflow [11] and Keras [12]. Results are visualized with Tensorboard and the Matplotlib Python library, and pre-processed into video using Python Image Library and the FFMPEG command-line interface. Our parameter settings are described in Table 24.1.

## 24.6 Testing and Validation

Our experiment requires measuring several different metrics to gauge the success of different components of the system.

**Table 24.1** KP-RNN hyperparameter description

| Parameter | Setting |
|---|---|
| Dimensions of input vector | $25 \times 2$ |
| Dimensions of output vector | $25 \times 2$ |
| Number of LSTM layers | 2 |
| Size of LSTM layers | 64 |
| Number of dense hidden layers | 1 |
| Size of dense hidden layers | 64 |
| Activation | Sigmoid |
| Dropout probability | 0.3 |
| Loss function | MSE |
| Optimizer | SGD |
| Learning rate | $3 \times 10^{-3}$ to $1 \times 10^{-3}$ |
| Momentum | 0.2 |
| Maximum epochs | 1200 |

First, we evaluate the efficacy of the KP-RNN architecture using root mean squared error. Root mean squared error is measured as: the distance between the predicted set of keypoint coordinates and the actual set of keypoint coordinates; a low root mean squared error means that the model is predicting a sequence of movements that are close to the ground truth. The other potential measurement we consider is prediction accuracy; however, for evaluating KP-RNN, root mean squared error is deemed a more useful metric than prediction accuracy. Given that motion prediction is a form of regression task, KP-RNN is being applied to a regression task and not a classification task. Prediction accuracy scores each case in an all-or-nothing manner, but this removes information compared to root mean squared error and is therefore less relevant to the problem space. For example, when using prediction accuracy, a prediction that is significantly different from the ground truth is treated the same as a prediction that is completely correct except it is offset by one pixel from the ground truth. For this reason, a low root mean squared error is our best indication of success for predicting movements.

Equation **24.1**. Formula for Root Mean Squared Error

$$\text{RMSE} = \sqrt{\left(y - \hat{y}\right)^2}. \tag{24.1}$$

We evaluate the success of the EDN component of our system based on scoring the difference between the generated video and the actual ground truth video. This requires scoring the visual similarity between two images that are not identical, which is not a trivial task. There are several metrics that are used for this purpose. The first such metric is the feature matching loss (Figs. 24.4a and 24.5a). Feature matching is calculated by running both the generated image and the actual image for a given pose through the image discriminator neural network, comparing the similarities between the vector outputs of intermediate layers in the discriminator network, and optimizing to improve this similarity [13]. The second metric is perceptual reconstruction loss (Figs. 24.4b and 24.5b), which takes a similar approach, except instead of using the intermediate layers of the GAN discriminator, it observes the intermediate layers of a completely different image recognition neural network architecture, specifically VGG19 [5]. Note that, for this metric, EDN records the differences between intermediate layers rather than the similarities, so while a higher value indicates better accuracy for feature matching, the opposite is true for perceptual reconstruction loss. In conclusion, a high feature matching value and a low perceptual reconstruction loss implies that the sequence of images synthesized by the system are visually similar to the ground truth images, which indicates that the system is working as intended.

Recording the training loss values for the actual generator and discriminator neural networks of EDN is comparatively not useful for the purposes of evaluation. By construction, they exist in a zero-sum game, and in a successful GAN they will remain approximately matched. Because the loss metrics for the GAN, by design, contain virtually no meaningful insight into whether the system is working or not, they are omitted as evaluation metrics in this work.

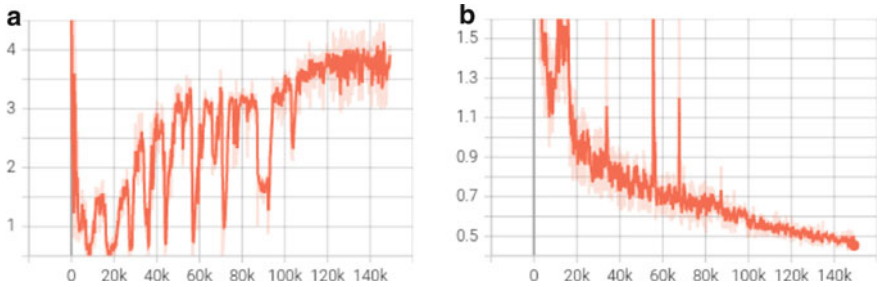



**Fig. 24.4** **a** EDN feature matching with its own data. **b** EDN perceptual reconstruction loss with its own data

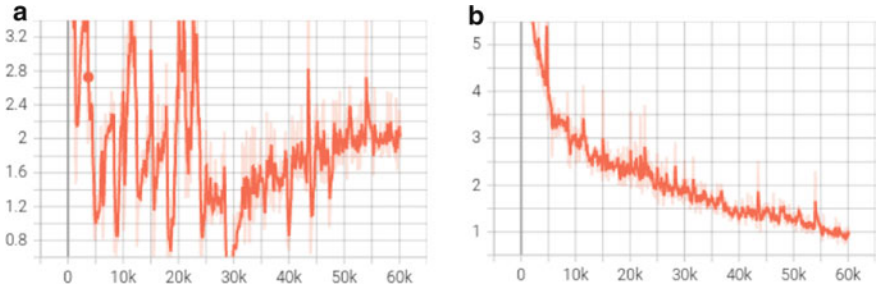



**Fig. 24.5** **a** EDN feature matching with TTL data. **b** EDN perceptual reconstruction loss with TTL data

Finally, EDN also produces video outputs, which may be evaluated qualitatively. See Figs. 24.6b and 24.7b for sample frames from the generated video, based on the inputs in Figs. 24.6a and 24.7a. Figure 24.6a is from the Everybody Dance Now demonstration dataset, and uses a more complex skeleton figure with individual keypoints for hands and face, whereas Fig. 24.7a is from the Take The Lead dataset, and uses a simplified skeleton with fewer keypoints.

### *24.6.1 Project Goal Evaluation*

Here is how we accomplished our original goals, along with some technical details:

- Implement an instance of OpenPose [3] for processing human skeletal poses from video and the Everybody Dance Now system [4] for video generation.

This was mostly accomplished. However, for EDN, we observed an issue where the model does not recognize more than one person, which posed an issue when using the TTL dataset. We built on the previous work by implementing a method to pre-process new videos of multiple people and focus on a single person throughout

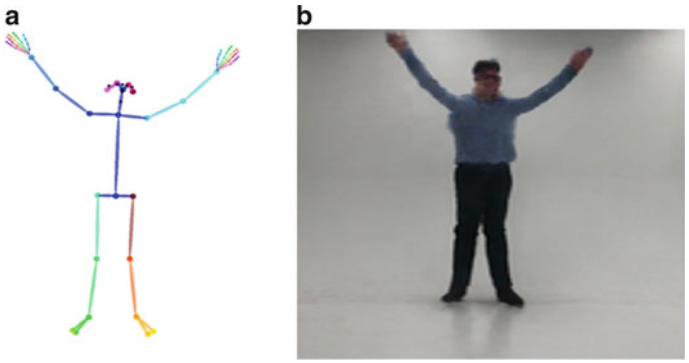



**Fig. 24.6** **a** Skeleton input image from EDN's data. **b** Synthesized reconstructed image from EDN's GAN

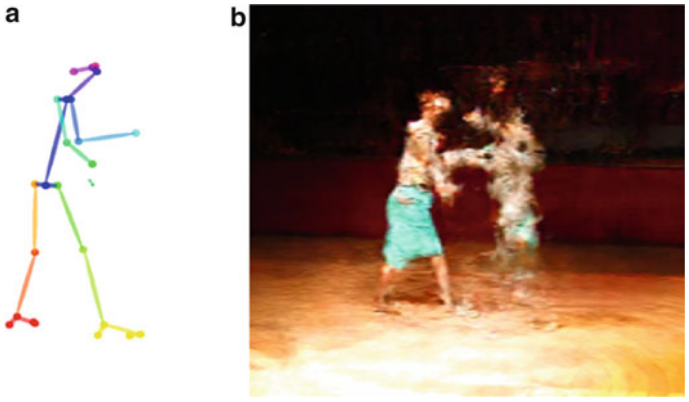



**Fig. 24.7** **a** Skeleton input image from TTL's data. **b** Synthesized reconstructed image from EDN's GAN

the video sequence. This complies with the limitation of EDN being only able to process one person's sequence.

- Design a custom deep learning model for human motion prediction, specifically tailored for performance art.

We now have our own LSTM-based architecture, KP-RNN, for predicting human motion sequences, which we developed in Google Colab Pro. Our final root mean squared error was 0.2328, which indicates that our system is functional. Given that we are using a new dataset of different structure to commonly studied motion datasets, it would be difficult to compare our error score with other prediction methods. This work can serve as a basis for future work to develop new methods for this dataset to compete with this score. Our code and empirical model weights are publicly available on this GitHub repository: https://github.com/patrickrperrine/comp-choreo.

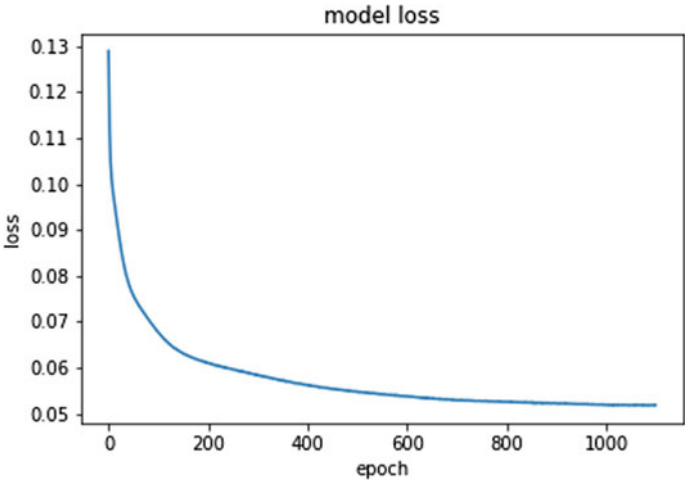



**Fig. 24.8** KP-RNN training error over time

- Apply a new dataset to all these models.

We were interested in creating a brand new dataset of performance art to be used for prediction along with our existing TTL dataset. We found this to be difficult and too time-constrained for the scope of this project.

## 24.7 Future Work

We found that creating a new, usable dataset can be rather difficult. Acquiring new performance art data as originally proposed could be a fruitful endeavor. Also, having more computational power to try and build larger, custom architectures could lead to novel results in human motion prediction. To move past 2D motion, there could be an exploration of the Human 3.6M dataset [14], which is popular in 3D human motion prediction models [7]. Also, an in-depth exploration of the theoretical foundations of deep learning could provide explanation for how these methods can be useful in a general case, allowing them to be useful in a more esoteric one such as ours.

## 24.8 Conclusion

We offer a novel approach to human motion prediction with our lightweight neural network, and its ability to integrate nicely with existing image processing pipelines. We have effectively combined EDN and TTL along with our own deep neural network to produce a new system for dance motion prediction, image-to-image translation,

and video generation. Our results indicate that our LSTM-based prediction network functions effectively on a new video dataset of human performance art. We admit that our results are ultimately qualitative, however, we argue that the evidence shown of the functionality of our system suffices to inspire new and specific applications of deep learning. Our overall processing system could inspire innovations in spaces, such as virtual reality, that are concerned with visualizing complex forms of human motion.

**Acknowledgements** This work was partially supported by a Cal Poly Graduate Assistant Fellowship to Patrick Perrine. We thank Professors Franz Kurfess and Jonathan Ventura for their support of this work.

## References

1. Chiara, L.F., Coscia, P., Das, S., Calderara, S., Cucchiara, R., Ballan, L.: Goal-driven self-attentive recurrent networks for trajectory prediction. In: CVPR Workshops, pp. 2517–2526 (2022)
2. Amamra, A.: Smooth head tracking for virtual reality applications. SIViP **11**(3), 479–486 (2017)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2019)
4. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 5933–5942 (2019)
5. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
6. Aggarwal, J.K., Cai, Q.: Human motion analysis: a review. Comput. Vis. Image Underst. **73**(3), 428–440 (1999)
7. Battan, N., Agrawal, Y., Rao, S.S., Goel, A., Sharma, A.: Glocalnet: class-aware long-term human motion synthesis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 879–888 (2021)
8. Hermes, L., Hammer, B., Schilling, M.: Application of graph convolutions in a lightweight model for skeletal human motion forecasting (2021). arXiv preprint arXiv:2110.04810
9. Farris, T.: Take the Lead: Toward a Virtual Video Dance Partner. Cal Poly Digital Commons (2021)
10. Fu, A., Ding, E., Hosseini, M.S., Plataniotis, K.N.: P4AI: Approaching AI Ethics through Principlism (2021). arXiv preprint arXiv:2111.14062
11. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265–283 (2016)
12. Chollet, F., et al.: Keras: the python deep learning library. In: Astrophysics Source Code Library, pp. ascl-1806 (2018)
13. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2019)
14. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2013)

# Chapter 25
# AI-Supported XR Training: Personalizing Medical First Responder Training

**Daniele Pretolesi, Olivia Zechner, Daniel Garcia Guirao, Helmut Schrom-Feiertag, and Manfred Tscheligi**

**Abstract** Extended Reality (XR) technologies have become an increasingly popular tool for training medical first responders (MFR) and simulation of mass casualty incidents (MCI). The possibility of training scenarios that would be highly complex and potentially dangerous in reality provides a considerable opportunity for adopting this technology. However, the level of automatization of performance monitoring and personalization of training content is still a challenge that has not been solved. This paper introduces an innovative AI-supported XR training approach to address the challenges faced by current simulation training solutions. The proposed approach incorporates wearable and sensor technology to collect physiological metrics and utilizes machine learning algorithms to identify stressors and key performance indicators (KPIs) unique to each trainee. The collected data is then analysed and translated into personalized recommendations for scenario adaptation. We discuss the advantages and challenges encountered during the development of this framework and propose methodologies for its implementation and evaluation.

## 25.1 Introduction

Medical first responders (MFR) and mass casualty incidents (MCI) require prompt and effective actions, often in high-pressure and complex situations. For this reason, training plays a vital role in ensuring that first responders are adequately prepared to

Daniele Pretolesi and Olivia Zechner have contributed equally to this research.

D. Pretolesi (✉) · O. Zechner · H. Schrom-Feiertag
AIT—Austrian Institute of Technology, Vienna, Austria
e-mail: daniele.pretolesi@ait.ac.at

D. G. Guirao
IDENER, Sevilla, Spain

M. Tscheligi
Department of Artificial Intelligence and Human Interfaces (AIHI), Salzburg, Austria



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_25

handle such scenarios. However, traditional training methods have limitations, such as high costs and a lack of personalization and adaptability to individual trainees' needs. Thus, it can be challenging to replicate real-life threatening situations in training environments.

Extended Reality (XR) simulation training is one solution that has gained popularity in recent years [46, 48, 64]. However, current XR simulation training solutions have limitations in personalizing the training experience for individual trainees based on their behaviour. Such limitations can impact the effectiveness of training, as not all trainees learn at the same pace, and individual differences can affect performance.

To address these challenges, this concept paper presents an innovative Artificial Intelligence (AI)-supported XR training approach. The proposed solution aims to personalize the training experience for individual trainees based on their behaviour and expertise. The approach incorporates wearable and sensor technology to collect physiological metrics, and machine learning algorithms to identify stressors and key performance indicators unique to each trainee. The collected data is then analysed and translated into personalized recommendations for scenario adaptation.

The objective of this paper is to introduce an innovative XR training approach for MFR and MCI that offers personalized learning experiences for individual trainees by leveraging AI and data-driven techniques. The paper discusses the challenges faced by current simulation training methods and proposes a solution that overcomes these limitations by personalizing the training experience. The contributions of this work are threefold: Introducing an innovative AI-powered XR training approach for MFR and MCI that offers personalized learning experiences for individual trainees. Describing the proposed system and how it addresses the limitations of current simulation training methods. Discussing the challenges associated with the suggested approach and how they can be overcome.

The following chapters provide a comprehensive overview of the presented solution, the challenges associated with it, and its potential impact on the field of medical training.

## 25.2 Related Work

XR technology has emerged as a promising approach for medical first responder training. This section aims to explore the current state of MFR training for MCI, focusing on simulation-based training including virtual reality (VR) and mixed reality (MR). Furthermore, we explore the literature on adaptive virtual environments (AVEs) and the use of Artificial Intelligence (AI) for adaptive training because of their potential to further enhance XR training solutions by incorporating novel approaches to leverage data collected during the exercise.

### 25.2.1 Simulation-Based Training for MFRs

Simulation-based training has become increasingly popular in MFR training for MCIs, as it allows for a safe and controlled environment in which to practice critical skills [2, 3]. High-fidelity manikins have been used to mimic human physiology and responses in a realistic manner [32]. These systems incorporate immersive environments and realistic scenarios, allowing participants to acquire knowledge more quickly and develop high-quality abilities [34]. Gunshin et al. [11] present an integrative literature review of 21 studies evaluating the role of high-fidelity simulation technologies in disaster response and preparedness, highlighting the cost-effectiveness compared to real-world training but also the need for further assessment across different disaster phases (e.g. triage and treatment) to maximize their potential. Koutitas et al. [21] showcase an example of how performance evaluation could look in XR training solutions.

### 25.2.2 Adaptive Virtual Environments

AVEs have emerged as a promising approach to training, grounded in Kelley's [18] concept of adaptive training, which involves altering the problem, stimulus, or task contingent on the trainee's performance and training goals. AVEs can be utilized for (a) adapting virtual content [16, 26, 35] or (b) providing personalized feedback [27], thereby tailoring the training experience to each trainee's needs and abilities [4]. The adaptive process in AVEs typically depends on various inputs, including physiological signals, performance measures, and profile data, which are utilized to tailor the training environment and feedback [62]. AVEs have found applications in diverse industries, including game-based training [49], rehabilitation [12], promotion of physical activity [26, 35] medical training [39] and strategic decision-making [7], offering personalized, effective, and engaging learning experience that caters to individual learning styles and cognitive abilities.

### 25.2.3 AI-Supported Training

AI and machine learning (ML) are rapidly transforming the field of personalization in training [29, 37]. With the increasing amount of data collected from individuals, organizations are leveraging AI/ML to tailor training programmes that meet the unique needs and preferences of each learner [53].

Personalization of training refers to the customization of training programmes to meet the specific needs of individual learners [44, 47]. This approach recognizes that different individuals have different learning styles, preferences, and skill

levels. By providing personalized training, learners can receive content that is relevant, engaging, and effective, leading to improved performance and productivity [5, 25, 60].

AI can be used to personalize training in several ways. One approach is to use machine learning algorithms to analyse learner data, such as learning history, skill level, and performance data [1, 42, 51]. These algorithms can then use this data to recommend personalized content and activities that are tailored to each learner's needs. For example, an AI system may recommend specific modules or resources to learners based on their learning style or skill level [30, 43, 52].

Another way AI can personalize training is through adaptive learning [8, 19, 24, 58]. Adaptive learning systems use machine learning algorithms to adjust the difficulty and content of training based on the learner's performance. For instance, if a learner is struggling with a particular concept, an adaptive learning system may adjust the content and provide additional support to help them grasp the material [14, 31, 57].

In the context of XR training, research has been conducted to create adaptive solutions by modifying the simulated scenario [20, 28, 50, 58]. However, as identified in the review of [63], most training approaches are still not adaptive to individual capabilities and skill sets. In particular, XR-based training is still limited as scenarios are not capable of dynamic adaptation [41].

## 25.3 MED1stMR Training System

MED1stMR is a European Horizon 2020 project that aims to better prepare MFRs for complex disaster situations by developing a new generation of mixed reality (MR) training with haptic feedback for enhanced realism. The project's objectives include developing a pioneering MR training approach and integrating patient simulation manikins and medical equipment into virtual environments for a richer sensory experience. To ensure a higher adherence to the user-centred design principles, an Agile End User Centred research methodology has been applied, allowing for the development of effective training scenarios and curricula [17].

At the beginning of the project, several activities were conducted within the project's eight end-user organizations, including training observation, contextual interviews, co-creation workshops, and focus groups, to determine the most important factors of current MCI training and requirements for a novel XR training platform. Results were clustered into training approach and learning goals, training scenarios and required physiological measurements to classify stress and cognitive workload.

The analysis of the feedback collected showed high heterogeneity amongst organizations when it comes to learning goals, training methods, and desired training scenarios. For example, while some organizations wanted to focus on common MCIs (e.g. car or train accidents) others were keen on exploring more challenging environments (e.g. CBRNE-related accidents, earthquakes or forest fires). This highlighted the need for the final XR solution to be highly flexible regarding scenario design

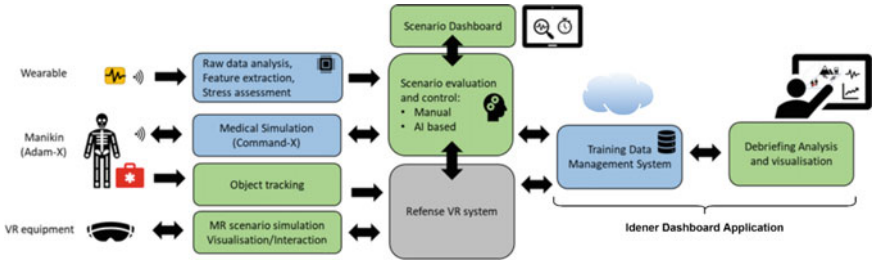



**Fig. 25.1** High-level system architecture based on end-user requirements

and adaptations as well as performance measurement metrics. Discussions about behaviour feedback loops and smart scenario control based on biosignals concluded the most suitable wearables and analysis to be based on heart rate (HR), heart rate variability (HRV), and electrodermal activity (EDA).

### *25.3.1 MR Training System*

An MR training system has been developed that utilizes state-of-the-art technologies to create realistic training scenarios for first responders and medical professionals. The system is developed using the Unity game engine[1] [54] and runs on Windows 10/11.[2]

There are four core components of this system: (a) the base system developed by REFENSE[3] [45] including a scenario editor and training simulator, (b) a biosignal measurement suite (PLUX)[4] [6] to monitor trainee's biosignals, (c) a realistic patient simulator (ADAM-X)[5] [33] that provides haptic feedback and is compatible with genuine medical devices, and (d) a centric data platform and dashboard for uploading, analysis, and debriefing purposes. Figure 25.1 presents a high-level overview of the system architecture.

**MR Base System**. To provide a fully immersive training experience, the system includes object tracking technology[6] [40], which uses LED sensors to provide real-time tracking with low latency, and motion-capturing technology-based Optitrack's Cameras and Motive application. The TrueVRSystems management system[7] [56] is

[1] https://unity.com/.

[2] https://www.microsoft.com/de-de/software-download/windows10.

[3] https://www.refense.com/.

[4] https://www.pluxbiosignals.com/.

[5] https://medical-x.com/product/adam-x/.

[6] https://optitrack.com/.

[7] https://www.truevrsystems.com/.

used to integrate all components into a single training scenario. To ensure compatibility with future hardware developments, all communications and interfaces are carried out via the experience host server.

The scenario editor is built for users with little prior experience with MR applications to make the system user-friendly for trainers and operators from MFR organizations. Scenarios can also be adapted live by an operator (e.g. for directing non-player characters (NPCs) or voice-acting patients).

Despite its advanced capabilities, the system has some limitations. For example, precise finger tracking for skills training with small objects such as needles and scissors is currently not possible. Even expensive gloves with optical finger tracking cannot provide precise tracking for multiple users. Additionally, small objects like needles cannot be tracked using the LED-based object tracking system. Furthermore, current wireless technologies have limitations in terms of bandwidth and latency, which must be taken into consideration when designing such systems.

**Biosignal Measurement Suite**. Finding the optimal level of stress to optimize learning performance is a challenge in simulation training [13] as it can vary from person to person. Consequently, personalized stress monitoring and scenario adaptation are significant milestones in the development of further personalized training.

To derive a robust indication of trainee's stress levels a multi-biosignal measurement approach was chosen. HR [61] and HRV [23] will be derived from electrocardiogram (ECG) electrodes worn as belts or sticky electrodes placed on the upper left part of the chest. In addition, EDA will be measured with sticky electrodes placed on the back of trainees. The most common and recommended location for EDA electrodes placement is the hand [38], however, MFR need free hands to treat patients and handle precision equipment. The signal will be further divided into Skin Conductance Response (SCR), the short-term, event-related changes in skin conductance that occur in response to specific external or internal stimuli and Skin Conductance Level (SCL), the underlying, relatively stable level of skin conductance that is present in the absence of specific stimuli. It reflects the general arousal state of the individual and can be influenced by factors such as anxiety, fatigue, or chronic stress levels. The sensor data is processed and streamed to the MR server via the wearable PLUX biosignals sensor system.

**Patient Simulator (ADAM-X)**. The patient simulator manikin ADAM-X adds to the immersive experiences of trainees and provides a tangible tool to practice skills that require haptic feedback (e.g. intubation or resuscitation). The manikin is a high-fidelity reproduction of the human's skeletal and anatomical structure. The manikin's response to treatment and medical data includes blood pressure, heart rate, respiratory rate, chest expansion, cyanosis in face and fingers, pulse, electrocardiogram, CPR frequency, depth and release time, tears, drooling, sweating, eyelids open, closed, blinking, access points touched, and treatments. This data is streamed wirelessly to the VR server system for real-time monitoring and stored for debriefing purpose.

**Trainer Dashboard**. The Trainer Dashboard supports the trainer during the live training, the debriefing process, and setting up new scenarios. During the training session, live feedback on each trainee is limited to current stress levels, vital signs of the patient simulator and selected KPIs due to the limited capacity of processing power within the MR system and the need for offline processing for more sophisticated features. Once the training is completed all data is transferred to the Centric Data Platform for further post-training processing and storage.

The Dashboard Application must be accessible at all times, regardless of whether it is deployed in the cloud or on a server at the training site. However, the data stored on the Centric Data Platform is highly sensitive because it contains personal and health-related information about trainees. Therefore, its design will be container-based and technologies such as Docker[8] will be used to run applications in isolated virtual machines (containers) to reduce overhead, improve portability, consistency, efficiency, and service migration between cloud-based and on-premises deployments. Stored data can be utilized to extract advanced analytics from each training session for visualization in the dashboard's interface and provide recommendations for improved scenario design, the so-called Smart Scenario Design feature, in future training. This involves training machine learning algorithms to enable personalized recommendations for each user.

### *25.3.2 AI Integration*

The proposed system uses machine learning algorithms, including recurrent neural networks, to process data from multiple sources such as manikins, simulations, and wearable biosignal sensors. The unique approach of this AI-driven analysis is to identify correlations and patterns in time-series data and provide personalized insights and recommendations tailored to each individual trainee.

The AI-generated data, such as predictions, recommendations, and explanations are stored in the data architecture, which can be used to improve scenario design and building and smart scenario control. Smart scenario control refers to the use of AI algorithms to manage and control complex systems and scenarios in real-time. It enables the automation of decision-making processes and the optimization of outcomes by analysing vast amounts of data and making predictions based on that data. The AI-powered training can detect interactions between different data sources that may not be visible to the human eye, resulting in better training outcomes.

The objective of the proposed AI implementation is to deliver sophisticated performance insights utilizing cutting-edge AI techniques, but the challenges inherent in its development constrain the selection of methods employed. For example, supervised learning approaches are significantly hindered by the scarcity of training data, as no complete MR training has taken place thus far. Therefore, the focus should shift towards unsupervised learning methods, where clusterization techniques for

[8] https://www.docker.com.

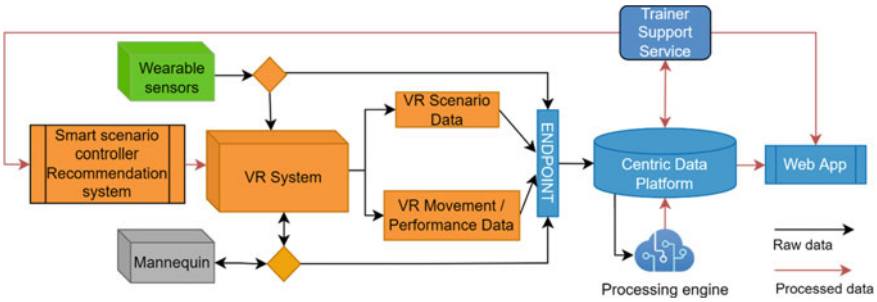



**Fig. 25.2** AI integration in MR training system

anomaly detection can identify unusual events or outliers in the data. For instance, if an anomaly is detected in a trainee's performance data, the system can automatically adjust the difficulty level of the training scenario, provide additional resources or support, or offer personalized recommendations to help the learner improve.

Techniques like Multi-Scale Convolutional Recurrent Encoder Decoder (MSCRED) [65] will be explored to detect and diagnose anomalies in time-series data, focusing on the severity and root cause identification while accounting for potential correlations due to proximity to key events. Correlations may also be identified employing association rule mining algorithms [22]. Other promising approaches focus on sentiment analysis, as implementing speech-to-text algorithms would enable processing communications during training to gauge the sentiment, emotions, or stress levels of the MFRs. Additionally, transfer learning techniques [55] will be applied to leverage pre-trained models and make the most out of the limited available data, accelerating the development of AI tools, especially when the first training data arrives. Building on the aforementioned AI techniques and approaches, the data collected from processing inputs can also be fed back into the MR base system, creating a recommendation system for scenario creation, and live training management (see Fig. 25.2).

To mitigate any potential risks associated with a self-directed AI-based smart scenario control system that has the potential to overwhelm trainees with too challenging and stressful situations, it currently serves as a recommendation system only. The human operator or trainer has the final say in accepting or declining AI-generated suggestions for adaptations of the virtual environments.

### *25.3.3 Explainability*

The efficacy of the presented AI systems relies on the ability of the trainer to understand and follow the suggestions provided by the models. To ensure a dependable relationship between the models and the user, explainable design solutions were implemented using works such as [9, 36, 59]. For instance, a dashboard with the


**Fig. 25.3** Example visualization of the smart scenario control dashboard interface

smart scenario control is provided to the trainer (Fig. 25.3) which presents recommendations to modify the scenario based on a correlation found between a specific event during the simulation and the trainees' response to it. This explanation, together with compelling visual feedback, helps the trainer understand the rationale behind the system's suggestion and decide whether to accept or reject it. Additionally, the output, such as recommendations, explanations, and predictions, is stored in the data architecture for the trainer's reference. These design solutions provide the trainer with transparency and insights into the models' decision-making process, enhancing the trainer's trust and confidence in the AI system.

### *25.3.4 Data Privacy and Ethical Considerations*

The collection and processing of sensitive personal physiological data, as is the case in our MR training system, requires significant ethical and privacy considerations. The protection of trainee privacy and compliance with data protection regulations are paramount in the design and operation of such a system.

In European countries, the General Data Protection Regulation (GDPR) [10] governs the handling of personal data, including its collection, storage, and processing. It requires organizations to obtain informed consent from individuals before collecting their personal data. This means trainees must be informed about how and why their data will be used, must agree to this usage and be able to revoke this agreement at any point.

The GDPR is considered one of the strictest and most comprehensive data protection laws in the world. While other countries and regions have their own data protection laws, many of them have used GDPR as a benchmark or inspiration when creating or updating their own regulations due to its stringent standards and wide scope.

Furthermore, IEEE's Ethically Aligned Design (EAD) [15] guidelines can significantly bolster the ethical handling of personal data in XR training systems. EAD principles prioritize human well-being and data privacy, advocating for transparency and accountability in the design and operation of autonomous systems. By adhering to these guidelines, designers can ensure XR training systems respect users' rights and privacy, fostering trust, and ensuring compliance with data protection norms.

### *25.3.5 Evaluation*

End users will be involved throughout the entire development process, and the system will be tested iteratively to scientifically evaluate their technology experiences and to integrate the results back into the development process. Various evaluation activities will be conducted to assess the system's performance, usability, and validity.

The development and evaluation process will be divided into three phases. In the first phase, we will focus on prototyping and validating the smart wearables for physiological measurements. Only parts of the system will be tested separately in smaller studies, and we will continue with iterative integration in the training system. In the second phase, we will develop the MR training environment during several iterations, accompanied by evaluations by MR experts. The training system will be shipped to Switzerland, Spain, Germany, Sweden, and Austria, where local partners will conduct numerous training sessions with different training scenarios under scientific observation for evaluation. We will evaluate the system's functionality and intuitive deployability in the virtual environment, and at the end of the second phase, the last prototype will be created and tested in a final evaluation.

The final phase of the development and evaluation process will consist of a large-scale field trial to assess the validity of the MR training system. This trial will involve the deployment of the last prototype in a real-world setting, where participants will receive MR-based training and will be evaluated on their performance. This study will provide insights into the development and evaluation of a technology-based training system that combines smart wearables for physiological measurements and a mixed reality training environment and will highlight the importance of involving end users in the development and evaluation process.

## 25.4 Conclusion

This work demonstrates how the integration of AI may enhance the performance of XR training systems. By utilizing AI algorithms to analyse data collected from trainees, the system can support trainers with advanced training performance analysis, including correlations between events, physiological signals, trainee behaviour, and KPIs that may not have appeared obvious to the trainer during the training. Furthermore, a smart scenario control system can provide recommendations on future improvements for scenario design, that is personalized to trainees' previous performance and training progress.

