



**Brunel**  
University  
London

DOCTORAL THESIS

---

**Fingers Micro-gesture Recognition  
based on Holographic 3D Imaging System**

---

*A Thesis submitted to Brunel University London  
in accordance with the requirements  
for award of the degree of Doctor of Philosophy*

*in*

*the department of Electronic and Computer Engineering*

Yi Liu

March 3, 2020

## Declaration of Authorship

I hereby declare that this thesis is entirely my research, except where otherwise indicated, describes my research or publications. No part of this thesis has been previously presented here has been published or distributed by anyone other than the author.

SIGNED: ..... Yi Liu ..... DATE: ..... 9 Mar 2020 .....  
(Signature of student)

# Abstract

Micro-gesture recognition has been widely researched in recent years, in particular there has been a great focus on 3D micro-gesture recognition which consists of classifying the micro-gesture movements of the fingers for touch-less control applications. Holographic 3D imaging system mimics fly's eye technique to capture true 3D scene which is rich in both texture and motion information. As a result, holographic 3D imaging system shall be a suitable approach for robust recognition application. This PhD research focuses on innovative 3D micro-gesture recognition based on holographic 3D system which delivers robust and reliable performance with precision for 3D micro-gestures. Indeed this can be applied to other wide range of applications such as Internet of things (IoT), AR/VR, robotics and other touch-less interaction.

Due to lack of holographic 3D dataset, a comprehensive 3D micro-gesture dataset (HoMG) includes both holographic 3D images and videos is prepared. It is a reasonable size holographic 3D dataset which is captured with different camera settings and conditions from 40 participants. Innovative 3D micro-gesture recognition is proposed based on 2D feature extraction methods with basic classification methods, the recognition accuracy can reach around 50.9%. For video-based data, the 3D feature extraction methods are achieved 66.7% recognition accuracy over 50.9% accuracy for micro-gesture images as the initial investigation. HoMG database held a challenge in IEEE International automatic face and gesture 2018, and 4 groups from the international research institutes joined the challenge and contributed many new methods as further development where the proposed method was published.

The holographic 3D dataset further enriches innovative micro-gesture 3D recognition system is proposed and its performance is evaluated by carrying out like to like comparison with state of the art methods. In addition, a fast and efficient pre-processing algorithm for H3D images to extract the element images. Simplified viewpoint image extraction method are presented. A pre-trained CNN model with the attention mechanism is implemented based on VP image for the predicted probabilities of gesture. The proposed approach is further improved using voting strategy. The proposed approach achieves 87% accuracy, which outperforms all existing state of the art methods on the image-based database.

Advanced 3D micro-gesture recognition is investigated based on sequence video database, the end-to-end model has been used on effective H3D based micro-gesture recognition system. For front-end network, there are two method of traditional viewpoint image extraction and novel pseudo viewpoint image extraction have been used and evaluated. The pseudo viewpoint (PVP) front-end has been created, which used to deep learning networks understanding the implied 3D information of H3D imaging system. The viewpoint (VP) front-end follows the traditional H3D image method to extract and reconstruct the multi-viewpoint images. Both front-end have been feed in four popular advanced deep networks using for learning and classification. This experiments evaluated the performance of 2D/3D convolutional, mixing 2D and 3D convolutional and LSTM on the HoMG video database, which is beneficial to H3D imaging system using deep learning network. Finally, in order to obtain the high accuracies, the majority voting has been applied for further improve. The final results show that the performance is not only better than the traditional methods, but also superior to the existing deep learning based approaches, which clearly demonstrates the effectiveness of the proposed approach.

Keywords: Micro-gesture Recognition, Gesture Recognition; Human-computer Interaction; Machine Learning; Deep Learning;

# Acknowledgements

I would like to take this opportunity to thank my principle supervisor Dr. Rafiq Swash, as this achievement would not be possible without his given opportunities and support. He inspired me to realise my potential to the fullest. He has been very supportive in my research work at Brunel and was always available for discussions and supervision when I needed it the most.

I would like to thank my second supervisor Dr. Hongying Meng who taught me how to be a good researcher, his patience and knowledge supported me to overcome research difficulties and through challenges to extend my knowledge further.

I want to thank Prof. Shiguang Shan, Dr. Shuang Yang and his group in Key laboratory of Intelligent Processing Institute of Computing Technology Chinese Academy of Science. I had a good study experience of deep learning.

To my friends and colleagues, many thanks for their help. I am grateful to the support given by family members and my boyfriend throughout my PhD studies.

They all deserve special thanks, as I would not have been able to face the challenges of this research work without their support.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Overview	1
1.2	Background	2
1.2.1	Human-computer Interaction	2
1.2.2	Hand Gesture	3
1.2.3	Holoscopic 3D Imaging System	4
1.2.4	Machine Learning	5
1.3	Research Motivation	5
1.4	Research Aim and Objectives	6
1.5	PhD Contributions	7
1.6	The Outlines of PhD Thesis	8
1.7	The Author's Publications	10
<b>2</b>	<b>Literature Review</b>	<b>11</b>
2.1	Human Computer Interaction	11
2.1.1	Body Language and Gesture	12
2.1.2	2D vs. 3D Gesture Interaction	13
2.1.3	3D Micro-gesture	17
2.1.4	3D Micro-gesture Technologies and System	17
2.2	3D Imaging Principles	20
2.2.1	Stereoscopic 3D Imaging System	20
2.2.2	ToF 3D Imaging System	22
2.2.3	Holoscopic 3D Imaging System	23
2.2.4	Holoscopic 3D Camera System and Reinstallation	27
2.2.5	H3D Image Processing and Viewpoint Extraction	29
2.3	Machine Learning Algorithms	32
2.3.1	Machine Learning	32
2.3.2	Deep Learning Models	34
2.3.3	Decision Fusion	36
2.4	Summary	37

<b>3</b>	<b>H3D Micro-gesture Database Creation and Validation</b>	<b>38</b>
3.1	Micro-gesture Interaction . . . . .	38
3.2	Holoscopic 3D Imaging for Micro-gesture . . . . .	39
3.3	H3D Dataset Preparation and Creation . . . . .	40
3.3.1	Micro-gesture Design . . . . .	40
3.3.2	Design Process . . . . .	43
3.3.3	H3D Imaging System . . . . .	44
3.3.4	Recording Setup . . . . .	46
3.3.5	Participants . . . . .	49
3.3.6	HoMG Database Structure . . . . .	49
3.4	Micro-Gesture Recognition Validation . . . . .	58
3.4.1	H3D Video Subset for Micro-gesture Recognition . . . . .	58
3.4.2	H3D Image Subset for Micro-gesture Recognition . . . . .	60
3.5	Conclusion . . . . .	62
<b>4</b>	<b>CNN and Decision Fusion for Image-based Micro-gesture Recognition</b>	<b>64</b>
4.1	State-of-the-Art Methods . . . . .	65
4.2	System Development . . . . .	67
4.2.1	Micro-gesture Recognition System . . . . .	67
4.2.2	H3D Pre-processing . . . . .	68
4.2.3	H3D Viewpoint Extraction based on Shifting Method . . . . .	70
4.2.4	Convolutional Neural Network Model . . . . .	73
4.2.5	Decision Level Fusion . . . . .	76
4.3	Experiments and Evaluation . . . . .	79
4.3.1	H3D Image Subset of HoMG Database . . . . .	79
4.3.2	H3D VP Extraction Parameters . . . . .	79
4.3.3	Implementation Details of CNN Model and Mixture of Experts . . . . .	80
4.3.4	Experimental Results and Comparison . . . . .	81
4.4	Conclusion . . . . .	82
<b>5</b>	<b>Pseudo Viewpoint and Deep Learning for Video-based H3D Micro-gesture Recognition</b>	<b>84</b>
5.1	Methodology . . . . .	85
5.1.1	Proposed Video-based H3D Micro-gesture Recognition System . . . . .	85
5.1.2	The VP based Front-end . . . . .	86
5.1.3	The PVP based Front-end . . . . .	88
5.1.4	Deep network Architecture Based Back-end . . . . .	90
5.2	Experiments . . . . .	100
5.2.1	Database . . . . .	100

5.2.2	Parameters Setting . . . . .	100
5.2.3	H3D Pre-processing . . . . .	101
5.2.4	Experiment Results . . . . .	102
5.2.5	Comparison with State-of-Art Method . . . . .	105
5.3	Conclusion . . . . .	106
<b>6</b>	<b>Conclusion and Future Work</b>	<b>109</b>
6.1	Conclusions . . . . .	109
6.1.1	HoMG dataset and baseline . . . . .	109
6.1.2	Image-based Micro-gesture recognition . . . . .	110
6.1.3	Video-based Micro-gesture recognition . . . . .	110
6.2	Future Work . . . . .	110
	<b>Appendices</b>	<b>131</b>
	<b>A Participant Information Form</b>	<b>131</b>
	<b>B End User License Agreement</b>	<b>135</b>



# List of Figures

1.1	Leap motion sensor captures hand motion by the skeleton. . . . .	3
1.2	Atheer AR glasses. . . . .	3
1.3	Google Soli project using Radar sensor[1]. . . . .	4
1.4	Leap Motion sensor capture hand motion by the skeleton[2] . . . . .	4
1.5	H3D micro-gesture recognition approach. . . . .	6
2.1	The diagram represents the distribution of the different gesture styles that appear in the literature review[3]. . . . .	14
2.2	Continuous gestures [2] . . . . .	16
2.3	Stereoscopic 3D glasses [4]. . . . .	20
2.4	Anaglyph stereoscopic 3D scene[4] . . . . .	21
2.5	Linear polarisation and circular polarisation [5] . . . . .	21
2.6	Time Division[6]. . . . .	22
2.7	ToF imaging system principle [7]. . . . .	23
2.8	(a) demonstrated the microlens arrays structures. (b) spherical microlens array with omnidirectional and full parallax. . . . .	25
2.9	(a) is the unidirectional holoscopic 3D for the single direction. (b) is Omnidirectional holoscopic image with full parallax. . . . .	26
2.10	Holoscopic 3D camera and imaging system,(a) Microlens arrays. . . . .	26
2.11	Assembled holoscopic 3D camera at Brunel University. (a) Main components, (b) The 3D VIVANT project installed details. . . . .	27
2.12	(a) Camera removed all the lens and prepare for checking the cracks and damage, (b) Clean up the camera. . . . .	28
2.13	The Microsoft surface supervised H3D camera interface. . . . .	28
2.14	The H3D camera re-installation for H3D micro-gesture capture. . . . .	29
2.15	The H3D imaging system principle:(a) recording and (b) display process[8].	30
2.16	Holoscopic 3D capturing systematic[9]. . . . .	31
2.17	Illustration of (a) UH3DI viewpoint image extraction, (b) OH3DI viewpoint image extraction[6]. . . . .	32
2.18	Top1 vs. operation, size and parameters [10]. . . . .	35

2.19	Multilayer perceptions and fully connection. . . . .	36
3.1	(a) The touchscreen micro-gestures, (b) The 2D micro-gestures movement tracking [11]. . . . .	41
3.2	(a) The 3D gesture movement tracks [2], (b) The 3D micro-gesture in an application [12]. . . . .	41
3.3	Comparison of manipulate gestures. . . . .	42
3.4	Google Soli project gesture [1]. . . . .	43
3.5	Three key types of finger 3D micro-gestures studied in HoMG database. . . . .	44
3.6	The process of gesture design process. . . . .	44
3.7	Principle of the holoscopic 3D camera. The microlens array is placed between objective and relay lens to produce fly eye style images [8]. . . . .	45
3.8	Assembled holoscopic 3D camera[13]. . . . .	45
3.9	H3D Data acquisition configuration. . . . .	46
3.10	H3D Data acquisition positions,(1)(2) are close positions of 50cm for left and right hand. (c)(d) are far positions of 90cm for left and right hand record. . . . .	47
3.11	Debrief for participants before record data. . . . .	47
3.12	Button gesture movement tracking. . . . .	50
3.13	Dial gesture movement tracking. . . . .	51
3.14	Button gesture movement tracking. . . . .	52
3.15	Frame numbers of each video in training set. . . . .	53
3.16	Fame numbers of each video in development set. . . . .	53
3.17	Fame numbers of each video in test set. . . . .	54
3.18	Six types of gestures on every class. . . . .	55
3.19	Video validation. . . . .	55
3.20	Video-based subset data overview. . . . .	56
3.21	Green colour background data of image-based subset overview. . . . .	57
3.22	White colour background data of image-based subset overview. . . . .	57
3.23	Video-based subset micro-gesture recognition process. (a) shows the video data have been resized, (b) use LBP-TOP and LPQ-TOP to extract the feature, (c) use the algorithms of k-NN, Support Vector Machines (SVM) and Naive Bayes classifiers to obtain the final results. . . . .	59
3.24	Image-based subset micro-gesture recognition process. (a)shows the image data is resized, (b)use LBP and LPQ to extract the feature, (c)use the algorithms of k-NN, SVM and Naive Bayes classifiers to classify . . . . .	61
4.1	Holoscopic 3D micro-gesture image capture. (a) Recording setting, (b) Obtained H3D image, (c) Three types of micro-gestures. . . . .	65

4.2	Block diagram of the proposed micro-gesture recognition system. There are four stages: (a) Pre-processing; (b) Viewpoint image extraction; (c) Deep learning for prediction on viewpoint images; (d) Decision level fusion.	67
4.3	H3D micro-gesture image is consist of multiple 2D Element Images (EIs). Here 9 EIs are enlarged from the original H3D image. . . . .	68
4.4	The minimal values of the summarized rows of a H3D image. The 38 points marked with small red triangles are the selected boundaries of the EIs. . . . .	70
4.5	The minimal values of the summarized columns of a H3D image. The 68 points marked with small red triangles are the selected boundaries of the EIs. . . . .	70
4.6	The horizontal and vertical lines of the H3D images are adjusted for the EI extraction. . . . .	71
4.7	The relationship between EIs and focus layers of the holoscopic 3D capturing system. (a) micro-lens array recording system, (b) Orthographic viewpoint images from different perspectives. . . . .	72
4.8	Illustration of the principle of H3D image viewpoint image extraction. (a) The $3 \times 3$ pixels under each micro-lens, (b) One viewpoint image extracted from same position under different micro-lenses, (c) Nine viewpoint images extracted from $3 \times 3$ EIs. . . . .	73
4.9	This is one EI image with the resolution of $27 \times 27$ . The boundary pixels are avoided, only the central pixels are selected. Each patch has $3 \times 3$ pixels. In the end, only 16 patch areas are selected from this EI. . . . .	74
4.10	The general structure of a CNN. The inputs are the VP images extracted. There are a few pair of convolution layer (C) and Sub-sampling layer (S). Finally, it is fully connected layers with the number of outputs being the number of class. . . . .	75
4.11	Attention-based residual block that was integrated to the CNN architecture.	76
4.12	First row and second row represent the feature maps learned by ResNet-50 and attention based ResNet-50, respectively at res2b layer (low-level), res4b layer (middle-level) and res5c layer (high level). . . . .	77
4.13	The 16 viewpoint images extract from a H3D image and its associated recognition accuracy based on CNN models. . . . .	80
5.1	The video-based H3D micro-gesture recognition pipeline: (a) Sample video data of HoMG (b) is front-end, which consists the PVP extraction and pre-processing (c) described the back-end, which consist the deep network models and majority voting. . . . .	85