In future work, we plan to use the proposed system in a series of field trials to collect sufficient data to test the various ML approaches. By presenting data and insights in a clear and compelling manner, the system has the potential to highlight trainees' strengths and improvement opportunities as well as the effectiveness of the trainer's current teaching methods and potential improvements through the recommendation system. Enhancing XR training with AI has the potential to improve the quality and acceptance of such relatively novel training systems. Although we discussed a very specific application area, our approach may apply beyond MFR training and could be relevant to a variety of training and performance-enhancing XR applications.

**Funding** This paper was created as part of the MED1stMR project. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 101021775.

## References

1. Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., Radi, N.: Machine learning approaches to predict learning outcomes in massive open online courses. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 713–720. IEEE (2017)
2. Andreatta, P.B., Maslowski, E., Petty, S., Shim, W., Marsh, M., Hall, T., Stern, S., Frankel, J.: Virtual reality triage training provides a viable solution for disaster-preparedness. Acad. Emerg. Med. **17**(8), 870–876 (2010). https://doi.org/10.1111/j.1553-2712.2010.00728.x
3. Baetzner, A.S., Wespi, R., Hill, Y., Gyllencreutz, L., Sauter, T.C., Saveman, B.I., Mohr, S., Regal, G., Wrzus, C., Frenkel, M.O.: Preparing medical first responders for crises: a systematic literature review of disaster training programs and their effectiveness. Scand. J. Trauma Resuscit. Emerg. Med. **30**(1), 76 (2022). https://doi.org/10.1186/s13049-022-01056-8, http://www.ncbi.nlm.nih.gov/pubmed/365662270A, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC9789518
4. Baker, C., Fairclough, S.H.: Adaptive virtual reality. Curr. Res. Neuroadap. Technol. 159–176 (2021). https://doi.org/10.1016/B978-0-12-821413-8.00014-2
5. Bhutoria, A.: Personalized education and artificial intelligence in United States, China, and India: a systematic review using a human-in-the-loop model. Comput. Educ. Artif. Intell. 100068 (2022)
6. Biosignals, P.: Plux Biosignals. https://www.pluxbiosignals.com/

7. Cesta, A., Cortellessa, G., De Benedictis, R.: Training for crisis decision making—an approach based on plan adaptation. Knowl. Based Syst. **58**, 98–112 (2014)
8. Durlach, P.J., Lesgold, A.M.: Adaptive Technologies for Training and Education. Cambridge University Press (2012)
9. Gedikli, F., Jannach, D., Ge, M.: How should I explain? A comparison of different explanation types for recommender systems. Int. J. Hum. Comput. Stud. **72**(4), 367–382 (2014)
10. Gdpr Archives. https://gdpr.eu/tag/gdpr/
11. Gunshin, M., Doi, K., Morimura, N.: Use of high-fidelity simulation technology in disasters: an integrative literature review. Acute Med. Surg. **7**(1) (2020). https://doi.org/10.1002/ams2.596, https://onlinelibrary.wiley
12. Heloir, A., Nunnari, F., Haudegond, S., Havrez, C., Lebrun, Y., Kolski, C.: Design and evaluation of a self adaptive architecture for upper-limb rehabilitation. In: ICTs for Improving Patients Rehabilitation Research Techniques: Second International Workshop, REHAB 2014, Oldenburg, Germany, May 20–23, 2014, Revised Selected Papers 2, pp. 196–209. Springer (2015)
13. Hernando-Gallego, F., Artés-Rodríguez, A.: Individual performance calibration using physiological stress signals. ArXiv abs/1507.03482 (2015)
14. Hubalovsky, S., Hubalovska, M., Musilek, M.: Assessment of the influence of adaptive e-learning on learning effectiveness of primary school pupils. Comput. Hum. Behav. **92**, 691–705 (2019)
15. IEEE. Ethically Aligned Design: Version 2—For Public Discussion, pp. 1–263. IEEE Standards (2017). https://standards.ieee.org/industry-connections/ec/ead-v1/
16. Iván, C., Reyes, A., Wozniak, D., Ham, A., Zahabi, M.: An Adaptive Virtual Reality-Based Training System for Pilots, pp. 1962–1966 (2022). https://doi.org/10.1177/1071181322661063
17. Jeon, S.G., Han, J., Jo, Y., Han, K.: Being more focused and engaged in firefighting training: applying user-centered design to VR system development. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology. VRST (2019). https://doi.org/10.1145/3359996.3364268
18. Kelley, C.R.: What is adaptive training? Human Factors **11**(6), 547–556 (1969). https://doi.org/10.1177/001872086901100602
19. Kerr, P.: Adaptive learning. Elt J. **70**(1), 88–93 (2016)
20. Kizony, R., Katz, N., Weiss, P.L.: Adapting an immersive virtual reality system for rehabilitation. J. Vis. Comput. Animat. **14**(5), 261–268 (2003)
21. Koutitas, G., Smith, S., Lawrence, G.: Performance evaluation of AR/VR training technologies for ems first responders. Virtual Real. **25**, 83–94 (2021)
22. Kumbhare, T.A., Chobe, S.V.: An overview of association rule mining algorithms. Int. J. Comput. Sci. Inf. Technol. **5**(1), 927–930 (2014)
23. Laborde, S., Mosley, E., Thayer, J.F.: Heart rate variability and cardiac vagal tone in psychophysiological research—recommendations for experiment planning, data analysis, and data reporting. Front. Psychol. **8**, 213 (2017)
24. Landsberg, C.R., Astwood, R.S., Jr., Van Buskirk, W.L., Townsend, L.N., Steinhauser, N.B., Mercado, A.D.: Review of adaptive training system techniques. Mil. Psychol. **24**(2), 96–113 (2012)
25. Lang, Y., Wei, L., Xu, F., Zhao, Y., Yu, L.F.: Synthesizing personalized training programs for improving driving habits via virtual reality. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 297–304. IEEE (2018)
26. Lima, R., Asif, M., Sousa, H., Bermu´dez i Badia, S.: Adaptive control of cardiorespiratory training in a virtual reality hiking simulation: a feasibility study, 91–99 (2022). https://doi.org/10.5220/0011004400003123
27. Lüddecke, R., Felnhofer, A.: Virtual reality biofeedback in health: A scoping review. Appl. Psychophysiol. Biofeedback **47**(1), 1–15 (2022). https://doi.org/10.1007/s10484-021-09529-9
28. Ma, M., McNeill, M., Charles, D., McDonough, S., Crosbie, J., Oliver, L., McGoldrick, C.: Adaptive virtual reality games for rehabilitation of motor disorders. In: Universal Access in Human-Computer Interaction. Ambient Interaction: 4th International Conference on Universal

Access in Human-Computer Interaction, UAHCI 2007 Held as Part of HCI International 2007 Beijing, China, July 22–27, 2007 Proceedings, Part II 4. pp. 681–690. Springer (2007)
29. Maity, S.: Identifying opportunities for artificial intelligence in the evolution of training and development practices. J. Manage. Dev. (2019)
30. Matos, P., Rocha, J., Gonçalves, R., Almeida, A., Santos, F., Abreu, D., Martins, C.: Smart coach—a recommendation system for young football athletes. In: Ambient Intelligence—Software and Applications—, 10th International Symposium on Ambient Intelligence, pp. 171–178. Springer (2020)
31. McCarthy, J.E.: Military applications of adaptive training technology. In: Technology Enhanced Learning: Best Practices, pp. 304–347 (2008)
32. McFetrich, J.: A structured literature review on the use of high fidelity patient simulators for teaching in emergency medicine. Emerg. Med. J. **23**(7), 509–511 (2006). https://doi.org/10.1136/emj.2005.030544
33. Medical-X. Adam-x Patient Simulator—Medical-x (2021). https://medical-x.com/product/adam-x/
34. Mills, B., Dykstra, P., Hansen, S., Miles, A., Rankin, T., Hopper, L., Brook, L., Bartlett, D.: Virtual reality triage training can provide comparable simulation efficacy for paramedicine students compared to live simulation-based scenarios. Prehospital Emerg. Care **24**(4), 525–536 (2020). https://doi.org/10.1080/10903127.2019.1676345 (PMID: 31580178)
35. Munoz, J.E., Cao, S., Boger, J.: Kinematically adaptive exergames: Personalizing exercise therapy through closed-loop systems. In: Proceedings—2019 IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR 2019, pp. 118–125 (2019). https://doi.org/10.1109/AIVR46125.2019.00026
36. Nunes, I., Jannach, D.: A systematic review and taxonomy of explanations in decision support and recommender systems. User Model. User Adap. Inter. **27**, 393–444 (2017)
37. Paranjape, K., Schinkel, M., Panday, R.N., Car, J., Nanayakkara, P., et al.: Introducing artificial intelligence training in medical education. JMIR Med. Educ. **5**(2), e16048 (2019)
38. Payne, A.F., Dawson, M.E., Schell, A.M., Singh, K., Courtney, C.G.: Can you give me a hand? A comparison of hands and feet as optimal anatomical sites for skin conductance recording. Psychophysiology **50**(11), 1065–1069 (2013)
39. Pham, T., Roland, L., Benson, K.A., Webster, R.W., Gallagher, A.G., Haluck, R.S.: Smart tutor: a pilot study of a novel adaptive simulation environment. Stud. Health Technol. Inf. **111**, 385–389 (2005)
40. Point, N.: Motion Capture Systems. https://optitrack.com/
41. Pretolesi, D.: Personalised training: integrating recommender systems in XR training platforms. In: Marky, K., Gru¨nefeld, U., Kosch, T. (eds.) Mensch und Computer 2022—Workshopband. Gesellschaft für Informatik e.V., Bonn (2022)
42. Purwoningsih, T., Santoso, H.B., Hasibuan, Z.A.: Online learners' behaviors detection using exploratory data analysis and machine learning approach. In: 2019 Fourth International Conference on Informatics and Computing (ICIC), pp. 1–8. IEEE (2019)
43. Qomariyah, N.N., Fajar, A.N.: Recommender system for e-learning based on personal learning style. In: 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 563–567. IEEE (2019)
44. Rasulova, N.Y., Karimova, M.: Adaptive training systems as a tool for personalized training implementation in universities. Int. J. Conf. Ser. Educ. Social Sci. **1** (2021) (Online)
45. REFENSE: Refense–VR training solutions for professionals. https://www.refense.com/
46. Regal, G., Murtinger, M., Schrom-Feiertag, H.: Augmented CBRNE responder-directions for future research. In: 13th Augmented Human International Conference, pp. 1–4 (2022)
47. Santos, O.C.: Training the body: the potential of AIED to support personalized motor skills learning. Int. J. Artif. Intell. Educ. **26**(2), 730–755 (2016)
48. Schneeberger, M., Paletta, L., Wolfgang Kallus, K., Reim, L., Sch¨onauer, C., Peer, A., Feischl, R., Aumayr, G., Pszeida, M., Dini, A., Ladst¨atter, S., Weber, A., Almer, A., Wallner, D.: First responder situation reporting in virtual reality training with evaluation of cognitive-emotional stress using psychophysiological measures. Cogn. Comput. Internet Things **43** (2022). https://doi.org/10.54941/ahfe1001841

49. Schwaninger, A., Hofer, F., Wetter, O.E.: Adaptive computer-based training increases on the job performance of x-ray screeners. In: 2007 41st annual IEEE International Carnahan Conference on Security Technology, pp. 117–124. IEEE (2007)
50. Siu, K.C., Best, B.J., Kim, J.W., Oleynikov, D., Ritter, F.E.: Adaptive virtual reality training to optimize military medical skills acquisition and retention. Mil. Med. **181**(Suppl 5), 214–220 (2016)
51. Spoon, K., Beemer, J., Whitmer, J.C., Fan, J., Frazee, J.P., Stronach, J., Bohonak, A.J., Levine, R.A.: Random forests for evaluating pedagogy and informing personalized learning. J. Educ. Data Mining **8**(2), 20–50 (2016)
52. Srivastava, R., Palshikar, G.K., Chaurasia, S., Dixit, A.: What's next? A recommendation system for industrial training. Data Sci. Eng. **3**(3), 232–247 (2018)
53. Taguma, M., Feron, E., Lim, M.H.: Future of education and skills 2030: conceptual learning framework. In: Organization of Economic Co-operation and Development (2018)
54. Technologies, U.: Unity–unity (2023). https://unity.com/
55. Torrey, L., Shavlik, J.: Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pp. 242–264. IGI global (2010)
56. TrueVR: Virtual reality. https://www.truevrsystems.com/
57. Vanbecelaere, S., Van den Berghe, K., Cornillie, F., Sasanguie, D., Reynvoet, B., Depaepe, F.: The effectiveness of adaptive versus non-adaptive learning with digital educational games. J. Comput. Assist. Learn. **36**(4), 502–513 (2020)
58. Vaughan, N., Gabrys, B., Dubey, V.N.: An overview of self-adaptive technologies within virtual reality training. Comput. Sci. Rev. **22**, 65–87 (2016)
59. van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating XAI: A comparison of rule-based and example-based explanations. Artif. Intell. **291**, 103404 (2021)
60. Wray, R.E., Woods, A.: A cognitive systems approach to tailoring learner practice. In: Proceedings of the Second Annual Conference on Advances in Cognitive Systems ACS, vol. 21, p. 18 (2013)
61. Wunsch, K., Kasten, N., Fuchs, R.: The effect of physical activity on sleep quality, well-being, and affect in academic stress periods. In: Nature and Science of Sleep, pp. 117–126 (2017)
62. Zahabi, M., Abdul Razak, A.M.: Adaptive virtual reality-based training: a systematic literature review and framework. Virtual Reality **24**(4), 725–752 (2020). https://doi.org/10.1007/s10055-020-00434-w
63. Zahabi, M., Abdul Razak, A.M.: Adaptive virtual reality-based training: a systematic literature review and framework. Virtual Reality **24**, 725–752 (2020)
64. Zechner, O., Kleygrewe, L., Jaspaert, E., Schrom-feiertag, H., Hutter, R.I.V., Tscheligi, M.: Enhancing Operational Police Training in High Stress Situations with Virtual Reality: Experiences, Tools and Guidelines (2023)
65. Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., Chawla, N.V.: A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 1409–1416 (2019)

# Chapter 26
# A Study on the Integration of Bim and Mixed Reality in Steel-Structure Maintenance


**Yi-Jao Chen, Hong-Lin Chiu, and Tzu-Hsiang Ger**

**Abstract** In recent years, there have been numerous successful cases of three-dimensional (3D) multimedia, augmented reality (AR)/virtual reality (VR), and mixed reality (MR) applications in information communication. Additionally, building information modeling (BIM) has become the primary tool for integrating and disseminating project information within the construction industry. However, there are still many challenges related to information access and accuracy at construction sites. This study combines BIM and MR technologies to address these limitations and focuses on complex steel-structure maintenance engineering. The project information compiled in the BIM model is transformed through information conversion and made accessible to head-mounted displays using the MR application developed in this study. Case validation has demonstrated that the combination of BIM and MR technologies can reduce the time required to create MR content in construction projects and solve the problem of difficult information access at construction sites. Furthermore, the simulation analysis of steel-structure deterioration and environmental factors presented through MR can serve as decision support for steel-structure maintenance management.

## 26.1 Introduction

In recent years, with the rapid development of building information modeling (BIM), the application and support of BIM as the core of project life cycle information integration in the design and construction stages have been effective. Through the component attribute information and visualization features of the BIM architecture information model [1], it can present the geometric appearance, spatial concept, and

Y.-J. Chen (✉) · H.-L. Chiu
National University of Kaohsiung, 700, Kaohsiung University Rd., Kaohsiung 811, Taiwan
e-mail: yjchen@go.nuk.edu.tw

T.-H. Ger
National Science and Technology Museum, No. 720, Jiuru 1st Rd., Kaohsiung City 807, Taiwan



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_26

attribute data of rough objects such as facilities, windows, and other building components. Although many studies have confirmed the benefits of BIM for various stages of the construction project life cycle, in practical applications, due to the limitations of the construction project operating environment and equipment, the construction site still cannot meet the needs of information access immediacy, accuracy, and convenience [2]. In addition, the related applications of BIM still focus on the design and construction stages, and there are few studies on the feasible solutions for BIM in maintenance engineering [3].

Taking steel-structure engineering as an example, in practice, traditional methods are still used for recording painting degradation in steel-structure maintenance engineering. The integration and convenience of management information construction and acquisition in the benefits assessment of operation and maintenance stages are lacking, and due to the difficulty in accessing management information, preventive maintenance cannot be performed through historical records interpretation, making it difficult to achieve the best state in maintenance cost and effectiveness.

To solve the problems that BIM cannot provide on-site information collaboration and cannot reflect the actual operating conditions, related research began to try to use AR, VR, MR, and other technologies to perceive information about the real environment through the sensors of immersive technology equipment, and provide the effect of virtual and real mixed reality [4, 5]. Mixed reality (MR) is a comprehensive technology that combines elements of both virtual reality (VR) and augmented reality (AR). It creates a seamless integration between the virtual and real world, allowing users to interact with and perceive both digital content and physical surroundings simultaneously. In MR experiences, virtual objects or digital information are overlaid onto the real world, enabling users to interact with and manipulate virtual elements within their physical environment. Related research puts the BIM model information into the real environment, allowing operators to interact with the virtual model to assist operators in correctly interpreting engineering information and achieve the ability of BIM to be applied in real-time on-site [6]. However, in practice, there are fewer analyses based on operating scenarios to construct back-end databases that meet relevant operations [7]. At the same time, there is still a lack of data exchange processes between mixed reality equipment and modeling software, so the problem of poor information transmission between BIM and the site has not been effectively solved.

This study proposes to combine BIM architecture information modeling with MR mixed reality technology, taking steel-structure maintenance engineering as an example, to establish a steel-structure maintenance engineering support system. It is expected to expand the application mode of maintenance management under the BIM development architecture, and achieve the completeness, mobility, and effectiveness of engineering information access through immersive technology assistance, as well as the immediacy, visibility, and convenience of information presentation, to overcome the inconvenience of information access on the construction site and reduce maintenance costs and improve efficiency.

## 26.2 Research Method

To effectively assist the execution of steel-structure maintenance projects, this research began with an information requirement analysis. Through procedures such as literature review, and expert and operator interviews, the execution processes and information content of each step were summarized, and steel-structure BIM components were developed. Relevant maintenance project data were described in the property panel, recording information such as the number, type, degree of rust, and rust photos of each steel component (see Fig. 26.1). Each component in the steel-structure BIM model is an individual independent component, and the damage of each part is presented separately in the property panel. The degree of steel corrosion is recorded in five levels: "0. None," "1. Partial paint loss," "2. Severe paint loss," "3. Partial rust," and "4. Severe rust."

The BIM model established in this study is a valuable aid for steel-structure maintenance management, offering various functions to enhance the overall process. Firstly, the BIM model is utilized to extract steel-structure component information and estimate the quantity of maintenance work, thus improving the accuracy of material estimation. Secondly, by conducting 4D scheduling simulations, the model facilitates engineering operation coordination and helps to minimize construction operation conflicts. Finally, the BIM model enables physical environment simulation analysis, which explores the correlation between steel-structure building deterioration and climate and environmental factors. This analysis provides crucial decision support for long-term maintenance management.

In addition, to enhance on-site operations in steel-structure maintenance, this study integrates MR technology with the BIM framework. The aim is to create an MR-assisted interactive content platform that combines the advantages of BIM and

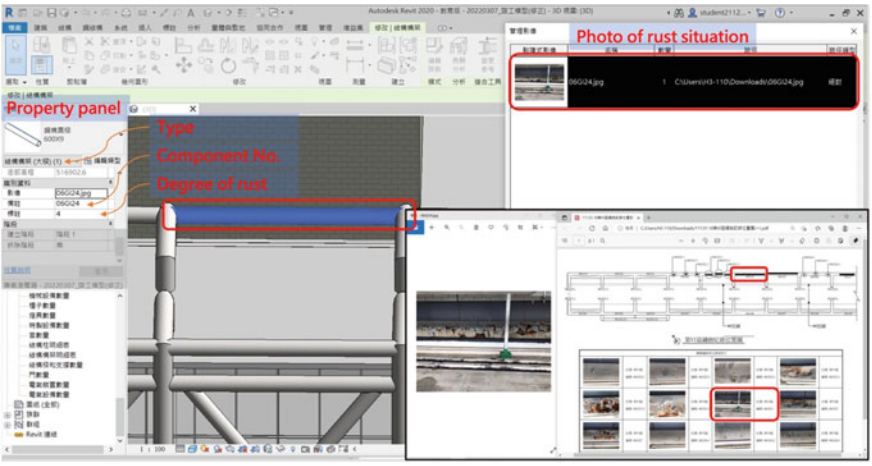



**Fig. 26.1** Attributes recording for each steel BIM component

immersion technologies. This study utilizes Microsoft HoloLens as the MR device for presenting BIM information content. The developed platform offers a range of functionalities specifically tailored for steel-structure maintenance operations. Firstly, the BIM model is exported as separate FBX and project data files. The open development plug-in MRTK is then employed to build MR operation functions. These functions enable seamless interaction with the MR-assisted platform, providing a user-friendly interface for on-site personnel.

Key features of the MR-assisted platform include a steel-structure maintenance system operation interface, QR code positioning, and model download function. These features facilitate easy access to specific components for maintenance purposes. The platform also enables the manipulation of the model, allowing users to turn components on or off and initialize the model as needed. Additionally, component selection and color-changing functions enhance visual identification and assist in maintenance operations. Detailed component information can be accessed through the platform, providing comprehensive data for decision-making. Furthermore, the platform includes a rust image function, enabling the tracking and documenting of deterioration over time. The construction video function provides guidance and reference during maintenance operations. Environmental factors are also considered, with sun radiation level grading and coloring functions to analyze the impact on steel-structures. Rust degree grading and coloring functions assist in assessing the condition of the structures. Lastly, the platform allows users to take photos and record videos, enabling the capture of important moments and documentation of maintenance activities.

By integrating these functionalities, the MR-assisted platform enhances on-site operations in steel-structure maintenance, improving efficiency and accuracy. It is a valuable tool for decision-making, data analysis, and comprehensive documentation in maintenance management.

## 26.3 System Development and Case Validation

Current building steel-structure maintenance management still relies heavily on manual paper-based recording and management, which not only lacks efficiency but also prone to errors. As a result, the decision-making support provided by operation and maintenance is limited. In large-scale steel-structure maintenance, it is difficult to provide suitable repair methods according to different levels of deterioration to maintain the normal operation of the building. Therefore, this study expects to use the BIM building information model as the core for collecting building steel-structure component information. In addition to providing functions such as maintenance project quantity estimation, project 4D scheduling simulation, and steel-structure deterioration analysis, this study further uses mixed reality (MR) to establish MR-assisted interactive functions for steel-structure maintenance engineering. It is hoped that under the current development structure of BIM, the feasible mode of combining and extending BIM and immersive technology can be further expanded to enhance

the application value of BIM models in the construction site by allowing engineering personnel to interact with virtual image information at the construction site.

Taking the exterior steel-structure maintenance project of a museum in southern Taiwan as an example (see Fig. 26.2), this study conducted project management for corrosion and painting deterioration maintenance of the exterior steel-structure. Based on the results of the project information needs analysis, BIM components were used to record various attribute information, such as location, angle, cross-section specification, length, component code, and corrosion level. The components were also linked to information such as corrosion deterioration photos and maintenance project implementation videos to assist the maintenance unit in making decisions for routine and preventive maintenance. The BIM model of the case study is shown in Fig. 26.3.

**Fig. 26.2** Case study of exterior steel-structure maintenance project

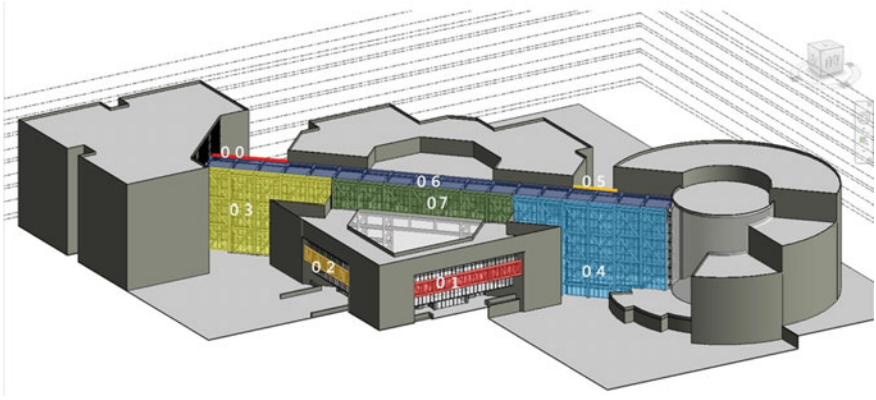



**Fig. 26.3** The BIM model of the case study

### 26.3.1 BIM Model for Assisting Project Management

The project BIM model in this study was built according to the Level of Development (LOD) development specification published by BIM FORUM in 2019. The scope of the steel-structure maintenance project in the project belongs to "Metal Building Systems—Primary Framing and Support." According to the information requirements of the maintenance project, LOD300 was used as the standard for component construction. The BIM component attribute table describes individual steel-structure components, including component codes, models, corrosion conditions, and photos of corrosion conditions. The corrosion condition of the steel-structure is recorded in five levels based on the degree of corrosion: "0. None," "1. Local peeling," "2. Severe peeling," "3. Local corrosion," and "4. Severe corrosion." The project BIM model is divided into two parts: the main structure and the steel-structure. The main structure presents a volumetric model of the building's appearance, and the steel-structure is divided into eight areas based on the partition of the steel-structure maintenance project (see Fig. 26.3). After completing the information construction of the BIM model, in order to avoid the loss of non-geometric attribute information in the BIM components, this study used Autodesk Dynamo to write data extraction rules and establish a project management database (see Fig. 26.4), providing functions for engineering quantity calculation, 4D scheduling simulation, and deterioration analysis of steel-structure components to assist in steel-structure maintenance project management. This research utilized Dynamo to extract BIM component information in Revit software to utilize BIM model information with MR devices fully. Rules for data extraction were programmed using Dynamo, and the extracted BIM component information was used as the information source for developing MR application tools using Unity programming in later stages.

### 26.3.2 Application of Mixed Reality in Steel-Structure

After conducting interviews with maintenance engineering personnel and reviewing relevant literature, the study identified several challenges in information management and construction operations during the maintenance and operation stages. These challenges can be summarized as follows:

1. Insufficient intuition in traditional drawing review and discussion: There needs to be a more intuitive understanding when reviewing and discussing traditional drawings, which can hinder effective decision-making and problem-solving.
2. Time-consuming use and comparison of maintenance information on-site: Accessing and comparing maintenance information on-site is time-consuming, resulting in delays and potential errors in decision-making.
3. Need for BIM models to be overlaid with the site environment: Integrating BIM models with the actual site environment is necessary to provide a comprehensive

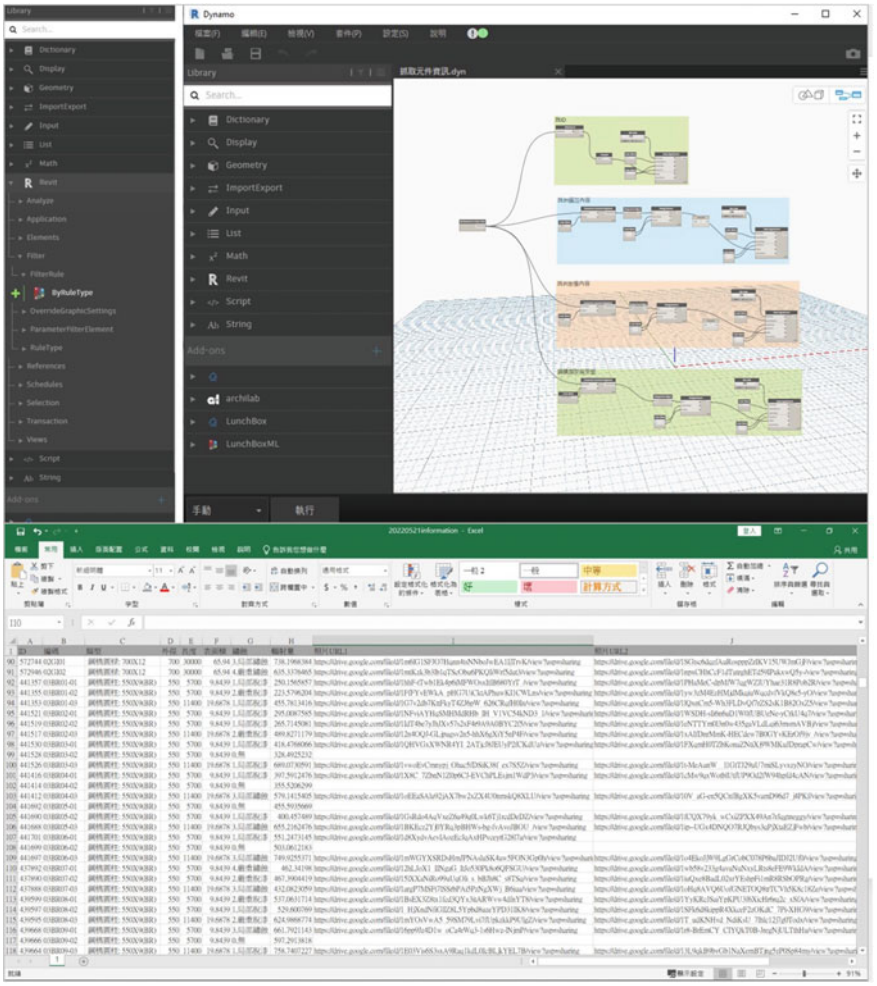


**Fig. 26.4** Using dynamo to extract and export model information

and accurate representation of the project, but it poses challenges in alignment and synchronization.

4. Complex transmission of construction technology and experience gap leading to construction errors: It is difficult to effectively transmit and bridge the experience gap between different personnel, resulting in construction errors and inconsistencies.
5. Poor communication and coordination in remote locations: Remote locations often need help with communication and coordination, leading to delays, miscommunications, and reduced productivity.

By recognizing and understanding these challenges, the study aims to develop solutions that address these issues and improve the efficiency and effectiveness of maintenance engineering processes.

To smoothly transmit project data from a cloud database to an MR device, this study converts .fbx files exported from BIM models into AssetBundle files, sets the QR code corresponding location, and stores them in the cloud database. AssetBundle is a resource compression file that can store data such as models, materials, sound files, and scenes for use in the Unity environment. Furthermore, a script is written to transmit the steel-structure files in the cloud database to the MR device after scanning the QR code, thereby reducing the memory and storage limitations of the device.

Based on these needs, this study established a collaborative model for applying mixed reality in maintenance engineering. Unity was used to set up the Holoans environment, and scene and function scripts were built using Unity and Visual Studio to apply Hololens devices in steel-structure maintenance engineering. The system includes:

1. QR code positioning and model download (see Fig. 5a).
2. Model on/off and initialization (see Fig. 5b).
3. Building model interaction (see Fig. 5c).
4. Steel-structure component information (see Fig. 6a).
5. Component photos (see Fig. 6b).
6. Degree of corrosion (see Fig. 6c).
7. Rust treatment construction videos.
8. Radiation level simulation.