5.2	Pseudo viewpoint(PVP) frame and viewpoint frame. (a) obtained PVP frame, and the resolution is 340 by 185.(b) VP frames extraction. . . . .	88
5.3	Pseudocode for Pseudo viewpoint(PVP) extraction algorithm. . . . .	89
5.4	The principle of viewpoint extraction . . . . .	89
5.5	PVP16 and PVP 25 extraction from horizontal and vertical. . . . .	90
5.6	The four network architectures. (a) VGG-16,(b) LSTM-1,(C) D3D,(d) 3D+2D . . . . .	91
5.7	VGG-16 is classic deep learning model, which consists of the 13 2D convolutional layers, 5 pooling layers, 3 fully connect layers and softmax layer. . . . .	94
5.8	2D convolutional layer generally used on Image data, the kernel slides are along 2 dimensions on the data. . . . .	96
5.9	The kernels are applied in 3D convolution, t is temporal. . . . .	97
5.10	The D3D network(DenseNet in 3D version) (a) The D3D model (b) The structure of each Dense Layer (c)The structure of each Trans Block . . . .	98
5.11	The architecture of mixing 2D and 3D convolutional layers (a) The 3D+2D model, (b) The structure of each Residual block Layer . . . . .	99
5.12	Initial testing on multilayer perception and LSTM. (a) is based on the VGG-16 last four layers (b) is based on the LSTM last four layers. . . . .	102
5.13	Each VP accuracies input the majority vote to fusion. (a)VGG-16 (b)LSTM (c)2D+3D (d)D3D . . . . .	104
5.14	Classification accuracy (%) comparison between the proposed method with all the existing methods on the testing subset of HoMG dataset. . . .	107
A.1	Explanation of the micro-gesture movements. . . . .	132
A.2	Data recording and classification. . . . .	132
A.3	Participant information sheet . . . . .	133
B.1	End User License Agreement . . . . .	136

# List of Tables

3.1	The summary recording length of videos. . . . .	43
3.2	Data acquisition details. . . . .	48
3.3	Participants genders and ages. . . . .	49
3.4	The summary recording length of videos. . . . .	49
3.5	The summary recording length of videos. . . . .	49
3.6	The summary of the HoMG database. . . . .	58
3.7	Recognition accuracy (%) of video based micro-gesture recognition on development (Dev.) and testing sets using k-NN, SVM and Naive Bayes classifiers. . . . .	60
3.8	Recognition accuracy (%) of image based micro-gesture recognition on development (Dev.) and testing sets using k-NN, SVM and Naive Bayes classifiers under different distance conditions. . . . .	62
4.1	Number of samples in each partition of the HoMG database. "B" stands for Button, D stands for Dial and S stands for Slide. . . . .	79
4.2	Classification accuracy(%) of CNN models on each VP images at the testing set of HoMG database. . . . .	81
4.3	Classification accuracy (%) comparison between the proposed method with all the existing methods on the testing subset of HoMG dataset. "A" means attention block. "M.V." means "majority voting". . . . .	82
5.1	Classification accuracy (%) initial testing on multiplayer perception and LSTM layers. "M.V." means "majority voting". . . . .	103
5.2	Classification accuracy (%) comparison between the proposed methods and Majority Vote on the testing dataset of HoMG dataset. + M.V. means the model results used Majority Vote. . . . .	103
5.3	Classification accuracy (%) comparison between the proposed methods on the development dataset of HoMG dataset. . . . .	104
5.4	Classification accuracy (%) comparison between the proposed methods on the testing dataset of HoMG dataset. . . . .	105
5.5	Classification accuracy (%) development dataset of HoMG dataset. . . . .	106

5.6	Classification accuracy (%) comparison "M.V." means "majority voting".	106
5.7	Classification accuracy (%) comparison between the proposed method with all the existing methods on the testing subset of HoMG dataset. "M.V." means "majority voting" GRU means "Gated Recurrent Unit". "n.SVM" means "non-Linear SVM" . . . . .	107

# List of Acronyms

AR	Augmented Reality
AI	Artificial Intelligence
CCD	Charge-Coupled Device
CNN	Convolutional Neural Network
D3D	DenseNet in 3D Version
EI	Element Images
HMMs	Hidden Markov Models
KNN	K-Nearest Neighbour
LSTM	Long Short Term Memory
OH3DI	Omnidirectional Holoscopic 3D Image
OCR	Optical Character Recognition
HCI	Human Computer Interaction
H3D	Holoscopic 3D
GUI	Graphical User Interface
LBP	Local Binary Pattern
LBPTOP	Local Binary Pattern – Three Orthogonal Planes
LCD	Liquid Crystal Display
LPQ	Local Phase Quantization
LPQTOP	Local Phase Quantization - Three Orthogonal Planes
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
PVP	Pseudo View Point
RGB	Red, Green and Blue Colour Model
SDK	Software Development Kit
SI	Sub-Image
SVM	Support Vector Machine
UH3DI	Unidirectional Holoscopic 3D Image
UGI	Unicursal Gesture Interface
VAE	Variational Autoencoder
VGG	Visual Geometry Group
VR	Virtual Reality
VP	View Point
3D+2D	mixing 3D and 2D convolutional layers

# Chapter 1

## Introduction

### 1.1 Overview

Human-computer interaction (HCI) is a multidisciplinary field of research, which focuses especially on the design of computer technology [14]. In particular, the interaction between humans and computers, such as body, hand, language, face and so on. Hand gesture is a non-verbal body language and ubiquitous in our life. Hand gesture recognition has achieved increasing research interest in computer vision, human-computer interaction and pattern recognition in recent years.

Hand gesture is a type of expression of human thoughts, since the beginning of human history, hand gesture has been used to manipulate, expression and so on, which is early than speech communication [15]. Because hand gestures feature natural, ubiquitous, and meaningful characters, that enable the combination with sound to achieve a tightly integrated system during human cognition [16]. Human-computer interaction is inspired by human interaction, therefore gesture, speech, and vision constitute the most efficient and powerful method of HCI [17].

With the rapid development of technologies such as augmented reality (AR) and virtual reality (VR) technology, HCI has been greatly improved for gaming interaction of AR and VR control. Finger micro-gesture is a hot research focus due to the growth of the Internet of Things (IoT) and wearable technologies. Recently Google has developed a radar based micro-gesture sensor which is Google Soli [1]. Also, there are a number of finger micro-gesture techniques that have been developed using Time of Flight (ToF) imaging sensors for wearable 3D glasses such as Ather mobile glasses [18]. Note that the finger



gesture is only between the two or three fingers movements, and it does not include the arm and body. The principle of holoscopic 3D (H3D) imaging mimics fly's eye technique that captures a true 3D optical model of the scene using a microlens array, however, there is limited progress of holoscopic 3D systems due to the lack of high quality and publicly available database.

There is a need of benchmark database for holoscopic 3D micro-gesture recognition that shall be made a publicly available and this makes the innovative holoscopic 3D imaging system accessible for researcher to continue innovation in H3D micro-gesture recognition. The baseline method with a comprehensive results as a benchmark are made available for other researchers to explore their methods. And results as a benchmark are made available for other researchers to explore their methods. Outstanding opportunity was given to organize the first Holoscopic Micro-Gesture Recognition Challenge (HoMGR 2018) [19] in IEEE conference which attracted researchers all over the world to take part to put their efforts to this research [20] [21] [22] [23]. Although significant progress has been made on the performance of micro-gesture recognition based on H3D imaging, there are still some challenges that have not been tackled. During in this PhD research, further developed H3D micro-gesture recognition, and achieves a robust H3D micro-gesture recognition system. Therefore, based on the each subset data, different methodologies are proposed in this thesis to tackle problems related to H3D information used in machine learning models. Additionally, the proposed methodologies achieved improved performance and accuracy compared to existing peer works in the field. a research area that includes computer vision, psychology HCI and H3D imaging technologies etc.

## **1.2 Background**

### **1.2.1 Human-computer Interaction**

User interaction is an activity between a certain device and the user. User interaction and user interface have many similarities. User interaction not only includes design interface, it also contains internet interaction, gesture, recognition and so on. Generally, user interface design is designed to enhance visuals and user-friendliness of the interface. It can increase the satisfaction of using the product [24]. Interaction not only enhances the experience of the user, but also uses many biological and electronic technologies to add more value. Click and touch are two traditional and basic methods of user interaction. Recently, gesture and speech methods became two popular research topics in the human-computer interaction filed. For new wearable designs, languages and gestures are the

primary methods. More and more products are beginning to use these new ways of interaction. For 2D interaction, many companies used large size screen display to improve the user experience, however, the disadvantages are obvious. Therefore, interaction method of 2D interface combining with multi-touch has been used to new kind of devices, such as Iwatch, pad and so on.

For the past couple of decades, the mouse and touch-based techniques are main interaction methods [25]. In recent years, AR/VR applications has been extend some traditional industries, like manufacturing. Digital manufacturing can help companies time and energy [26]. Moreover, wearable computer and AR have grown to a \$200 billion industry at a time when there are more and more business establishments [27]. Therefore, the wearable devices will gradually infiltrate the daily lives of people and become assistants [28]. Usually immersing technology is based on the 3D as well as AR and VR. Virtual reality or augmented realities is also known as immersive multimedia or computer-simulated reality. a computer technology that replicates an environment, real or imagined, and simulates a user's physical presence and environment to allow for user interaction. Virtual realities artificially create sensory experience, which can include sight, touch, hearing and smell.



Figure 1.1: Leap motion sensor captures hand motion by the skeleton.



Figure 1.2: Atheer AR glasses.

## 1.2.2 Hand Gesture

Gesture is a type of body language that can be used to communicate a specific and single command [29]. Gesture and body languages are not same with the sign language because, sign languages are full complex languages like other spoken languages and have their complex grammar systems, as well as being able to exhibit the essential properties that exist in all languages. Body languages are natural activities of humans and animals that used to express feelings. These feelings can be categorised into facial expressions,

body posture, touch as well as the use of space, gestures and eye movement [30] [31]. This project does not focus on the other type of body language apart from gesture.

Gesture can be include of the movement of speech, face and other parts of the body. This gesture can be grouped as descriptive gestures, emphatic gestures, suggestive gestures and prompting gestures. Most gestures express the natural reaction of hearing, visualization, touch, and feel. 3D gesture can be more naturally described as the behaviour and thinking of a user. To achieve the best user experience, gesture control started to mimic more natural movements with higher precision. Meanwhile, a sensor with the same capabilities is required to support this function. Hence, an increasing number of commercial companies and research groups are working on this topic such as Google Soli Project [1], Leap Motion [2] and so on. Even though multiple types of sensors support novel technologies of gesture recognition, most of these technologies presented are still in the exploratory stage and are lacking in reliability and precision aside from high cost of production.

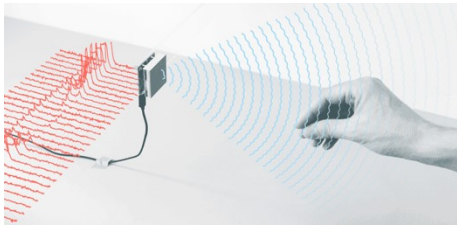


Figure 1.3: Google Soli project using Radar sensor[1].



Figure 1.4: Leap Motion sensor capture hand motion by the skeleton[2]

With the development of new techniques, human computer interaction has started to multiply in professional research and applications. Gesture and speech support the human semantic intent and manipulate activity. Recently, in order to improve human conversation interaction, many research and application proposed the method of combining gesture and speech to produce multi-model interaction.

### 1.2.3 Holographic 3D Imaging System

Lippmann and Ives first implemented the holographic 3D imaging system in 1908 [32]. The holographic 3D imaging system mimics viewing system of the fly's eye technique and uses special optical units to capture true 3D scene. Holographic 3D imaging system has more widely view coverage than 2D imaging technology. This technique uses unique optical components, that create and represent the naked eye object's optical model in the

form of the planar intensity distribution.

The 3D imaging system is popular uses in academia and industry for decade years, its especially contributed to 3DTV and display area. Nonetheless, the holoscopic 3D camera is enabled to offer to embed 3D information high-quality RGB image and video, which satisfy the gesture recognition of high demand.

## 1.2.4 Machine Learning

### Classification Models

This section is briefly introduce machine learning techniques of classification tasks.

- **Support Vector Machine**

Support vector machine (SVM) is part of the supervised learning algorithms that are widely use for classification and regression analysis of data analysis. It used hyperplanes to classes and separate the different classes of patterns. It is achieved by mapping the given features into a high dimensional feature space that enable to separate the feature from different classes.

- **K-Nearest Neighbour**

K-Nearest Neighbour(KNN) is a efficient machine learning algorithm that enable to calculated distance between an input sample with a set of training samples. The classification is based on a plurality vote of its neighbours, then the object will be assign to the class , which is most common among its K-nearest neighbour.

- **Deep Learning**

Deep learning is part of machine learning methods. And the learning is classified to supervised, semi-supervised, and unsupervised. deep neural networks, deep belief networks, recurrent neural networks (CNN) and convolutional neural networks are very popular architecture enable to using for computer vision, natural language processing, audio recognition, speech recognition and so on. The principle of deep learning is using multiple layers to progressively extract higher level features from raw input. With developing of high demand, the trend of the networks are going to deep such as ResNet 152 [33] and DenseNet 161 [34].

## 1.3 Research Motivation

There has been a lot of work done for gesture recognition based on various sensors and technologies, such as Kinect, Leap motion and Google soli. With the rapidly developing

of virtual reality (VR) and augmented reality (AR), the gesture controlling is required high precision and convenience to achieve the interaction. However, there are many particular problem which still have not yet been solved. Specifically, outstanding performance and manipulation of high demands are turning into the researchers' spotlight issue for discussing. The holoscopic 3D camera has widely viewing to capture and tracking the object's movement, which is as a novel robust sensor enables to more possibilities for gesture recognition. As a single aperture camera with micro-lens array enable to capture the multi-view image with 3D information in 2D format. However, there is no available public data enable to use for 3D micro-gesture recognition. Therefore, this research is first proposed using the H3D camera as a capture sensor to recognize the 3D micro-gesture. Then hold a international challenge in order to inspire more researchers and companies join this new research.

Machine learning, especially deep learning is a hot topic in the pattern recognition area. Since Convolutional Neural Networks has been proved a great success for large-scale ImageNet dataset. It created many miracles on different areas. Many researchers are based on deep learning to produce the gesture recognition, for example Tas [35] proposed using first layer of CNN to recognition Kinect-based data. Using CNN architectures and LSTM to recognition the dynamic hand gesture by Chinmaya [36]. Although the data are based on different technologies, the performance has been surpassed before. Based on peer works, the motivation is to produce advanced techniques to improve the performance of H3D 3D micro-gesture recognition. The Holoscopic 3D micro-gesture(HoMG) dataset has been created and public to encourage more researchers work for this area.

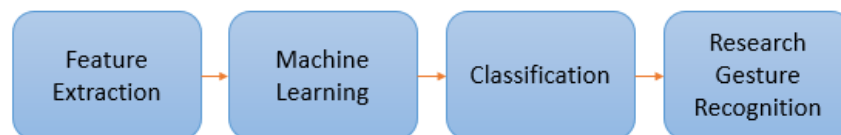


Figure 1.5: H3D micro-gesture recognition approach.

## 1.4 Research Aim and Objectives

The main aim of the research is to use advanced techniques to improve the 3D micro-gesture controlling for VR/AR immersive devices. Since H3D imaging system provides enrich high accuracy, high-resolution image-based and video-based data with 3D information, and it has many advantages for micro-gesture recognition.

Good performances have been made on H3D display technology and image processing. However, this is first time based on the H3D image system to recognize the micro-gestures, and the main challenge that is a combination of different subjects such as gesture interaction, H3D imaging system, image processing, and deep learning and so on. Therefore, learned many experience for different area to try on this new topic.

During the research, holoscopic 3D gesture dataset was developed that will be made available public for a better research dissemination which is Holoscopic micro-gesture(HoMG) dataset. Based on this dataset, it had a fundamental result on image-based subset and video-based subset. Moreover, we hold an international challenge to public this dataset and baseline then encourage more people to join this new area research. After this dataset got more and more attention, the research will investigate efficient H3D micro-gesture system, which includes holoscopic 3D image processing, image-based micro-gesture recognition and video-based gesture recognition for deep network.

Investigate Holoscopic 3D content analysis for 3D feature identification and selection.

The research finding will be evaluated based on different novel methodologies. List of main objectives of the research as follows.

- Carry out a literature review on gesture interaction, 3D imaging system and machine learning algorithms.
- Investigate Holoscopic 3D imaging system for 3D micro-gesture application.
- Design protocol for data collection on holoscopic 3D imaging system based micro-gesture recognition.
- Identify and develop a suitable machine-learning algorithm for micro-gesture recognition.
- Evaluate the whole system and provide further improvement.

## 1.5 PhD Contributions

Chapter 3 based on the human gesture conventionality to designed three micro-gestures to use on holoscopic 3D micro-gesture (HoMG) recognition dataset set up. HoMG datasets is consist of the image-based subset and video-based subset, 50 participants have been

joined this dataset recording, including 17 female and male participants from different races. 40 subjects have been on HoMG and 10 subjects as backup. In order to data has diversity and rich, four positions and two background colours for each participant has been used recorded. Image-based subset contains 16763 training images, 6560 development images, and 7291 testing images. Video-based subset have 240 training videos, 240 development videos, and 240 testing videos. For image-based subset, it used LBP and LPQ to extract the feature for classification. For video-based subset, used LBPTOP and LPQTOP for feature extraction. Then used three common classifiers of k-NN, SVM and Bayes on baseline results to classify the three H3D micro-gestures. HoMG dataset as a first H3D micro-gesture dataset has been made publicly available for micro-gesture recognition competition (HoMGR 2018, <http://3dvie.co.uk/>) and the baseline was published on the challenge workshop in 2018.

Chapter 4 creates automatically detected the edges of element images using advanced the H3D image pre-processing algorithms and extract 16 viewpoint images from image-based subset data. Then uses convolutional neural network with attention-based residual block to learn the micro-gestures on each viewpoint, along with the finger movements and different angles. Finally, bagging classification tree decision level fusion was used to combine the predictions together. Innovative methods are designed to combine image processing, deep learning and fusion to achieved pre-processing easily-obtainable viewpoint image, used the 3D information from H3D data and gained the best accuracy on the image-based subset.

Chapter 5 develops a novel end-to-end system that obtains the 3D information from original H3D frames then input to deep learning networking to training for classification. Therefore, the PVP based front-end has been created to assist the deep Networks' learning the H3D features. First, the PVP based front-end used the classic deep network architecture to evaluate feasibility. The result surpassed all the before result on the video-based subset and proven its effectiveness and robustness. Then, the PVP based front-end has been produce in four popular deep network architectures. And based on the input particular questions, the experiments were shown the performance and comparison.

## **1.6 The Outlines of PhD Thesis**

Chapter 1 is an introductory chapter, which briefly introduces background of the research, the aim objectives, the research contributions and the thesis outlines.

Chapter 2 provides a detailed review of current technology and techniques in the area of human hand gesture, gesture recognition, holoscopic 3D imaging and deep networks. The chapter also contains the details of creating efficient system based on holoscopic 3D imaging for micro-gesture recognition, such as deep network.

Chapter 3 proposed a novel holoscopic 3D micro-gesture (HoMG) dataset. It consists of image-based dataset and video-based subset. And the baseline has published using to speed up the research in this area.

Chapter 4 presents a micro-gesture recognition system based on the image-based subset data. It contains a fast, robust method for multiple viewpoint image extraction, CNN mode with an attention-based residual block, and Bagging classification tree decision level fusion.

Chapter 5 proposed a micro-gesture recognition system based on the video-based subset data. It contains easily-obtainable pseudo viewpoints extraction, deep network and majority voting to compare each method performance.

Chapter 6 presents conclusions of the research findings and development. In addition, it discusses the potential future work in the area of H3D micro-gesture recognition.



## 1.7 The Author's Publications

### Journal Article

[1] **Y. Liu**, M. Peng, M. R. Swash, T.Chen, R. Qin, H. Meng, “Holoscopic 3D Micro-gesture Recognition by Deep Neural Network Model on Viewpoint Images and Decision Fusion”, IEEE Transactions on Human-Machine Systems, 2019. (Submitted on Jan. 2019, and revised version submitted on Aug. 2019.)

### Conference Article

[1]**Y. Liu**, S. Yang, H. Meng, M. R. Swash, S. Shan. “Novel pseudo viewpoint based Holoscopic 3D Micro-gesture Recognition”, In 2020 IEEE International Conference on Multimedia and Expo (ICME)(Submitted).

[2] R. Qin, **Yi. Liu**, M. R. Swash, M. Li, H. Meng, T. Lei and T. Chen, “A Fast Automatic Holoscopic 3D Micro-gesture Recognition System for Immersive Applications”. ICNC-FSKD 2019: Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery, pp. 696-703, 2019.

[3] **Y. Liu**, H. Meng, M. R. Swash, Y. F. A. Gaus, and R. Qin, “Holoscopic 3D Micro-Gesture Database for Wearable Device Interaction”, 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition. (FG 2018), pp. 802-807, 2018.

# Chapter 2

## Literature Review

### 2.1 Human Computer Interaction

Human-Computer Interaction (HCI) is an interdisciplinary research between computer technology and user interaction design and appeared as early as 1983 [37]. HCI tends to improve the interaction between users and computers, which gives the user more flexibility in commanding and controlling computer functions. This multi-disciplinary subject involves computer science, media studies, design, behavioural science and several other relevant field subjects.

There are many interaction methods between people and computer systems - many devices are used to assist in HCI. In the early stages, most methods belong to Graphical User Interface (GUI), but also have some methods of using speech recognition and synthesizing systems, which is referred to as Voice User Interfaces (VUI). In addition, interaction method of multi-modal and user interfaces permit the user have freedom nature body languages. For example, VUI uses speech recognition and synthesizing systems to achieve a complete system. Therefore, the development and applications of HCI can be found almost everywhere, from Internet browsers, handheld computers to desktop applications.

Hand gesture recognition is a subject that part of the HCI. As a robust, natural interaction method, gesture interaction benefit sign language recognition and the gaming industry. Although the technology in these fields have yet to achieve technology-readiness levels that can be applied independently to various industries, the revolution and impact it brings to the interactive industry cannot be underestimated.

HCI has a history of more than 40 years, and different types of devices have also determined different ways of interaction [38]. User interaction principle is human-centred, which follow design concepts based on human natural behaviour. It is a perceptible type of user interface that can be used to guide users with implicit and explicit information [39]. It interacts with the sensor through vision, hearing, touch and taste. For example, Google Soli Project uses the sense of touch to interact with the sensor.

### **2.1.1 Body Language and Gesture**

Humans and animals naturally use body language to express thoughts and feelings. These body languages can be further categorised into facial expressions, body posture, touch and the use of space, gestures, and eye movement. This research focuses on gestures as a form of expression. Gesture and body languages are not same with the sign language, because sign languages are full complex languages like other spoken languages and have their own complex grammar systems, as well as being able to exhibit the essential properties that exist in all languages. This gesture can be grouped into descriptive gestures, emphatic gestures, suggestive gestures and prompting gestures. Most gestures express reaction to hearing, visualization, touch and feeling. 3D gesture can be specifically described as the behaviour and thoughts of a user.

Based on previous works, several gesture varieties can be defined. For instance, Quek et al. [3] proposed gesture types, which are based on manipulation, gesticulation, deictic and language gestures and semaphores to classify the current gestures.

- **Deictic Gestures**

Deictic gestures contain pointing to establish the identity or spatial location of an object within the context of the application domain. The application domain can include desktop computer, virtual reality applications and mobile devices.

- **Manipulative Gestures**

The intended purpose of manipulation gestures is to control some entity by applying a tight relationship between the actual movements of the gesturing hand or arm with the entity being manipulated.

Manipulations can occur both on the desktop in a 2-dimensional interaction using a direct manipulation device such as a mouse or stylus, as a 3-dimensional interaction involving empty-handed movements to mimic manipulations of physical objects as

in virtual reality interfaces or by manipulating actual physical objects that map onto a virtual object in tangible interfaces.

- **Semaphoric Gestures**

Flags, lights or arms are commonly as the signalling used for Semaphore gesture system [3]. Additionally, we define semaphoric gestures to be any gesturing system that employs a stylized dictionary of static or dynamic hand or arm gestures. Semaphoric approaches may be referred to as "communicative", in the sense that gestures serve as a universe of symbols to be communicated to the machine.

- **Gesticulation**

For the past couple of decades, many researchers believe that the application of gestures is only used for theoretical research [40]. However, with rapid developing of technologies, this gesture also known as descriptive or iconic gestures, is designed to increase the clarity of speech recognition and orally describe the physical shape or form through gestures.

### **2.1.2 2D vs. 3D Gesture Interaction**

A gesture is a form of non-verbal communication in which visible bodily actions communicate particular messages, either in place of, or in conjunction with speech. Gestures include movement of the hands, arm, or other parts of the body [41]. Thus, developing sensor uses different method to capture gesture that is based on the different type sensor device such as the Wii. The popular and easy to control display is Microsoft Kinect that fiducially markers to capture visual tracking system. Meanwhile, the gestural interaction following the increasing 3D displays to rapidly developing. Immersive 3D user experience is used for the gaming environment, so freehand interaction and no hands-on input is a mainstream [42]. Although touchless gestures bring convenience, like no need to hold and touch the screen, they lose some advantages, like clicking buttons and tapping surfaces. Nevertheless, this project uses the simulate touch gesture and the interactions feel physical and responsive. Users can feel the finger's haptic sensation [43]. The research development provides more natural and human-centered methods of interacting with computers. This is also a kind of the perceptive user interfaces, which support the perceptive, as same time effect the implicit and explicit information between the user and the environment. Vision is a non-intrusive and low cost method, therefore, imaging capture became a popular and modality method for user face.

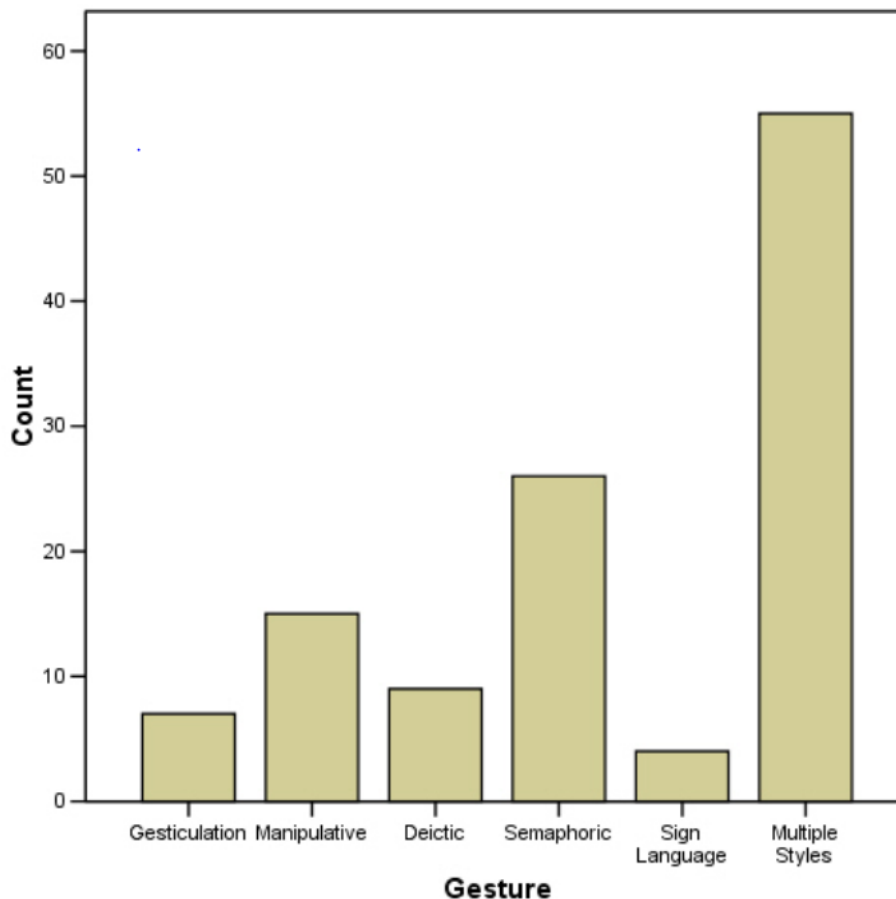


Figure 2.1: The diagram represents the distribution of the different gesture styles that appear in the literature review[3].