When used at the construction site, users first use QR code positioning and download the building model, overlaying the virtual model with the real object to enhance intuition in obtaining information and identifying components. By loading the model with the URL of the cloud model library, the load caused by large model files can be effectively reduced (see Fig. 26.5). At the same time, users can use gesture

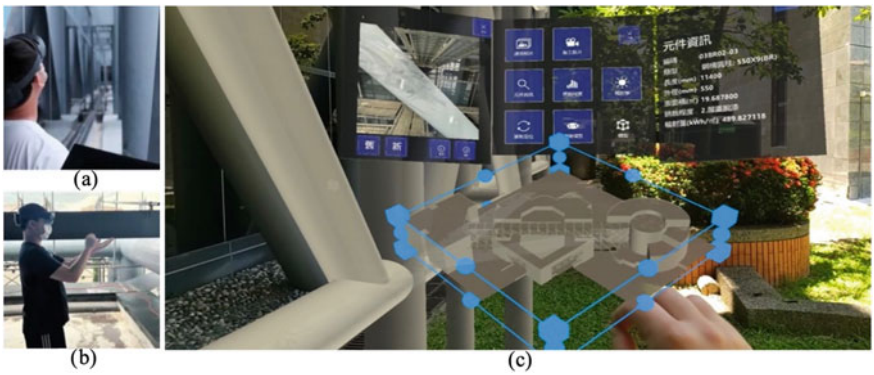



**Fig. 26.5** Combining BIM and MR application in HoloLens steel-structure environment. **a** QR code positioning and model download. **b** Model on/off and initialization. **c** Building model interaction

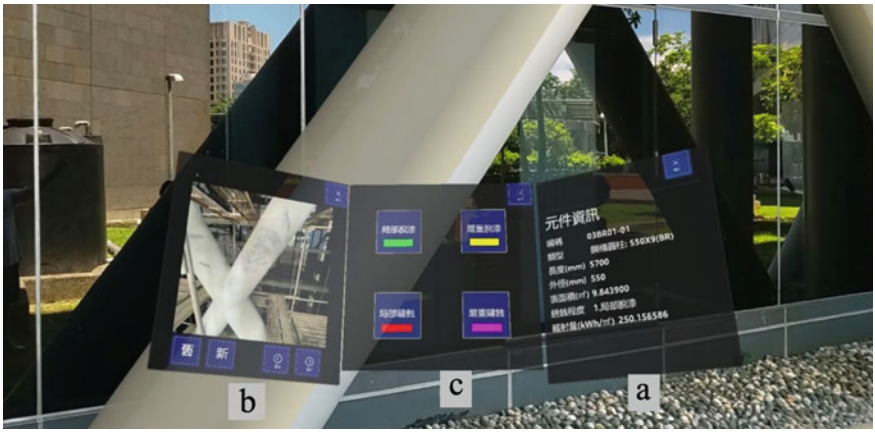



**Fig. 26.6** Mixed reality assists steel-structure maintenance and engineering management. **a** Steel-structure component information. **b** Component photos. **c** Degree of corrosion

sensing to perform model interactive operations, assisting in the interpretation of the overall environment. In terms of information display assistance, users can query steel-structure component size specifications, corrosion levels, solar radiation levels, and other information by clicking on components (see Fig. 26.6), as well as provide information such as steel-structure maintenance record videos (see Fig. 26.7). In addition, to understand the impact of environmental factors on the rust severity of components, this study examines the correlation between solar irradiance levels and the surface rust severity of steel components. The rust severity is classified into different levels, represented by the colors green, yellow, red, and magenta, indicating partial peeling, severe paint loss, partial rusting, and severe rusting, respectively (see Fig. 8a). Solar irradiance levels are depicted using a color spectrum to represent the intensity of sunlight (see Fig. 8b). The graphical representation demonstrates a correlation between rust severity and solar radiation levels, where higher solar radiation leads to more severe rusting. Therefore, utilizing BIM model-based solar radiation simulation can serve as a decision-making reference for future maintenance strategies of steel components.

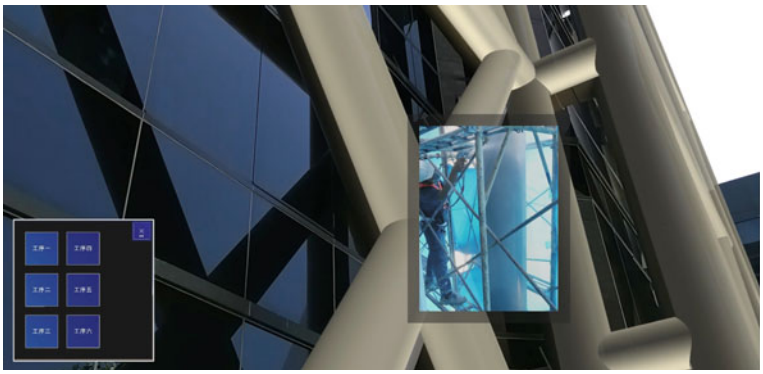
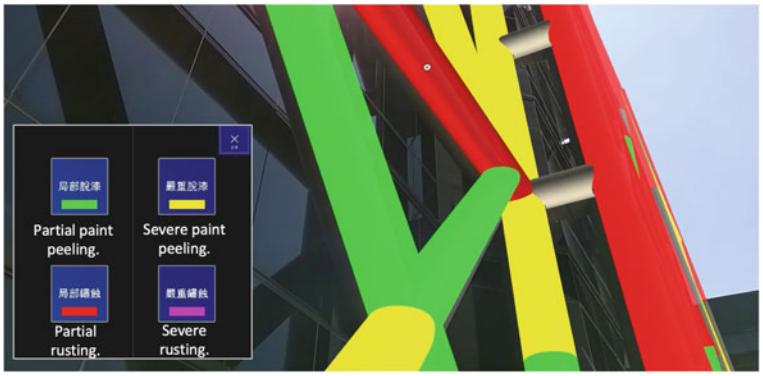
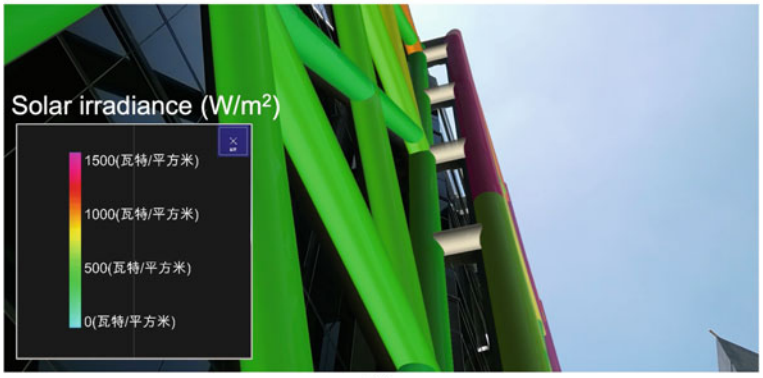


**Fig. 26.7** Steel-structure maintenance record videos displayed in MR device




(a)




(b)

**Fig. 26.8** Analysis of the correlation between solar irradiance levels and rust severity

## 26.4 Conclusion

To conclude, this study demonstrated the successful integration of BIM and MR technology in the development of an assistance system for steel-structure maintenance engineering. The system effectively addressed the challenges of real-time and accurate information access at construction sites by utilizing MR devices to present virtual images and physical objects. This enhanced on-site coordination and information delivery. Furthermore, the simulation and analysis capabilities of the system provided valuable decision-making support for steel-structure maintenance management, particularly regarding deterioration and environmental factors. Moving forward, future research should focus on the advancement of BIM transfer software to improve the efficiency of combining BIM and MR applications. Additionally, exploring collaborative functions among multiple MR devices would enhance the practical utility of the system, further empowering maintenance engineers in their tasks. By continuously refining and expanding the capabilities of this assistance system, the field of steel-structure maintenance engineering can benefit from improved efficiency, accuracy, and decision-making capabilities.

**Acknowledgements** This work was supported by the National Science and Technology Council, Taiwan under Grant MOST 111-2221-E-390-008.

## References

1. Eastman, C.: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers, and Contractors (2011)
2. Chu, M., Matthews, J., Love, P.E.D.: Integrating mobile building information modelling and augmented reality systems: an experimental study. Autom. Constr. **85**, 305–316 (2018). https://doi.org/10.1016/j.autcon.2017.10.032
3. Wang, Y., Wang, X., Wang, J., Yung, P., Jun, G.: Engagement of facilities management in design stage through BIM: framework and a case study. Adv. Civ. Eng. **2013**, 1–8 (2013). https://doi.org/10.1155/2013/189105
4. Bae, H., Golparvar-Fard, M., White, J.: High-precision vision-based mobile augmented reality system for context-aware architectural, engineering, construction and facility management (AEC/FM) applications. Visualization Eng. **1**(1), 3 (2013). https://doi.org/10.1186/2213-7459-1-3
5. Hou, L., Wang, X., Bernold, L., Love, P.E.D.: Using animated augmented reality to cognitively guide assembly. J. Comput. Civ. Eng.Comput. Civ. Eng. **27**(5), 439–451 (2013). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000184
6. Alizadehsalehi, S., Hadavi, A., Huang, J.C.: From BIM to extended reality in AEC industry. Autom. Constr. **116**, 103254 (2020). https://doi.org/10.1016/J.AUTCON.2020.103254
7. El Ammari, K., Hammad, A.: Remote interactive collaboration in facilities management using BIM-based mixed reality. Autom. Constr. **107**, 102940 (2019). https://doi.org/10.1016/J.AUTCON.2019.102940

# Chapter 27
# Pathway-Based Analysis Using SVM-RFE for Gene Selection and Classification

**Nurazreen Afiqah A. Rahman, Nurul Athirah Nasarudin, and Mohd Saberi Mohamad**

**Abstract** The pathway-based analysis is one method for selecting and classifying genes by incorporating pathway information. Integration of pathway knowledge into microarray data significantly advances researchers in the analysis of complex diseases. Microarray data involves thousands of genes to be selected, and therefore, a suitable method to eliminate noisy and uninformative genes is needed. Selecting significant genes for a specific disease is crucial to identifying genes highly related to disease production. Previous research shows that using both pathway information and gene expression data is more significant in disease identification. Therefore, pathway-based analysis using Support Vector Machine Recursive Feature Elimination (SVM-RFE) is introduced in this research to identify significant genes associated with analyzing the targeted phenotype. The datasets involved in this research are lung cancer and gender dataset. The results from the proposed method performed better than previous work as the significant genes are selected from the highest rank of genes in the highest rank of the pathway. The performance of the proposed method was evaluated using 10-fold cross-validation in terms of accuracy. Finally, a biological validation was conducted on selected genes in the top 5 pathways based on biological literature.

N. A. A. Rahman
Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor, Malaysia

N. A. Nasarudin · M. S. Mohamad (✉)
Health Data Science Lab, Department of Genetics and Genomics, College of Medicine and Health Science, United Arab Emirates University, P.O. Box 17666, Al Ain, Abu Dhabi, United Arab Emirates
e-mail: saberi@uaeu.ac.ae



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_27

## 27.1 Introduction

Gene selection is the process of selecting informative genes from a high-dimensional dataset and eliminating irrelevant and redundant genes [1]. It is mainly used to select significant genes which enhance the classification method. Gene selection aims to select a small subset of genes from a larger pool, rendering not only good classification performance and biologically meaningful [2].

SVM-RFE is a wrapper-based method of gene selection. This method starts with all the potential genes involve and remove one at a time in a sequential backward elimination procedure. At each round, the genes are ranked based on their features. SVM-RFE is performed to achieve better classification performance and lead to better generalization compared to previously unseen patterns.

Pathway analysis has ushered in a fresh era in genomic research by presenting more comprehensive biological process information when compared to the traditional approach of analyzing individual genes. Despite its advantages, this method poses challenges to researchers, particularly regarding the quality of pathway data itself. Typically, pathway data is defined without considering a specific biological context, resulting in only a few genes within the pathways being responsible for the corresponding cellular process. This can lead to the inclusion of uninformative genes in some pathways while excluding informative ones. Furthermore, numerous algorithms used in pathway analysis must address these limitations by considering all the genes within pathways as significant.

Identifying significant genes for specific cancer is crucial to tackling and modifying genes highly related to cancer cell production. Since there are thousands of genes associated with a certain disease interconnected to one another pathway, a gene mapping method is proposed. This method is also known as pathway-based analysis, where genes are grouped according to their pathway.

Without gene mapping, a more significant number of potential genes might be selected. This is because the genes selected did not undergo gene ranking and error generalization processes during analysis. In general, a better gene selection method can be done by analyzing genes according to their pathway and selecting genes according to their rank.

Gene selection and classification methods are required to select informative genes and produce output with less noisy data. Integrating SVM-RFE with pathway-based analysis can produce a more accurate result as this method eliminates the less significant genes step by step and ranks genes accordingly by each pathway. This research aims to develop a pathway-based analysis using SVM-RFE method for gene selection and classification to identify significant genes associated with specific cancer diseases. By identifying the significant genes from the highest rank of the pathway in this newly proposed method, highly associated genes to specific cancer could be identified.

## 27.2 Methods and Materials

### *27.2.1 Pathway-Based Analysis*

Pathway-based analysis has been introduced to overcome the limitation of GSA methods by considering the pathway structure. This approach combines the advantages of GSA while leveraging information from gene–gene interactions within the pathway database. Two hypothesis tests are employed in this analysis: first, it tests entire pathways for differential expression, and second, it identifies informative paths that carry substantial information regarding the differential expression for the entire pathway. Consequently, researchers can accurately pinpoint the associated pathways linked to a biological condition relevant to the targeted phenotype. Prior research demonstrates that the pathway structure offers valuable biological insights and contributes to the understanding of higher-order functions within biological systems.

### *27.2.2 SVM-RFE*

According to [3], the Support Vector Machine Recursive Feature Elimination (SVM-RFE) technique was initially suggested for gene selection in cancer classification. It follows a sequential backward elimination process, where subsets of features are selected and one feature is removed at a time, resulting in a ranking of all the feature variables. The feature ranking criterion is based on the coefficients of the weight vector w from a linear SVM at each step.

To enhance speed, the algorithm can be adapted to eliminate multiple features in each step. Nonetheless, this approach of removing several features at once may negatively impact the classification performance [3]. Consequently, if the elimination process removes only one gene at each step, the resulting subset of genes that remains should be the most informative ones [1].

With this method, we can identify nested subsets of genes that are suitable for a model selection technique aimed at finding the optimal number of genes. This suggests that RFE is significantly more resistant to data overfitting compared to other approaches, such as combinatorial search [4].

### *27.2.3 Proposed Method*

Based on the flowchart in the Fig. 27.1, the improvement made for this research is by adding gene mapping into pathways. The additional step enables a better selection of genes from every pathway. Furthermore, the new flowchart proposed includes

performance measurement using a tenfold CV to choose the top five highest ranks of the pathway based on accuracy.

Gene mapping is a process of reading pathway information containing gene ID and simultaneously searching the genes according to their ID from microarray data. Gene mapping into pathways is done using R, and output is translated into a folder and Microsoft Excel 2010. To measure the performance, an error generalization method is performed whereby the error rate for each pathway is calculated. The error generalization used is tenfold cross-validation. In tenfold CV accuracy, the samples

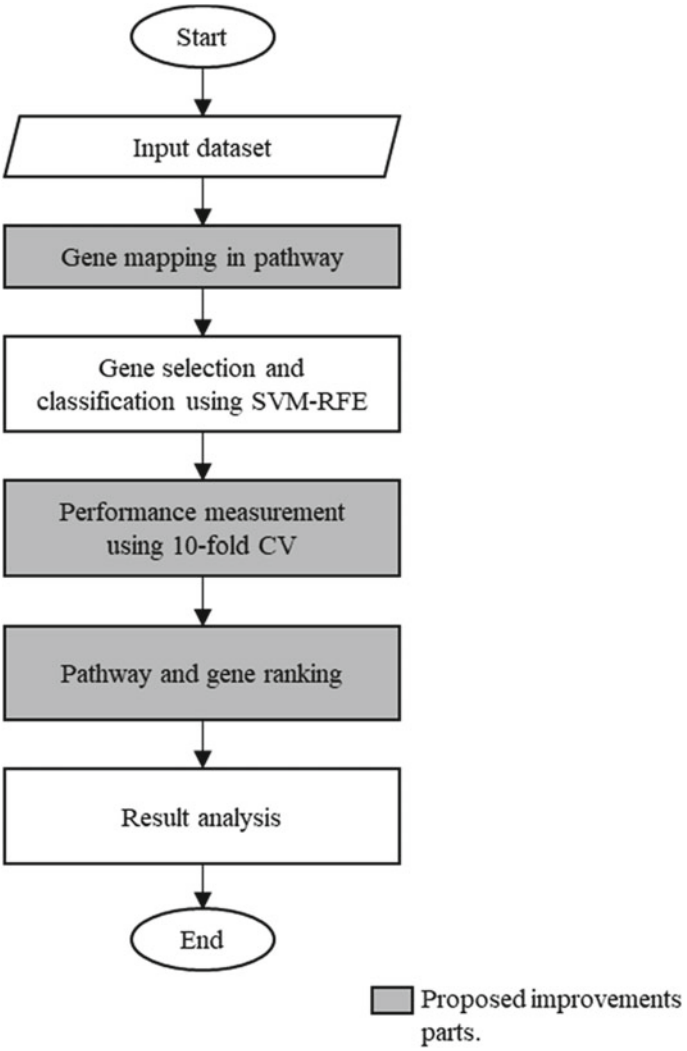



**Fig. 27.1** Flowchart of the proposed method

are divided into ten partitions, in which nine out of ten partitions of samples are treated as a training set while the left one is used as a validation set. This experiment is executed ten times. Tenfold CV accuracy is executed with the equation as stated below:

$$10 - \text{foldsCVAccuracy} = \frac{1}{10}\sum_{1}^{10} \text{Acc}_i, \tag{27.1}$$

where $\text{Acc}_i$ is the single accuracy executed in $i$th experiment. Next, pathway and gene ranking is performed where the top five highest-ranking pathways are selected based on their accuracy, and informative genes in each pathway are chosen. The highest-ranking gene in each pathway is the most significant gene ranked by SVM-RFE. To prove that the gene selected is highly associated with the cancer disease, biological validation is provided whereby a reference paper related to the gene selected is attached.

## 27.3 Dataset

Pathway sets of genes are collected from previous research obtained from the Molecular Signature Database (MSigDB). In the MSigDB database, lung and gender datasets were retrieved. Each dataset consists of two classes, which are normal and tumor, as well as male or female. Both datasets must go through the transpose process using RStudio to set as input data format. This is because SVM-RFE can only start processing when the rows presented are 'Sample' and the column presented are 'Genes'. Table 27.1 shows the summary of datasets used for this research.

**Table 27.1** Summary of datasets

| Dataset | No. of genes | No. of samples | Class | Link | References |
|---|---|---|---|---|---|
| Lung | 7129 | 86 | Normal/tumor | https://www.gsea-msigdb.org/gsea/datasets.jsp | [5] |
| Gender | 22,283 | 32 | Male/female | https://www.gsea-msigdb.org/gsea/datasets.jsp | Unpublished |

## 27.4 Results and Discussion

### *27.4.1 Performance Measurement*

This section used tenfold CV accuracy results as a performance measurement of proposed method. Each pathway's accuracy was calculated by using the formula (1 − error rate × 100%). With that, this research can produce a better and more accurate result. The top five pathways and informative genes were selected during pathway and gene ranking according to their accuracy result produced. In average, the lung dataset obtained 70.91% accuracy while the gender dataset obtained 44.52%, whereby lung dataset performed 26.39% higher than the gender dataset. Regarding computational time, the gender dataset used 2 h and 49 min longer than the lung dataset for analysis. This condition happens because the gender dataset has a more significant dimension compared to lung dataset. Besides, it has a larger number of genes for each pathway to analyze although it contains a smaller sample size. Table 27.2 shows the comparison of lung and gender performance.

Originally, the lung dataset had a total of 439 pathways. Five pathways were excluded from this research as it contains only one gene for each pathway which is unsuitable for this method. This is because SVM-RFE conducts gene selection and classification, and ranks genes accordingly. With other genes, gene selection, classification, and feature ranking are able to be executed, and thus no significant genes will be identified. For the gender dataset, all pathways were included in the analysis.

Furthermore, this method has proved to have the highest performance compared to the previous method. The result comparison of the average top ten pathways is presented in Table 27.3. Overall, pathway-based analysis using SVM-RFE outperforms another method in terms of higher accuracies and lower error rates.

**Table 27.2** Comparison of lung and gender dataset performance

| Dataset | No. of pathway | Average tenfold accuracy (%) | Time taken |
|---|---|---|---|
| Lung | 434 | 70.91 | 8 h and 9 min |
| Gender | 480 | 44.52 | 10 h and 58 min |

**Table 27.3** Comparison of average tenfold CV accuracy from top ten pathways

| Method | Lung dataset (%) | Gender dataset (%) |
|---|---|---|
| Pathway SVM-RFE (proposed method) | **75.11** | **83.75** |
| Pathway RF[a] | 71.00 | 81.75 |
| $L_1$-SVM[a] | 55.14 | 80.76 |
| SVM-SCAD[a] | 53.50 | 77.96 |

Values in **bold** are the highest accuracy
[a] *Result published by* Misman et al. [35]

The table above shows that pathway-based SVM-RFE performs better than Random Forest. This is because SVM-RFE selects the highest feature among all genes, while RF randomly selects genes for classification and has a higher risk of selecting insignificant genes. Plus, pathway-based analysis performs better than another method that does not run by pathway. Genes selected according to their pathway is more helpful in identifying cancer disease as they bring along useful information for a researcher to investigate further the pathway selected. This is also because the pathway selected usually contained significant genes which interrelated to other diseases.

### *27.4.2 Biological Validation*

The list of ranked pathways and genes from the result is used for biological validation to show biological relevance to the target disease. The top five higher-accuracy pathways and informative genes were selected and validated with literature. Basically, the genes and pathways selected are represented by specific IDs. Before the biological validation process starts, all the significant genes and pathways must be converted to identify their original name and further investigate their importance to cancer diseases. The full name of the pathway can be searched on this platform BioCarta Pathways (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) and KEGG (http://www.genome.jp/kegg/).

**Lung Dataset**

The top five pathways and selected genes for each pathway of the lung dataset were identified. The selected genes are validated based on literature proving the significance and importance of lung cancer development. The result is presented in Table 27.4.

In the lung dataset, genes RCC1 and RANBP1 appeared in both pathways, cycling of Ran in nucleocytoplasmic transport and the role of Ran in mitotic spindle regulation. According to [17], RANBP1 involves as Ran-binding protein I, and this proves the presence of the genes in the pathway related to its functions.

**Gender Dataset**

The top five pathways and selected genes for each pathway of the gender dataset were identified. The selected genes are validated based on literature proving the significance and importance of ovarian, prostate, or breast cancer development. The result is presented in Table 27.5.

In the gender dataset, gene RPS4X appeared in both pathways, SIG regulation of the actin cytoskeleton by Rho GTPases and Willard intact. According to [25], RPS4X serves as a reliable indicator of cisplatin resistance in breast cancer patients. Additionally, the YB-1/RPS4X complex was identified in ovarian cancer cells as well. Therefore, it is proved that RPS4X significantly correlated in both ovarian and breast cancer.

**Table 27.4** Selected genes in top the five pathways from the lung dataset

| Pathways | Tenfold CV accuracy (%) | No. of genes | Genes selected |
|---|---|---|---|
| Flavonoids stilbene | 77.92 | 7 | PRDX6 [6], CYP24A1 [7], LPO [8], MPO [9] |
| Hypoxia-inducible factor in the cardiovascular | 75.28 | 19 | ARNT [10], ASPH [11], COPS5 [12], CREB1 [13] |
| Protein export | 75.28 | 7 | OXA1L [14], SRP14, SRP54 [15], SRP72 [12] |
| Cycling of Ran in nucleocytoplasmic transport | 75.00 | 5 | RCC1 [16], RANBP1 [17], RANBP2 [18] |
| Role of Ran in mitotic spindle regulation | 74.86 | 7 | RCC1 [16], KPNA2 [19], KPNB1 [19], RANBP1 [17] |

**Table 27.5** Selected genes in the top five pathways from the gender dataset

| Pathways | Tenfold CV accuracy (%) | No. of genes | Genes selected |
|---|---|---|---|
| GNF female genes | 100.00 | 116 | RPL21 [20], RPL24 [21], HNRPA1 [22], RPS27A [23] |
| Testis genes from xhx and netaffx | 93.33 | 111 | ACE [24], ADAM2 [25], ADAM29 [26] |
| RAP down | 90.00 | 434 | COX7B [27], DDX3X [28], PTBP1 [29], WNT10B [36], RAB1A [30] |
| SIG regulation of the actin cytoskeleton by rho GTPases | 88.33 | 67 | RPS4X [31], ACTG1 [32], WASL [25], PAK4 [33] |
| Willard inact | 86.67 | 31 | RPS4X [31], UBE1 [34], ATP6AP2 [37] |

## 27.5 Conclusion

This research indicates that pathway-based analysis using SVM-RFE can achieve higher accuracy and lower error rates if compared to traditional SVM-RFE without gene selection. Plus, its performance was higher than other methods, as described in Table 27.2. Three improvements have been made to the previous method, which is gene mapping into pathways, performance measurement, and pathway and gene ranking. The informative genes selected by pathway SVM-RFE are analyzed and further validated biologically. Most of the genes selected from the top five pathways are proven informative from previous studies.

However, there are limitations to this research where five pathways were excluded since it contains only one gene in the pathway. This is because pathway-based analysis SVM-RFE cannot perform gene ranking and gene selection on pathway with only

one gene. Using a method that can analyze, calculate a single gene, and determine their significance is suggested. Plus, it is advised to try pathway-based analysis on another method than SVM-RFE, which requires less computational time.

**Acknowledgements** This work was sponsored by the United Arab Emirates University through Strategic Research Program (Grant # 12R111) and the Research Start-up Program (Grant # 12M109).

## References

1. Tang, Y., Zhang, Y.Q., Huang, Z.: Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM Trans. Comput. Biol. Bioinf.Comput. Biol. Bioinf. **4**(3), 365–381 (2007)
2. Mundra, P.A., Rajapakse, J.C.: SVM-RFE with MRMR filter for gene selection. IEEE Trans. Nanobiosci.Nanobiosci. **9**(1), 31–37 (2009)
3. Duan, K.B., Rajapakse, J.C., Wang, H., Azuaje, F.: Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans. Nanobiosci.Nanobiosci. **4**(3), 228–234 (2005)
4. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**, 389–422 (2002)
5. Beer, D.G., Kardia, S.L., Huang, C.C., Giordano, T.J., Levin, A.M., Misek, D.E., Hanash, S.: Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med. **8**(8), 816–824 (2002)
6. Ho, J.N., Lee, S.B., Lee, S.S., Yoon, S.H., Kang, G.Y., Hwang, S.G., Um, H.D.: Phospholipase A2 activity of peroxiredoxin 6 promotes invasion and metastasis of lung cancer CellsRole of PLA2 activity of PRDX6 in cancer. Mol. Cancer Ther. **9**(4), 825–832 (2010)
7. Kim, B., Lee, H.J., Choi, H.Y., Shin, Y., Nam, S., Seo, G., Lee, S.: Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data. Can. Res. **67**(15), 7431–7438 (2007)
8. Petruzzelli, S., Hietanen, E., Bartsch, H., Camus, A.M., Mussi, A., Angeletti, C.A., Giuntini, C.: Pulmonary lipid peroxidation in cigarette smokers and lung cancer patients. Chest **98**(4), 930–935 (1990)
9. Kiyohara, C., Yoshimasu, K., Takayama, K., Nakanishi, Y.: NQO1, MPO, and the risk of lung cancer: a HuGE review. Genet. Med. **7**(7), 463–478 (2005)
10. Lin, P., Hu, S.W., Chang, T.H.: Correlation between gene expression of aryl hydrocarbon receptor (AhR), hydrocarbon receptor nuclear translocator (Arnt), cytochromes P4501A1 (CYP1A1) and 1B1 (CYP1B1), and inducibility of CYP1A1 and CYP1B1 in human lymphocytes. Toxicol. Sci.. Sci. **71**(1), 20–26 (2003)
11. Zhao, X., Weir, B.A., LaFramboise, T., Lin, M., Beroukhim, R., Garraway, L., Beheshti, J., Lee, J.C., Naoki, K., Richards, W.G., Sugarbaker, D., Meyerson, M.: Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. Cancer Res. **65**(13), 5561–5570 (2005)
12. Chitale, D., Gong, Y., Taylor, B.S., Broderick, S., Brennan, C., Somwar, R., Golas, B., Wang, L., Motoi, N., Szoke, J., Reinersman, J.M., Ladanyi, M.: An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. Oncogene **28**(31), 2773–2783 (2009)
13. Linnerth, N.M., Baldwin, M., Campbell, C., Brown, M., McGowan, H., Moorehead, R.A.: IGF-II induces CREB phosphorylation and cell survival in human lung cancer cells. Oncogene **24**(49), 7310–7319 (2005)

14. Phelan, J., MacCarthy, F., Feighery, R., O'Farrell, N., Lynam-Lennon, N., Doyle, B., O'Toole, D., Ravi, N., Reynolds, J., O'Sullivan, J.: Differential expression of mitochondrial energy metabolism profiles across the metaplasia–dysplasia–adenocarcinoma disease sequence in Barrett's oesophagus. Cancer Lett. **354**(1), 122–131 (2014)
15. Gao, W., Xu, J., Liu, L., Shen, H., Zeng, H., Shu, Y.: A systematic-analysis of predicted miR-21 targets identifies a signature for lung cancer. Biomed. Pharmacother.Pharmacother. **66**(1), 21–28 (2012)
16. Zhong, L., Peng, X., Hidalgo, G., Doherty, D., Stromberg, A., Hirschowitz, E.: Identification of circulating antibodies to tumor-associated proteins for combined use as markers of non-small cell lung cancer. Proteomics **4**(4), 1216–1225 (2004)
17. Rohrbeck, A., Neukirchen, J., Rosskopf, M., Pardillos, G., Geddert, H., Schwalen, A., Gabbert, H., von Haeseler, A., Pitschke, G., Schott, M., Kronenwett, R., Haas, R., Rohr, U.: Gene expression profiling for molecular distinction and characterization of laser captured primary lung cancers. J. Transl. Med. **6**(1), 69 (2008)
18. Gerber, D., Minna, J.: ALK inhibition for non-small cell lung cancer: from discovery to therapy in record time. Cancer Cell **18**(6), 548–551 (2010)
19. Li, X., Jia, L., Shi, M., Li, X., Li, Z., Li, H., Wang, E., Jia, X.: Downregulation of KPNA2 in non-small-cell lung cancer is associated with Oct4 expression. J. Transl. Med. **11**(1), 232 (2013)
20. Sherman-Baust, C., Weeraratna, A., Rangel, L., Pizer, E., Cho, K., Schwartz, D., Shock, T., Morin, P.: Remodeling of the extracellular matrix through overexpression of collagen VI contributes to cisplatin resistance in ovarian cancer cells. Cancer Cell **3**(4), 377–386 (2003)
21. Wilson-Edell, K., Kehasse, A., Scott, G., Yau, C., Rothschild, D., Schilling, B., Gabriel, B., Yevtushenko, M., Hanson, I., Held, J., Gibson, B., Benz, C.: RPL24: a potential therapeutic target whose depletion or acetylation inhibits polysome assembly and cancer cell growth. Oncotarget **5**(13), 5165–5176 (2014)
22. Yu, J., Sieuwerts, A., Zhang, Y., Martens, J., Smid, M., Klijn, J., Wang, Y., Foekens, J.: Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. BMC Cancer **7**(1), 182 (2007)
23. Kosaka, Y., Kataoka, A., Yamaguchi, H., Ueo, H., Akiyoshi, S., Sengoku, N., Kuranami, M., Ohno, S., Watanabe, M., Mimori, K., Mori, M.: Vascular endothelial growth factor receptor-1 mRNA overexpression in peripheral blood as a useful prognostic marker in breast cancer. Breast Cancer Res. **14**(5), R140 (2012)
24. Jethon, A., Pula, B., Piotrowska, A., Wojnar, A., Rys, J., Dziegiel, P., Podhorska-Okolow, M.: Angiotensin II type 1 receptor (AT-1R) expression correlates with VEGF-A and VEGF-D expression in invasive ductal breast cancer. Pathol. Oncol. Res.. Oncol. Res. **18**(4), 867–873 (2012)
25. Rodenhiser, D., Andrews, J., Kennette, W., Sadikovic, B., Mendlowitz, A., Tuck, A., Chambers, A.: Epigenetic mapping and functional analysis in a breast cancer metastasis model using whole-genome promoter tiling microarrays. Breast Cancer Res. **10**(4), R62 (2008)
26. Purrington, K.S., Slager, S., Eccles, D., Yannoukakos, D., Fasching, P.A., Miron, P., Carpenter, J., Chang-Claude, J., Martin, N.G., Montgomery, G.W., Kristensen, V., Couch, F.J.: Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. Carcinogenesis **35**(5), 1012–1019 (2014)
27. Ning, Q., Wu, J., Zang, N., Liang, J., Hu, Y., Mo, Z.: Key pathways involved in prostate cancer based on gene set enrichment analysis and meta-analysis. Genet. Mol. Res. **10**(4), 3856–3887 (2011)
28. Xia, T., Wang, G., Ding, Q., Liu, X., Zhou, W., Zhang, Y., Zha, X., Du, Q., Ni, X., Wang, J., Miao, S., Wang, S.: Bone metastasis in a novel breast cancer mouse model containing human breast and human bone. Breast Cancer Res. Treat. **132**(2), 471–486 (2011)
29. He, X., Arslan, A., Ho, T., Yuan, C., Stampfer, M., Beck, W.: Involvement of polypyrimidine tract-binding protein (PTBP1) in maintaining breast cancer cell growth and malignant properties. Oncogenesis **3**(1), e84 (2014)

30. Sun, T., Wang, X., He, H., Sweeney, C., Liu, S., Brown, M., Balk, S., Lee, G., Kantoff, P.: MiR-221 promotes the development of androgen independence in prostate cancer cells via downregulation of HECTD2 and RAB1A. Oncogene **33**(21), 2790–2800 (2013)
31. Tsofack, S., Meunier, L., Mes-Masson, A., Lebel, M.: Abstract 2864: RPS4X, a new prognostic and predictive biomarker of ovarian and breast cancer. Can. Res. **74**(19 Supplement), 2864–2864 (2014)
32. Cicatiello, L.: A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone-responsive transcriptome. J. Mol. Endocrinol. **32**(3), 719–775 (2004)
33. Wells, C., Whale, A., Parsons, M., Masters, J., Jones, G.: PAK4: a pluripotent kinase that regulates prostate cancer cell adhesion. J. Cell Sci. **123**(10), 1663–1673 (2010)
34. Jazaeri, A.: Gene expression profiles of BRCA1-Linked, BRCA2-linked, and sporadic ovarian cancers. Cancer Spectrum Knowl. Environ. **94**(13), 990–1000 (2002)
35. Misman, M.F., Mohamad, M.S., Deris, S., Mohd. Hashim, S.Z.: A group-specific tuning parameter for hybrid of SVM and SCAD in identification of informative genes and pathways. Int. J. Data Mining Bioinform. 9 **10**(2), 146–161 (2014)
36. Bui, T., Rankin, J., Smith, K., Huguet, E., Ruben, S., Strachan, T., Harris, A., Lindsay, S.: A novel human Wnt gene, WNT10B, maps to 12q13 and is expressed in human breast carcinomas. Oncogene **14**(10), 1249–1253 (1997)
37. Debily, M., Marhomy, S., Boulanger, V., Eveno, E., Mariage-Samson, R., Camarca, A., Auffray, C., Piatier-Tonneau, D., Imbeaud, S.: A functional and regulatory network associated with PIP expression in human breast cancer. PLoS ONE **4**(3), e4696 (2009)

# Chapter 28
# An AI-Assisted Skincare Routine Recommendation System in XR


**Gowravi Malalur Rajegowda, Yannis Spyridis, Barbara Villarini, and Vasileios Argyriou**

**Abstract** In recent years, there has been an increasing interest in the use of artificial intelligence (AI) and extended reality (XR) in the beauty industry. In this paper, we present an AI-assisted skin care recommendation system integrated into an XR platform. The system uses a convolutional neural network (CNN) to analyse an individual's skin type and recommend personalised skin care products in an immersive and interactive manner. Our methodology involves collecting data from individuals through a questionnaire and conducting skin analysis using a provided facial image in an immersive environment. This data is then used to train the CNN model, which recognises the skin type and existing issues and allows the recommendation engine to suggest personalised skin care products. We evaluate our system in terms of the accuracy of the CNN model, which achieves an average score of 93% in correctly classifying existing skin issues. Being integrated into an XR system, this approach has the potential to significantly enhance the beauty industry by providing immersive and engaging experiences to users, leading to more efficient and consistent skincare routines.

## 28.1 Introduction

Skin care has been an important aspect of personal hygiene and beauty for centuries, and often involves several steps in order to maintain and improve the texture of the skin. With the advent of new technologies in the medical and pharmaceutical industries, there has been an increase in the number of skin care products available in the market. However, choosing the right skin care product that suits an individual's

G. M. Rajegowda · V. Argyriou
Department of Network Ands and Digital Media, Kingston University, London, UK

Y. Spyridis
Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, UK

B. Villarini (✉)
School of Computer Science and Engineering, University of Westminster, London, UK
e-mail: b.villarini@westminster.ac.uk



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_28

skin type and concerns can be overwhelming, especially with the ever-increasing number of products available. In recent years, the use of artificial intelligence (AI) in skin care recommendation systems has gained popularity as it helps to personalise skin care recommendations based on an individual's skin type, concerns, and lifestyle.

The motivation behind developing an AI-assisted skin care recommendation system arises from the need to help individuals choose the right skin care products that are suitable for their skin type, taking into consideration their personal needs, while avoiding visits to medical experts in case there are non-clinical skin concerns. The use of AI can help to reduce the guesswork and time spent on researching and selecting skin care products, which can be often time-consuming and inefficient. By using an AI-assisted skin care recommendation system, individuals can get personalised recommendations based on their own needs and treat mild abnormalities in an efficient way.

In recent years, there has been an increase in research on recommendation systems using AI. Several studies have shown that AI models can provide accurate and personalised recommendations based on an individual's needs [1]. For instance, a study conducted by Kumar et al. [2] proposed an AI-assisted college recommendation system that provides personalised recommendations after building a user profile based on certain questions. The system then maps this profile to college profiles scraped from the web based on a content-based approach. Skin recognition algorithms have also been widely used in recent years to help in medical applications. These systems use machine learning models to analyse an individual's skin type and abnormalities and classify it to certain categories. Using microscopic images, Saidah et al. [3] developed a convolutional neural network (CNN) to classify the given image into normal, dry, oily, or combined skin, with a very high accuracy. While there is a lack of extensive research on extended reality (XR) recommendation systems, there are a few studies that aim to provide personalised recommendations to users in an immersive and interactive setting. For instance, Lin et al. [4] proposed a recommender system that employs virtual reality, which serves as a platform for retrieving historical interior design drawings from a database and recommending a prototype drawing to the consigner. Such a system can be quite useful for designers as it enables them to store historical drawing items, extract pertinent design features, and guide consigners from the query system to the historical database, thus accessing most appropriate design drawing that match their interests and requirements.

In this paper, we present a skin care recommendation system, which utilises deep learning models to analyse the skin type of users, and along with targeted inputs from specialised questions, provides specific suggestions with respect to skin care products. Furthermore, a dataset for skin analysis and classification is part of the contributions of this work. To make the skin care recommendation experience more immersive and interactive, we have integrated the AI-assisted skin care recommendation system into an XR platform. By incorporating XR technology into the AI-assisted skin care recommendation system, users can experience a personalised skin care journey, allowing them to see the effects of the recommended products on their skin over time. This can lead to a more engaging experience that encourages individuals to adhere to their personalised skin care routine.

The rest of the paper is organised as follows. Section 28.2 discusses the related literature on skin issue detection and recommendation systems. Section 28.3 presents the methodology that was followed for the dataset creation and the training process of the CNN model, while Sect. 28.4 discusses the system evaluation. Finally, Sect. 28.5 concludes the paper.

## 28.2 Literature Review

### *28.2.1 Skincare Issues and Importance of Product Ingredients*

In a survey conducted by El-Essawi et al. [5], it was found that a large group of individuals in the USA suffer from multiple skin issues. Table 28.1 illustrates the distribution of the most common skin concerns identified.

It is observed that roughly 55% of the participants suffer from conditions such as uneven skin tone, discolouration, and about 50% suffer from acne or dry skin. Other issues such as wrinkles, redness, rashes, and oily skin are also quite common, while it should be noted that most participants suffer from multiple skin issues. To solve these conditions, one of the proven scientific methods is the adaption to a carefully curated skin routine. However, identifying such a routine can be complex, because it involves relying upon multiple skin care products, examining their ingredients, their proportion, and impact [5].

In a study by Rodan et al. [6] it was found that daily skincare routines may have statistically significant long-term effects on the overall quality of one's skin health and complexion. Basic skin care needs involve processes of skin protection, issue prevention, cleansing, and moisturising. Ingredients such as either zinc oxide or avobenzone for example can help block UV-B and UV-A effects on the skin. While the skincare routine is complex, cosmetic stores still recommend products based on popularity with less regard to each individual's skin conditions and issues they are

**Table 28.1** Distribution of common skin issues

| Skin issue | Percentage of population in the sample (%) |
|---|---|
| Uneven skin tone | 56.4 |
| Skin discoloration | 55.9 |
| Dry skin | 51.9 |
| Acne | 49.4 |
| Wrinkles | 39.4 |
| Moles | 32.4 |
| Rashes | 30.9 |
| Oily skin | 23.2 |

seeking to solve [7]. Since active ingredients in the product and their knowledge are hard to procure due to the vast assortment of alternatives, it is important to enable a solution that assists users in recommending products based on ingredients and alternatives and to personalise this recommendation based on their skin issues and conditions.

### *28.2.2 Skin Detection Models*

Arifin et al. [8] developed an automated dermatological diagnostic system that detects and identifies skin anomalies using high-resolution colour images and patient history. The system uses colour image processing techniques, clustering, and neural networks to achieve high accuracy rates of 95.99% for detecting diseased skin and 94.016% for identifying diseases. ALEnezi [9] proposed an image processing-based method to detect skin diseases of the affected skin area in provided images. The approach works on the inputs of a colour image and resizes the image to extract features using a pretrained CNN. The system then classifies the features using a multiclass SVM and shows the results to the user, including the type of disease, spread, and severity.

With respect to non-clinical skin issues, Alamdari et al. [10] compared image segmentation methods to detect acne lesions and several machine learning methods to distinguish between different types of such lesions. Two-level k-means clustering was found to outperform other techniques with an accuracy of approximately 70% for detecting acne lesions. The accuracy of differentiating between acne scarring and active inflammatory lesions was 80% and 66.6% for fuzzy-c-means and the SVM method, respectively. The performance accuracy of classifying the normal skin from detected acne lesions was 100% using fuzzy-c-means clustering.

### *28.2.3 Skincare Recommendation Systems*

A study by Hsia et al. [11] utilised machine learning to process and classify multiple features of skin quality and acne status to provide recommendations for facial skincare products. The proposed system, which relies on an SVM classifier was evaluated on 15 subjects and resulted in a consumer satisfaction index of 80%.

Similarly, Lin et al. [12] proposed a new business model of facial skincare products that utilises computer vision technology. The framework consists of a finger vein identification system, a skincare product recommendation system, and an electronic payment system. Experimental results showed that the finger vein identification system had the lowest equal error rate and shortest response time, while the skin type classification accuracy was the highest. Lee [7] proposed a skincare product recommendation system that uses content-based filtering to suggest products based on a user's skin type and desired beauty effect. The system analyses the chemical composition of products to find those with similar ingredient compositions.

## 28.3 Methodology

The implementation workflow of our AI-assisted skincare recommendation system is presented in Fig. 28.1. The system first relies on a two-stage process to analyse provided facial images, extract relevant features, and identify existing skin issues. The outcome of this process is then fed to the skincare routine recommendation algorithm, which considering the similarities between ingredients in the product catalog, suggests a series of products that target the specific needs identified in the previous process.

### *28.3.1 Facial Landmark Detection Subsystem*

This subsystem is used to detect important characteristics in the provided facial images, such as eyes, forehead, cheeks, and chin, that are considered predominant areas in which non-clinical skin issues can be observed. The developed algorithm relies on the "Haar Cascade" classifier, which is an object detection algorithm commonly used in face detection applications, using a machine learning approach to identify objects in image streams. The algorithm is based on the "Haar Wavelet"

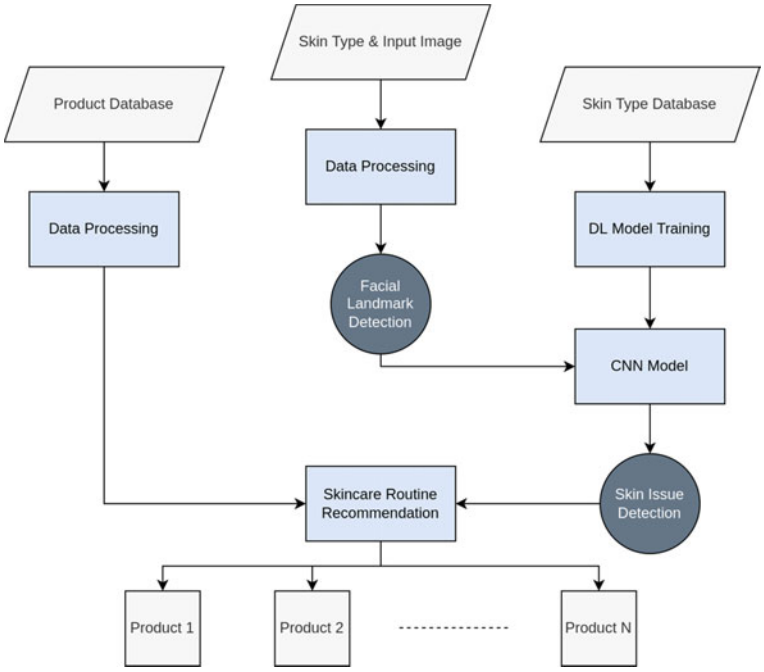



**Fig. 28.1** Implementation workflow

[13] function, which is used to extract key features from an image, then combine them to detect objects of interest.

More specifically the system utilises, the "Eye Cascade", a specific "Haar Cascade" classifier, designed to detect eyes in the provided image. This classifier is trained on a set of positive and negative images and learns to differentiate between the ones that contain eyes, thus extracting features that are important for eye detection. The location of the rest of the facial characteristics is determined using the "Shape Predictor 68 Face Landmark", which is a machine learning model, capable of predicting the locations of 68 specific facial landmarks on a human face.

Following the prediction of the eye detection and the facial landmark models, four main regions of the face are spliced and stored to be provided as input to the Skin Issue Recognition subsystem. This process is implemented using the "OpenFace" face recognition library, which relies on a combination of convolutional neural networks (CNN) and recurrent neural networks (RNN) to extract the facial features from the given images.

### *28.3.2 Skin Issue Recognition Subsystem*

#### Dataset Collection and Data Preprocessing

Due to the lack of large publicly available datasets for skin issues, a new dataset was developed by collecting free images from different sources on the Internet. The dataset contains approximately 3500 images of four labelled classes: (a) Acne, (b) Pigmentation, (c) Wrinkles, and (d) Clear Skin. The images were processed so that the background noise was removed, and the exposure was adjusted to normal values to ensure robust training.

The final images were subjected to the Facial Landmark Detection system presented in Sect. 28.3.1, so that the relevant skin patches from each image is extracted and saved along with the correct label. Finally, a process of data augmentation was followed whereby a data generator was utilised to apply image augmentation options such as zoom, shear, flip, and brightness adjustment. Through the augmentation, the amount and diversity of data in the training dataset was increased, resulting in a model more robust to variations in the input data, and capable of generalising to a wider range of input conditions.

#### Training Process

For the training process, we utilised transfer learning using the VGG16 model, due its relatively small number of parameters, consistent block structure, and modular nature, which allows it to be easily adapted for other tasks. The VGG16 architecture consists of 16 layers of convolutional and pooling layers, followed by three fully connected layers for classification. The convolutional layers use small filters ($3 \times 3$) with a stride of 1 pixel and padding to preserve spatial resolution, while the pooling

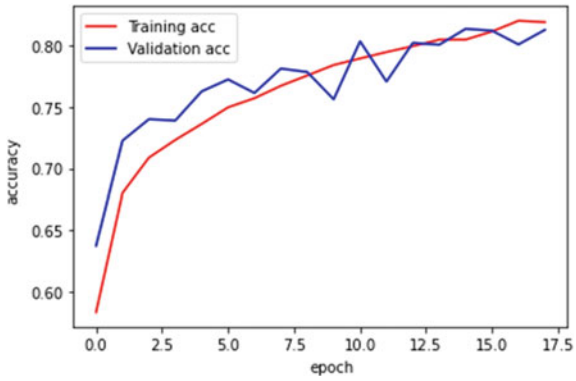



**Fig. 28.2** Training and validation accuracies

layers use max pooling with a $2 \times 2$ filter and a stride of 2 pixels to reduce the spatial dimensions.

After splitting the dataset into 80%, 15%, and 5% for training, testing, and validation respectively, we trained the model utilising the VGG16 pretrained weights, to classify the input image into one of the four labels. We used SGD-momentum as the optimiser which is an extension of the gradient descent algorithm. In SGD-momentum, a momentum term is introduced that accumulates the gradient over previous iterations and adds it to the current update, to smooth out the updates and reduce oscillations in the parameter space. The update for a parameter $\theta$ at iteration $t + 1$ is given by:

$$\theta_{t+1} = \theta_t - V_t, \tag{28.1}$$

$$V_t = m * V_{t-1} + \text{lr} * \text{gradient}, \tag{28.2}$$

where $V_t$ is the momentum vector at iteration $t$, $m$ is a hyperparameter that controls the contribution of the previous velocity to the current update, and lr is the learning rate. Using the momentum update helped improve the accuracy and solve over-fitting problems, while allowing for a higher learning rate. Ultimately, this accelerated the convergence without destabilising the optimisation process, as demonstrated in Fig. 28.2. The training and validation losses throughout the training process are shown in Fig. 28.3.

### *28.3.3 Product Recommendation Engine*

The skincare recommendation system presented in this paper utilises a content-based approach to product recommendations, specifically focusing on ingredient similarity

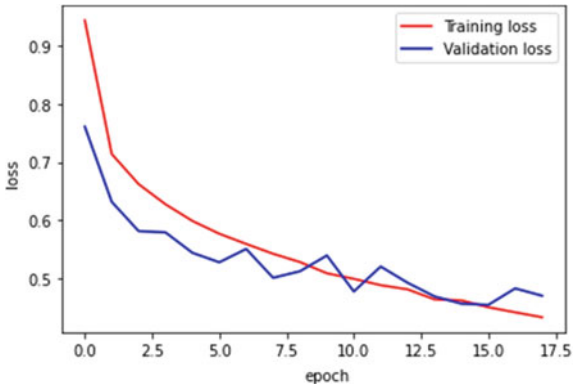



**Fig. 28.3** Training and validation losses

within the same product category. This method distinguishes itself from other recommendation systems, as skincare requires personalised care and attention based on individual skin types. As a result, a review-based recommendation system may not be appropriate in this context and therefore the developed recommendation system is designed based on the content-based approach necessitated by the complexity of skincare. By leveraging ingredient information and comparing products within the same category, we provide a more personalised and accurate recommendation for users. The presented approach accounts for the unique needs and characteristics of each user's skin type, as designated by the Skin Issue Recognition subsystem, resulting in a more effective and efficient skincare recommendation process.

### Dataset Preparation

For the skincare product recommendation engine, we utilised an existing dataset [14] that focuses on day-to-day skincare routines. The primary categories considered in this dataset include cleanser, serum, treatment, moisturizer, and sunscreen products. The data was sourced from a popular skincare website and included information such as ingredients, price, colour, brand, and chemical components. The dataset consists of 1472 unique items and 17 columns. The utilisation of this dataset allowed us to focus the analysis on the most relevant skincare categories and ensured that the results were based on a comprehensive and representative set of data. A sample of the dataset is presented in Table 28.2.

**Table 28.2** Sample of the skincare product dataset

| ID | Label | Issue | Brand | Name | Ingr | Combin | Dry | Oily |
|---|---|---|---|---|---|---|---|---|
| 1 | Moistu. | Acne | LA MER | Crème… | Algae… | 1 | 1 | 1 |
| 2 | Moistu. | Acne | SK-II | Facial… | Galact… | 1 | 1 | 1 |
| 6 | Moistu. | Acne | DRUNK. | Protin… | Dicapr… | 1 | 1 | 1 |
| 11 | Moistu. | Acne | BELIF | The Tru | Diprop… | 1 | 0 | 1 |

The dataset was adjusted with the specific goal to address the challenges associated with predicting skincare products based on past usage in mind. Reliance solely on past usage to make recommendations can be unreliable and unpredictable due to the vast domain of skincare and the personalised nature of cosmetic recommendations. Instead, the developed recommendation system focuses on identifying the ingredients in each product and comparing them to identify similarities between products. Preprocessing steps were conducted to prepare the dataset for analysis, including removing duplicate items and cleaning the text data in the "Ingredient" column. Additional information, such as compositions and ingredients, was also collected to ensure that the final dataset was comprehensive and representative.

The analysis focused on products that were classified based on different skin types and skin concerns to provide personalised recommendations. Specifically, the main dataset includes five recommended products in each of the five categories: cleanser, moisturizer, treatment, mask, and sunscreen. By mapping these categories to specific skin types, such as "Cleanser" for "Dry Skin", targeted recommendations are provided that are customised to each user's individual needs and preferences. The content-based approach which focuses on ingredient similarity between products, allows for reliable and accurate recommendations despite the personalised nature of skincare.

**Determining Similarity of Products**

To determine the similarity of ingredients between products, the t-SNE technique is employed, leading to the reduction of the dimensionality in the data. By preserving the similarities between instances, t-SNE effectively visualizes high-dimensional data on a two-dimensional plane. Similarities are calculated based on the distances between data points, and cosine similarity is used to find similarities between non-zero vectors. Unlike distance-based measures, cosine similarity captures more information about vector direction.

In the developed system, this technique is applied when the user selects a known brand on the recommender system. The system then analyses ingredient similarities based on skin type and skin concern and recommends a complete skincare routine with up to five products in each category, based on t-SNE and cosine similarity.

**Recommendation Through Matrix Factorisation**

The skincare industry can be overwhelming for users due to the vast number of products available. To simplify the process, the developed recommendation system has been designed to consider user input before suggesting products. The system considers two inputs—brand name or desired product and the skin concern detected by the CNN model of the Skin Issue Recognition subsystem—using the Matrix Factorisation method. This method is ideal as it is non-biased towards sparse data.

The Matrix Factorisation method calculates two factors: user input features and product similarity based on predicted skin issues from the products dataset. The system then reduces dimensions based on skin issues, brands, skin type, and ingredients to provide recommendations. For instance, if the user has acne problems, a

product that suits their needs is recommended based on the comparison of the similarity scores among relevant products. Therefore, the system suggests 5 products that are nearest to the selected product based on their ingredients. The recommendation process contains a few different steps, which are followed in an XR setting, as detailed in the following subsection.

### XR Integration

The XR platform was developed to offer an immersive experience to users when going through the skincare recommendation process. The platform offers an interface of widgets through which the application communicates with the deployed AI models, so that the individual can get a personalised recommendation for their skincare routine. The first such widget allows users to upload their facial image, which is provided as input to the Facial Landmark and Skin Issue Recognition subsystems. The recognised faces are processed and uploaded in a queue (Fig. 28.4). The faces are analysed by the deep neural network, which provides the skin classification. Using this information, the recommendation engine provides the suggestions for the skincare routine as a list of products. In the final widget, the user can select alternatives brands, and the recommendation engine suggests corresponding products based on the similarity of the ingredients identified. The system also keeps track of uploaded images, allowing users to view the results of the followed skincare routine over time.


**Fig. 28.4** The automatic recognised faces are uploaded in the queue ready to be analysed by the deep neural network, which provides the skin classification

## 28.4 System Evaluation

### *28.4.1 Facial Landmark Performance*

The system's performance in detecting skin patches was evaluated under varying lighting conditions and was found to be effective in detecting the eyes and mouth in the uploaded image. The accuracy of the skin patches was assessed using confusion metrics, which compared the classification results of already classified images to the live classification. The results of the analysis are presented in the confusion matrix of Fig. 28.5.

The findings reveal that the face landmark detection system was highly accurate, with a success rate of over 95% in correctly categorising the spliced area based on facial landmarks for all cases. However, there were a few instances in which the system failed to correctly separate the face landmarks, but these were due to the forehead, chin, or cheek not being fully visible in the image. This issue was slightly more pronounced for the chin, particularly when the person was not facing straight. Despite this minor limitation, the overall recommendation system designed in this work was not significantly impacted by this issue. Nevertheless, based on these results users are advised to show their full face as much as possible during usage, to minimise the occurrence of this issue.

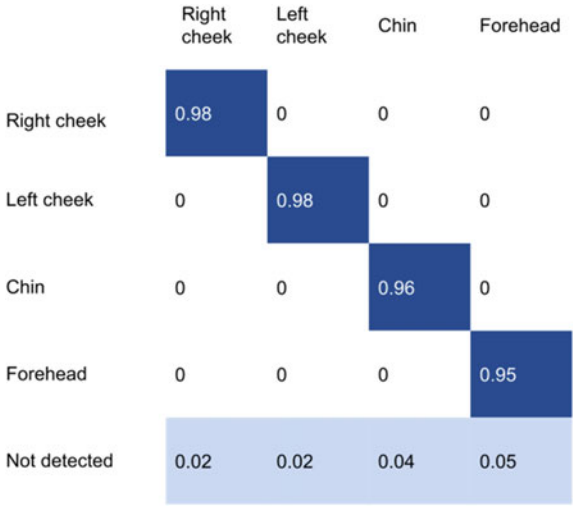



**Fig. 28.5** Face landmark detection confusion matrix

### 28.4.2 Skin Issue Recognition Performance

The evaluation of the CNN model utilised by the Skin Issue Recognition subsystem involved the use of several metrics, such as (a) Validation accuracy, (b) Precision, (c) Recall, and (d) $F1$-score. The validation accuracy was determined during the training process by computing the ratio of correctly classified images to the total number of guesses made by the model. Precision and recall are metrics that are computed independently of validation accuracy, and their scores are determined after the model is trained with an unbiased set of data. Precision is a measure of the accuracy of a model's positive predictions, while recall measures the actual positives of the model. The formulas for precision and recall are as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}, \tag{28.3}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{28.4}$$

$F1$-score is a function of both recall and precision and is computed as an average weighted by precision and recall. This is especially important when there is an uneven class distribution, which is the case in our model. The formula for $F1$-score is as follows:

$$F1 = 2*\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{28.5}$$

The confusion matrix of the success rate is shown in Fig. 28.6, while the evaluation of the above metrics in our modified VGG16 model are presented in Table 28.3.

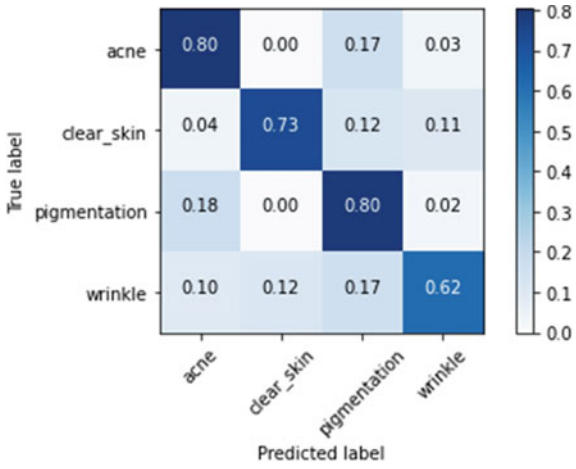



**Fig. 28.6** Skin issue recognition confusion matrix

**Table 28.3** Precision, recall, and $F1$-score metrics

| Skin issue | Precision | Recall | $F1$-score | Accuracy |
|---|---|---|---|---|
| Acne | 0.83 | 0.76 | 0.79 | 0.96 |
| Clear skin | 0.84 | 0.91 | 0.87 | 0.94 |
| Pigmentation | 0.65 | 0.77 | 0.71 | 0.91 |
| Wrinkles | 0.88 | 0.52 | 0.66 | 0.89 |

As observed in the table, the model achieved a very high accuracy across all classes, achieving the top performance in the "Acne" case with a score of 96%, and in general being capable of correctly classifying most of the samples in the dataset. However, due to the imbalanced structure of the dataset, it is important to examine the rest of the metrics to get a more complete understanding of the performance. When investigating the precision, we still observe a high percentage of correctly identified positive samples as true positives, except for the "Pigmentation" case.

A similar trend is observed in the recall metric, but in this case the low score affects the "Wrinkles" class, indicating the difficulty of the model in identifying true positive samples in this category. These metrics also reflect the performance in $F1$-score, which is the harmonic mean of precision and recall and in our case depict a balanced performance in the "Acne" and "Clear Skin" classes, but a lower score in the other two classes.

## 28.5 Conclusion

Skincare is a crucial aspect of personal care that enables individuals to maintain healthy skin. This practice typically involves the use of products prescribed by dermatologists, cosmetologists, or sourced from various online platforms. Skincare encompasses a wide range of topics that require extensive research, such as ingredients, skin types, and skin concerns in order to identify the appropriate products for one's use case. As the skincare industry continues to grow, research on cosmetics has shown significant impacts on the overall appearance and health of the skin. With the rise of the global pandemic, individuals have increasingly invested more time and financial resources in self-care, turning to online sources for guidance in a process that is often inefficient.

Skincare products contain varying ingredients, and selecting the appropriate ones for different skin types and concerns can be challenging and complex. This paper aimed to enhance and automate the process of diagnosing these concerns, analysing skin types and finally offering suitable product recommendations tailored to an individual's skincare needs. Towards this goal, a comprehensive recommendation system was designed and developed, including a Facial Landmark Detection subsystem, a Skin Issue Recognition algorithm, and finally the skincare product recommendation engine. Key focus was especially given in the CNN model that was built to analyse

the users' skin type. The model relies on a modified version of the VGG16 architecture, which in combination with the Facial Landmark Detection subsystem is able to recognise and classify the skin type into one of four classes. The model's performance is highly adequate for this use case, achieving an overall average accuracy of 93%. This output is then used by the skincare recommendation engine along with related user questions to recommend a series of products that target the specific needs of the individual.

Through the integrated platform, users are able to employ a one-stop solution for treating their skin issues, by just using an image and providing minimal input to the system, which then automatically recommends an appropriate skincare routine. By incorporating the process in an XR environment, users can interact with the system in an immersive and engaging manner that plays a vital role in identifying and adhering to the routine consistently. While the system displays high potential in the skincare domain, there are still limitations with respect to the performance of the Skin Issue Recognition algorithm. More specifically, as suggested by the precision and recall metrics, there is room for improvement in identifying tricky skin issues, such as pigmentation or wrinkles. In addition, the current system supports four categories of skin issues, therefore limiting the extent of the target audience. The above issues could be solved by employing a more extensive dataset that includes additional classes and poses, further fine-tuning and retraining of the CNN model.

## References

1. Zhang, Q., Lu, J., Jin, Y.: Artificial intelligence in recommender systems. Complex Intell. Syst. **7**, 439–457 (2021)
2. Kumar, K., Sinha, V., Sharma, A., Monicashree, M., Vandana, M.L., Vijay Krishna, B.S.: AI-assisted college recommendation system. In: Intelligent Sustainable Systems: Proceedings of ICISS 2022, pp. 141–150. Springer Nature Singapore, Singapore (2022)
3. Saidah, S., Fuadah, Y.N., Alia, F., Ibrahim, N., Magdalena, R., Rizal, S.: Facial skin type classification based on microscopic images using convolutional neural network (CNN). In: Proceedings of the 1st International Conference on Electronics, Biomedical Engineering, and Health Informatics: ICEBEHI 2020, 8–9 Oct, Surabaya, Indonesia, pp. 75–83. Springer Singapore (2021)
4. Lin, K.S., Ke, M.C.: A virtual reality based recommender system for interior design prototype drawing retrieval. In: New Trends in Intelligent Information and Database Systems, pp. 141–150. Springer International Publishing (2015)
5. El-Essawi, D., Musial, J.L., Hammad, A., Lim, H.W.: A survey of skin disease and skin-related issues in Arab Americans. J. Am. Acad. Dermatol. **56**(6), 933–938 (2007)
6. Rodan, K., Fields, K., Majewski, G., Falla, T.: Skincare bootcamp: the evolving role of skincare. Plastic Reconstr. Surgery Global Open **4**(12 Suppl) (2016)
7. Lee, G.: A Content-Based Skincare Product Recommendation System (2020)
8. Arifin, M.S., Kibria, M.G., Firoze, A., Amini, M.A., Yan, H.: Dermatological disease diagnosis using color-skin images. In: 2012 International Conference on Machine Learning and Cybernetics, vol. 5, pp. 1675–1680. IEEE (2012)
9. ALEnezi, N.S.A.: A method of skin disease detection using image processing and machine learning. Procedia Comput. Sci. **163**, 85–92 (2019)

10. Alamdari, N., Tavakolian, K., Alhashim, M., Fazel-Rezai, R.: Detection and classification of acne lesions in acne patients: a mobile application. In: 2016 IEEE International Conference on Electro Information Technology (EIT), pp. 0739–0743. IEEE (2016)
11. Hsia, C.H., Lin, T.Y., Lin, J.L., Prasetyo, H., Chen, S.L., Tseng, H.W.: System for recommending facial skincare products. Sens. Mater **32**, 3235 (2020)
12. Lin, T.Y., Chan, H.T., Hsia, C.H., Lai, C.F.: Facial skincare products' recommendation with computer vision technologies. Electronics **11**(1), 143 (2022)
13. Stanković, R.S., Falkowski, B.J.: The Haar wavelet transform: its status and achievements. Comput. Electr. Eng.. Electr. Eng. **29**(1), 25–44 (2003)
14. Skincare Product Dataset on Github. https://github.com/jjone36/Cosmetic/tree/master/data

# Chapter 29
# The Role of Artificial Intelligence in Improving Failure Mode and Effects Analysis (FMEA) Efficiency in Construction Safety Management

**L. Hezla, R. Gurina, M. Hezla, N. Rezaeian, M. Nohurov, and S. Aouati**

**Abstract** This study presents an analysis of the introduction of the artificial intelligence to the FMEA method. The traditional function of the FMEA only evaluates the impact and causes of system failure which takes a long time. Therefore, it was necessary to find a way to develop and improve the performance of FMEA. The goal is to develop FMEA technique into an electronic application in which artificial intelligence is used to identify, calculate, and evaluate risks at the same time. For achieving the set goal, this paper introduces a real field study on how to develop the FMEA from a traditional method to a more developed program in which artificial intelligence is used as an alternative to the work team in calculating and evaluating various risks. In this study, we took construction as a primary field for evaluating the performance of the modern FMEA program, where the work team takes two risks from the field hazards as a sample for use ("Construction project delivery failed on time" and "Workers falling from high floors"). In this study, 5 factors were put as causes for each risk with an evaluation of 50 possible cases for the occurrence of each risk. Our dataset of 50 cases for each risk has been loaded into our software. This program analyzed the dataset for each hazard in the $50 \times 8$ dimension. In order to reduce the number of outputs for the (D, O, S) computation, our team used multi output regressor outputs instead of single output regressor, and focused on three types (multiple output regressor algorithms), namely: Linear Regression, Training with the Random Forest Regressor Model, Training with the Decision Tree Regression Model. The results showed that the best metric algorithm MAE = 0.36 turned out to be an algorithm Random Forest Regressor. By using the artificial intelligence feature, we found that the features that play a major role in training the model for failure to deliver the Construction project on time is (work suspension due to the Corona pandemic (two weeks) for 20% of the total number of workers) and that the least influential feature in the model is (The acute shortage of funding for raw materials such as iron and concrete leads to a delay in implementation). As for the workers

L. Hezla (✉) · R. Gurina · M. Hezla · N. Rezaeian · M. Nohurov · S. Aouati
Peoples Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow 117198, Russian Federation
e-mail: Lokmanehezla@gmail.com



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_29

who fall from high floors, the feature that plays a big role is (falling from the stairs) and the feature that has the least impact is (falling from the stairs). The research employs qualitative, analytical, statistical and comparative methods of FMEA. The acquired results can be used by different construction companies to identify points of failure, anticipate risks, and calculate them in an innovative and smooth way. Also, it allows companies to avoid notifications in advance. This contributes to the improvement of institutions' reputation and the quality of their products.