Gestures have been used for over 40 years in HCI, namely for mouse control, keyboard and stylus [44] in graphical user interface (GUI). However, not all the technology of output and input can permit gesture interaction and the immersive user experience demand more natural and non-intrusive ways to improve the interaction between human and computers. With the fast-growing smart mobile phone market, multi-touch interaction became a significant component of gesture interaction. More and more research in this day and age focus on the touch screen gesture control [45] [46] [47] [48]. Aoki et al. [46] proposed 2D gesture use for touch screen interface of TV controllers, which is designed for 2D finger gesture movement on Unicursal Gesture Interface (UGI). Then, Bragdon et al. [47] considered the use of the environment to supplement several design elements for touch-screen gestures and highlighted the limitation of dedicated screen space. They compared hard button-initiated gestures [49], Bezel gestures [50] and soft buttons [51], and describes physical button user experience, advantages and disadvantages of each gesture in detail. However, the evaluation proofed the touch screen gesture limitations for

new devices. Therefore, manipulative gesture has been used in new technology applications, such as wearable device, Internet of Things(IoT) and so on. Although Yepez et al. [52] proposed using projection to support the interface of the gesture assistant, the free manipulation gesture is yet to be solved due to technical limitations.

3D gesture system is beyond the traditional interaction and is a more complex version of HCI. In the last few decades, it is clear that 2D gestures are ubiquitous and significant for HCI, such as a mouse, pad and so on. Since 2006, WiiMote has proposed a method of hand movement in 3D space [32]. Xbox designed by Microsoft Kinect has made a major breakthrough in 2010, which has driven users to get a 3D motion control experience [37], especially using Kinect as a sensor for 3D interactive systems. The Software Development Kit (SDK) also breaks the tedious 2D gesture system and provides a more accurate display, such as RGB-D cameras and microphones. Later, due to the development of 3D gesture systems, more and more displays can be used in augmented reality interface. Most 3D gesture manipulations become non-touch, natural and flexible, which allows the users to express more freedom in defined 3D space. Users can truly use their body or finger to control displays. The trend of the 3D gesture system is heuristics, which means the user can use it based on their instinctive activities [38].

However, there are many challenges in the area of 3D gesture design. Most marking gestures are based on the previous 2D input devices to design. G.Ren [42] proposed four main challenges in 3D gesture designs:

- Ensure the user can see all the menu items and correctly perceive their 3D position in order to select them. Hence, rendering the 3D marking menu needs very carefully.
- How to ensure users can easily choose an accurate option. Thus, in addition to facilitating the perception of the correct position of each menu item, an effective 3D hand gesture selection technique should also facilitate efficient selection of the correct menu item in the presence of such an interfere.
- How to coordinate the user's eyes and hands to effectively to control their hand movement in a defined physical 3D space and measure the relationship between the virtual and physical 3D space.

- How to balance the traditional user experience of user selection of objects. Due to the freehand gestural interaction are no button and physical touch interface [53].

However, User experience of 3D gesture system is a complex process as it demands high speed motion capture, however at the same time a simple interface platform needs to be presented to the users to enable usage with minimal training. Recently, many productions have functions to capture gesture. For high-level precision gestures, most sensors essentially lack flexibility and accuracy in gesture recognition.

Limitations exist in many devices of this sort. Primarily, the RGB data need to be layered and then supplemented with the use of certain tools and accessories. The advantages and disadvantages of each sensor are discussed in detail in the following sections.

3D micro-gesture is ideal for people with disabilities, as it has touchless function and does not require any additional interfaces. This new type of design also matches new technologies such as Google Soli Project's 3D micro-gesture recognition by radar [1]. This type of design is based on the human centered principle that collects common human finger gesture movements.

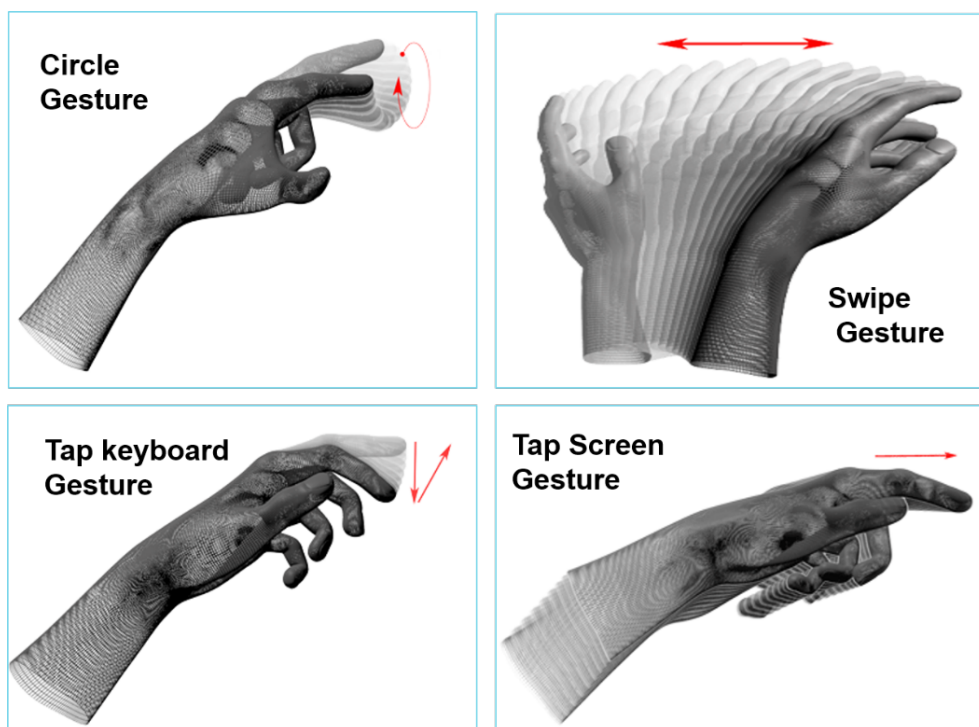


Figure 2.2: Continuous gestures [2]

### 2.1.3 3D Micro-gesture

Gesture recognition is part of computer vision, human-computer interaction and pattern recognition. Three-dimensional (3D) hand gesture recognition, because of the emerging depth sensors, assist in developing gesture recognition diversification. In this section, will introduce 3D depth sensors based on gesture recognition, which support more advantages beyond traditional 2D domain sensors.

3D gesture system is more advanced than traditional interaction, and therefore it becomes much better than previous HCI methods. In the last few decades, it is clear that 2D gestures are ubiquitous for HCI, such as mouse, pad and so on. Since 2006, WiiMote proposed the hand movement in the 3D space [43]. This is a significant breakthrough for Xbox, designed by Microsoft Kinect in the 2010s, which enabled users to have 3D motion control experience [54]. Specifically, Kinect uses skeleton capturing to develop their 3D interaction system. Software Development Kits (SDKs) also broke the tedious 2D gesture system and achieved high-accuracy display used in RGB-D cameras and microphones. Therefore, most displays can be used for augmented reality interfaces, because of the development of 3D gesture system. Most 3D gesture features are touchless, natural and flexible, giving users more freedom to express themselves within a defined 3D space. The user is able to truly use their body or finger to control the displays. 3D gesture system design is commonly heuristic, user can use it based on their instinctive activities [41].

### 2.1.4 3D Micro-gesture Technologies and System

#### Microsoft Kinect

Kinect as a motion device by Microsoft is used for Xbox 360 and Xbox. It is able to link video game consoles and Windows PCs. The webcam-style device is an add-on peripheral of the console and computer, which uses gestures and spoken commands to control games through a natural user interface [55]. The first-generation of Kinect was first introduced in November 2010 in an attempt to broaden Xbox 360's audience beyond its typical gamer base [56]. A version for Windows was released on February 1, 2012 [57]. Kinect is in competition with several motion controllers on other home consoles, such as Wii Remote Plus for Wii and Wii U, PlayStation Move/PlayStation Eye for PlayStation 3, and PlayStation Camera for PlayStation 4. Microsoft released the beta version of the Kinect SDK for Windows 7 on June 16, 2011 [58] [59]. This SDK was meant to allow developers to write Kinect apps in C++/CLI, C, or Visual Basic [58].

#### RGB-D Cameras



RGB-D cameras is type of depth sensing devices. In the past few years, used in novel camera systems such as Microsoft Kinect and Asus Xtion [60]. The sensors provide both colour and dense depth images, which are able to be used in real-time applications. Those systems will lead to a boost of new applications in the field of 3D perception. Particularly, it is suited to robots operating in unstructured environments and under real-world conditions [59]. RGB-D sensors are also used for 3D mapping and location determination, path planning, navigation, object recognition and the tracking of people. RGB-D cameras estimate depth using structured light techniques. However, the optical power of the projected pattern reflects insufficient information back to the sensor. Another drawback with RGB-D is that the active triangulation method requires a baseline for depth estimation.

### **Leap Motion**

Leap Motion is a company developing motion-sensing technology. It creates the no touch gesture interactions to enable computer hardware to recognize hand and finger motions as a form of command input. It has similar functions to a mouse but does not require a cable connection, which brings about more convenience to the user. It is very prominent in visual aspects as it uses 3D technology to present virtual reality scenes and provide an immersive experience for the user. Users can use their hands to interaction with the scene. It is compact and easily portable, on top of that, it is flexible and can identify swipes, grabs, pinches and so on. Leap motion can also work with a traditional mouse and keyboard. Nowadays, due to the ease of use in order to achieve immersive experiences, Leap Motion has hundreds of applications [2].

### **Google Soli project**

Google Soli Project is a new technology that uses radar to capture and recognise micro-gestures. Google Soli project designs powerful micro-gesture interaction systems that are based on human intuition.

Google Soli project's gestures are designed for two finger (thumb and forefinger) operations to minimize interference with the outside world. This is of worthy potential for wearable technologies, phone, computer, car and IoT digital displays. The radar has characteristically high recognition accuracy - an advantage over most sensors, which typically have low precision recognition. Although the radar is limited by distance, there is no 2D interface as an auxiliary operation [1].

Micro-gesture not only is minimal in size, they can also accurately control displays. Intricate gestures has a variety of advantages, such as movement privacy, no limitation in space and are free from sound or speech barriers. Nonetheless, it is difficult to track small movements accurately. Additionally, in noisy environments most reasonably low

cost technology cannot meet satisfactory requirements. Moreover, there are many sensors that cannot track acceleration movement [42]. This project uses holoscopic 3D camera as the sensor, which is a promising technology that can record full vertical and parallax of a scene. The holoscopic 3D camera offers the 3D precision gesture system more freedom and higher accuracy [1].

Cheng et al. [61] proposed the classification of 3D hand gesture into four categories. These are 3D hand modelling, static hand gesture recognition, hand trajectory gesture recognition and continuous hand gesture recognition. The team intensive analysed 2D hand gesture recognition approaches during past decades. Based on the type of sensor used in the data collection process, there are three main types of gesture data. For example the first type is based on the sensor's accelerometers or gyros to capture hand and finger movement. The second type is multi-touch screen sensors, which are widely used in mobile devices. Thus, the limitation of sensor is obvious, in the sense that they are not able to support touch-less interaction. The third type is vision-based sensors, which has significant advantages over the other two types. Compare this to another two type of sensors, vision-based sensors do not need physical contact with the users and support a much larger working distance. Due to the high demand for hand detection and tracking, they used gloves or coloured markers in the early stages, although their approach can aid the algorithm, it compromises user experience. With the development of commercial 3D sensor technologies, the new sensors provide more robust approaches for gesture recognition. Moeslund et al. [62] classified their recognition types into static recognition and dynamic recognition. For static recognition, the aim is to capture and recognise movements. Therefore the data is based on spatial or frame, and recognition method is compared with prestored information of the current image and data formed can generate normalized silhouettes, templates, postures, or transformed templates. For dynamic recognition, the data includes spatial-temporal templates [58] and motion templates [63], which uses temporal characteristics to achieve the recognition task.

### **3D Micro-gesture system**

Sang et al.[64] proposed a system that uses the ultrasonic active sensing to recognize the micro hand gesture. In this system, they use the end-to-end neural network model to learn radar signal processing and time-sequence pattern by machine learning. This system achieved the accuracy of 96.32% and supported the real-time prototype. The Pyro[65] is a micro thumb-tip gesture recognition technology which is based on the thermal infrared signals radiating of the finger movements. The sensor advantages are low-power, compact. It is suitable for wearable device and mobile applications. This system includes the software for signal processing and machine learning, infrared pyroelectric sensor, a custom sensing circuit. And it achieved 84.9% from ten-participant user data without the

light conditions, background motion and hand temperatures.

## 2.2 3D Imaging Principles

### 2.2.1 Stereoscopic 3D Imaging System

This technique was first proposed by Charles Wheatstone in 1838 [66] [67]. In the early stages, stereoscopic images required complicated and bulky optical components to show the perception of 3D depth effect for the viewer. With further developments in research, the pair of 3D glasses replaced large optical components. There are three methods of stereoscopic glasses, as shown in Figure 2.3 below, which enable the viewers binocular visual system to perceive slightly different views.

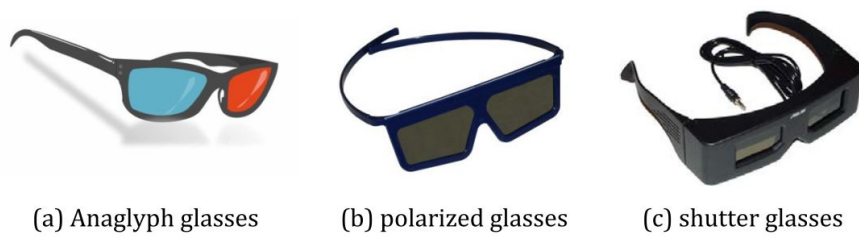


Figure 2.3: Stereoscopic 3D glasses [4].