## 29.1 Introduction

Risk management is one of the fundamental matters that falls within the scope of what is known as informational systems research, since it helps to protect the assets of any company's information system [1]. Risk management is a strategic management application and a practical systematic procedure that allows companies to identify, analyze, and control risk processes. This helps to ensure good quality and reducing the risk of product or service failure [2]. Companies are capable of managing the common risks that can be found in field activities, by applying a risk system that allows them to manage their activities in more effective ways. This system enables them to obtain better results at a lower cost and in a short period of time [3]. There are several methods that can be utilized for risk analysis, such as those presented in ICH Q9 on Quality Risk Management like Failure Mode Effects and Analysis (FMEA), Failure Mode Effects and Critical Analysis (FMECA), Fault Tree Analysis (FTA), Hazard Analysis and Critical Control Points (HACCP), Primitive Risk Analysis (PHA), and Risk Ranking and Filtering [4]. FMEA has been an important risk management technology to understand and use. Also, it has been used in the field of management and quality, including personnel and general aspects of operations activities of industrial companies.

The Failure Modes and Effects Analysis (FMEA) is a proactive method which systematically identifies, analyzes, and mitigates potential product and process failures. As well, it assists in developing of test methods and troubleshooting strategies. Moreover, it offers a base for qualitative reliability, maintainability, safety, and logistics analyses, and estimates of system critical failure rates [5]. This method can be applied through determining the potential occurrence of the product or the process, its fundamental causes, consequences, and effects [6]. Identifying and recognizing the consequences and effects of these failures on the level of performance and safety, proper actions should be taken to remove or to reduce them at least. Subsequently, FMEA is an essential reliability tool that plays a major role in evading costs resulted from product or process failure and liability. The uses of FMEA's were developed for the first time by the US Military at the end of the 1940's due to their displeasure with failures in the production of munitions. This resulted in the development of a technique that would minimize all the potential causes of malfunctioning of munitions. The technique was documented: MIL-P-1629, and showed such a high effectiveness that it was used in the aerospace and nuclear industry [7–9]. In the early 70s, FMEA

was applied in the automotive industry as an effective method to improve product quality. Ford Motor Company presented the first integration of FMEA in automotive industry, by implementing it in their design process. In the present day, it is commonly used in risk assessment and quality improvement in various industries, including medicine, military, automobile, and semiconductor [10]. The FMEA-organization is an applicable to all levels of the business process, starting with the first level, which includes the management system, information system, production system, individual system, marketing system, and financial system, and ending with the last level, which involves organizing a work task [11].

The application of FMEA includes creating a risk factor known as the Risk Priority Number (RPN). RPN offers a quantitative score in order to evaluate failures where each failure is actually converted into a numerical value. Accordingly, RPN evaluates the severity of the potential failure of three parameters: severity, occurrence, and detestability. While severity is the result of customer, occurrence is the possibility of occurrence of the failure, and detestability is the possibility to detect the failure before it reaches the customer [12]. This implies that detecting a higher RPN value indicates a higher priority of risk [13]. Proper corrective actions are generally recommended depending on RPN threshold value. If this threshold value is reached, a risk mitigation procedure is applied correspondingly [14].

On the other hand, FMEA is widely criticized for several conceptual features. The most prominent weakness of this technique is the qualitative and narrative way of its structure. To explain more, FEMA documents are usually developed by experts, through using subjective linguistic terms that are formulated on the personal evaluation of the product or the process. The RPN parameters' values which are estimated by experts may contain uncertainty and ambiguity [15]. What is more, the parameters used in FEMA are introduced by (1–10) crisp scale which is considered as being an undependable representation of real-application cases [16]. As well, the RPN assessment has been criticized by Chang et al. in [17] for the heterogeneous morphologic associations between severity, occurrence, and detestability, which represent the three parameters as mentioned above. This criticism is formulated on the fact that each of three parameters is attained and linearly multiplied by the other with an indistinguishable scale. This process is completed regardless of the actual effect of each independent parameter and the dissimilar qualitative interpretation of the scale.

Traditionally, FMEA only considered the impact of system failure system in addition to the assessment of its root causes, which usually took a long period of time. This is why it was necessary to reconsider this method, especially when it comes to defining and calculating risks in order to come up with useful and unified results in a short period of time. Accordingly, FMEA documents have to be interactively updated on a regular basis. To put it in another way, employing new technologies is essential to overcome these deficiencies. Thus, this paper aims to introduce the feature of artificial intelligence in the FMEA method. It focuses on the development of FMEA to an electronic application, in which artificial intelligence is used to simultaneously identify, calculate, and evaluate risks. This paper presents a real field study on how to develop the FMEA from a traditional method, in which risks are assessed in an individual and personal way, to a modern developed program, in

which artificial intelligence is used as an alternative to the working team in terms of calculating and evaluating various risks. In this study, the field of construction was taken as a primary area for evaluating the performance of the modern FMEA program. The work team took two field hazards in construction as a sample to use in this study. 5 factors have been suggested as being the causes of each risk, with an assessment of 50 possible cases of each risk's occurrence. These data are used in the training of artificial intelligence. That is to come up with results that allow the modern FMEA program to calculate and evaluate the risks on a data basis, making the results close to the results provided by experts, using the traditional FMEA. The integration of artificial intelligence in FMEA method is an advanced qualitative transition in the field of risk management, since it allows companies to identify points of failure, anticipate risks, and calculate them in an innovative and smooth way, avoiding risks in advance. This contributes in the improvement of institutions' reputation and the quality of their products.

## 29.2 Research Methodology

FMAE is a systematic method and a technique for analyzing a system to track probable failure modes, their reasons, and how they might affect the performance of a particular system [18, 19]. In this scientific study, we introduced the artificial intelligence feature to the FMEA method, by developing the FMEA into an electronic application. This application uses artificial intelligence to identify, calculate, and evaluate risks at the same time. In other words, it presents an accurate field study on how to develop the FMEA from a traditional method, in which risks are assessed in an individual and personal way, to a modern and developed program. Thus, the artificial intelligence is used as an alternative to the work team in calculating and evaluating various risks.

In the chosen field of construction, we have identified various risks and relevant causes to each risk. Two risks, "Construction project delivery failed on time" and "Workers falling from high floors," were identified by experts and were included in this data, along with five factors affecting each risk, as shown in Tables 29.1 and 29.2.

A dataset with significant data was built for each case based on the figures provided by our expert. The essence of this work is to measure the number of deviations in the forecasts explained by the dataset. Simply put, it represents the difference in results between the samples in the dataset and the predictions made by the model.

Risk Priority Number (RPN) is considered to be a crucial index in the FMEA. It is the composite of the ratings for occurrence (O), severity (S), and detection (D), as indicated in equation $RPN = O \times S \times D$ [20]. O stands for "occurrence of failure," indicating the likelihood that the failure mode will manifest itself as a result of a particular cause; S stands for "severity," indicating how seriously the potential failure mode will affect the process once it has manifested, and D stands for the likelihood that a potential failure will be detected [21]. The deficiency lies especially

**Table 29.1** The application of FMEA on risk—construction project delivery failed on time

| Construction project delivery failed on time | | | | | | | |
|---|---|---|---|---|---|---|---|
| Work has been suspended due to the Corona pandemic (2 week for 20% from of the total number of workers) | Technical error in the structure of the building require reconstruction in some places | The pace of work is slow in the winter due to the harsh weather conditions (−40 below zero) | Some officials' lack of knowledge of the goals and dates set by the company | A severe shortage of funding for raw materials such as iron and concrete leads to delays in implementation | O | D | S |
| 0 | 0 | 0 | 0 | 0 | **0** | **0** | **0** |
| 1 | 0 | 0 | 0 | 0 | **3** | **1** | **3** |
| 0 | 1 | 0 | 0 | 0 | **1** | **1** | **2** |
| 0 | 0 | 1 | 0 | 0 | **2** | **1** | **3** |
| 0 | 0 | 0 | 1 | 0 | **3** | **1** | **1** |
| 0 | 0 | 0 | 0 | 1 | **3** | **1** | **3** |
| 3 | 10 | 5 | 3 | 2 | **5** | **4** | **5** |
| 2 | 5 | 3 | 2 | 4 | **5** | **3** | **5** |
| 0 | 4 | 2 | 0 | 6 | **4** | **2** | **4.5** |
| 5 | 7 | 4 | 5 | 9 | **5** | **5** | **5** |
| 2 | 0 | 3 | 0 | 3 | **3.5** | **1.5** | **4.5** |

**Table 29.2** The application of FMEA on risk—workers falling from high floors

| Workers falling from high floors | | | | | | | |
|---|---|---|---|---|---|---|---|
| Falling roof structures and columns | Falling through existing openings | Falling from stairs | Work at height without safety equipment | Lack of a safety system for heights | O | D | S |
| 0 | 0 | 0 | 0 | 0 | **0** | **0** | **0** |
| 1 | 0 | 0 | 0 | 0 | **2** | **3** | **1** |
| 0 | 1 | 0 | 0 | 0 | **3** | **1** | **1** |
| 0 | 0 | 1 | 0 | 0 | **1** | **2** | **1** |
| 0 | 0 | 0 | 1 | 0 | **2** | **4** | **1** |
| 0 | 0 | 0 | 0 | 1 | **3** | **4** | **1** |
| 3 | 5 | 4 | 7 | 3 | **5** | **5** | **4** |
| 4 | 3 | 7 | 0 | 0 | **4.5** | **3** | **2** |

in conducting RPN calculations individually. The potential value of RPN was not fixed; rather, it varied according to the work team. In addition, there was a duplicate RPN value [22].

## 29.3 Case Study

### *29.3.1 Project Background*

The project under consideration is a local study that was started by a significant well-known firm in Russia. This project was accomplished within 14 months and it is considered to be one of the biggest industrial structures in the world. It is located in Blagoveshchensk, the Amur Region. The two building yards make up approximately 32,000 $m^2$ of the entire construction space. Since concrete is the primary building material, each yard has three structures, each of which has six floors.

The teamwork is made up of 12 experts, including a customer representative, a project manager, a worker officer, four occupational health, safety engineers, three quality engineers, an architect, and a design engineer. They are the ones who constantly oversee work being done on the building site. Consequently, they might offer significant information that supports the case study. The team cooperated in order to collect the required data to accomplish the report.

## 29.4 Result

This work focuses; as it is mentioned above, on introducing the artificial intelligence to the FMEA method to quantify the number of forecast variations that the dataset can account for. Simply expressed, it is the variation between the model's predictions and the dataset's samples. First and foremost, we have determined two risks and 5 some factors that affect each risk. Based on the numbers produced by our expert for each case, a dataset with important data was prepared.

Next, our dataset (listing-1) which includes 50 cases for each risk was loaded into our program (listing-2). This program analyzed the dataset (shown by Figs. 29.1 and 29.2) for each risk that was in dimension of 50 by 8. It means that we analyzed 50 cases with 5 factors affecting 3 outputs O, S, and D. The diagram below represents how each factor affects D (yellow), O (green), S (blue) based on the 50 cases. We can see that our *x*-axis in the diagram represents the scale of the specific factor for the risk, as the following: 10, 14, 12, 7, 12, 9, 13, 11, 13, 8 for every factor in each risk, respectively. Our *y*-axis represents maximum value of 5 that D, O, S can get as an output.

For each of the graphs below (shown by Figs. 29.1 and 29.2), we can see sort of cloud which represents the strong correlation. As well, we can see some dots

that stay out of the "cloud." These dots might be trimmed out for analysis to give better results. One example of such deviation could be the result stated in (12, 2) for Detection value at "A severe shortage of funding for raw materials, such as iron and concrete leads to delays in implementation" factor of "Construction project delivery failed on time" risk.

Focusing on calculating some correlation measures, we first looked at an important statistical building block called covariance [11]. Variables can be related by a linear relationship. In Fig. 29.3 we can see two matrices for each risk, respectively. These matrices represent correlation. The main diagonal of the matrix contains the covariance between each variable and itself. The other values in the matrix represent the covariance between the two variables.

The coefficient returns a value between − 1 and 1. It represents the limits of correlation from full negative correlation to full positive correlation. A value of 0 means no correlation. Thus, the value should be interpreted, where often a value below − 0.5 or above 0.5 indicates a marked correlation, and values below these

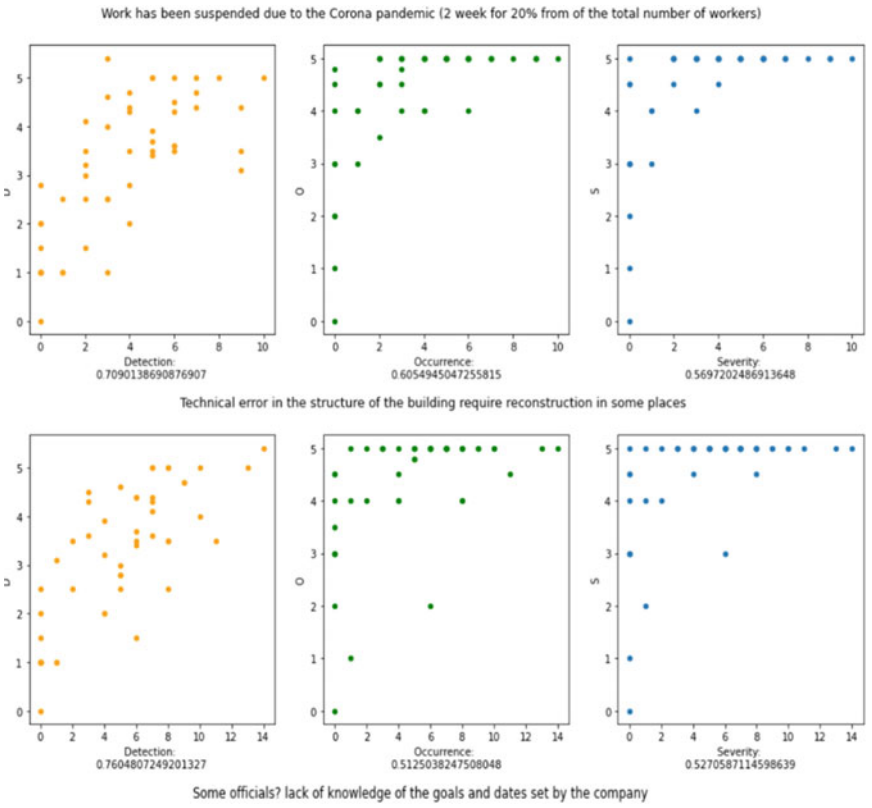



**Fig. 29.1** Scatterplot of clustered data to show clusters and centers—construction project delivery failed on time

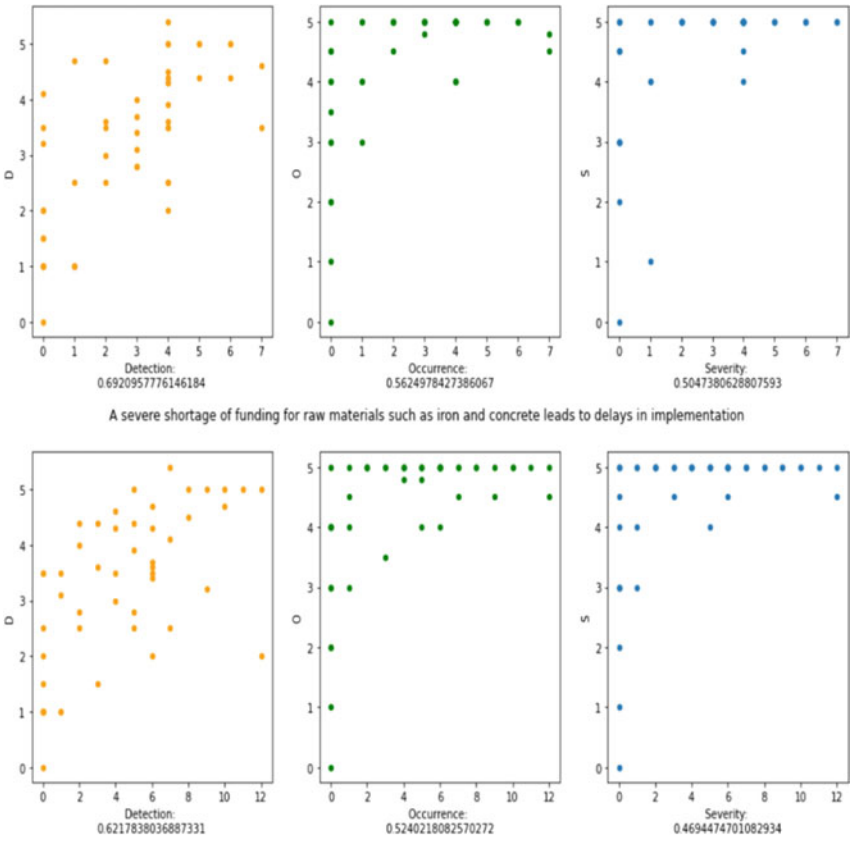



**Fig. 29.1** (continued)

values indicate a less marked correlation. As an example, for the first risk, we can see that correlation between "Technical error in the structure of the building requires reconstruction in some places" and "Work has been suspended due to the Corona pandemic" is (0.4); it is considered as one of the least marked correlation. Same could be said about "Failing from stairs," "Falling roof structures and columns" of the risk of Workers falling from high floors is (0.36). On the other hand, we can see strong correlation between "Technical error in the structure of the building require reconstruction in some places" of the risk "Construction project delivery failed on time" and Detection (D). Strong correlation in risk 2 can be said about the factor of "Falling through existing openings" and Detection of the risk. The result is stated in next two Figs. 29.3 and 29.4.

If we must consider many fields of life where we can implement our model for each field, we have variety of risks consisting of number of factors. Also, we will have a big number of models to calculate our outputs (D, O, S) for. For the reason of shortening the number of models for this case and for future models, our team

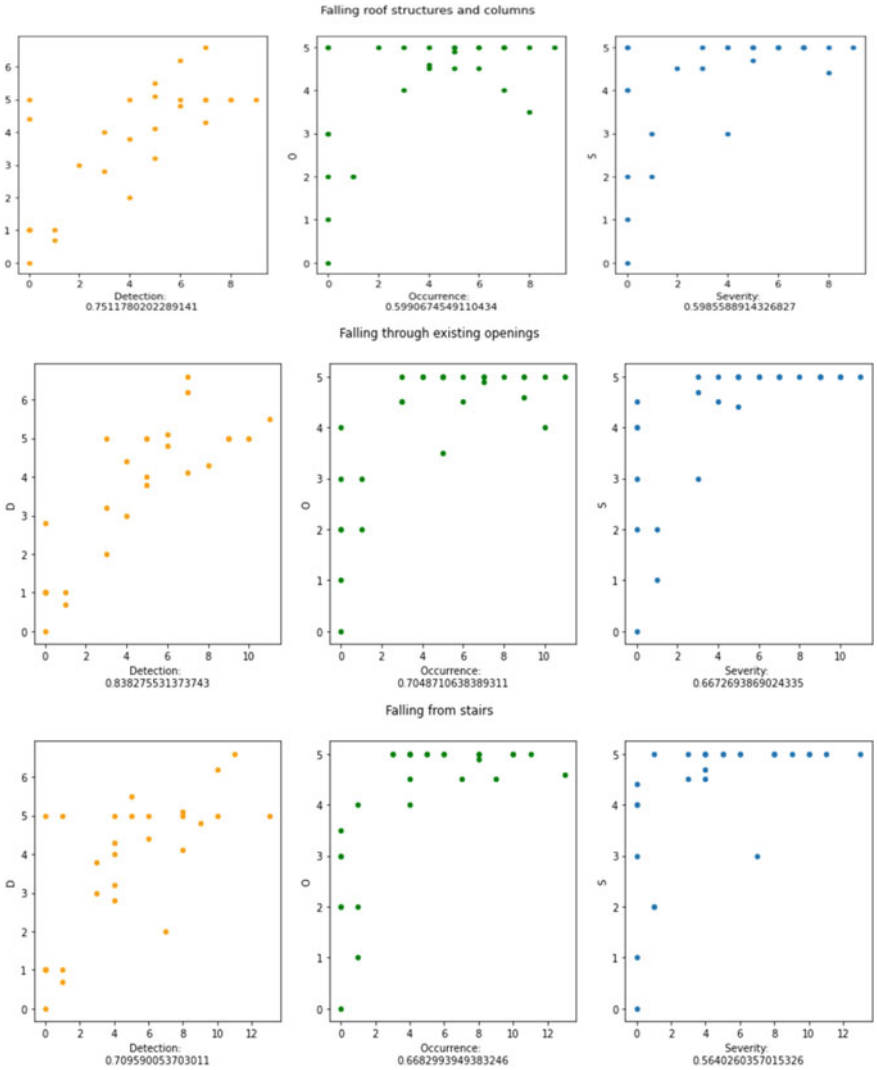



**Fig. 29.2** Scatterplot of clustered data to show clusters and centers 2—workers falling from high floors

has decided to use multi output regressor outputs instead of single output regressor algorithm. The number of models to train will decrease threefold due to our decision which will optimize our model. In order to train our models, we focused on 3 types (multi output regressor algorithms), namely:

1. Linear Regression.
2. Training with the Random Forest Regressor Model.
3. Training with the Decision Tree Regression Model.

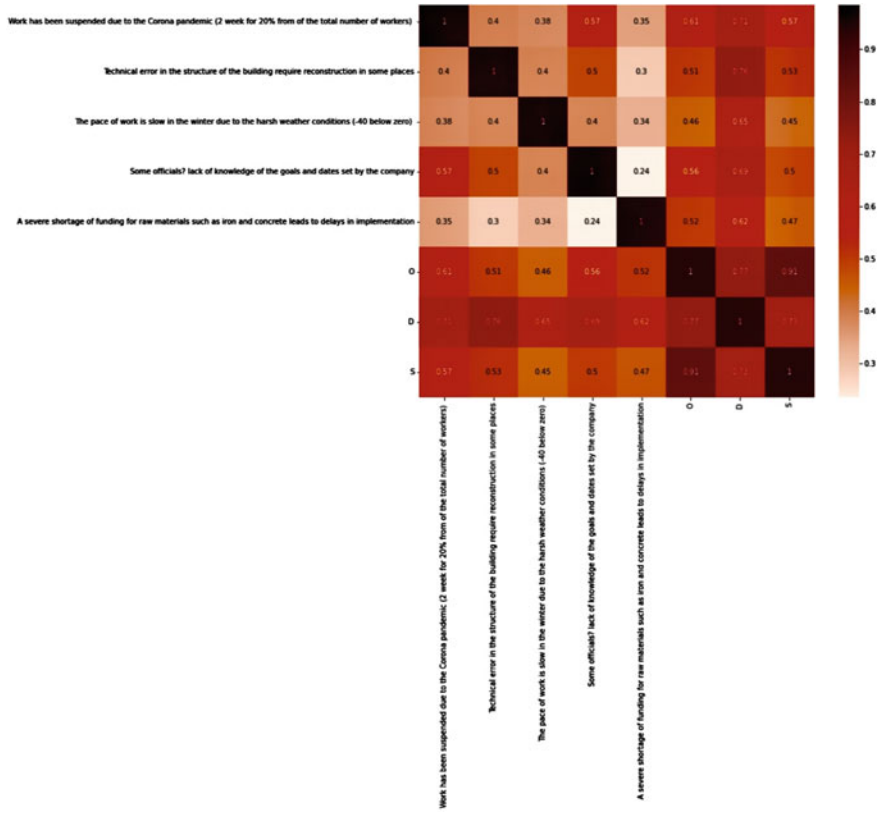



**Fig. 29.3** Correlation matrix 1—construction project delivery failed on time

The results, produced by for each model are given in Table 29.3.

The $R^2$ score is a crucial metric that is used to evaluate the performance of a machine learning model based on regression. It is pronounced R squared and it is also known as the coefficient of determination [23]. The essence of this work is to measure the number of deviations in the forecasts explained by the dataset:

- First, we make an 80/20 split into training and test samples.
- Further, the following algorithm was used to train the model.
- Linear Regression Random Forest Regressor Decision Tree Regressor.

In all the above algorithms, Grid Search CV was used to find the optimal learning parameters.

As a result of the work, the worst result was shown by the Linear Regression algorithm (linear regression), since there is no linear relationship between the features. And the best metric algorithm MAE = 0.36 turned out to be an algorithm Random Forest Regressor.

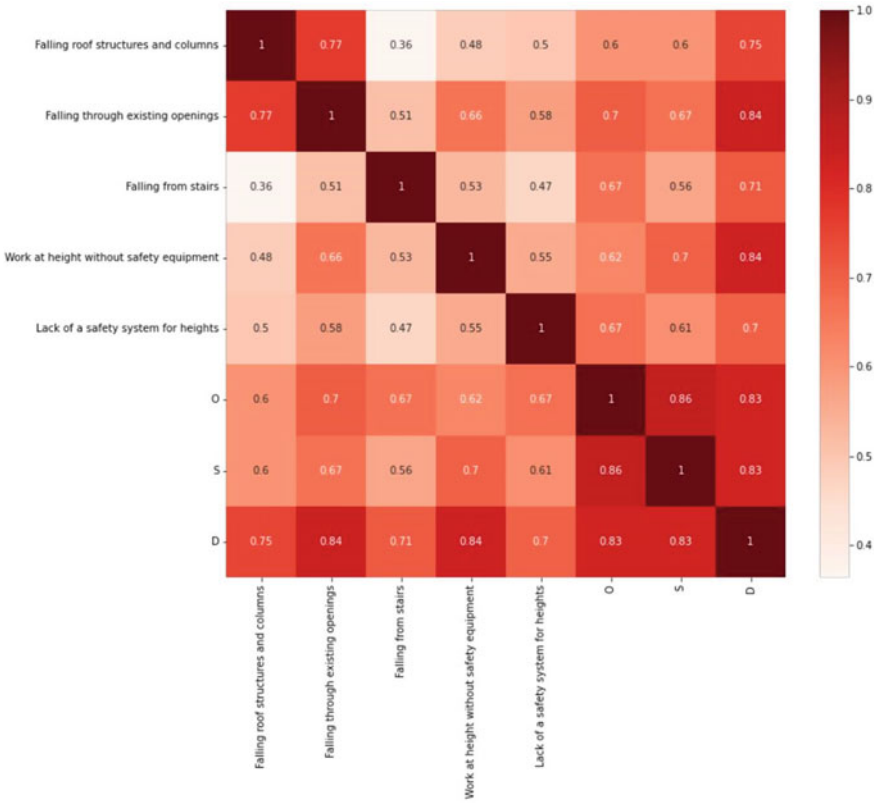



**Fig. 29.4** Correlation matrix 2—workers falling from high floors

**Table 29.3** Model training (multi output regressor algorithms)

| | Linear regression | Random forest regressor | Decision tree regressor |
|---|---|---|---|
| Trian_Time | 00:00:00:02 | 00:00:00:42 | 00:00:00:02 |
| Trian_MAE | 0.521294 | 0.365142 | 0.520833 |
| Test_MAE | 0.517319 | 0.4324 | 0.62 |
| $R^2$ | 0.405628 | 0.801885 | 0.739649 |

Next, we checked the importance of features stated bellow. The features that play a big role in training the model for Construction project delivery failed on time are presented bellow:

1. Work has been suspended due to the Corona pandemic (2 weeks for 20% from of the total number of workers).
2. Technical error in the structure of the building require reconstruction in some places.

3. Least affecting feature in the model is "A severe shortage of funding for raw materials such as iron and concrete leads to delays in implementation."

For Workers falling from high floors risk the feature playing a huge role is:

1. Falling from stairs

For the same risk, least affecting feature is falling from stairs as can be seen from Fig. 29.5.

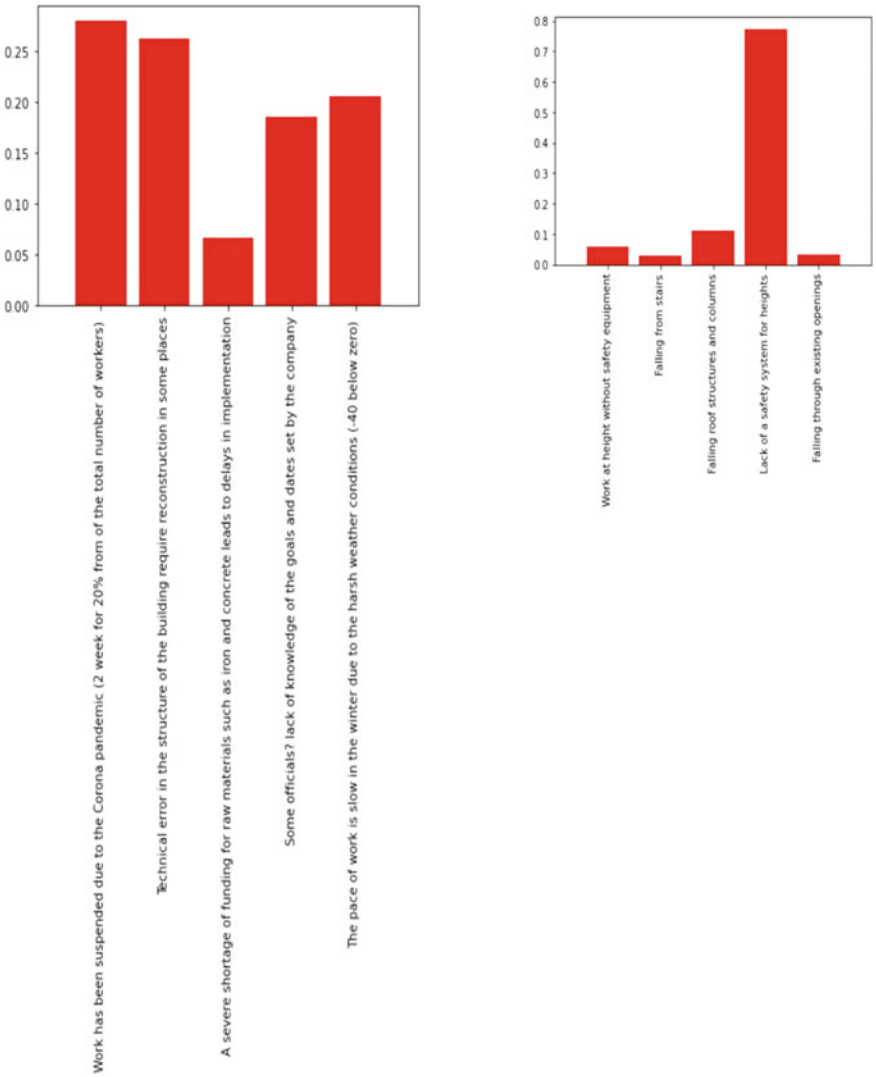



**Fig. 29.5** The graph of random forest regressor algorithm result

## 29.5 Discussion

The FMEA as a methodology, in these stated articles [24–29], was only used to identify, categorize, and evaluate the risks that have persisted in diverse domains vaguely. In this matter, researchers put out a number of strategies to enhance the use of FMEA and the establishment of RPN. To address the shortcomings of FMEA, a novel risk assessment strategy using a variety of fuzzy methodologies was investigated. In their article [13], Haktanır and Kahraman have compiled a number of fuzzy methodologies and the grey theory, along with the interval-valued neutrosophic (IVN) sets-based FMEA, to get rid of human evaluations and subjective conclusions that are inaccurate. Also, a new risk analysis technique has been proposed by Ayber and Erginelto get rid of the ambiguity of the linguistic terms; this technique is called Single-Valued Neutrosophic (SVN) Fuzzy FMEA [12]. In addition to that, cloud model theory and the hierarchical TOPSIS method were utilized by Liu et al. to improve FMEA effectiveness, eliminate human judgment bias, and make it easier to convert qualitative concepts to numeric values [7]. In order to enhance the traditional way of computing the RPN, Keskin and Zkan introduced a fuzzy adaptive resonance theory (ART) method for FMEA modeling, which ultimately lowered the cost and effort needed to respond to corrective action alerts [10].