The first method is anaglyph, which uses colour coded image projection to allow viewers to perceive slight different views in each eye. Specifically, the blue tint and red tint of the anaglyph glasses filters view respectively by allowing only compatible views to be perceived by the viewer. This system is simple and popular due to the low-cost and ease of production. However, one crucial problem is the weakly rigging or misalignment of its recording system, that lead to ghosting effect and various degrees of cross-talk. Additionally, the blue and red filters leak over and produce eyestrain, and thus have a high risk of an unsatisfactory 3D effect. Figure. 2.4 shows the stereoscopic concept of red and blue filters from the left and right eye.

The second is polarisation, which is widely used in industry and entertainment. This technique depends on perpendicular polarisation, which projects the two separate views simultaneously, then uses varying polarized lenses to filter out an unwanted view, resulting to each visual system (eye) to perceive the correct view [68]. Techniques of linear polarisation and circular polarisation both support stereoscopic 3D imaging system, Fig-

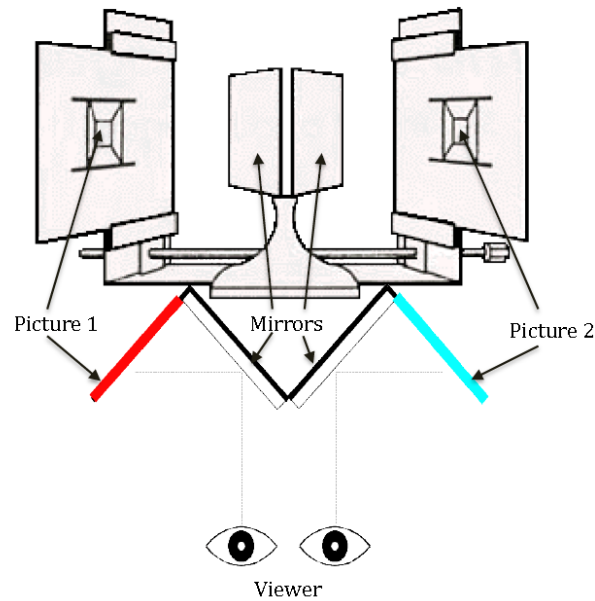


Figure 2.4: Anaglyph stereoscopic 3D scene[4]

Figure 2.5 shows the difference between linear polarisation and circular polarisation. As circular polarisation is more flexible in angles than linear polarisation, which do not have any limitations in terms of the viewing angle and produce more natural user experiences. Compared with the anaglyph technique, the benefit of polarisation is that it offers full-colour images, while the disadvantage is the reduction of resolution by half, which is still a significant limitation of the technique [69].

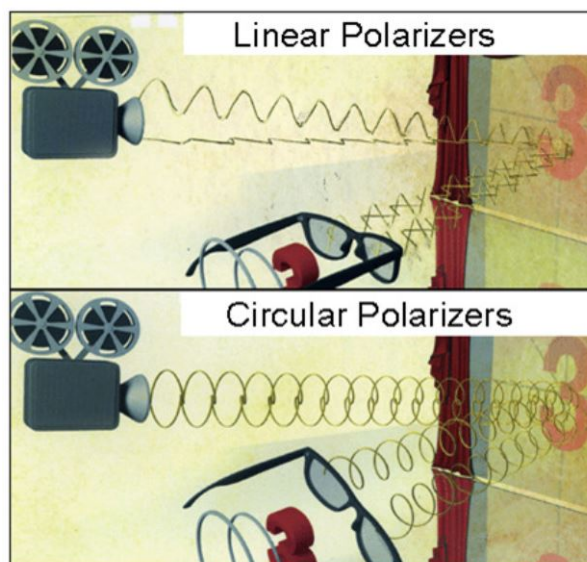


Figure 2.5: Linear polarisation and circular polarisation [5]

The third method is time division, also called “active shutter”, which displays the left and right views at a high frame rate. Therefore, the shutters of the lens need to rapidly open and shut, which shown in Figure. 2.6 below, alternating images of the screen display to the left and right eyes. All the methods mentioned above have its disadvantages and for active shutter, the time delay between the two images, and visual eye strain is inevitable after a considerable duration of usage.

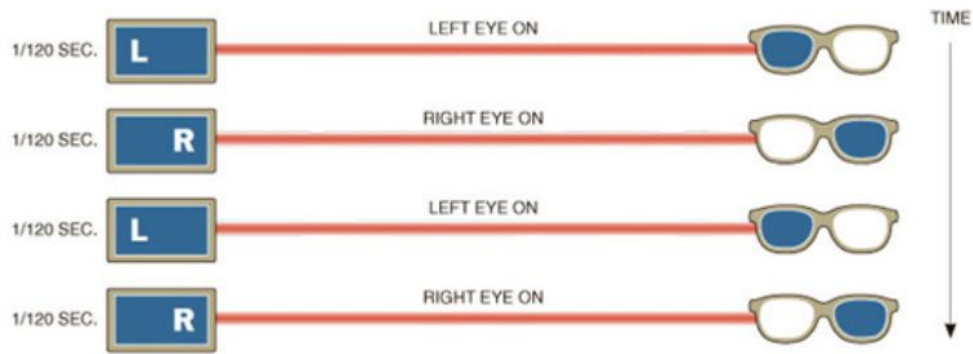


Figure 2.6: Time Division[6].

The major drawbacks of stereoscopic 3D imaging is that wearers of the glasses over a period of time will experience eyestrain, headache and other fatigue related symptoms. However, the principle behind the most commercial displays is based on the stereoscopic imaging system, mostly used to display true 3D modelling and improve on current 2D viewer experience with the introduction of depth.

### 2.2.2 ToF 3D Imaging System

Time-of-flight (ToF) camera is a popular imaging system due to its fast, accurate, easy depth measurement principle, and relatively low-cost, compared to other active imaging systems. This imaging system is used in HCI and computer vision applications such as object recognition, 3D modelling, mixed reality, obstacle avoidance and gesture recognition [70] [71]. The principle behind the Time-of-flight imaging system evolves around measurement of the time it takes a projected light source from the image system to reflects off an object in a given scene and back the imaging sensor, shown in Figure. 2.7 below.

Figure. 2.7 above presents the general principle behind the ToF imaging system, the following system usually consist of four major components namely: i) a recording lens ii) an integrated light source that is emitted either as a pulse or static frequency, iii) an imaging sensor and iv) interface. The ToF imaging system is mainly grouped in two categories

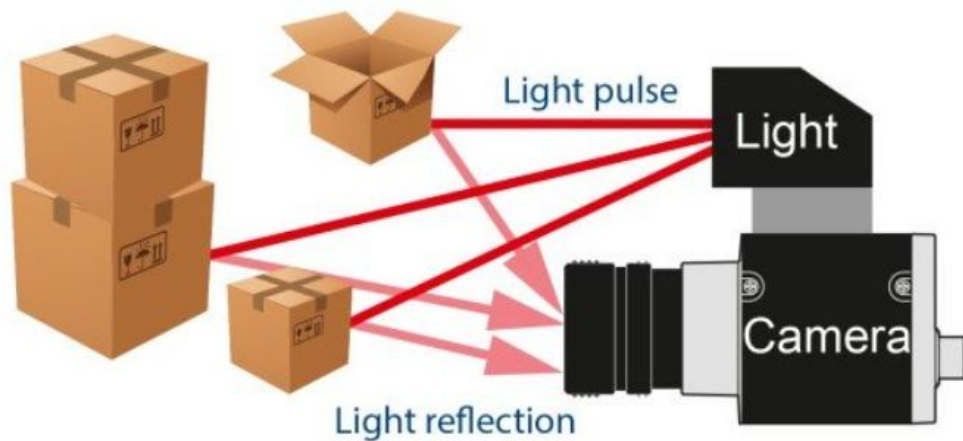


Figure 2.7: ToF imaging system principle [7].

based of their light emission process, knows as pulse-light and modulated-light ToF systems [72]. The pulse-light ToF imaging system measures the time taken for a light to reflect of an object in a scene and back to the sensor. As the speed of the light is already know, the distance is easy calculated with the acquired time data. However, in the case of modulated-light ToF imaging system, a constant light is emitted, therefore phase shift between the emitted and reflected light is calculated to determine the time, where distance is easily calculated. Based on the imaging principles of the ToF imaging system, they are commonly used in game devices and implemented in touch-less controllers, as a result a popular system used to capture hand movement [70].

### 2.2.3 Holographic 3D Imaging System

Lippmann and Ives first implemented the holographic imaging system in 1908 [32]. The holographic 3D imaging system mimics the fly's eye viewing system and uses special optical units to record the angular and spatial information of any scene. Holographic 3D imaging system has wider view coverage over the 2D imaging technology as this technique uses unique optical components, mainly the micro lens array, that can create and represent the naked eye object's optical model in the form of planar intensity distribution. The 3D imaging system has been widely applied in academic research and industry during the past decades, such as 3DTV and performing arts.

Development of the holographic 3D imaging system was initiated by Prof Gabriel M. Lippmann from 1908 [32] when he used a series of lenses to record an image with parallax

in all directions. An array of spherical convex lenses was used to record and play back the images. Emphasis was placed on the electronic image sensor, which captured all 3D information into a single aperture. Many sub aperture 2D images of the recorded scene are overlapped through the micro-lenses, known as elemental images (EI). Lippmann's method has limitations, as a result, Ives [73] presented the technique called "two-step integral photography" to solved the image problem known as the pseudoscopic image. The rectification of the pseudoscopic "inverted depth" problem, which enable to use for commercial applications [74]. However, this this resulted to image degradation. Chutjian and Collier of Bell Labs [75] proposed a more effective method in the year 1968, which uses computer generation to produce the correct 3D orthoscopic depth image. After, in order to achieve depth orthoscopic image without distortion and high degradation, Vil-lums [76] and Hain [77] further improved the method with the use of colour transparency and explored the technique of diffractive lens arrays to produce integral imaging.

With the rapid development of micro-lens manufacturing, optoelectronic sensors and display devices, and increasing number of research groups and companies have utilised integral imaging the principle to enhance sensor functionality, and tackled many crucial problems, for example, using the curved lens array to expand the viewing angle drastically. However, that could not get rid the image distortion caused by the lens arrays.

In 2010, H3D imaging has been further developed by 3D VIVANT in the research group of Brunel University London. The initial research aim is the capture and display of 3D content for H3DI, therefore the H3D content has been enhanced by the prototype of a 2D camera. The single aperture ultra-high H3D camera is capable of a high resolution RGB data. Following the early publications, they proposed many novel methods to tackle many critical problems, which achieved immersive ultra-high resolution 3D content. Most works are based on the display function to investigate the 3D optical work. Recently, research works have explored other related fields such as imaging processing, gesture recognition, facial recognition, and other similar virtual applications.

The holoscopic 3D imaging also known as integral imaging offers a technique that enable to record and replay the true spatial optical model of the 3D scene in the form of a planar intensity distribution by using optical components [78] on a much bigger sensor compared to competing Holoscopic industries like Lytro. This eliminates the need for a coherent source and limits darkness, making this method a more practical method to capture and display in real-time [79].

The holoscopic 3D imaging captures varying 3D scene's views via lens arrays. And the lens arrays is made up of hundreds of micro-lenses, which are 1D lenticular sheet or 2D

micro-lens array (see Figure. 2.8 below). H3DI have two main forms: unidirectional technique and omnidirectional technique. The unidirectional technique uses a lenticular sheet to form a single direction 3D depth and motion parallax image. The omnidirectional technique uses a 2D microlens array of based on the fly's eye technique to offer an image, that has full parallax 3D depth and motion parallax [8]. The Figure. 2.8 (a)(b) demonstrated the microlens arrays structures. (c)(d) spherical microlens array with single direction and full parallax. The microlens arrays has been spaced in tube and shown in Figure. 2.10(a). (b) shows holoscopic 3D camera setup, which can capture the larger overall objective or 3D scenes, and also easily capture optical geometry from microlenses arrays. Because the each small lens of microlens arrays are capable of acquiring the scene from different viewpoint. (c) shows the holoscopic 3D camera record the H3D micro-gesture image, which view the object at the a slight different angle to its neighbour and produce elemental images.

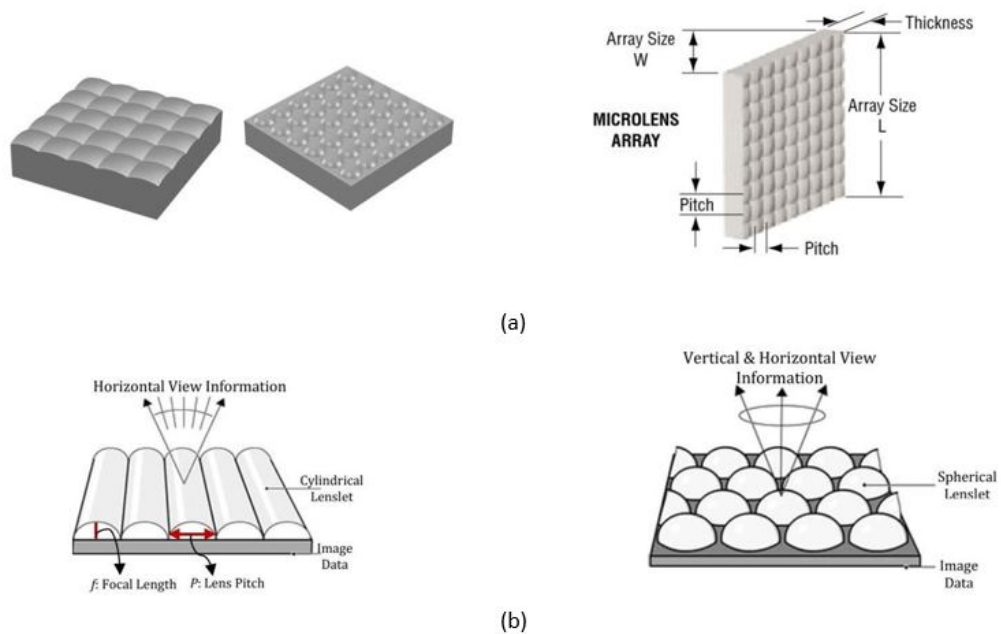


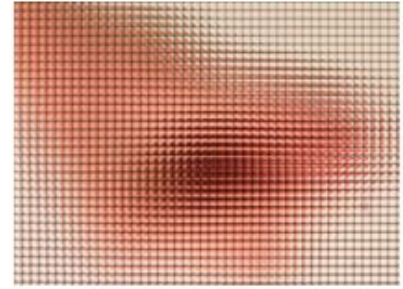
Figure 2.8: (a) demonstrated the microlens arrays structures. (b) spherical microlens array with omnidirectional and full parallax.