## 29.6 Conclusion

This scientific study introduces the FMEA to an electronic application, in which artificial intelligence is used to simultaneously determine, calculate, and evaluate risks. It presents an actual case study of how to transform the FMEA from an approach in which risks are assessed on an individual and personal basis to a more advanced program that uses artificial intelligence as an alternative to the work team, when calculating and evaluating various risks. Using FMEA, the study identified a number of risks in addition to the pertinent causes for each risk in the chosen construction. These risks are "Construction project delivery failed on time" and "Workers falling from high floors." The extracted data were used to train artificial intelligence to come up with results that allow the modern FMEA program to automatically calculate and evaluate risks. Eventually, these results were close to those presented by experts when using traditional FMEA.

The outcomes allow us to raise the level of the business to an effective, higher, strong performance. Also, it increases productivity, effective health management of the employees, and enhance the reputation of the company.

The main focus of our future research is to widen the function of the developed FMEA by introducing another technique that states not only the risks and their causes in an effective way, but also suggests and provides the most suitable solutions to avoid it.

## References


1. Reed, A.H., Mark, A.: Risk management usage and impact on information systems project success. Int. J. Inform. Technol. Project Manage. **9**(2), 1–19 (2018). https://doi.org/10.4018/IJITPM.2018040101
2. Zhao, X., Bai, X.: The application of FMEA method in the risk management of medical device during the lifecycle. In: 2nd International Conference on E-business and Information System Security (EBISS2010), pp. 455–458 (2010)
3. Cicek, K., Celik, M.: Application of failure modes and effects analysis to main engine crankcase explosion failure on-board ship. Saf. Sci. **51**, 6–10 (2013)
4. Sharma, K.D., Srivastava, S.: Failure mode and effect analysis (FMEA) implementation: a literature review. Adv. Res. Aero Space Sci. **5**(1&2), 1–17 (2018)
5. Hezla, L., Avdotin, V., Gurina, R., Derouiche, L.: The role of organizational failure mode, effects and criticality analysis (FMECA) in correcting, implementing and achieving corporates' goals in HSE. Int. J. Sup. Chain. Mgt. **10**(3) (2021). file:///C:/Users/benz/Downloads/5816-16943-1-PB%20(3)%20(1).pdf
6. Claxton, K., Campbell-Allen, N.M.: Failure modes effects analysis (FMEA) for review of a diagnostic genetic laboratory process. Int. J. Qual. Reliab. Manag. **34**(2), 265–277 (2017)
7. Liu, H.-C., Wang, L.-E., Li, Z., Hu, Y.-P.: Improving risk evaluation in FMEA with cloud model and hierarchical TOPSIS method. IEEE Trans. Fuzzy Syst. **27**, 84–95 (2019)
8. Hezla, L., Avdotin, P., Plyuschicov, V.G., Sambros, N.B., Hezla, N., Derouiche, L.: The role of organizational failure mode, effects & analysis (FMEA) in risk management and its impact on the company's performance. Association for Computing Machinery, Manchester, United Kingdom, 15–17 May 2020
9. MIL-P-1629: Procedure for performing a failure mode effect and criticality analysis. United States Military Procedure, 9 Nov 1949
10. Keskin, G.A., Özkan, C.: An alternative evaluation of FMEA: fuzzy ART algorithm. Qual. Reliab. Eng. Int. **25**, 647–661 (2009)
11. Hezla, L., Avdotin, V., Derouiche, L., Plushikov, V., Norezzine, A., Kucher, D., Khomenets, N., Poddubsky, A., Gadzhikurbanov, A., Ivanov, N., Dokukin, P., Rebouh, N.Y.: The relationship of organization failure modes and effects analysis with the safety quality for supply chain risk management. Int. J. Sup. Chain. Mgt. **9**(2) (2020). file:///C:/Users/benz/Downloads/4713-13830-1-PB%20(4)%20(1).pdf
12. Ayber, S., Erginel, N.: Developing the Neutrosophic Fuzzy FMEA Method as Evaluating Risk Assessment Tool, pp. 1130–1137. Springer, Cham, Switzerland (2020). ISBN 9783030237554
13. Haktanır, E., Kahraman, C.: Failure Mode and Effect Analysis Using Interval Valued Neutrosophic Sets, pp. 1085–1093. Springer, Cham, Switzerland (2020). ISBN 9783030237554
14. Al-Khafaji, M.S., Mesheb, K.S., JabbarAbrahim, M.A.: Fuzzy multicriteria decision-making model for maintenance management of irrigation projects. J. Irrig. Drain. Eng. **145**, 04019026 (2019)
15. Yang, C., Zou, Y., Lai, P., Jiang, N.: Data mining-based methods for fault isolation with validated FMEA model ranking. Appl. Intell. **43**, 913–923 (2015)
16. Stamatis, D.H.: Failure Mode and Effect Analysis: FMEA From Theory to Execution. ASQC Press, New York (2003)
17. Chang, C.L., Wei, C.C., Lee, Y.H.: Failure mode and effects analysis using fuzzy method and grey theory. Kybernetes **28**, 1072–1080 (1999)
18. Gandhi, O.P., Agrawal, V.P.: FMEA—a diagraph and matrix approach. Reliab. Eng. Syst. Saf. **35**(2), 147–158 (1992)
19. Sader, S., Husti, I., Daróczi, M.: Enhancing failure mode and effects analysis using auto machine learning: a case study of the agricultural machinery industry. Processes **8**(2), 224 (2020). https://doi.org/10.3390/pr8020224
20. Tay, K.M., Lim, C.P.: Fuzzy FMEA with a guided rules reduction system for prioritization of failures. Int. J. Qual. Reliab. Manage. **23**(8), 1047–1066 (2006)

21. Zeng, S.X., Tam, V.W.Y., Tam, C.M.: Integrating safety, environmental and quality risks for project management using a FMEA method. Eng. Econ. 44–52 (2010)
22. Braglia, M., Frosolini, M., Montanari, R.: Fuzzy criticality assessment model for failure modes and effects analysis. Int. J. Qual. Reliab. Manage. **20**(4), 503–524 (2003). https://doi.org/10.1108/02656710310468687
23. Turney, S.: Coefficient of Determination (R2)|Calculation and Interpretation (2022). https://www.scribbr.com/statistics/coefficient-of-determination/
24. Wessiani, N.A., Sarwoko, S.A.: Risk analysis of poultry feed production using fuzzy FMEA. Procedia Manuf. **12**, 270–281 (2015)
25. Renua, A., Visotskya, D., Knackstedt, S., Mockoa, G., Joshua, D., Schulteb, J.: In: 6th CIRP Conference on Assembly Technologies and Systems (CATS). Procedia CIRP **6**, 157–162 (2016)
26. Balaraju, J., Raj, M.G., Murthy, C.S.: Fuzzy-FMEA risk evaluation approach for LHD machine—a case study. J. Sustain. Min. **12**, 257–268 (2019)
27. Patil, R.B., Kothavale, B., Waghmode, L.: Failure mode effect and criticality analysis (FMECA) of manually and electrically operated butterfly valve. In: 2nd SRESA National Conference on 'Reliability and Safety Engineering (NCRS'15), p. 6 (2015)
28. Patel, M.T.: A case study: a process FMEA tool to enhance quality and efficiency of manufacturing industry. Bonfring Int. J. Ind. Eng. Manage. Sci. **4**(3), 145–152 (2014)
29. Hecht, H., Baum, D.: Failure propagation modeling in FMEAs for reliability, safety, and cybersecurity using SysML. In: 17th Annual Conference on Systems Engineering Research (CSER), vol. 153, pp. 370–377 (2019)

# Chapter 30
# Rescue Decision Support for Marine Wrecked Ships Based on Multi-agent Modeling and Simulation

**Lu Yang, Hu Liu, YuanBo Xue, YongLiang Tian, and Xin Li**

**Abstract** Maritime search and rescue (MSAR) play an important role in the development of the maritime business. In particular, the rescue for marine wrecked ships is one of the current hot research topics in the field of MSAR. During the response of rescue for marine wrecked ships, one of the key issues is how to develop an optimal rescue plan after an accident occurs. To provide the decision support for the real rescue, this paper proposed a simulation system by using the multi-agent modeling and simulation method. Specifically, the mission scene is modeled for marine wrecked ships including main agents involved, information flows, rescue process. Based on that, a multi-agent model is built considering the main individuals in the rescue mission. Moreover, the evaluation indicator system is established to evaluate various rescue plan by using the analytic hierarchy process. After that, a simulation system is designed with multiple agents and the evaluation model. Finally, a case study is carried out to verify the decision support capability of the simulation system. Three different rescue plans are developed to describe rescue mission scenes and analyses the rescue effectiveness. In the case study, it could be concluded that S-76C++ has better rescue capability than S-76D for marine wrecked ships. By using the proposed models and evaluation system, the simulation system can be served as a decision support system for potential maritime rescue applications.

## 30.1 Introduction

With the rapid development of marine business, the frequency of marine accidents has increased significantly. Having said that, maritime search and rescue (MSAR) plays an important role in the development of the maritime business. It is important to develop a rescue plan quickly and efficiently when an accident occurs [1]. For marine wrecked ships, as an important proportion of maritime accidents, usually

L. Yang · H. Liu · Y. Xue · Y. Tian (✉) · X. Li
Beihang University, Beijing 100083, China
e-mail: tianyongliang_buaa@163.com



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_30

presents unique characteristics: large number of people in distress, multiple equipment collaboration rescue, difficulties in rescue mission planning, and unable to predict the effectiveness of the rescue plan. Therefore, our question is to choose a preferable rescue plan for the collaboration of multiple equipment.

For maritime rescue mission planning algorithms, many research papers have been conducted. Based on the possible drift range of the target in distress, Kratzke proposes an algorithm to determine the location and size of the optimal search area for search and rescue forces [2]. The folding search algorithm is used to determine the optimal search area for search and rescue forces by Agbissoh [3]. For the problem of SAR force deployment, selection and allocation, a model of maritime SAR helicopter allocation based on a hybrid optimization approach was proposed by Koester [4]. Chen proposed a global optimal planning method of mission area for helicopter MSAR missions, and proposed a simulation evaluation method of helicopter MSAR missions considering uncertainty factors [5]. Xiong proposed a helicopter MSAR mission planning method on a minimum bounding rectangle and k-means clustering [6].

However, these algorithms are specific to a particular mission in the whole rescue mission which cannot simulate the whole operations in the rescue process for wrecked ships. Therefore, we are motivated to develop a system which can simulate the whole process for the maritime rescue and provide decision support for the rescue plan. Regarding of the decision support system for marine rescue, the Search and Rescue Planning Program (SARP) was launched by the US Coast Guard in 1970 which is the world's first decision support system for MSAR mission planning [7]. The system supports drift trajectory prediction of distress targets based on environmental data. After that, the Search and Rescue Planning Program (SARP) is established, which could provide functions including personnel survival time calculation, available search and rescue force deployment information display, search mission planning and program evaluation [8]. The National Maritime Search and Rescue Support System of China is built in 2016. And it realizes a complete set of decision support services for drift prediction calculation, search and rescue force selection, search and rescue mission planning, and search and rescue result prediction [9]. But the maritime search and rescue force supported by the system only contains rescue ships.

In that case, a simulation system is expected to provide decision support for multi-aircraft collaborative rescue of marine wrecked ships. The simulation system could simulate different rescue plans and predict the effect of the rescue plan when an accident occurs. Then, the simulation system chooses the optimal rescue plan for actual rescue. Besides, the simulation system could prepare ourselves unconventional accidents that have never been observed in the real world. Therefore, a rescue decision support for marine wrecked ships is built in this paper.

This paper is organized as follows: The following section will introduce the mission scene model of the wrecked ship where the multi-agent model, the evaluation indicator system, and the simulation system are given; a case study is studied in Sect. 30.3 for the comparison and analyses of the multiple rescue plans; finally, this paper is finally concluded in Sect. 30.4.

## 30.2 Model and Simulation System

### *30.2.1 Mission Scene Model*

A maritime accident usually includes rescue bases, rescue equipment, and resettlements. Compared to other maritime accidents, the rescue time for wrecked ship is longer with the large number of people in distress, which makes the rescue mission more difficult. Having said that, more rescue equipment is required to be put into the rescue mission, such as, fixed-wing aircraft, helicopters, and rescue ships. As a consequence, it is important to choose the combination of rescue equipment and make rescue decision. Besides, the influence of the environment on the unit in distress cannot be ignored.

Considering that, the mission scene model of wrecked ship is shown in Fig. 30.1. The mission is initiated with a distress message from the distress unit. After receiving the distress message from the distress units, the rescue mission is assigned to fixed-wing aircraft, rescue helicopters, and rescue ships by rescue base. Throughout the rescue mission, fixed-wing aircraft perform search tasks and helicopters are responsible for rescue tasks. In a rescue mission, helicopters transit distress units to rescue ships and resettlements, and rescue ships and resettlements supply helicopters. And considering the actual distress situation, distress units are in different situations, including on the ship, on the lifeboats, overboard. The people overboard are influenced by environment, including winds, water temperature, etc.

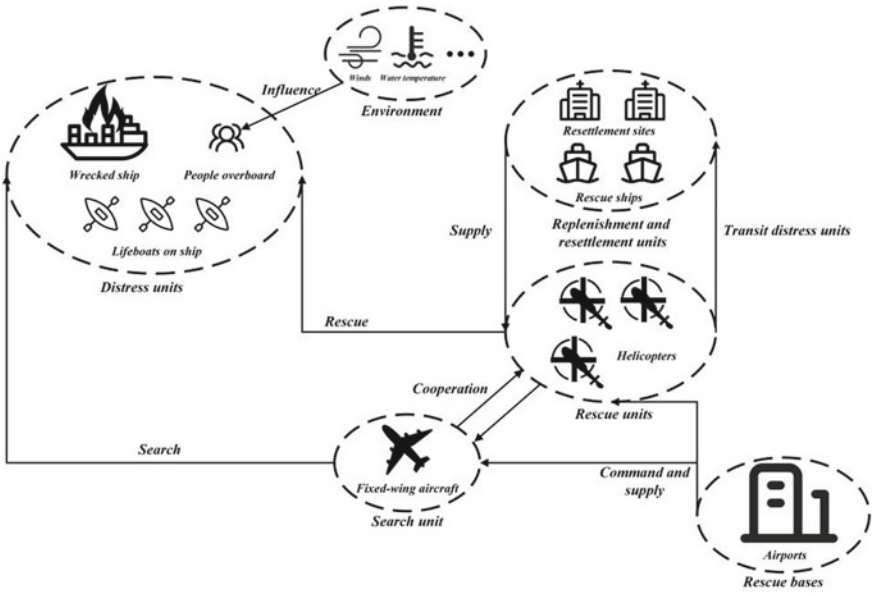



**Fig. 30.1** The mission scene model for marine wrecked ships

### *30.2.2 Multi-agent Model*

It can be seen from the mission scene model of wrecked ship that it is of great significance to model units in rescue missions. Obviously, there are different behavioral logics for different units. In addition, there is information exchange between different units and they can influence each other's behavioral decisions. Therefore, the rescue mission is modeled by many interaction and communication behaviors between units with explicit behavioral logic and state migration characteristics.

The multi-agent modeling and simulation is adopted to model and simulate the rescue mission of wrecked ship [10]. The main units in the rescue mission are modeled based on the mission scene model. The multi-agent model includes environment agent, distress people agent, rescue ship agent, resettlement site agent, fixed-wing aircraft agent, helicopter agent, and airport agent. And a detailed description of each type of agents is given in Table 30.1.

Based on the mission scene model and the description of the multi-agent model, the detailed modeling of the agent in a rescue mission is as follows.

A. **Environment Agent**

The function of the environment agent in a rescue mission is affecting the state, decision and behavior of other agents. In this paper, the parameters of the environment agent include wind speed ($wp$), water temperature ($wt$), wave height ($wh$), and visibility ($vis$). Assuming that the parameters of the environment agent would not change over time.

B. **Distress People Agent**

Distress people agent is generated to simulate the distress people in a rescue mission. Based on the actual rescue process, the parameters of the distress people agent at a

**Table 30.1** The description of the multi-agent model

| Agent type | Agent name | Description of this agent |
|---|---|---|
| Environment agent | A. Environment agent | The agent for simulating the real influence of environment to other agents |
| Distress agent | B. Distress people agent | The rescued target in a rescue mission |
| Resettlement agent | C1. Rescue ship agent | The agent for receiving the distress people and supplying the helicopters |
| Resettlement agent | C2. Resettlement site agent | The agent for receiving the distress people and supplying the helicopters |
| Search agent | D. Fixed-wing aircraft agent | The agent for performing search missions |
| Rescue agent | E. Helicopter agent | The agent for performing rescue and transit missions |
| Rescue base agent | F. Airport agent | The agent not only for receiving the distress people and supplying the helicopters, but also for deploying the helicopters and fixed-wing aircraft |

certain moment of the simulation ($C_{\mathrm{dp}}$) are defined as Eq. (30.1):

$$C_{\mathrm{dp}} = \{cl, ds, inj, rt\}, \tag{30.1}$$

where $cl$ is the current location of the distress people, the $ds$ is the situations of the distress people, $inj$ is the injuries of the distress people, $rt$ is the remaining live time of the distress people.

Assuming that the location of the distressed unit does not change over time. Ats mentioned, distress units are in different situations, including on the ship, on the lifeboats, overboard. Considering the actual situation, only overboard people are at risk of death. And maximum survival time ($T_{\max}$) of overboard people is related to the water temperature in the environment. In addition, the influence of the injuries of overboard people could not be ignored. Then, $T_{\max}$ could be given by Eq. (30.2):

$$T_{\max} = \gamma \cdot 5.75 \cdot e^{0.1 \cdot wt}, \quad 0.5 \leq \gamma \leq 1, \tag{30.2}$$

where $\gamma$ is injury factor, which is used for indicate injuries to overboard people. Then, $rt$ could be given by Eq. (30.3):

$$rt = \begin{cases} T_{\max} - t, & t < T_{\max} \\ 0, & t \geq T_{\max} \end{cases}, \tag{30.3}$$

where $t$ is the time after people overboard.

C1. **Rescue Ship Agent and C2. Resettlement Site Agent**

Both rescue ship agent and resettlement site agent are used for receiving the distress people and supplying the helicopters (could not refuel the helicopters). The parameters of the rescue ship agent and resettlement site agent are defined as Eq. (30.4):

$$C_r = \{rsl, pa, hcp\}, \tag{30.4}$$

where $rsl$ is the location of the rescue ship agent/resettlement site agent, $pa$ is the number of people currently available for resettlement, and $hcp$ is the number of working helicopters at the rescue ship agent/resettlement site agent.

D. **Fixed-Wing Aircraft Agent**

In a rescue mission, the search mission for distress people agents is performed by fixed-wing aircraft agent. In a simulation, performance parameters ($C_f^p$) and state parameter ($C_f^s$) are used to model the fixed-wing aircraft agent. $C_f^p$ is defined by Eq. (30.5), and $C_f^s$ is defined by Eq. (30.6).

$$C_f^p = \left\{fa, fvc, fm^f, fr^f\right\} \tag{30.5}$$

$$C_f^s = \{fcl, fv, fn^f, ctw^f, npf\} \tag{30.6}$$

In Eq. (30.5), $fa$ is the airport of the fixed-wing aircraft agent, $fvc$ is the cruising speed of the fixed-wing aircraft agent, $fm^f$ is the maximum fuel weight of the fixed-wing aircraft agent, and $fr^f$ is the fuel consumption rate of the fixed-wing aircraft agent.

In Eq. (30.6), $fcl$ is the current location of the fixed-wing aircraft agent, $fv$ is the speed of the fixed-wing aircraft agent, $fn^f$ is the remaining fuel of the fixed-wing aircraft agent, $ctw^f$ is the current target way point of the fixed-wing aircraft agent, and $npf$ is the number of the distress people agents found by fixed-wing aircraft agent.

Considering the no-fly zone in the actual rescue process, for $ctw^f$, the Rapidly-exploring Random Tree algorithm* is used to plan the route for the fixed-wing aircraft agent [11].

Other than that, the behavior model of the fixed-wing aircraft agent is shown in Fig. 30.2. In Fig. 30.2, for the behavior model D5, the search area size depends on the winds. And the behavior model D6 is judged by $npf$. Continue if $npf$ is less than the number of distress people, otherwise end. In addition, the behavior model D9 is judge by Eq. (30.7). Continue if satisfied, otherwise perform D3.

$$fn^f \geq fr^f \times \left(\frac{L}{fvc} + \Delta t\right), \tag{30.7}$$

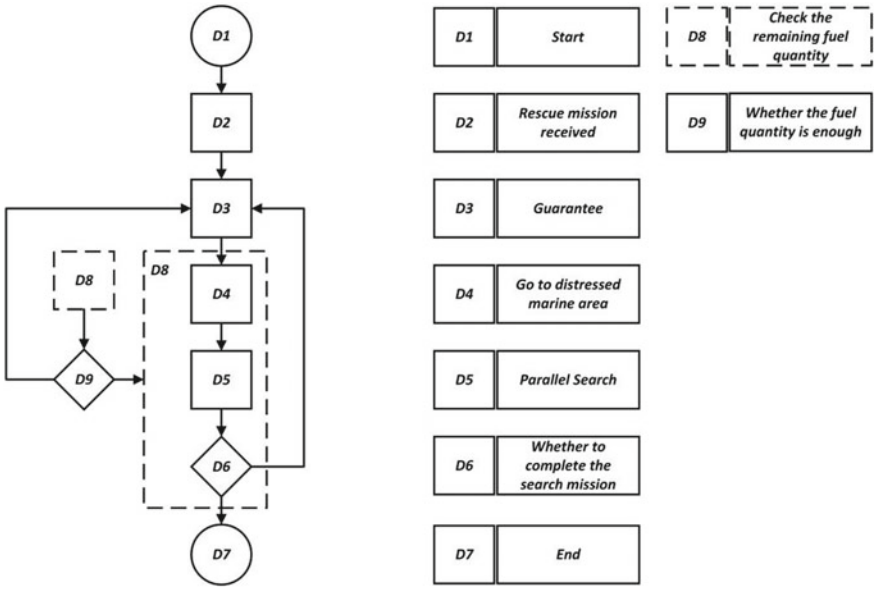



**Fig. 30.2** The behavior model of the fixed-wing aircraft agent

where $L$ is the distance between $fcl$ and the airport, and $\Delta t$ is the required time to perform the remaining search mission.

E. **Helicopter Agent**

In a rescue mission, the rescue and transit missions for distress people agents is performed by helicopter agent. In a simulation, performance parameters ($C_h^p$) and state parameter ($C_h^s$) are used to model the helicopter agent. $C_h^p$ is defined by Eq. (30.8), and $C_h^s$ is defined by Eq. (30.9).

$$C_h^p = \left\{ha, hvc, hmp, fm^h, fr^h\right\} \tag{30.8}$$

$$C_h^s = \left\{hcl, hv, fn^h, ctw^h, hnp\right\} \tag{30.9}$$

In Eq. (30.8), $ha$ is the airport of the helicopter agent, $hvc$ is the cruising speed of the helicopter agent, $hmp$ is the maximum number of people carried of the helicopter, $fm^h$ is the maximum fuel weight of the helicopter agent, and $fr^h$ is the fuel consumption rate of the helicopter agent.

In Eq. (30.9), $hcl$ is the current location of the helicopter agent, $hv$ is the speed of the helicopter agent, $fn^h$ is the remaining fuel of the helicopter agent, $ctw^h$ is the current target way point of the helicopter agent, and $hnp$ is the number of the distress people agents carried by helicopter agent at a certain moment of the simulation.

Considering the no-fly zone in the actual rescue process, for $ctw^h$, the Rapidly-exploring Random Tree algorithm* is used to plan the route for the helicopter agent [11].

Other than that, the behavior model of the helicopter agent is shown in Fig. 30.3.

In Fig. 30.3, the behavior model E10 is judged by Eq. (30.10). Perform E12 if satisfied, otherwise perform E5/E11.

$$hnp \leq hmp \tag{30.10}$$

The behavior model E15 is judged by Eq. (30.11). Perform E17 if satisfied, otherwise perform E16.

$$\begin{cases} pa \geq 0 \\ hcp \leq 2 \end{cases} \tag{30.11}$$

The behavior model E19 is judged by Eq. (30.12). Perform E3 if satisfied, otherwise perform E4.

$$fn^h \geq fr^h \times \left(\frac{L_3 + L_4}{hvc}\right) + 200 \tag{30.12}$$

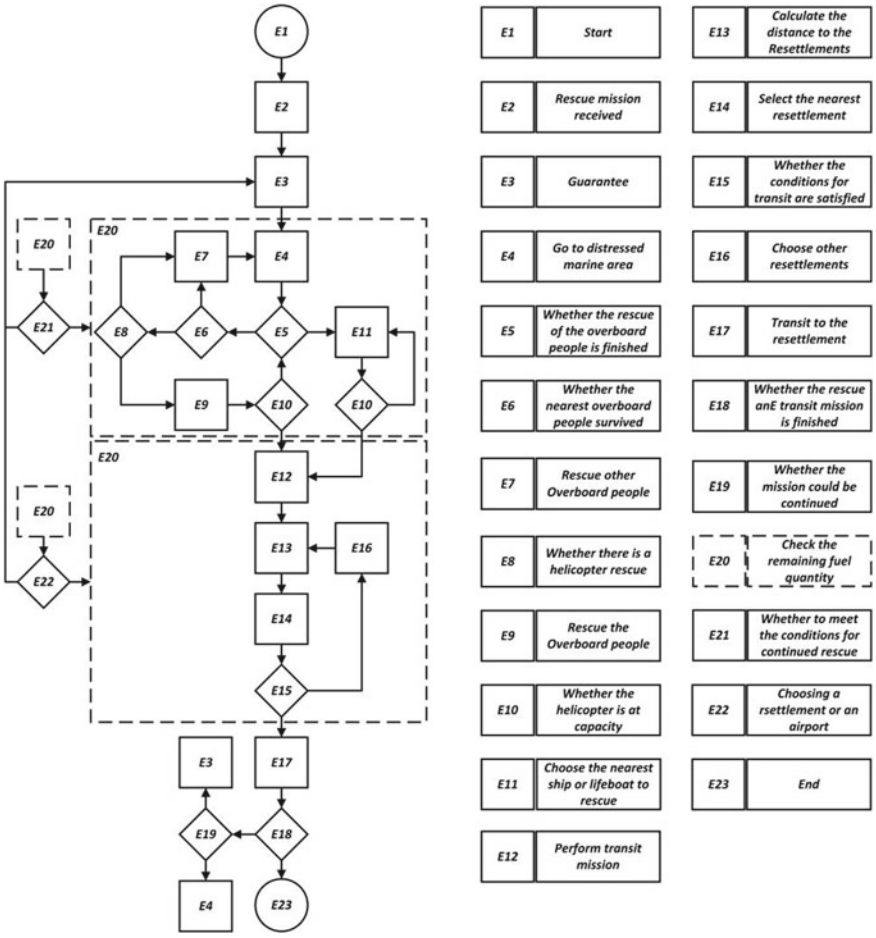



**Fig. 30.3** The behavior model of the helicopter agent

where $L_3$ is the distance between $rsl$ and distress marine area, $L_4$ is the distance between distress marine area and the airport, and 200 indicates that for safety reasons, the helicopter has to leave 200 kg fuel.

The behavior model E21 is judged by Eq. (30.13). Continue if satisfied, otherwise perform E3.

$$fn^h \geq fr^h \times \left( \frac{2L}{hvc} + hnp \times \Delta t_h \right) + 200, \tag{30.13}$$

where $L$ is the distance to the airport, and the $\Delta t_h$ is the required time for rescue a distress person. $\Delta t_h$ could be calculation by Eq. (30.14).

$$\Delta t_h = \Delta t_{\text{one}}/(\mu), \tag{30.14}$$

where $\Delta t_{\text{one}}$ is a constant value, and $\mu$ depends on $wp$, $wh$, and $vis$.

The behavior model E22 is judged by Eq. (30.15). Continue if satisfied, otherwise perform E3.

$$fn^h \geq fr^h \times \left(\frac{L_3 + L_5}{hvc}\right) + 200, \tag{30.15}$$

where $L_5$ is the distance between $rsl$ and the airport.

F. **Airport Agent**

The airport agent not only for receiving the distress people and supplying the helicopters (could refuel the helicopters), but also for deploying the helicopters and fixed-wing aircraft. The parameters of the airport agent are defined as Eq. (30.16):

$$C_p = \{acl, apa, hcp, df, fh\}, \tag{30.16}$$

where $acl$ is the location of the airport agent, $apa$ is the number of people currently available for resettlement, $hcp$ is the number of working helicopters the airport agent, $df$ is the deployed fixed-wing aircraft at the airport agent, and $fh$ is the deployed helicopter at the airport agent.

Based on the detailed modeling of the agent in a rescue mission, the multi-agent model of the rescue mission for the wrecked ship is shown in Fig. 30.4. The multi-agent model shows the interaction between agents. Besides, the behavior logic of the rescue mission is shown in Fig. 30.4.

### *30.2.3 Evaluation Model*

It is justified to formulate multiple rescue plans for a rescue mission. Therefore, it is essential to evaluate the effectiveness of rescue plans and choose the best for real rescue. The effectiveness reflects the ability of the system for accomplishing the specified missions. In this paper, to evaluate the effectiveness of different plans, several effectiveness indicators of the plan are selected (shown in Table 30.2).

$C_1$ is the security of the rescue mission, which considers the security of the distress units, and the availability of rescue equipment. $C_2$ is the efficiency of rescue mission, which considers the availability of rescue equipment, the helicopter route planning, the influence of the environment, and utilization of rescue resources.

In the actual rescue process, the relative importance of each indicator varies. Therefore, in this paper, the analytic hierarchy process (AHP) [12] is used to calculate indicator weights by using Eq. (30.17).

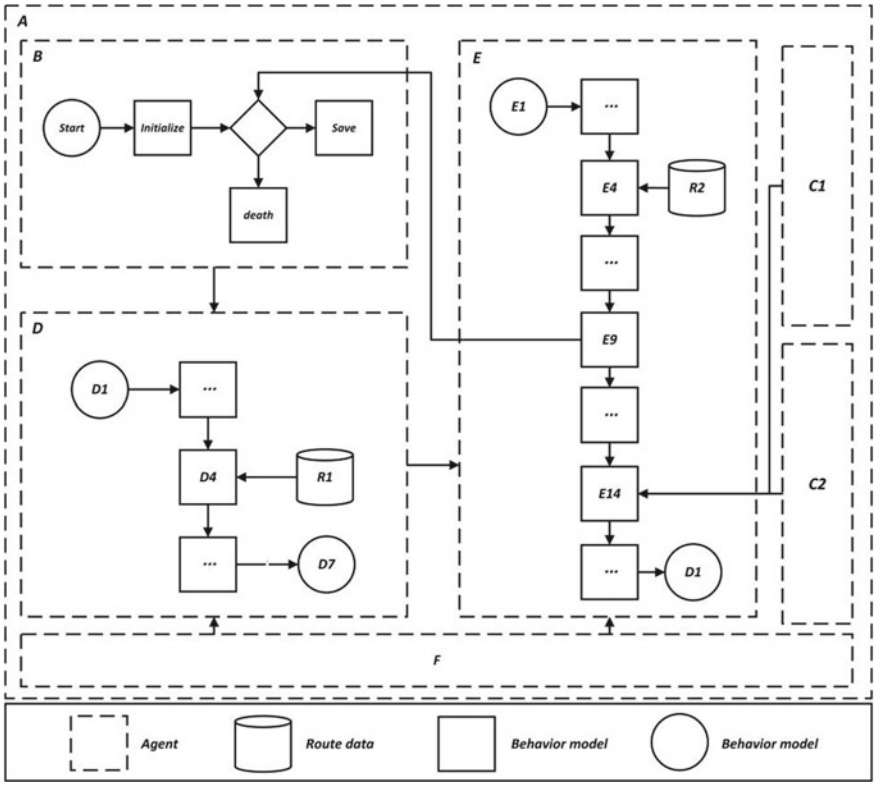



**Fig. 30.4** The multi-agent model of the rescue mission for the wrecked ship

$$A^k = \left(a_{ij}^k\right)_{N\times N} = \begin{bmatrix} a_{11}^k & a_{12}^k & \cdots & a_{1N}^k \\ a_{21}^k & a_{22}^k & \cdots & a_{2N}^k \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1}^k & a_{N2}^k & \cdots & a_{NN}^k \end{bmatrix}, \quad a_{ij}^k = \frac{1}{a_{ji}^k}, \tag{30.17}$$

where $A^k$ is the combined judgment matrix after multiple expert ratings, and $a_{ij}^k$ is importance of the $i$th indicator relative to the $j$th indicator.