In this study, the omnidirectional images have been recorded with omnidirectional lens array. Therefore, the H3D image have vertical and horizontal depth information. Furthermore, H3D imaging system extraction and reconstruction in this research is based on the omni-directional 3D information. Computationally, the Holoscopic 3D image can be rendered in lens design software like POV-Ray. The rendering method depend on the relationship between the lens array and the camera clock model [80], where M by N or-





(a) Unidirectional holoscopic image



(b) Omnidirectional holoscopic image

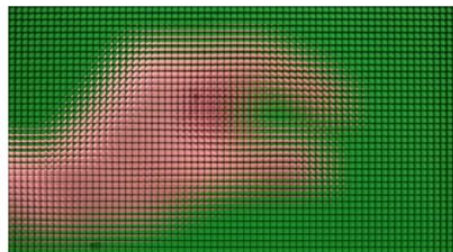
Figure 2.9: (a) is the unidirectional holoscopic 3D for the single direction. (b) is Omnidirectional holoscopic image with full parallax.



(a)



(b)



(d)

Figure 2.10: Holoscopic 3D camera and imaging system,(a) Microlens arrays.

thographic cameras represent the row number of per lens, and  $N$  is column number of per lens, supporting the full H3D spatial and angular data structure. As mentioned earlier the Holoscopic 3D imaging system offers high accuracy, full colour and multidirectional viewpoint image, therefore used as the primary recording device for this research.

## 2.2.4 Holographic 3D Camera System and Reinstallation

The holographic 3D camera is a single aperture camera with a micro lens array, which uses optical components to record and replay the true spatial optical model of the 3D scene in form of a planar intensity distribution. Figure. 2.11 (a) shows each main component, namely the relay lens, micro-lens array and camera. (b) shows the parameters of the 3D VIVANT project installation. Two relay lens as a group inverts an image and extends the optical tube, which is used to refract the 3D object. The micro-lens array is mounted between two relay lens, which capture the positions in the scene from different perspectives.

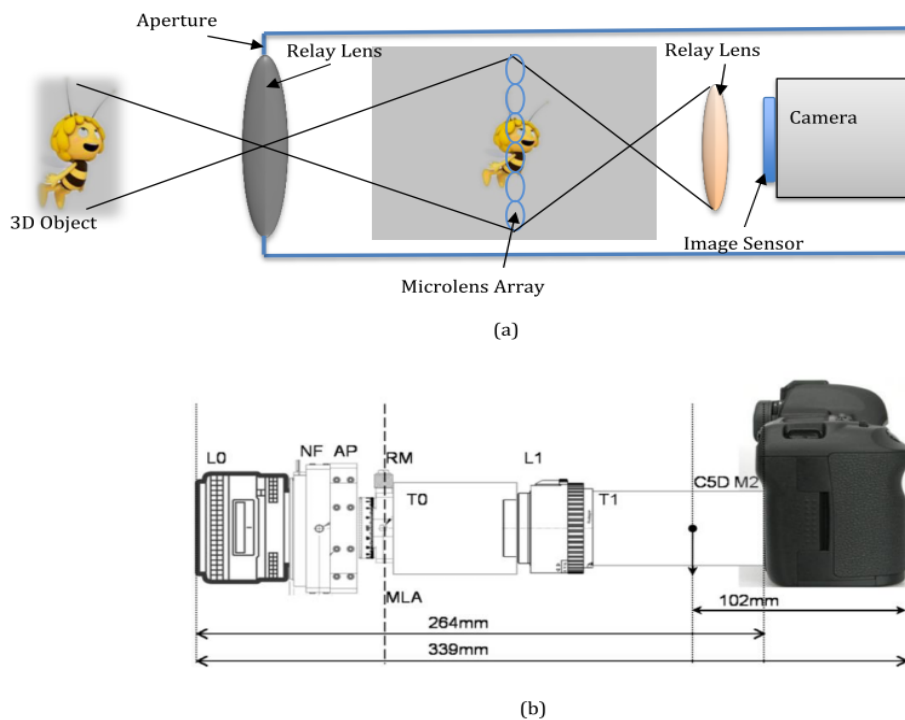


Figure 2.11: Assembled holographic 3D camera at Brunel University. (a) Main components, (b) The 3D VIVANT project installed details.

In this research, the camera is assembled for the 3D VIVANT project at Brunel University London. In the current study, the Canon 5D Mark2 (C5D M2) DSLR camera was reinstalled as the sensor to achieve the purpose of development and capture of H3D micro-gesture movement. The main Holographic 3D components consist of the micro-lens array, relay lens and digital camera. The installation work is based on the previous H3D system theory and work to re-installed the sensor, then the modifying part is depend on the micro-gesture capture requirement to adjusted the new focus lens. For the setup process, firstly, the inside of the camera should be cleaned in order to make sure that there is

no dirt or dust to affect the image. Figure. 2.12 shows the camera removal and cleaning process. All previous components are removed from original camera, then checking the cracks and damages, which ensure the encapsulation and precision of the optics. The installation and assembly process is based on the principle in Figure. 2.11 (a) Figure. 2.11 (b) illustrates the installation starts from right to left. The digital camera connect the tube and relay lens. The tubes of attaching the lenses are length and size of specially tailored, which is according to the needs of the camera principle. All the tailored tubes and holders are made in Brunel University London’s 3D printing studio. For better visual feedback during capture, H3D camera is connected with a Microsoft Surface tablet, which is used to supervise the conditions of the recording. It limits motion within a set frame, which eliminates video recording issues outside of camera’s viewing area. Figure. 2.13 shows the surface interface recording of the H3D scene. Figure. 2.15 shows the reinstated camera that used for recording the H3D micro-gestures.

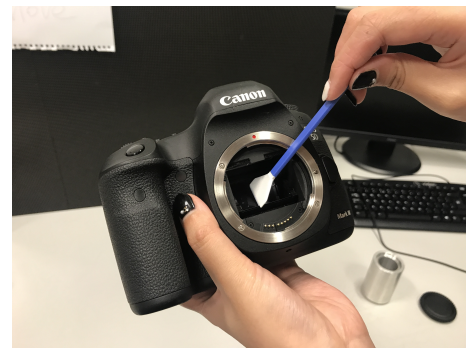


Figure 2.12: (a) Camera removed all the lens and prepare for checking the cracks and damage, (b) Clean up the camera.



Figure 2.13: The Microsoft surface supervised H3D camera interface.



Figure 2.14: The H3D camera re-installation for H3D micro-gesture capture.

### 2.2.5 H3D Image Processing and Viewpoint Extraction

The initial purpose of correcting geometric distortions is to avoid the banding and moire effects on the replayed image, caused by placement of the microlens array is placed in front of the camera sensor. The distortions lead to rotational and scaling errors became the main disadvantages. However, the methods require specialist solutions unique to tackle the different integral image establishment [81] [9] [82] [42]. A. Aggoun [83] proposed based on the Hough Transform to calculate the deviation's angle, which use to correct geometric distortion. Swash et al. [84] proposed detect introduced dark borders and remove them to correct the distortion of H3D imaging system.

The H3D imaging technology has been successfully applied to the 3D cinema, 3D-capable televisions and broadcasters. The H3D camera used here is built from the 3D Vivant Project (3D Live Immerse Video-Audio Interactive Multimedia) [8] and the purpose is to capture high quality 3D images. The developed camera includes micro-lens array, relay lens, and digital camera sensors. The principle of the holoscopic 3D imaging is shown in Figure. 2.11 (a)The 3D holoscopic image's spatial sampling is determined by the number of lenses. It shows that the captured 2D lenslet array views are slightly different angled than its neighbour [8]. The detailed parameters of the camera installation are shown in Figure. 2.11(b).

The fly's-eye lenses have a slightly different angle to other individual micro lens and hence multiple viewpoint of the scene can be capture in a single shot. Although, these methods have some limitations for depth field. The nowadays replay device of the 3D Holoscopic image can place the microlens array on top of the re-encoded planar intensity distribution from the diffuse light illumination from behind. The object will be constructed in space by the intersection of ray bundles emanating from each of the lenslets as shown in figure 2.7(b) [85]. In replay, the reconstructed image is pseudoscopic (inverted in depth). In the

last two decades, optical and digital techniques to convert the pseudoscopic image to an orthoscopic image have been proposed [8] [86] [87] [88].

Holoscopic 3D camera sensor has unique optical components which support the continuous parallax RGB image system, contain the depth information viewpoint images. The Figure shows the H3D imaging having the full colour with full parallax. H3D imaging is comprised of the 2D array of micro images. The H3D sensor is a crucial requirement for the capture of the objects. This database uses the H3D imaging system to support the dynamic and static RGB data. It can also record the continuous motion, and the varying views can be extracted through the repetitive lens array. The system brings more potentials for innovation in the field of gesture capture and recognition.

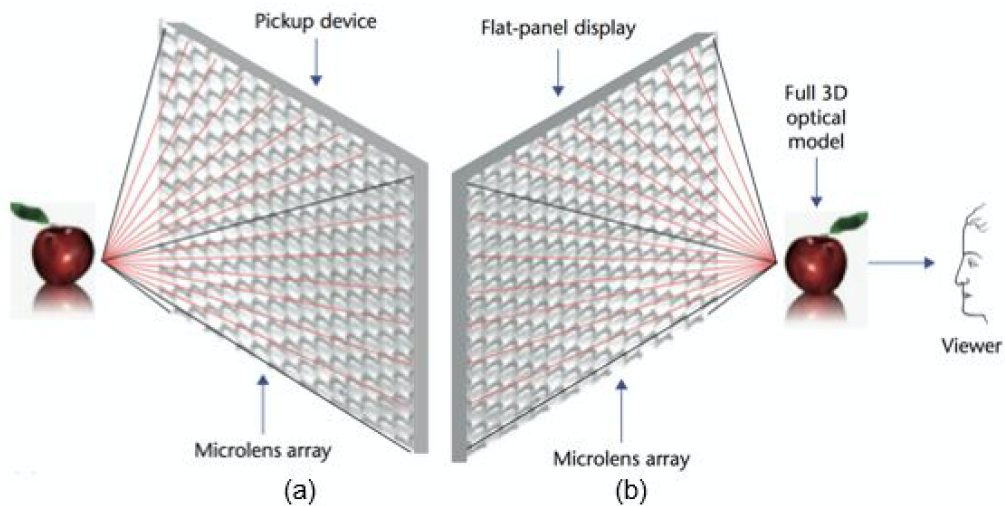


Figure 2.15: The H3D imaging system principle:(a) recording and (b) display process[8].

The 3D gesture system presented in this research is based on the Holoscopic camera technology capture multiple viewing angles of finger motion gesture, resulting to the creating of H3D dataset. This system can completely capture and accurately recognize micro-gesture, even in the case of micro-gesture. And immersive design made user can freedom enjoy virtual reality and augmented reality user experience. Holoscopic camera has widely viewing and uses view point map to recognize every elements images.

Holoscopic 3D imaging system is records 3D information in 2D format. During to the record step, the 3D scenes includes rich depth information. The viewpoint extraction is a popular method used for reconstructing multiple 2D viewpoint image from a single Holoscopic image.

The real-world light rays are captured via a microlens array as shown in Figure. 2.16 (a). The rays marked n1, n2, n3, n4, and n5 represent different perspective views of the same scene. Since a microlens array is involved in the recording stage, the local pixel position under each microlens contains directional view of the scene presented in Figure. 2.16(b).

A single orthographic (VPI) is reconstructed by sampling of all pixels in the same location under different microlenses, this creates a perspective image that portrays one directional view of the scene. The construction of viewpoint images in both UH3DI and OH3I are graphically illustrated in Figure. 2.17. It is also mathematically expressed in 2.2.5 [6]:

$$O(i, j) = I(i + nk, j + mp) \quad (2.1)$$

The above equation describes the viewpoint sampling, the n and m are the pixel coordinates under micro-lens of j and i, where  $j = 1$  to  $k$ ,  $i = 1$  to  $p$ ,  $n = 1$  to  $N$  and  $m = 1$  to  $M$  are the horizontal and vertical positions of an Omnidirectional Image (OI)'s pixel respectively as shown in Figure. 2.16(b)

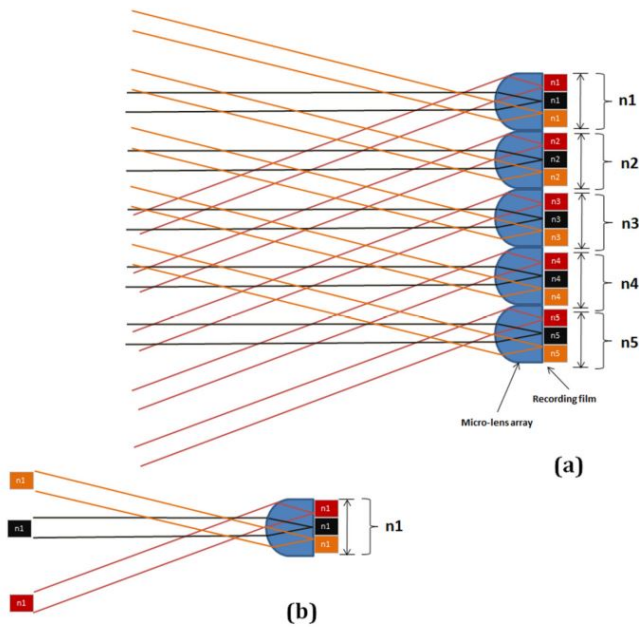
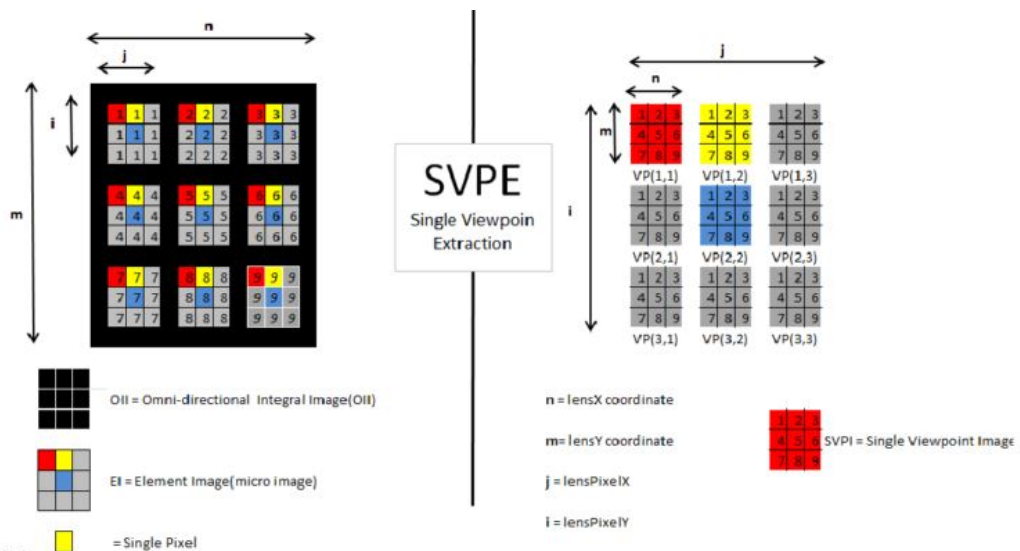


Figure 2.16: Holoscopic 3D capturing systematic[9].