After the evaluation indicator system is established, data collection based on evaluation indicator system is evaluated. In order to reduce the influence of the differences in indicator units on evaluation results, it is necessary to normalize the indicator values. In this paper, the normalization process is applied by the threshold method. For benefit-based metrics ($I_{11}$, $I_{12}$, $I_{23}$), Eq. (30.18) is used for normalization, and for cost-based metrics ($I_{13}$, $I_{14}$, $I_{21}$, $I_{22}$, $I_{24}$), Eq. (30.19) is used for normalization.

$$y_i = \frac{1}{2} \cdot \frac{x_i}{\max\{x_i\}} \tag{30.18}$$

**Table 30.2** Description of the evaluation indicator system

| Criteria | Criteria description | Indicator | Indicator description |
|---|---|---|---|
| $C_1$ | Security of mission | $I_{11}$ | Percentage of overboard people rescued. Ratio of successful rescue of overboard people to all overboard people |
| | | $I_{12}$ | Percentage of all distress people rescued. Ratio of successful rescue of distress people to all distress people |
| | | $I_{13}$ | Average rescue time of overboard people. Ratio of total time spent on rescuing overboard people to successful rescue of overboard people |
| | | $I_{14}$ | Rescue time of the mission. Time from the start of the rescue mission to the end of the rescue mission |
| $C_2$ | Efficiency of mission | $I_{21}$ | Average time for helicopters to reach the distress marine which reflects the results of route planning |
| | | $I_{22}$ | Average number of helicopter sorties. Ratio of the total number of departures of all helicopters at the airport to the number of helicopters |
| | | $I_{23}$ | Number of resettlements and airports on rescue mission |
| | | $I_{24}$ | Average time taken to rescue a person from a distress location to a helicopter |

$$y_i = \frac{1}{2} \cdot \frac{\max\{x_i\} + \min\{x_i\} - x_i}{\max\{x_i\}}, \tag{30.19}$$

where $x_i$ is the original data, and $y_i$ is the normalized data.

### *30.2.4 Simulation System*

In order to verify the validity and feasibility of rescue plan, the simulation system is built with AnyLogic based on the multi-agent model and evaluation model. In addition, to improve the universality of the simulation system, the parameters of the agents could be changed according to the rescue mission and the rescue plan. Then, the framework of the simulation system is shown in Fig. 30.5.

Firstly, for a simulation of the rescue mission, the distress information is edited which includes the parameters of the environment agent, the parameters of the distress people agents, and the number of the distress people agent. Secondly, the database is built in advance, which not only includes the parameters of the rescue ship agents, the resettlement site agents, and the airport agents, but also has the planning routes of the fixed-wing aircraft agent and the helicopter agent. Thirdly, the parameters and number of the fixed-wing aircraft agent and the helicopter agent is determined, which

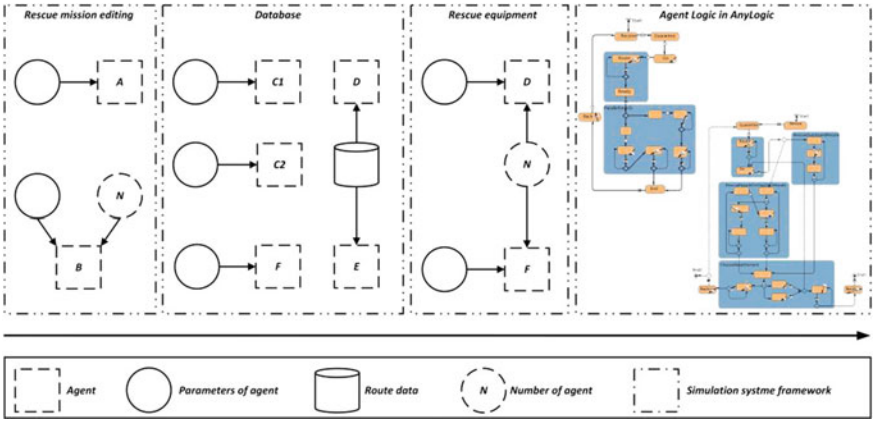



**Fig. 30.5** Framework of the simulation system

represents a rescue plan. Finally, the simulation is performed based on the behavior logic of the agents, and the evaluation data is exported.

The user interface of simulation system is shown in Fig. 30.6.

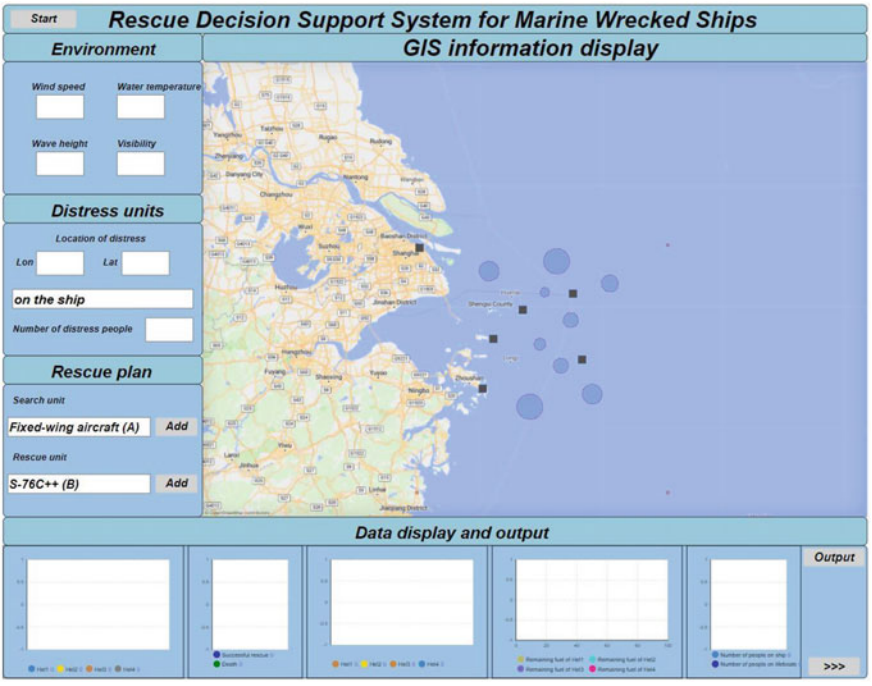



**Fig. 30.6** The user interface of simulation system

## 30.3 Case Study and Analysis

### *30.3.1 Rescue Mission Case*

To verify the decision support capability of the simulation system, a rescue mission case hypothesized a wrecked ship occurred at the East China Sea. The distress information of the rescue mission case is shown in Table 30.3. Available rescue equipment data is shown in Table 30.4. The airports information is shown in Table 30.5. The information of rescue ships and resettlement sites is shown in Table 30.6.

**Table 30.3** Distress information of the rescue mission case

| Distress information | Data |
|---|---|
| Location of distress | 30°30′ N, 124° E |
| Number of distress people | 750 people on the ship |
| | 450 people in the lifeboat |
| | 250 overboard people |
| Environmental information in distress | Water temperature 9 °C |
| | Wind speed 10 m/s |
| | Wave height 1.5 m |
| | Visibility 1.5 km |

**Table 30.4** Available rescue equipment data

| Rescue equipment | $vc$/Search speed (km/h) | $fm$ (kg) | $fr$ (kg/h) | Guarantee time (min) | $hmp$ | Number |
|---|---|---|---|---|---|---|
| Fixed-wing aircraft | 480/270 | 22,600 | 2880 | 30 | — | 1 |
| S-76D | 287 | 1120 | 374 | 20 | 12 | 5 |
| S-76C++ | 287 | 1063 | 342 | 20 | 13 | 2 |
| Modified S-76C++ | 287 | 1063 | 274 | 25 | 8 | 1 |
| H225 | 275 | 2553 | 576 | 30 | 19 | 1 |

**Table 30.5** The information of available airports

| Airport | $acl$ | $apa$ | Type |
|---|---|---|---|
| Airport A | 121°48′28″ E, 31°8′47″ N | – | Fixed-wing aircraft |
| Airport B | 121°37′31″ E, 31°20′5″ N | 1000 | Helicopter |
| Airport C | 122°21′23″ E, 29°55′53″ N | 1000 | Helicopter |

**Table 30.6** The information of rescue ships and resettlement sites

| Resettlement | *rsl* | *pa* | *hcp* |
|---|---|---|---|
| Rescue ship A | 123°24′3″ E, 30°52′50″ N | 120 | 2 |
| Rescue ship B | 123°30′21″ E, 30°13′22″ N | 150 | 2 |
| Resettlement site A | 122°49′12″ E, 30°43′12″ N | 1000 | 2 |
| Resettlement site B | 120°28′48″ E, 30°25′48″ N | 1000 | 2 |

### *30.3.2 Simulation Verification*

Based on the information of the rescue equipment, resettlements, and rescues bases, three rescue plans are developed for comparison. And the details of the rescue plans are shown in Table 30.7.

After the rescue plans are developed, the simulation and the evaluation analysis are initiated. Furthermore, based on the evaluation indicator system and AHP, the weight coefficient is determined by using judgment matrix after multiple expert ratings. Then, the weight coefficient and the normalized data of the indicators are shown in Table 30.8.

**Table 30.7** The details of the rescue plans

| Rescue plan | Search unit (Deployed airport) | Rescue units (Deployed airport) | | | |
|---|---|---|---|---|---|
| Rescue plan 1 | Fixed-wing aircraft (A) | S-76D (B) | S-76D (B) | S-76D (C) | S-76D (C) |
| Rescue plan 2 | Fixed-wing aircraft (A) | S-76D (B) | S-76C++ (B) | Modified S-76C++ (C) | H225 (C) |
| Rescue plan 3 | Fixed-wing aircraft (A) | S-76C++ (B) | S-76C++ (B) | Modified S-76C++ (C) | H225 (C) |

**Table 30.8** The weight coefficient and the normalized data of the indicators

| Indicator | Weight coefficient | Rescue plan 1 | Rescue plan 2 | Rescue plan 3 |
|---|---|---|---|---|
| $I_{11}$ | 0.3192 | 0.434 | 0.480 | 0.500 |
| $I_{12}$ | 0.1832 | 0.491 | 0.497 | 0.500 |
| $I_{13}$ | 0.1823 | 0.463 | 0.479 | 0.500 |
| $I_{14}$ | 0.065 | 0.463 | 0.500 | 0.499 |
| $I_{21}$ | 0.1098 | 0.500 | 0.494 | 0.494 |
| $I_{22}$ | 0.026 | 0.405 | 0.493 | 0.500 |
| $I_{23}$ | 0.037 | 0.417 | 0.500 | 0.417 |
| $I_{24}$ | 0.078 | 0.490 | 0.500 | 0.471 |

**Table 30.9** Simulation results of different rescue plans

| Rescue plan | Rescue plan 1 | Rescue plan 2 | Rescue plan 3 |
|---|---|---|---|
| Effectiveness value | 0.462 | 0.489 | 0.494 |

### *30.3.3 Analysis*

Based on the proposed evaluation model (see Sect. 2.3), the comprehensive effectiveness values of the three rescue plans are obtained after the simulation. Table 30.9 gives the simulation results of different rescue plans.

The comparisons and the merits of these plans are as follows:

(a) It can be noted that the effectiveness value of the Rescue Plan 1 is the lower than others, probably because Rescue Plan 1 has only one type of rescue helicopter, S-76D, which results in a weak collaboration.
(b) The effectiveness values of the Rescue Plan 2 and the Rescue Plan 3 are higher, and the Rescue Plan 3 is slightly higher than the effectiveness evaluation value of Rescue Plan 2. Comparing the two rescue plans, the analysis could be concluded that S-76C++ has better rescue capability than S-76D for marine wrecked ships.
(c) In addition to that described in (b), the percentage of overboard people rescued ($I_{11}$), the percentage of all distress people rescued ($I_{12}$), and the average rescue time of overboard people ($I_{13}$) in Rescue Plan 3, which are the indicators that experts are most concerned, is higher than it in Rescue Plan 2. Therefore, Rescue Plan 3 has more reference value for the actual rescue.

## 30.4 Conclusion

For marine wrecked ship rescue, this paper develops a simulation system based on multi-agent modeling. In particular, the mission scene model is modeled for marine wrecked ships including main agents involved, information flows, rescue process. Moreover, the evaluation indicator system is proposed to evaluate different rescue plans based on AHP. The security of the rescue mission and the efficiency of rescue mission is considered, and eight indicators is built to evaluate the effectiveness value of the rescue plans. Finally, a case study is designed to verify the simulation system.

The conclusions are as follows:

(a) The simulation system is capability to provide different combinations of rescue equipment for marine wrecked ships.
(b) Based on the effectiveness evaluation model to select the optimal combination of the rescue equipment for a rescue mission.
(c) The simulation system is portable and can be used for mission planning in any marine area with minor changes, including airports, resettlements, rescue equipment, no-fly zones, etc.

(d) The simulation system could provide the rescue decision support for the actual rescue. Besides, the simulation system is ability to conduct rescue practice for imagined accident.

In the future applications, to improve the reliability of the simulation system, the accuracy of rescue decisions compared to the actual situation should be verified by experts. In addition, as potential rescue missions evolving various rescue equipment, the issue at hand is to consider equipment collaboration, for example, amphibious aircraft. And it is expected for more accurate distress information, which could greatly improve the rescue success rate. Besides, it is essential to note the dynamic distress units in the actual rescue, which could influence the decision of marine wrecked ships. Furthermore, feedback from user experience of the simulation system should be considered to improve the interface design of the simulation system.

## References

1. Han, P., Li, Y., Jie, X.: Comparative research of the maritime search and rescue systems in developed countries and its enlightenment to China. J. Ocean Technol. 107–113 (2020)
2. Kratzke, T.M., Stone, L.D., Frost, J.R.: Search and rescue optimal planning system, pp. 1–8. IEEE (2010)
3. Otote, D.A., et al.: A decision-making algorithm for maritime search and rescue plan. Sustainability 2084 (2019). Multidisciplinary Digital Publishing Institute
4. Koester, R.J., et al.: Use of the visual range of detection to estimate effective sweep width for land search and rescue based on 10 detection experiments in North America. Wilderness Environ. Med. 132–142 (2014)
5. Liu, H., Chen, Z., Tian, Y., Wang, B., Yang, H., Wu, G.: Evaluation method for helicopter maritime search and rescue response plan with uncertainty. Chin. J. Aeronaut. 493–507 (2021)
6. Xiong, P., Liu, H., Tian, Y., Chen, Z., Wang, B., Yang, H.: Helicopter maritime search area planning based on a minimum bounding rectangle and K-means clustering. Chin. J. Aeronaut. **493–507**, 554–562 (2021)
7. Frost, J.R., Stone, L.D.: Review of search theory: advances and applications to search and rescue decision support (2001)
8. Kratzke, T.M., Stone, L.D., Frost, J.R.: Search and rescue optimal planning system. In: Proceedings of the 2010 13th International Conference on Information Fusion, pp. 1–8. IEEE (2010)
9. National Maritime Search and Rescue Support System. http://www.marinesar.cn/index.html. Last accessed 2022/8
10. Zeigler, B.P.: Theory of Modeling and Simulation. Wiley, New York (1976)
11. Tan, X., Li, G., Yi, J., Xue, C., Long, H.: Path planning of manipulator based on improved RRT algorithm. Comput. Integr. Manuf. Syst. (2022)
12. Saaty, T.L.: The Analytic Hierarchy Process. McGraw Hill, New York (1980)

# Chapter 31
# Development of an AI-Powered Interactive Hand Rehabilitation System


**Ryota Goto, Ari Aharari, and Farhad Mehdipour**

**Abstract** In recent years, serious game-based systems have emerged as valuable tools for improving hand rehabilitation. These systems utilize interactive virtual environments to enhance body mobility. This paper aims to develop a customized hand rehabilitation system tailored to individual rehabilitation needs. This system integrates an interactive interface with vision-based hand-tracking technology. By utilizing this approach, the proposed rehabilitation system enables users to participate in physical rehabilitation therapies actively. The interactive interface has been specifically designed to provide real-time feedback, including metrics such as finger bending, hand clapping, and thumb touch exercises. The evaluation results of the proposed system demonstrate its effectiveness in supporting hand rehabilitation and motivating individual users to engage in rehabilitation exercises.

## 31.1 Introduction

The global population is experiencing rapid aging, and Japan is facing particularly severe challenges. According to the Ministry of Health, Labor, and Welfare (MHLW) of Japan, the aging rate in the country is projected to reach 38.4% by 2065 [1]. As the population ages, the number of individuals requiring nursing care also steadily increases. As of July 2022, approximately 6.97 million people were certified as needing nursing care or support, which accounts for about 5.5% of Japan's total population [2].

Japan's overall population peaked at 128.08 million in 2008 and has since started to decline. It is expected to decrease further to 86.74 million by 2060. Conversely, the elderly population is projected to continue growing and is estimated to peak in 2042 [3]. Meanwhile, the healthcare landscape in Japan is facing significant challenges due

R. Goto · A. Aharari (✉)
Department of Computer and Information Sciences, SOJO University, Kumamoto, Japan
e-mail: info@ahrary.org

F. Mehdipour
Department of Information Technology, Otago Polytechnic—Auckland International Campus (OPAIC), Auckland, New Zealand



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_31

to the aging population and limited availability of medical services. The healthcare system in the country often imposes time limits on rehabilitation, resulting in patients being discharged before achieving full recovery. This premature discharge can have detrimental effects on patients' long-term outcomes, leaving them without the necessary support to regain their functional abilities. Furthermore, with the continuous aging of the population due to low birth rates, the strain on healthcare resources is expected to increase, making it even more challenging to provide adequate medical services in the future.

In light of these circumstances, there is a growing demand for home healthcare services in Japan. Specifically, focusing on hand rehabilitation, it becomes crucial to continue the rehabilitation process at home for patients who have been discharged before fully recovering. The recovery of hand function is known to be a slow and intricate process, and since hands are extensively utilized in daily activities, it is essential to provide post-discharge rehabilitation to ensure optimal functionality and quality of life. However, delivering effective rehabilitation in a home setting poses unique challenges that need to be addressed.

One significant challenge in home-based treatment is the absence of healthcare professionals. Unlike in clinical settings where healthcare providers can closely monitor and guide the rehabilitation process, patients at home lack direct supervision and guidance. This absence can hinder the progress of rehabilitation and lead to suboptimal outcomes. Additionally, the decreased frequency of intervention in home-based settings may result in reduced motivation and adherence to the rehabilitation program. Without regular interactions with healthcare professionals, patients may feel isolated and lack the necessary support and encouragement to continue their rehabilitation journey.

To overcome these challenges, one potential solution is the integration of assistive devices into the rehabilitation process. These devices can serve as substitutes for healthcare professionals, providing guidance, monitoring, and feedback to patients during their rehabilitation at home. By leveraging technological advancements, a rehabilitation system can be developed that combines interactive devices with intelligent algorithms to offer personalized and adaptive rehabilitation programs.

Such a rehabilitation system could include hand exoskeleton devices that assist in facilitating hand movements and tracking progress. These devices can be equipped with sensors to detect and analyze hand movements, providing real-time feedback to patients. The system can also incorporate virtual reality or augmented reality technologies, creating an immersive and engaging environment for rehabilitation exercises. Through these interactive interfaces, patients can receive visual and auditory cues, game-like challenges, and performance assessments, enhancing their motivation and engagement in the rehabilitation process.

Moreover, the rehabilitation system can leverage emerging technologies such as machine learning and artificial intelligence to adapt the rehabilitation program to individual patients' needs. By continuously analyzing and learning from patient data, the system can dynamically adjust the difficulty level of exercises, provide personalized recommendations, and track progress over time. This adaptive nature of

the system ensures that rehabilitation remains challenging yet achievable, promoting patient satisfaction and facilitating better outcomes.

The existence of such rehabilitation systems holds tremendous value in the current healthcare landscape, where the demand for home healthcare services is expected to rise. By enabling patients to independently carry out their rehabilitation at home with the assistance of these devices, the system addresses the limitations of traditional healthcare settings and empowers individuals to take control of their recovery. This approach not only improves access to rehabilitation services but also reduces the burden on healthcare resources, allowing healthcare professionals to focus on critical cases while still providing support remotely.

This paper focuses on the urgent need for effective home healthcare solutions in Japan due to the aging population and the limited availability of medical services. Specifically, hand rehabilitation requires continuous support and intervention to optimize functional recovery. The proposed rehabilitation system presented in this paper allows users to actively engage in physical rehabilitation therapies. The interactive interface has been meticulously designed to provide real-time feedback, including metrics such as finger bending, hand clapping, and thumb touch exercises. By integrating these rehabilitation systems into home-based treatment, the challenges posed by limited healthcare resources can be addressed, empowering patients to participate actively in their rehabilitation journey. Consequently, these systems have the potential to revolutionize the delivery of hand rehabilitation and improve outcomes for patients within the aging population.

## 31.2 Related Works

Real-time hand rehabilitation systems have undergone significant advancements in recent years, with a range of new technologies and approaches proposed for improving their accuracy, efficacy, and user-friendliness. In this session, we compare some of the latest technologies proposed in this field, including wearable sensors, haptic devices, virtual reality interfaces, and machine learning algorithms.

Wearable sensors are a popular choice for capturing hand movement and muscle activity in real-time hand rehabilitation systems. Recent advancements in sensor technology have led to the development of smaller, lighter, and more accurate sensors that can be worn comfortably on the hand. For example, the use of flexible sensors such as stretchable strain sensors or textile-based sensors can provide improved comfort and flexibility for the user. Liao et al. [4] introduced a novel type of strain sensor that exhibits exceptional sensitivity and remarkable stretchability. These sensors were developed using a combination of silver nanowires and a polydimethylsiloxane composite material. The key innovation of their approach was the incorporation of patterned percolating network microstructures, which significantly enhanced the sensitivities of the strain sensors.

Additionally, some systems combine multiple types of sensors, such as Inertial Measurement Unit IMUs and EMG sensors, to capture a more comprehensive picture of hand movement and muscle activity [5, 6].

Haptic devices are another important component of real-time hand rehabilitation systems, as they can provide feedback and guidance to the user during the rehabilitation process. Recent advancements in haptic technology have led to the development of devices that are more compact, portable, and user-friendly. For example, some systems use haptic gloves or sleeves that provide tactile feedback to the user based on their hand movement. Masmoudi et al. [7] introduced an innovative system utilizing virtual reality (VR) for fine motor rehabilitation. Recognizing the importance of the sense of touch in daily activities, the researchers incorporated haptic feedback into their system through the use of a vibrating glove. This glove was designed to aid patients in performing rehabilitation exercises effectively.

Other systems use exoskeletons or robotic devices that can assist or resist hand movement to provide a more challenging and engaging rehabilitation experience. Bouteraa et al. [8] introduced a novel exoskeleton device that serves not only as a mechanical system for rehabilitation but also incorporates an efficient tracking and traceability software solution. The device utilizes electromyography (EMG) signals captured during hand motion to detect the intention of hand opening or closing. Based on this detection, the exoskeleton device is activated to perform the desired rehabilitation task.

Virtual reality interfaces are also becoming increasingly popular in real-time hand rehabilitation systems, as they can provide an immersive and motivating environment for the user. Recent advancements in virtual reality technology have led to the development of more realistic and interactive virtual environments, as well as more user-friendly and accessible interfaces. For example, some systems use gamification techniques to make the rehabilitation process more engaging and rewarding for the user. Tuah et al. [9] focused on developing a classification framework for rehabilitation gamification. This framework is based on the requirements specific to rehabilitation gamification and the common characteristics of gamification used in rehabilitation applications. The primary contribution of this paper is the proposed classification, which offers valuable insights for researchers and practitioners in designing and implementing gamification techniques that enhance motivation and sustain engagement in rehabilitation treatment and care.

Machine learning algorithms are also an important component of real-time hand rehabilitation systems, as they can provide accurate and reliable analysis of hand movement and muscle activity. Recent advancements in machine learning have led to the development of deep learning models that can recognize and classify hand gestures with high accuracy and robustness. Additionally, some systems use reinforcement learning techniques to personalize the rehabilitation process based on the user's individual needs and progress. Mujahid et al. [10] introduced a lightweight model for gesture recognition that is based on the You Only Look Once (YOLO) v3 and DarkNet-53 convolutional neural networks. The proposed model aims to perform gesture recognition without the need for additional preprocessing steps, image filtering, or image enhancement techniques.

In this paper, we applied a vision-based framework that provides a set of pre-built ML models and tools that can be used to process a variety of media inputs, including images, video streams, and audio recordings. These models are designed to perform tasks such as object detection, pose estimation, facial recognition, and hand tracking, among others.

## 31.3 Human Pose Estimation

OpenPose [11, 12], PoseNet [13], and MediaPipe [14] are three prominent computer vision frameworks that have gained significant attention in the field of Human Pose Estimation (HPE) and tracking. While they share the common goal of extracting pose information from images or videos, each framework possesses unique characteristics and approaches. In this session, we will explore the key features, advantages, and limitations of OpenPose, PoseNet, and MediaPipe. These three HPE libraries specifications are summarized in Table 31.1. Among the three human pose estimation (HPE) libraries, a total of 17 keypoints are commonly detected. These keypoints provide information about the positioning of various body parts. Specifically, the head region consists of 5 commonly detected keypoints, including the ears, eyes, and nose. These keypoints are essential for accurately estimating the orientation and position of the head. The upper body region encompasses the shoulders, elbows, and wrists, and it is associated with six commonly detected keypoints. These keypoints play a crucial role in capturing the posture and movement of the upper limbs. By accurately estimating the positions of these keypoints, HPE libraries can provide valuable information for applications such as gesture recognition or upper body rehabilitation. Similarly, the lower body region comprises the hips, knees, and ankles, which are represented by six commonly detected keypoints. These keypoints are essential for analyzing lower body movements, such as walking, running, or performing exercises. By tracking the positions of the lower body keypoints, HPE libraries can provide valuable insights into gait analysis, sports performance, and lower body rehabilitation. The detection and tracking of these commonly detected keypoints enable comprehensive and detailed human pose estimation. By accurately estimating the positions of these keypoints, HPE libraries can facilitate a wide range of applications, including activity recognition, motion tracking, virtual reality, and healthcare. It's worth noting that the specific number and naming of keypoints may vary slightly between different HPE libraries, as they may adopt different keypoint definitions or conventions. However, the common goal is to accurately detect and track keypoints representing different body parts to enable robust and accurate human pose estimation.

**OpenPose**. Developed by Cao et al. [11, 12], is a widely used and comprehensive framework for multi-person pose estimation. It leverages deep learning techniques and a bottom-up approach to detect human body key points, including body joints and limb connections, in real time. OpenPose's strength lies in its ability to simultaneously track multiple individuals within a single frame, making it suitable

**Table 31.1** Specifications of each HPE library

| HPE libraries | Maximum number of keypoints | Underlying network |
|---|---|---|
| OpenPose [11, 12] | 135 | ImageNet with VGG-19 |
| PoseNet [13] | 17 | ResNet and MobileNet |
| MediPipe [14] | 33 | CNN |

for scenarios involving crowded scenes or group activities. Additionally, OpenPose provides robustness against occlusion and pose variations, allowing accurate estimation even when body parts are partially or fully hidden. However, the computational requirements of OpenPose can be high, limiting its real-time performance on resource-constrained devices.

**PoseNet**. Developed by Google [13], takes a different approach by employing a lightweight deep learning model for single-person pose estimation. Unlike OpenPose, PoseNet follows a single-person detection strategy known as "top-down," where a pre-trained model localizes the person in the image before estimating their pose. This approach makes PoseNet computationally efficient, enabling real-time inference on a variety of devices, including smartphones and web browsers. The simplicity and portability of PoseNet make it suitable for applications where single-person pose estimation is sufficient. However, the limitation of PoseNet lies in its inability to handle multiple people within a frame, which restricts its applicability in crowded or multi-person scenarios.

**MediaPipe**. Developed by Google [14], is a flexible framework that offers a wide range of computer vision and machine learning solutions, including pose estimation. MediaPipe's pose estimation module adopts a multi-person pose estimation strategy like OpenPose, utilizing a combination of deep learning models and efficient inference techniques. One notable advantage of MediaPipe is its emphasis on real-time performance and cross-platform compatibility. It provides optimized implementations for both mobile and desktop environments, enabling developers to deploy pose estimation applications on various devices. Additionally, MediaPipe offers a user-friendly interface, making it accessible to both researchers and developers. Nevertheless, the accuracy of pose estimation in MediaPipe may not be on par with OpenPose, especially in challenging scenarios with occlusion or complex poses.

OpenPose, PoseNet, and MediaPipe are influential frameworks in the field of human pose estimation, each with its unique strengths and limitations. OpenPose excels in multi-person pose estimation and robustness against occlusion but requires significant computational resources. PoseNet offers lightweight single-person pose estimation suitable for resource-constrained devices but lacks multi-person tracking capabilities. MediaPipe provides a flexible and real-time solution with cross-platform support but may sacrifice some accuracy compared to OpenPose. The selection of the most appropriate framework depends on the specific requirements of the application,

such as the number of people to be tracked, real-time performance constraints, and the target device's capabilities.

Considering the assumption that rehabilitation will be performed by one person and the need for a framework that offers both accuracy and speed, MediaPipe appears to be a suitable choice for our development.

## 31.4 Proposed System

The proposed system architecture, as illustrated in Fig. 31.1, provides an overview of the entire system. Rehabilitators actively engage in rehabilitation exercises while monitoring their progress on a screen. The system utilizes a camera to capture the activities of the care recipients, and MediaPipe is employed for accurate assessment of the rehabilitation exercises. The system then records the relevant data, including the number of repetitions, and stores it in Excel, enabling rehabilitation coaches to review the information at their convenience.

Figure 31.2 presents the flowchart of the system from the perspective of the rehabilitators. When the system is launched, an exercise menu is displayed on the screen, presenting a range of available exercises. The rehabilitators can select a specific exercise by instructing the care recipient to bring their index finger close to the corresponding number shown on the screen. Once the exercise is selected, the display switches and detailed instructions for the chosen exercise are presented.

The rehabilitators then proceed to perform the exercise according to the provided instructions. The system closely monitors the movements and actions of the rehabilitators through the camera, utilizing MediaPipe for accurate assessment of the exercise performance. If the exercise is executed correctly for the predetermined number of repetitions, the system acknowledges the completion and switches the display once again. The exercise menu is then presented once more, allowing for the selection of the next exercise.

This process is repeated throughout the rehabilitation session, ensuring that multiple exercises can be performed and recorded. The system keeps track of the exercises performed and the corresponding number of repetitions for each exercise. This data is systematically recorded and stored in Excel, enabling rehabilitation

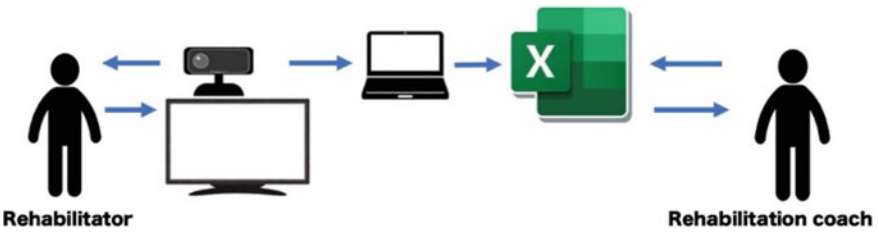



**Fig. 31.1** The proposed system architecture

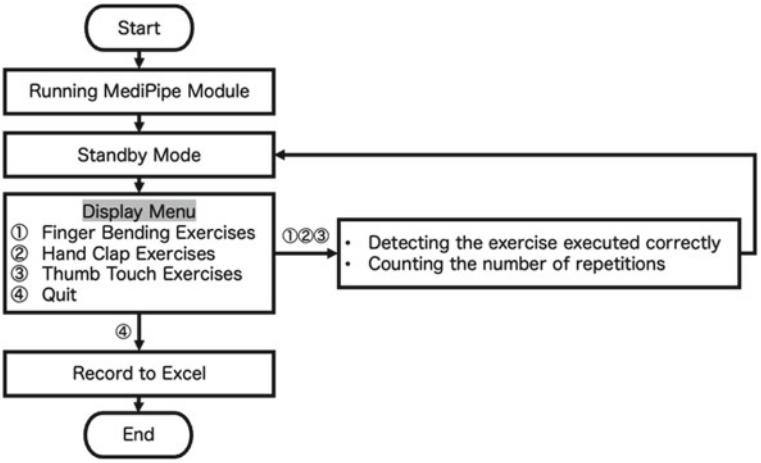



**Fig. 31.2** Flowchart of the proposed system from the perspective of the rehabilitators

coaches to conveniently review and analyze the progress and performance of the rehabilitators at any given time.

By adopting this system architecture, rehabilitation coaches can effectively monitor and support the rehabilitation process of rehabilitators. The interactive display and exercise selection mechanism provides a user-friendly interface, making it easier for both rehabilitators and rehabilitation coaches to navigate and engage with the system. The utilization of MediaPipe ensures accurate assessment and feedback on exercise performance, enhancing the effectiveness of the rehabilitation program.

Furthermore, the seamless integration of the system with Excel allows for efficient data management and analysis. Rehabilitation coaches can access the recorded data at their convenience, enabling them to evaluate the progress, identify areas for improvement, and tailor the rehabilitation program to meet the specific needs and capabilities of the care recipient.

This research focuses primarily on assessing hand movements. This rehabilitation is based on activities performed at our collaborator, the elderly care facility "Taiju," as well as rehabilitation recommended by the Ministry of Health, Labor and Welfare, and rehabilitation facilities.

**Finger Bending Exercises**. This rehabilitation involves alternating between making a fist and opening the hand. During the first position, it is necessary to bend all the fingers, while during the open hand position, all fingers should be fully extended. This rehabilitation is effective for brain activation and dementia prevention. Additionally, it helps prevent hypertension by repeatedly contracting and relaxing the muscles, and it also promotes metabolism by moving the fingers located at the extremities of the body, providing relief for conditions such as sensitivity to cold.