It is important to mention that each individual viewpoint can also be defined as  $VP_j, i(n, m) = OI(j, i, n, m)$  where  $n, m$  are the co-ordinates of parallel light rays that is different from



(b): For illustration purpose, suppose  $3 \times 3$  pixels under each microlens. Extracting one pixel from the same position under different microlenses and placing them in an orderly fashion to form one viewpoint image.

Figure 2.17: Illustration of (a) UH3DI viewpoint image extraction, (b) OH3DI viewpoint image extraction[6].

perspective image property. Hence, the final output  $O(i, j)$ 's image resolution is the  $(n \times m)$  pixels.

## 2.3 Machine Learning Algorithms

### 2.3.1 Machine Learning

Arthur Samuel first proposed machine learning in 1959 [89] which demonstrated the pattern recognition and computational learning theory in artificial intelligence area. Machine learning offers the algorithms of study and construction. As the algorithm development, machine learning achieved a range of computing tasks such as data breach, learning to rank and optical character recognition and computer vision. Furthermore, based on the machine-learning task, they classified two of the broad categories: supervised learning and unsupervised learning. In the supervised learning has semi-supervised learning, active learning and reinforcement learning.

In 1959, the Arthur Samuel explained what is the machine learning “gives computers the ability to learn without being explicitly programmed” [90] [91], which described the machine learning model learns and make the data through the algorithm. For example, an implementation of the predictions and decisions are through the program instruction to do

algorithms, sample input is through the building the model. Machine learning can work for program and design part of the computer work. It is involved the optical character recognition (OCR) [92], search engines, computer vision and so on. Machine learning is the latest development and prospect of AI (Artificial Intelligence) and is an important part of AI. With the rapid development of the AR and VR technology, machine machine learning will become a fundamental change in digital display technology.

Machine learning need consider computational statistics, and they have overlapping parts, and they all focus on the predictions used by computers. Sometimes machine learning overlaps with data mining [93], which focuses on data analysis and unsupervised learning. However, the function of unsupervised machine learning learn and built a summary of the baseline behaviour of various entities [94] and then use to find meaningful exceptions.

Exception the unsupervised, machine learning has various concepts such as supervised learning, semi-supervised learning and deep learning etc.

### **Supervised Learning**

This is inference branch of the labeled training data. The training data is including the training samples. And every sample has input object and supervisory signal. It is through analysis the training data and produces. The algorithm determine the class labels for unseen instances to achieve optimal scenario.

### **Semi-supervised Learning**

It is a class of supervised learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data.

A standard neural network (NN) consists of many simple, connected processors called neurons, each producing a sequence of real-valued activation. Input neurons get activated through sensors perceiving the environment, other neurons get activated through weighted connections from previously active neurons. Some neurons may influence the environment by triggering actions. Learning or credit assignment is about finding weights that make the NN exhibit desired behaviour, such as driving a car. Depending on the problem and how the neurons are connected, such behaviour may require long causal chains of computational stages, where each stage transforms (often in a non-linear way) the aggregate activation of the network. Deep Learning is about accurately assigning credit across



many such stages [95].

### **Support Vector Machine**

Support Vector Machine(SVM) also called support-vector networks,which is supervised learning models in machine learning. This models used to data classification and regression analysis. The algorithm is generalized linear classifier, which performs binary classification on data by a supervised learning manner. The decision boundary is maximum-margin hyperplane of solving the learning sample [96].

The SVM uses the hinge loss function to calculate the empirical risk and adds a regularization term to the solution system to optimize the structural risk. It is a classifier with sparsity and robustness. SVM can be nonlinearly classified by the kernel method, which is one of the common kernel learning methods [97].

### **K-Nearest Neighbour**

k-nearest neighbors algorithm (k-NN) is a simple and classic machine learning method, which is a popular use for classification and regression in pattern recognition. The principle of k-NN is if a sample has most of the k most similar samples in the feature space (ie, the nearest neighbour in the feature space) belongs to a certain category, then the sample also belongs to this category. In k-NN classification, the majority voting method has been used, which is the K samples closest to the sample features in the training set and the predicted sample are predicted to have the most categories of categories.

## **2.3.2 Deep Learning Models**

Deep Learning also called deep structured learning or hierarchical learning, and it is part of the machine learning method. Deep learning is based on the learning data representations, which is consist of the supervised, semi-supervised and unsupervised [98] [99] [100].

Deep learning have many different networks for various function studies. Deep learning models are vaguely inspired by information processing and communication patterns in biological nervous systems yet have various differences from the structural and functional properties of biological brains, which make them incompatible with neuroscience evidences [101] [102].

### **Deep Neural Networks (DNNs)**

Since the ImageNet classification challenge inspired many advance methods, Deep Neural Networks (DNNs) has been used in computer vision. In 2012, AlexNet has been proved

successful using DNN model. Due to this model achieved best the performance, the network architecture develop to deep and complexity. However, the useful part has been questioned. Canziani [10] et al. proposed a comprehensive analysis, which indicated: firstly, the power consumption is not affected by batch size and architecture. Secondly, the relationship between accuracy and inference time are in a hyperbolic. Thirdly, the upper bound on the maximum achievable accuracy and model complexity are based on the energy constraint. Then, the number of operations is a reliable estimate of the inference time. Figure. 2.18. provides a series of information, that benefit the use and design DNN model.

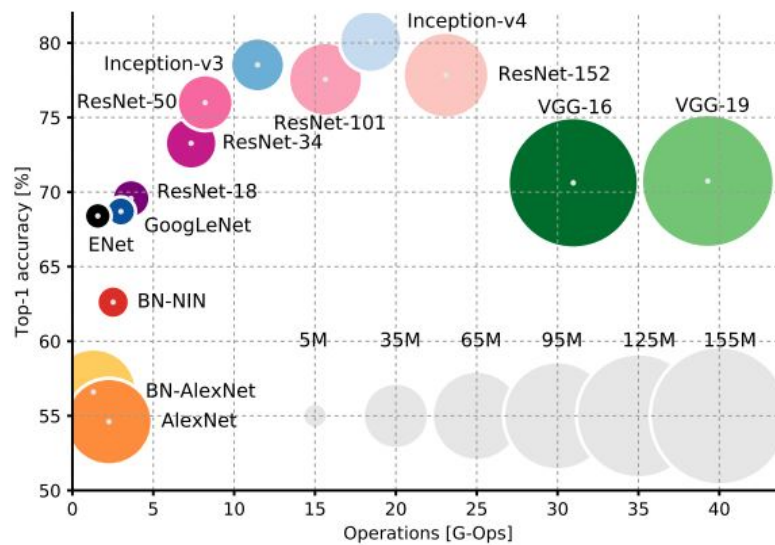


Figure 2.18: Top1 vs. operation, size and parameters [10].

### Convolution Neural Networks CNNs(CNN)

CNNs were inspired by biological processes, which are connection by neurons. Hence, multilayer is a significant characters, multilayer through each neuron to connected different layer and most networks are fully connected.

In recent years, CNN played a very important role in computer vision. The application of CNN enable to use anything image related works such as gesture recognition, object detection and face recognition. Laviola et al. [103] and Trindade et al. [104] have defined a number of joint descriptors that represent hand states in order to learn these descriptors. Then different vectors are used in this method [103]. Elmezain et al. [105] also used feature vector, which is consists of the geometric and physical characteristics of the hand motion. Reddy et al. [106] used a popular method is that uses local histogram feature descriptors. Many machine learning algorithm have been used to learning those feature

vectors such as SVM, HMM, and artificial neural networks [107].

CNNs are not only shown to be very powerful for solving image-based tasks. 2D CNN was popular using for gesture recognition. With more research proposed the issue of the 2D CNN is desirable to capture the motion in the video. And most methods are based on the conventional paradigm of pattern recognition. The Figure. 2.19 shows the structure of fully connect of multilayer perceptions.

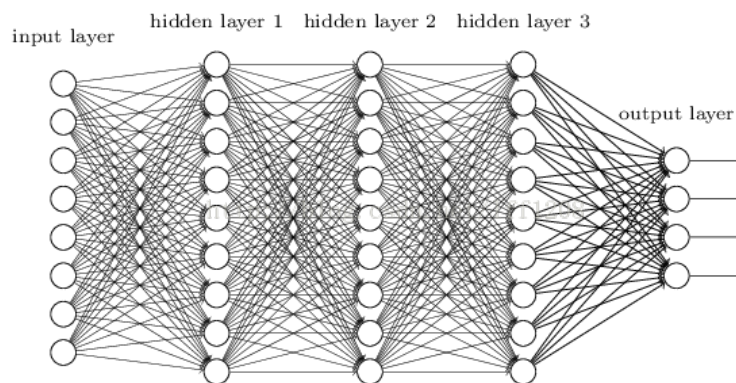


Figure 2.19: Multilayer perceptions and fully connection.

### 2.3.3 Decision Fusion

In the early step, seeking a second or third opinion before making a decision that commonly using for financial, medical and so on. This process obtains the final result through the weighing individual opinions, combing all the thoughts and reaching most informed one. However, the benefits of automated making decision applications were used to computational intelligence community in recent years. Many used this method systems have been produced better results comparing the single expert system, such as mixture of experts, multiple classifier systems and so on. These systems are hot topic, which work to current and future research directions for novel applications, Such applications data fusion, feature selection, confidence estimation, learning with missing features, and error correcting output code.

In this research, the application focus on the data fusion, which consists of the feature, classifier, and decision level. Sharma et al. [108]. proposed using individual decisions to solve feature vector of multiple-modal sensor's data in 1998. The classifier used to solve the recognition of complex movements [109]. In 2006, Polikar et al. [110] introduced classifier fusion.

Decision fusion called decision level fusion, which is part of the distributed detection systems[111] and it used for a broader area, especially having significant performance in classification. Generally the decision fusion combines the decisions of multiple classifiers in order to a common decision[112].

**Mixture of Experts (ME)** Mixture of Experts(ME) is popular use for classification, regression, and fusion application in recognition, healthcare, surveillance, and finance. The mixture of experts (ME) model has been combined many different classification and regression models such as hidden Markov models (HMMs), support vector machines (SVMs), Gaussian processes (GPs) to improved performance[113].

## 2.4 Summary

Chapter 2 is reviewed latest technologies of gestures. Although many companies have been introduced different sensors for micro-gesture interaction, there problems have many limitations such as immature technology, low recognition rate, etc. Holoscopic 3D imaging system mimic fly's eye, which uses coherent replication of light to construct a true 3D scene in space. Then, it offers high resolution full colour continuous video with 3D depth. Considering the unique advantages of H3D imaging system, in this study it was decided to use H3D as a sensor to implement micro-gesture interaction. Although H3D technology has been more mature research and application in the field of 3D display, it is unprecedented to use as a sensor to record high-precision micro-gesture. Because of it has great potential and guarantee, the use of H3D technology to implement micro-gesture interaction is a subject that should be developed for extensive research. Since this is a completely new study, in addition to the H3D imaging system technology and gesture interaction, a guaranteed database as the basis for the research is required. Therefore, designing and creating a micro-gesture database based on H3D imaging system has become the primary task of this research.

## Chapter 3

# H3D Micro-gesture Database Creation and Validation

### 3.1 Micro-gesture Interaction

However, with the development of the gaming interaction and wearable device, precise finger gesture has more advantages than body gesture, especially for control devices[114]. The finger movement is one of the micro-gestures that can accurately manipulate the device. However, the traditional devices are matched by the 2D touchscreen gesture not enable to working for a new trend. Since the Kinect and RGB-D camera were popular sensors for gaming in the Augmented Reality (AR) and Virtual Reality (VR) community with its low-cost been a major advantage, as well as its immersive user experience and usability [54]. Due to the people prefer to use more natural and conventional 3D gesture system, which mean the sensors enable to supported more free space flexible interaction [115]. However, these systems lack the ability to capture quality and accurate objects which could be seen as one of its major drawbacks [116].

Recently, some new research from Leap motion [117] and Google Soli project [118] created new techniques for 3D detection that has huge potentials for success. Holoscopic 3D (H3D) imaging system is a novel potential technique which can satisfy the higher demand for user interactive experience. Detection of precision 3D micro-gesture can make use of the wide view coverage of the Holoscopic 3D camera to capture accurate finger movement [13]. This addressed the significant issue of using additional gloves and rings

to improve accuracy, as well as contributed to the more perceptual interaction experience [119]. H3D system supports the RGB high quality dynamic and static data, and it is renowned for high accuracy and true 3D to excellence than traditional 3D capture devices. H3D system has been very successful in 3DTV and display area. However, this technology has not been used widely for capturing and recognition of the finger gesture. This chapter aims to use the H3D imaging system to create an unique 3D micro-gesture database, further to promote the gesture recognition.