**Hand Clap Exercises**. This rehabilitation involves bringing both hands together. It is necessary to join the fingertips with the palms of the hands, and when separating, they should be kept a certain distance apart.

**Thumb Touch Exercises**. This rehabilitation involves touching the thumb to the index, middle, ring, and little fingers in that order. It is important to keep the other fingers that are not being touched straight without bending. This rehabilitation, which utilizes finger movements, is effective for brain activation and dementia prevention.

## 31.5 Experimental Results

The experimental results for each exercise's assessment method are presented as follows.

**Finger Bending Exercises**: The assessment method involves counting the number of extended fingers. When all fingers are bent with no extended fingers, it is classified as a "fist" position. Conversely, when all five fingers are fully extended, it is categorized as an "open hand" position. The count is incremented by 1 when transitioning from the "fist" position to the "open hand" position. The assessment method is visually depicted using images, with the landmarks omitted. Figure 31.3 illustrate the screens displayed when selecting exercise option one from the menu.

In Fig. 31.3, all fingers are bent, indicating a "fist" position. Then, all five fingers are fully extended, indicating an "open hand" position. As stated previously, the count increases by one during the transition from "fist" to "open hand," causing the count on the screen to change from 0 to 1. This process is repeated accordingly.

**Hand Clap Exercises**: The hand clap exercise involves calculating the distance between the palms of the left and right hands to determine whether they are close together (indicating a hand clap) or far apart (indicating hands are separated). Initially, the development plan utilized the Hands tool due to its hand recognition capabilities.


**Fig. 31.3** Selecting exercise option one from the menu

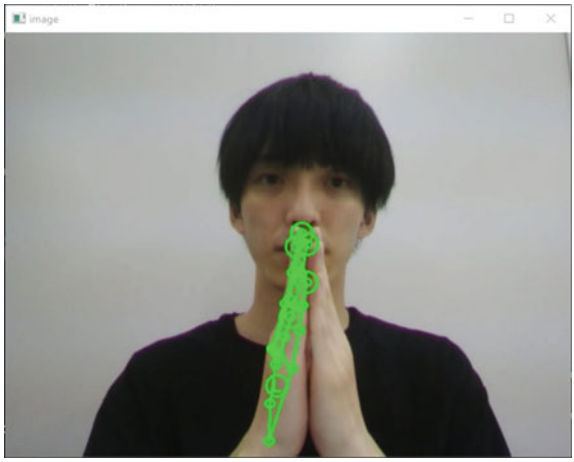



**Fig. 31.4** Losing recognition of one hand

However, the decision was made to use the Pose tool instead for the following reasons, accompanied by a detailed explanation of the assessment method using images.

The main reason for choosing the Pose tool over the Hands tool was the issue of the Hands tool frequently losing recognition of one hand when both hands are in close proximity (refer to Fig. 31.4). The original plan was to leverage the Hands tool's ability to provide 21 landmarks and their corresponding coordinates. The intention was to obtain the $x$-coordinate of each landmark for both hands and calculate the distance between corresponding landmarks (with the same $x$-coordinate) by subtracting their coordinates. However, due to the unreliable recognition in close proximity, the Pose tool was selected as a more suitable alternative.

Similar to the Hands tool, the Pose tool also allows for the acquisition of landmarks and coordinates. Moreover, when using the Hands tool, the recognition issue mentioned earlier does not occur. Consequently, with the Pose tool, it becomes possible to calculate the distance between the coordinates obtained from the blue circles shown in Fig. 31.5. The assessment is then performed by comparing the calculated distances with a predefined threshold.

**Thumb Touch Exercises**: The assessment method entails calculating the distance between fingers and comparing it against a predefined threshold. Additionally, the number of remaining extended fingers is counted. If three fingers are extended, the exercise is considered correctly performed. The assessment method is elucidated through images. As mentioned earlier, two conditions are imposed: ensuring firm contact between the thumb and other fingers and confirming that the remaining fingers are not bent. The $y$-axis distance between the thumb tip and the tips of the other fingers is calculated to satisfy the first condition. The second condition involves counting the number of extended fingers, and if three fingers are extended, it meets the criteria for correct performance. The count displayed at the top of the screen keeps track of the number of repetitions, and the landmarks are concealed. Figure 31.6

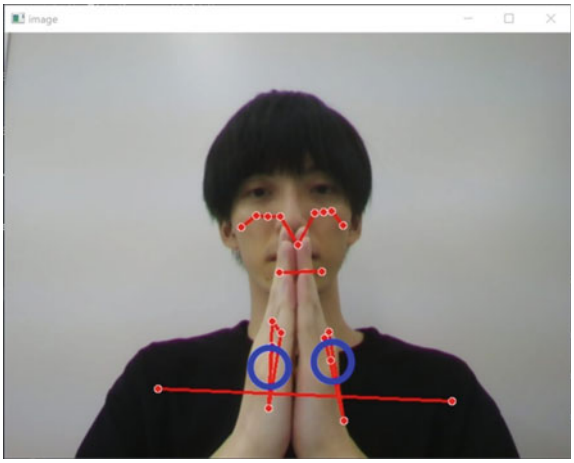


**Fig. 31.5** Correctly recognized of one hand

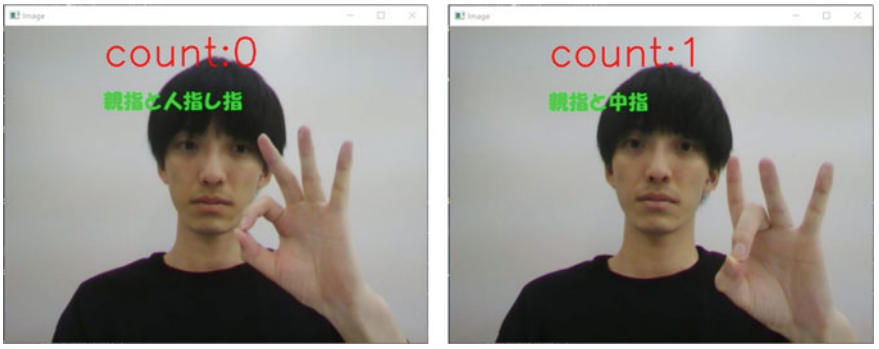



**Fig. 31.6** Demonstrating the thumb touching the index finger and the thumb touching the middle finger, respectively

depict actual screens demonstrating the thumb touching the index finger and the thumb touching the middle finger, respectively.

## 31.6 Conclusions

This paper presents the development of a customized hand rehabilitation system that is tailored to meet the individual rehabilitation needs of users. By integrating an interactive interface with vision-based hand-tracking technology, the proposed

system enables active participation in physical rehabilitation therapies. The interactive interface offers real-time feedback, incorporating metrics such as finger bending, hand clapping, and thumb touch exercises.

The evaluation results of the developed system highlight its effectiveness in supporting hand rehabilitation and fostering motivation among individual users to engage in rehabilitation exercises. The system's ability to provide real-time feedback enhances the user's awareness and understanding of their progress, facilitating a more interactive and engaging rehabilitation experience.

By combining technology and rehabilitation, this system contributes to the advancement of personalized and interactive rehabilitation approaches. The integration of vision-based hand-tracking technology enables accurate and precise tracking of hand movements, ensuring the system's reliability in assessing and monitoring rehabilitation progress.

Overall, the proposed customized hand rehabilitation system offers a promising solution to address individual rehabilitation needs. Its effectiveness in promoting active engagement and providing real-time feedback demonstrates its potential to improve the outcomes of hand rehabilitation therapies. Further research and development in this area have the potential to enhance the system's capabilities and expand its application in broader rehabilitation contexts.

**Acknowledgements** The authors would like to thank FuisonTech Inc. for their incredible support during this project.

## References

1. Tokyo: Cabinet Office Homepage. https://www8.cao.go.jp/kourei/whitepaper/w-2022/zenbun/pdf/1s1s_01.pdf. Last accessed 2023/06/12
2. Tokyo: Cabinet Office Homepage. https://www8.cao.go.jp/kourei/whitepaper/w-2021/html/zenbun/s1_2_2.html. Last accessed 2023/06/12
3. Tokyo: Ministry of Health, Labor and Welfare Homepage. https://www.mhlw.go.jp/wp/hakusyo/kousei/15/dl/1-00.pdf. Last accessed 2023/06/12
4. Liao, X., Zhang, Z., Kang, Z., Gao, F., Liao, Q., Zhang, Y.: Ultrasensitive and stretchable resistive strain sensors designed for wearable electronics. Mater. Horiz. **4**(3), 502–510 (2017)
5. Wang, Q., Markopoulos, P., Yu, B., Chen, W., Timmermans, A.: Interactive wearable systems for upper body rehabilitation: a systematic review. J. Neuroeng. Rehabil. **14**(1), 1–21 (2017)
6. Wittmann, F., Held, J.P., Lambercy, O., Starkey, M.L., Curt, A., Höver, R., Gonzenbach, R.R.: Self-directed arm therapy at home after stroke with a sensor-based virtual reality training system. J. Neuroeng. Rehabil. **13**, 1–10 (2016)
7. Masmoudi, M., Zenati, N., Benbelkacem, S., Hadjadj, Z., Djekoune, O., Guerroudji, M.A., Izountar, Y.: Low-cost haptic glove for grasp precision improvement in Virtual Reality-Based Post-Stroke Hand Rehabilitation. In: 2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP), pp. 1–3 (2021)
8. Bouteraa, Y., Abdallah, I.B., Elmogy, A.M.: Training of hand rehabilitation using low cost exoskeleton and vision-based game interface. J. Intell. Rob. Syst. **96**, 31–47 (2019)
9. Gmez-Portes, C., Lacave, C., Molina, A.I., Vallejo, D.: Home rehabilitation based on gamification and serious games for young people: a systematic mapping study. Appl. Sci. **10**(24), 8849 (2020)

10. Mujahid, A., Awan, M.J., Yasin, A., Mohammed, M.A., Damaševičius, R., Maskeliūnas, R., Abdulkareem, K.H.: Real-time hand gesture recognition based on deep learning YOLOv3 model. Appl. Sci. **11**(9), 4164 (2021)
11. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 7291–7299 (2017)
12. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2021)
13. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person poseestimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 4903–4911 (2017)
14. Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., Grundmann, M.: BlazePose: on-device real-time body pose tracking. arXiv 2020 arXiv:2006.10204 (2020)

# Chapter 32
# Formalization and Verification of Fuzzy Approximate Reasoning by Mizar


**Takashi Mitsuishi**

**Abstract** In this paper, we describe the verification and formalization of fuzzy optimal control problems using the proof checker Mizar. First, we provide a brief introduction to fuzzy inference. In order to achieve the optimization of fuzzy optimal control, the compactness of the family of membership functions and two types of continuity of fuzzy approximate reasoning are required. We have proven the existence of IT-THEN rules that minimize the performance function for fuzzy control. To verify these properties using Mizar, we formalize the set of membership functions as a generalization of IT-THEN rules. Additionally, we verify the properties of piecewise linear functions. Furthermore, we formalize the definition of the defuzzified value using the centroid method in the Mizar language. The paper also explains several theorems and definitions concerning fuzzy logic, which are written in the Mizar language.

## 32.1 Introduction

With the increasing complexity and diversification of social systems, research fields have become more specialized. However, interdisciplinary studies spanning multiple disciplines, including not only social and natural sciences, but also others, are actively being pursued [19]. The verification process involved in peer reviewing scientific and technological papers, which are the outcomes of these endeavors, requires extensive research, consuming significant time and human resources. Most of these complex social and natural science theories seek mathematical foundations for their formalization and systematization. Since the invention of computers, the demand for computer-aided verification of mathematical theorems and proofs has naturally increased.

T. Mitsuishi (✉)
Nagano University,386-1298 Ueda, Japan
e-mail: takashi-mitsuishi@nagano.ac.jp



K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4_32

In 1973, Andrzej Trybulec (University of Białystok, Poland) developed a system using the Mizar language, which was specifically designed to enable computer-assisted analysis of mathematics. This system allows for the description (formalization) of mathematical proofs and their logical validation through the use of a proof checker known as the Mizar system. He initiated the project. [6, 8]. Similar to Izabelle and Coq [1, 3], which are referred to as theorem proving assistants, Mizar is a software tool used for the verification of mathematical proofs [2]. Mathematical properties and their proofs described in the Mizar language are referred to as "articles." Numerous articles that have been verified for their validity can be referenced in subsequent proofs within the Mizar Mathematical Library (MML).

On the other, we studied fuzzy optimal control problem using functional analysis [9, 10, 16]. In this research, some theorems for fuzzy optimal control are formalized by Mizar language and verified [13].

## 32.2 Fuzzy Theory for Mizar

### *32.2.1 IF-THEN Type Fuzzy Rules and Fuzzy Approximate Reasoning*

Commonly and widely used IF-THEN type fuzzy rules are following [7, 14, 17, 18].

$$\begin{aligned}
&\text{RULE 1: } IF\ x_1\ is\ A_{11}\ and\ \ldots\ and\ x_n\ is\ A_{1n}\ THEN\ y\ is\ B_1\\
&\vdots\\
&\text{RULE } i\text{: } IF\ x_1\ is\ A_{i1}\ and\ \ldots\ and\ x_n\ is\ A_{in}\ THEN\ y\ is\ B_i\\
&\vdots\\
&\text{RULE } m\text{: } IF\ x_1\ is\ A_{m1}\ and\ \ldots\ and\ x_n\ is\ A_{mn}\ THEN\ y\ is\ B_m eALT
\end{aligned} \tag{32.1}$$

IF-THEN rules are able to be considered as a set of membership functions. Then we formalized it by Mizar language in this study.

When an input variable $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is given to the IF-THEN rules (32.1), then the crisp inference result value is calculated through the following process.

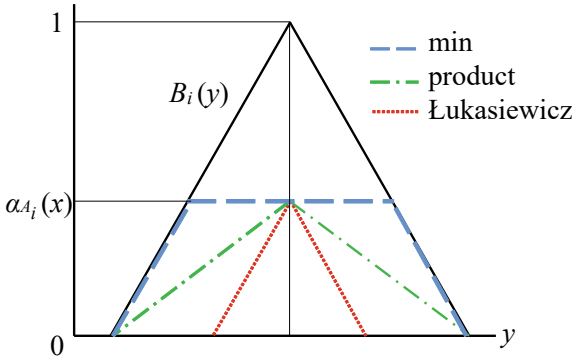



**Fig. 32.1** Minimum(min), product and bounded product (Łukasiewicz t-norm)

**1st process**: Results of fuzzifications and of $i$-th fuzzy operators $\alpha_{\mathcal{A}_i}$ by the input is calculated by

$$\alpha_{\mathcal{A}_i}(x) = \prod_{j=1}^{n} A_{ij}(x_j) \quad \text{or} \quad \bigwedge_{j=1}^{n} A_{ij}(x_j) \quad (i = 1, 2, \ldots, m).$$

Here, scaling down calculation "product $\prod$" or clipping calculation "minimum $\bigwedge$" are applied for implication method in this study [7, 18].

**2nd process**: Inference result of implication in $i$-th rule is output as a membership function.

$$\begin{aligned} \beta_{\mathcal{A}_i B_i}(x, y) &= \alpha_{\mathcal{A}_i}(x) \cdot B_i(y) \\ &\quad \text{or} \\ &= \alpha_{\mathcal{A}_i}(x) \wedge B_i(y) \\ &\quad \text{or} \\ &= (\alpha_{\mathcal{A}_i}(x) + B_i(y) - 1) \vee 0 \end{aligned}$$

The third operation named bounded product (Łukasiewicz t-norm) [17] pushes down the graph of membership functions in consequent part of IF-THEN rules $B_i(y)$ by $(1 - \alpha_{\mathcal{A}_i}(x))$. Three kind of inference results $\beta_{\mathcal{A}_i B_i}(x, y)$ are following Fig. 32.1.

**3rd process**: Aggregate all outputs and its defuzzification.

$$\rho_{\mathcal{AB}}(x) = \frac{\int y \sum_{i=1}^{m} \beta_{\mathcal{A}_i B_i}(x, y) dy}{\int \sum_{i=1}^{m} \beta_{\mathcal{A}_i B_i}(x, y) dy} \quad \text{or} \quad \frac{\int y \bigvee_{i=1}^{m} \beta_{\mathcal{A}_i B_i}(x, y) dy}{\int \bigvee_{i=1}^{m} \beta_{\mathcal{A}_i B_i}(x, y) dy}. \tag{32.2}$$

The inference result of each rule is integrated by sum operation or max operation. Then the centroid of the integrated function is the defuzzified value as crisp value. This is the final output of fuzzy inference.

Mamdani method uses minimum and maximum operator. The choice of operations in fuzzy approximate reasoning depends on the application. In addition, many defuzzification methods have been proposed. Which of these to apply also depends on the application.

### *32.2.2 Overview of Optimization Using Functional Analysis*

The authors have proved the existence of the IF-THEN rule that achieves optimal control in fuzzy control by utilizing the method of functional analysis. The main theorem is Weierstrass's theorem (also known as the extreme value theorem), which states that "continuous functions on compact sets have a maximum (minimum) value." To apply this theorem, the following properties were proved:

1. Compactness of the fuzzy set genus.
2. Lipschitz continuity of the inference calculation (ensuring the unique existence of the solution to the state equation) in feedback control.
3. Continuity as a functional of the inference calculation.

In this paper, in order to verify properties 1 and 2 using Mizar, various mathematical properties were formalized and verified as described later.

## 32.3 Formalization by Mizar

### *32.3.1 Definition of Set of Fuzzy Membership Function*

The concept of membership functions formalized and registered in the Mizar Mathematical Library is as follows.

**Listing 32.1** FUZZY_1 mode [15]

```
definition
 let C be non empty set ;
 mode Membership_Func of C is [.0,1.] −valued Function of C,REAL;
end;
```

This definition means that Membership_Func of C is a function from a nonempty set to a real number and takes the value from [0, 1]. In the proof checker Mizar, mathematical symbols like "'Membership_Func of C" are determined and registered by the users.

Although some studies define fuzzy set as pairs of elements and fuzzy values, Mizar equates the two. Fuzzy set and membership function are identified as the same in Mizar as following definition.

**Listing 32.2** FUZNUM_1 mode [5]

```
definition
 let C be non empty set ;
 mode FuzzySet of C is Membership_Func of C;
end;
```

In our research, we proved the existence of the optimal solution by using the compactness of the family of the set of membership function. Therefore, the set of membership functions was formalized by Mizar language as theorem FUZZY_5:def 1 [11]. The following defined set can be regarded as IF-THEN rules (32.1) as pairs of membership functions and also as a family of sets to which they belong.

**Listing 32.3** FUZZY_5:2 [11]

```
theorem :: FUZZY_5:2
Membership_Funcs (REAL)
={f where f is Function of REAL,REAL : f is FuzzySet of REAL};
```

Various membership functions are used in the practical application of fuzzy theory. Some of theorems in [11] show that these sets are a subset of Membership_Funcs defined in FUZZY_5:def 1. The theorem FUZZY_5:23 shows that for all function $g : \mathbb{R} \to \mathbb{R}$, put $f(x) = \min(1, \max(0, g(x))\ \forall x \in \mathbb{R}$, then the set of $f$ is the subset of the set of membership functions. In other words, $f$ is a membership function. AffineMap$(a, b)$ is a linear function in Mizar.

$$\text{AffineMap}(a, b).x = ax + b.$$

Among them are functions composed only of straight lines such as triangles, trapezoids, $S$-shapes, and $Z$-shapes (FUZZY_5:63). The theorem FUZZY_5:37 (about trigonometric functions) and FUZZY_5:59 (about Gaussian functions) set its range to [0, 1] for membership function [11].

### 32.3.2 Properties of Piecewise Linear Function

Since triangular, trapezoidal, $S$-shaped, and $Z$-shaped membership functions that are constructed with linear functions make fuzzy inference faster, they are used in this research and formalized by Mizar.

The following function $f + *g$ takes $f(x)$ or $g(x)$ depending on the domain of the variable. For example, if $x$ is in the domain of $g$, then $(f + *g)(x) = g(x)$.

$$(f + * g)(x) = \begin{cases} f(x) & (x \notin D(g)) \\ g(x) & (x \in D(g)) \end{cases}$$

In the MML, the composition of function mentioned above defined as follows.

**Listing 32.4** FUNCT_4:def 1 [4]

```
definition
 let f, g be Function;
 func f +* g -> Function means :Def1: :: FUNCT_4:def 1
 ( dom it = (dom f) \/ (dom g) &
 ( for x being object st x in (dom f) \/ (dom g) holds
 ( ( x in dom g implies it . x = g . x ) &
 ( not x in dom g implies it . x = f . x ) ) ) );
end;
```

This definition does not necessarily require that the two functions have an intersection.

A triangular membership function of three points $(a, 0)$, $(b, 1)$, $(c, 0)$ is formalized as FUZNUM_1:def 7 using symbol "+*" in [5].

### *32.3.3 Defuzzification*

The operations used in fuzzy reasoning are sum, product, minimum and maximum. Since these operations were already formalized in Mizar, they do not need to be formalized. Specific calculation in fuzzy inference is the procedure for calculating defuzzified crisp values from the composite function that is the inference result. This is defuzzification. Several calculation methods, such as height method, area method, first of maximum, are introduced [20]. The most widely used method is centroid method (32.2). This method calculates weighted average of membership function as centroid. Then the centroid is considered as defuzzified crisp value of result of approximate reasoning. In this paper, we formalized centroid method (center of gravity method) is formalized as following definition Listing 1.5.

$$\text{centroid}(f, A) = \frac{\int_A x f(x) \mathrm{d}x}{\int_A f(x) \mathrm{d}x}$$

**Listing 32.5** FUZZY_6:def 1 [12]

```
definition :: FUZZY_6:def 1
 let A be non empty closed_interval Subset of REAL;
 let f be Function of REAL,REAL;
 func centroid (f,A) -> Real equals
 integral((id REAL)(#)f,A)/integral(f,A);
end;
```

Here, (id REAL): $x \in \mathbb{R} \mapsto x \in \mathbb{R}$ is a identify function on real number.

The following function $f$ is a piecewise linear function that $ax + b$ and $px + q$ connect at $x = \dfrac{q - b}{a - p} \in A$.

$$f(x) = \begin{cases} ax + b & \left(x \leq \frac{q-b}{a-p}\right) \\ px + q & \left(x \geq \frac{q-b}{a-p}\right) \end{cases} \tag{32.3}$$

The centroid $x^*$ of $f$ in $A$ is calculated using centroid method as follows:

$$x^* = \frac{\int_A x f(x) \mathrm{d}x}{\int_A f(x) \mathrm{d}x}$$

$$= \frac{\frac{a}{3}(t^3 - s^3) + \frac{b}{2}(t^2 - s^2) + \frac{p}{3}(u^3 - t^3) + \frac{q}{2}(s^2 - t^2)}{\frac{a}{2}(t^2 - s^2) + b(t - s) + \frac{p}{2}(u^2 - t^2) + q(u - t)}$$

This centroid is formalized by Mizar as following theorem FUZZY_6:48.

**Listing 32.6** FUZZY_6:48 [12]

```
theorem :: FUZZY_6:48
for a,b,p,q,c,d,e being Real, f be Function of REAL,REAL st a <> p &
f | A = AffineMap (a,b) | [.lower_bound A,(q−b)/(a−p).]
    +* AffineMap (p,q) | [.(q−b)/(a−p),upper_bound A.] & (q−b)/(a−p) in A
holds
centroid(f,A) =
( 1/3*a*(((q−b)/(a−p))^3 − (lower_bound A)^3)
  + 1/2*b*(((q−b)/(a−p))^2 − (lower_bound A)^2)
  + 1/3*p*((upper_bound A)^3 − ((q−b)/(a−p))^3)
  + 1/2*q*((upper_bound A)^2 − ((q−b)/(a−p))^2) ) /
( 1/2*a*(((q−b)/(a−p))^2 −(lower_bound A)^2)
  + b*((q−b)/(a−p) − lower_bound A)
  + 1/2*p*((upper_bound A)^2 −((q−b)/(a−p))^2)
  + q*(upper_bound A −(q−b)/(a−p)) );
```

Here $a \neq p$. Put $s$=lower_bound A to be lower bounf of $A$, $u$ = upper_bound A to be upper bound of $A$ and $t = \dfrac{q-b}{a-p}$.

### 32.3.4 Isosceles Triangle and Isosceles Trapezoidal Membership Functions

The isosceles function $f$ as shown (32.4) was represented by max operation and absolute value symbol. The following FUZZY_5:65 shows that if $b = 1$, then $f$ is a isosceles triangular membership function. The next FUZZY_7:6 is formalization of isosceles trapezoidal membership function in Fig. 32.2 obtained by clipping the isosceles triangular membership function $f$ with height of 1.

$$f(x) = \max\left\{0, b - \left|\frac{b(x-a)}{c}\right|\right\} \tag{32.4}$$

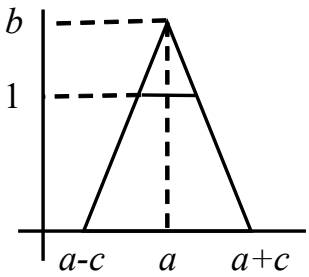



**Fig. 32.2** Isosceles triangle and isosceles trapezoidal membership function

**Listing 32.7** FUZZY_5:65, FUZZY_7:6 [11]

```
theorem TR6: :: FUZZY_5:65
for a, b being Real st b > 0 holds
for x being Real holds
TriangularFS ((a − b),a,(a + b))) . x = max (0,(1 − |.((x − a) / b).|);

theorem :: FUZZY_7:6
for a,b,c be Real, f be Function of REAL,REAL st
b > 1 & c > 0 &
( for x be Real holds f.x = min(1, max(0, b − |. b∗(x−a)/c .|)) )
holds
f is FuzzySet of REAL & f is trapezoidal FuzzySet of REAL &
f is normalized FuzzySet of REAL;
```

The theorems in the article FUZZY_7 are already verified by Mizar system and will be registered in the MML at a later date.

The centroid of a function with a symmetrical graph is clearly the midpoint of the bottom.

**Listing 32.8** FUZZY_7:30, 7:33, 7:51

```
theorem :: FUZZY_7:30 ::Th20:
for a,b,c be Real, f be Function of REAL,REAL st
b > 0 & c > 0 & ['a−c,a+c'] c= A &
(for x be Real holds f.x = max(0,b−|. b∗(x−a)/c .|))
holds centroid (f,A) = a;

theorem :: FUZZY_7:33
for a,b,c,d be Real st a < b & b < c & b−a = c−b & d <> 0 holds
centroid (d (#) TriangularFS (a,b,c),['a,c']) = b;

theorem :: FUZZY_7:51
for a1,c,a2,d be Real st c > 0 & d > 0 & a1 < a2 holds
centroid ( d (#) TrapezoidalFS (a1−c,a1,a2,a2+c), ['a1−c,a2+c']) = (a1+a2)/2;
```

The symbol (#) means the product of function with itself or with scalar.

### *32.3.5 Lipschitz Continuity*

If a part of state equation of fuzzy control is Lipschitz continuous on premise variables, then it has unique solution. Therefore Lipschitz continuity of the calculation

in fuzzy inference is required. The following theorems show Lipschitz continuity of various functions.

The maximum and minimum operation frequently used in Mamdani method [7] is continuous. Since the range of the membership function is [0, 1], given an appropriate value for $r$ and $s$, the formula in theorem FUZZY_5:22 [11] can be useful for inference calculations.

The followings shows Lipschitz continuity of composition of Lipschitz continuous functions. Since linear function is Lipschitz continuous obviously, we can formalized and verified following theorem by Mizar system.

**Theorem 1** *Assume that $a \neq p$. The piecewise function (32.3) is Lipschitz continuous.*

**Listing 32.9** FUZZY_6:26 [12]

```
theorem :: FUZZY_6:26
for c being Real, f,g,F be Function of REAL,REAL st
f is Lipschitzian & g is Lipschitzian &
f . c = g . c & F = (f | ].−infty,c.[) +∗ (g | [.c,+infty.[)
holds F is Lipschitzian;
```

The trigonometric function with periodicity can be used for fuzzy representation of periodic phenomenon like time, color, direction, and so on.The inequality in FUZZY_5:3 was verified to prove Lipschitz continuity of the sine function.

**Listing 32.10** FUZZY_5:3 [11]

```
theorem LmSin2: :: FUZZY_5:3
for x, y being Real holds |.((sin x) − (sin y)).| <= |.(x − y).|;
```

## 32.4 Conclusion

In this research, definitions, theorems, and various properties related to fuzzy theory were formalized and verified using the proof checker Mizar. The purpose of this study is to verify the functional analytical proofs of fuzzy optimal control problems using Mizar. To formalize an IF-THEN rule, which consists of a tuple (pair) of membership functions, we also formalize the definition of a set of membership functions.

The triangular and trapezoidal membership functions, which are commonly used in fuzzy approximate reasoning, are piecewise linear functions that consist of a combination of two or more linear functions. We have formalized their properties. Furthermore, we have defined the calculation for defuzzification using the centroid method and formalized the defuzzification values for the piecewise linear functions, triangular functions, and trapezoidal functions.

On the other hand, the Lipschitz continuity of a part of the calculation of fuzzy inference on premise variables is proved and verified by Mizar. The formalized and verified definitions, theorems, and properties registered in the Mizar Mathematical Libraries will be utilized as necessary theorems in the future verification process of optimization problems in fuzzy control.

## References


1. Isabell. https://isabelle.in.tum.de/index.html. Last accessed 31 May 2023
2. Mizar Home Page. http://mizar.org/. Last accessed 31 May 2023
3. The Coq Proof Assistant. https://coq.inria.fr/. Last accessed 31 May 2023
4. Byliński, C.: The modification of a function by a function and the iteration of the composition of a function. Formalized Math. **1**(3), 521–527 (1990)
5. Grabowski, A.: The formal construction of fuzzy numbers. Formalized Math. **22**(4), 321–327 (2014). https://doi.org/10.2478/forma-2014-0032
6. Grzegorz, B., Czesław, B., Adam, G., Artur, K., Roman, M., Adam, N., Karol, P.: The role of the Mizar mathematical library for interactive proof development in Mizar. J. Autom. Reasoning **61**(1-4), 9–32 (2017)
7. Mamdani, E.H.: Application of fuzzy algorithms for control of simple dynamic plant. Proc. IEE **121**(12), 1585–1588 (1974)
8. Matuszewski, R., Rudnicki, P.: Mizar: the first 30 years. Mech. Math. Appl. **4**, 3–24 (2005)
9. Mitsuishi, T., Shimada, N., Homma, T., Ueda, M., Kochizawa, M., Shidama, Y.: Continuity of approximate reasoning using fuzzy number under łukasiewicz t-norm. In: 2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), pp. 71–74 (2015). https://doi.org/10.1109/ICCIS.2015.7274550
10. Mitsuishi, T., Terashima, T., Shimada, N., Homma, T., Shidama, Y.: Approximate reasoning using LR fuzzy number as input for sensorless fuzzy control. In: 2016 IEEE Symposium on Sensorless Control for Electrical Drives (SLED), pp. 1–5 (2016)
11. Mitsuishi, T.: Some properties of membership functions composed of triangle functions and piecewise linear functions. Formalized Math. **29**(2), 103–115 (2021). https://doi.org/10.2478/forma-2021-0011
12. Mitsuishi, T.: Definition of centroid method as defuzzification. Formalized Math. **30**(2), 125–134 (2022). https://doi.org/10.2478/forma-2022-0010
13. Mitsuishi, T.: Formalization of centroid method in fuzzy approximate reasoning by Mizar. In: IEICE Technical Report. CS2022-26, vol. 122, pp. 53–56. Kagoshima (July 2022)
14. Mitsuishi, T.: Study on continuity of defuzzification using density moment method. In: 2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE), pp. 90–94 (2023). https://doi.org/10.1109/ISCAIE57739.2023.10165332
15. Mitsuishi, T., Endou, N., Shidama, Y.: The concept of fuzzy set and membership function and basic properties of fuzzy set operation. Formalized Math. **9**(2), 351–356 (2001)
16. Mitsuishi, T., Kawabe, J., Wasaki, K., Shidama, Y.: Optimization of fuzzy feedback control in $L^{\infty}$ space. In: Proceedings of the 10th IEEE International Conference on Fuzzy Systems, Melbourne, Australia, 2–5 Dec 2001, pp. 896–899. IEEE (2001)
17. Mizumoto, M.: Fuzzy conditional inference under max-$\odot$ composition. Inf. Sci. **27**(3), 183–209 (1982)
18. Mizumoto, M.: Improvement of fuzzy control (IV)-case by product-sum-gravity method. In: Proceedings of 6th Fuzzy System Symposium, pp. 9–13 (1990)
19. Perrin, W.F.: In search of peer reviewers. Science **319**(5859), 32–32 (2008)
20. Van Leekwijck, W., Kerre, E.E.: Defuzzification: criteria and classification. Fuzzy Sets Syst. **108**(2), 159–178 (1999)

# Author Index





K. Nakamatsu et al. (eds.), *AI Technologies and Virtual Reality*, Smart Innovation, Systems and Technologies 382, https://doi.org/10.1007/978-981-99-9018-4