HCI appeared early in 1983 [37], which use multiple modalities such as voice, gestures (e.g. body, hand, arm, finger). For example, Siri [120] is a very popular voice-based interface. However, the natural gesture is another way to interact with the computer. The trend of the HCI is user experience of intuitionistic and effective [115]. The gesture is a touchless, non-intrusive method for HCI, and it is represented as the diverse type of the gestures [121]. Manipulative type of gesture appears the most popular one from the previous literature. The aim is to control entity being manipulated through the actual movements of the gesturing hand and arm [38]. Hand as a direct input device is more and more popular, as one of the outstanding interaction methods.

The Kinect and RGB-D camera were very popular in recent years due to the benefits of Kinect and RGB-D camera that have low cost and wide availability as a sensor [42] to capture gestures. However, RGB-D camera suffers from the underside artifacts such as the edge inaccuracies, low object remission [116]. The Kinect sensor offers the information of the depth measurement and creates coordinates of the 3D objects. Although the abundant development toolkits can support the human body recognition, the weakness is its lacking ability to capture the flexible and robust mechanism to perform high-level gesture [122].

Leap Motion (LM) [123] is a device that can be used to detect the hand and finger dynamic movements through its API software. The API has the robust pre-processing function which can reduce the complexity of the user control. However, LM is a monocular video sensor which is a challenging for capturing the abundant dynamic hand gestures and finger micro movements [124].

## **3.2 Holographic 3D Imaging for Micro-gesture**

Holographic 3D camera is a single aperture sensor not only to represent the real-time and represents a true volume spatial optical model of the object scene but also to record the

viewing natural continuous parallax 3D objects within a wide viewing zone [125]. It provides a new way to capture micro-gestures.

Recently, there are numerous 2D and 3D gesture or body movement datasets, which have supported body gesture research. For example, the Cambridge Hand Gesture data set [126] is consist of 900 image sequence of 9 gesture classes. The Imperial College London [127] had success on Hands In the Million Challenge (HIM2017), they proposed 10 methods on three tasks of single frame 3D pose estimation, 3D hand tracking, and hand pose estimation during object interaction. Isaac et al. [115] presents a review summary of 21 gesture datasets from previous research and public datasets, in which 7 databases are for hands. ChaLearn[128] gesture database proposed in 2012, which contains 50,000 hand and arm gestures recorded by RGB and depth camera. Most gestures are signals used by divers, referees, marshalling, and so on. This dataset has been used a challenge for training a large number of sub-tasks. Most datasets are recorded using to the Kinect or RGB-D camera as the sensor.

In order to the support the diversification of the gesture recognition and encourage the development of the human computer interaction, we propose a new 3D gesture database included the three ubiquitous micro-gestures that are most the popular ones used in the Google Soli project. Those are intuitive and unobtrusive manipulative gesture. This database does not only include the continuous dynamic data but also contained the abundant static data to support the 3D micro-gesture recognition.

## **3.3 H3D Dataset Preparation and Creation**

### **3.3.1 Micro-gesture Design**

Micro-gesture is part of the hand gesture motivation, a notable question is the movements only between the fingers, it the highlight different between the macro-gesture and micro-gesture, which is a type of accessible design. The micro-gestures is consist of 2D and 3D micro-gestures, which use to the different type devices. The 2D micro-gestures are used for the touchscreen to assist the interface to manipulate. The Figure. 3.1 shown 2D micro-gestures, the movement tracks are direct touch gestures based on the interface, which have the issues of high visual attention and soft button of low perceived power [47].

The 3D micro-gestures are combined the 2D and 3D gesture's functions to summarised the control type gestures. The movement track shown Figure. 3.2. The movements are

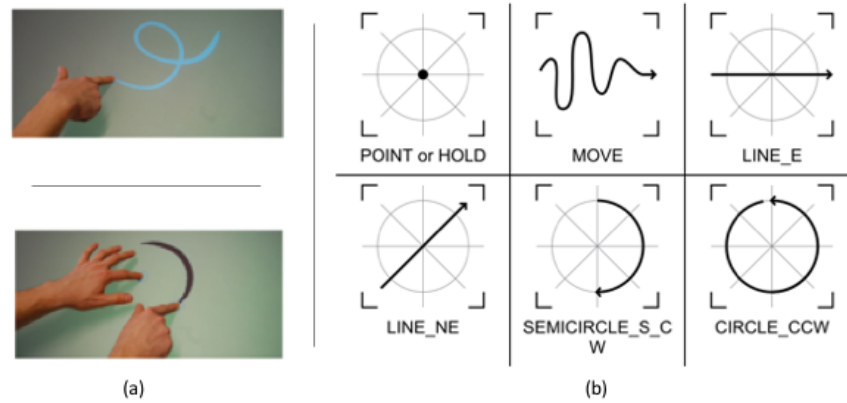


Figure 3.1: (a) The touchscreen micro-gestures, (b) The 2D micro-gestures movement tracking [11].

no longer restricted for direction and space, as well as the gesture manipulations are more flexible, which satisfying the non-barrier and adaptive design thinking [129] [130].

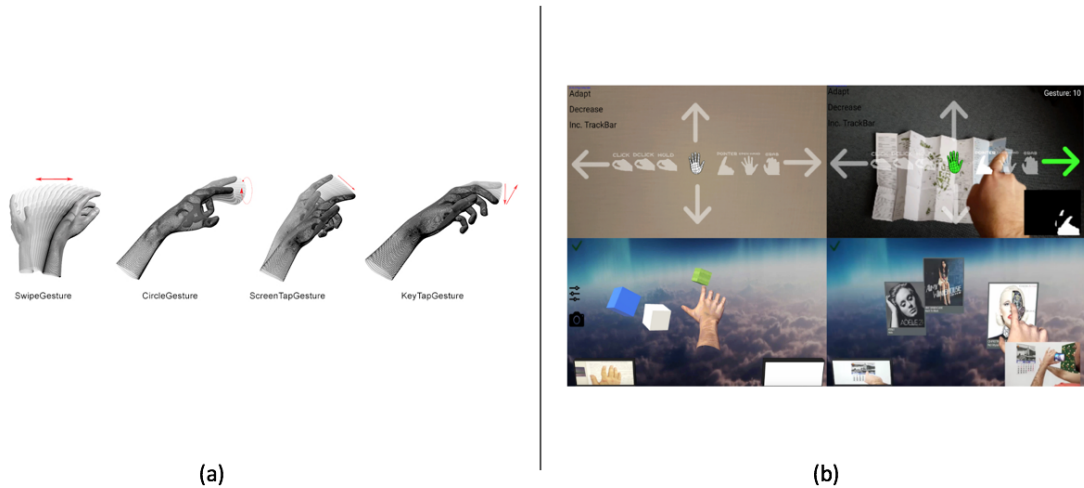


Figure 3.2: (a) The 3D gesture movement tracks [2], (b) The 3D micro-gesture in an application [12].

Meanwhile, the gestural interaction following the increasing 3D displays to rapidly developing. Immersive 3D user experience most uses to the gaming environment, so freehand interaction and no hands-on input is mainstream[42]. Although freehand gestures bring convenient like no need hold and touchscreen, lost some advantages like clicking buttons and tapping surfaces. Nevertheless, some researchers use to simulate touch gesture



to satisfy the physical interactions feel and responsive. With the developing of research is provided more natural and based human-centred method of interacting with the computer. Matching perceptual user interfaces of 3D micro-gesture are the product of new interaction, which supports the perceptual interaction, as same time affect the implicit and explicit information between the user and the environment[43].

In this research, the traditional 2D touch screen gesture and 3D gesture movements all have been considered. In order to follow the user's hobby and basic function for manipulation, we tried select to some control gestures that have the character of 2D and 3D, which is enabled to feel the control response. The micro-gesture movements design inspiration shown in Figure.3.3.

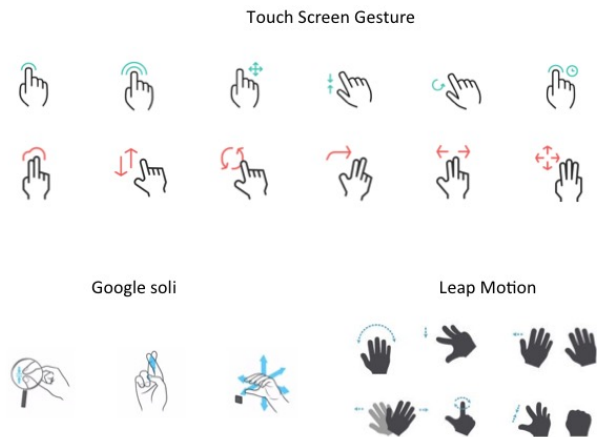


Figure 3.3: Comparison of manipulate gestures.

There are many micro-gestures that can be used for control in AR and VR applications. In this research, three intuitive micro-gestures are selected references to the Google Soli project as shown in Figure.3.4. The three gestures are based on the human intuitiveness when they try to control display. For instance, the button gesture executes the submission function, dial gesture shows that user wants to slightly adjust the current situation, and the slider gesture is to express the slide up or down to adjust the volume and options. This three gesture belong to the manipulative type of the gesture, which are used to touch-less control the devices or simulation console. In this research, we focus on mid-air gestures only between the fingers. Since the fingers are more flexible part of the body, the finger gestures are kaleidoscopic. For holoscopic 3D gesture design, I prefer considering the barrier-free design, which means the gesture movements as much as been decrease. To matching control devices, this research will be an initial design exploration, the ubiquitous

and conventional gestures which are enabled to use for wearable devices such as watch and glasses. Therefore, the three spontaneous control micro-gestures have been selected use to the HoMG dataset first time to shoot.

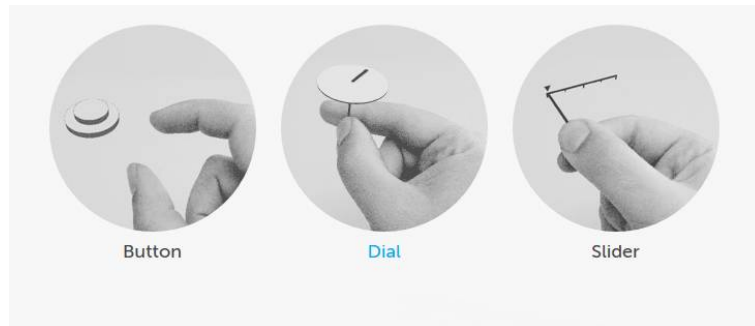


Figure 3.4: Google Soli project gesture [1].

In recent years, with the rapid development of new sensors, some researchers and companies have established related gesture databases for research. Figure. 3.1 shown four gesture database details. Some of them receive special attention, such as DVS128 of IBM [131] and Cambridge GestureDataset [132]. Table. 3.1 show four related gesture database. Most datasets are based on Kinect or Leap Motion to recorded the database, gestures are movements that involve hands and arms. DVS 128 have 11 type gestures with 29 subject [131] [133] created an infrared image based 10 type gestures with 10 subjects, which recorded by Leap Motion. These gesture database establishments have promoted the development of more gesture recognition algorithms, but since most of the gestures in these databases are macro-gestures, they are different from the application direction of this study. In addition, different database sensors have some disadvantages and shortcomings, which are difficult to fully refer and judge.

Table 3.1: The summary recording length of videos.

Database	Gesture Type	No. Subject	Sensor
DVS128[131]	11	29	Dynamic Vision Sensor
Cambridge-gesture database[132]	9	2	Camera
I.C.V.L./Hand[134]	25	180	6D sensors
Infrared Based [133]	10	10	Leap Motion

### 3.3.2 Design Process

The three types of gestures have been chosen: Button, Dial and Slider. Those three types gestures are ubiquitous, perceptual and conventional, the gesture design concept is easy

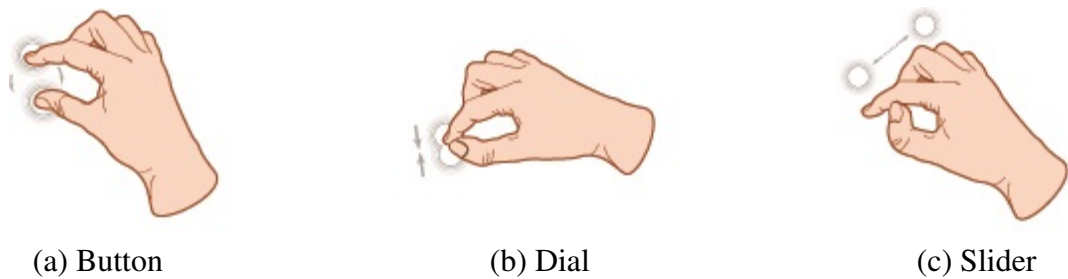


Figure 3.5: Three key types of finger 3D micro-gestures studied in HoMG database.

learning and underling the manipulation.

Fig.3.6 shows the gesture design process: Firstly, the investigate the exciting gesture types and feature for H3D micro-gesture. Secondly, the ethic course has been required in Brunel academic research. Thirdly, creating the H3D micro-gesture movements, that are verified in the research group and gained the standard for acquiring movements. Then, a series of forms and tutorial works have been finished, such as participants information sheets, tutorial video etc. I recruited 50 participants are join this database recording. The required data have been checked quality and random the order for promising the equity. The database has been uploaded to Google Drive for sharing the other researcher.

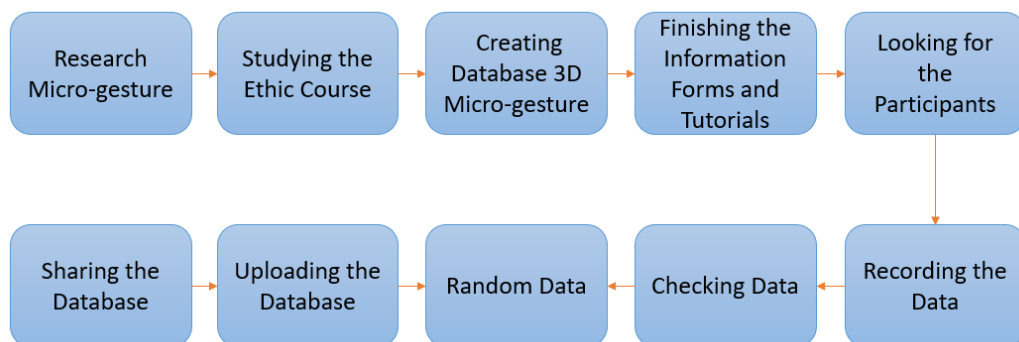


Figure 3.6: The process of gesture design process.

### 3.3.3 H3D Imaging System

H3D imaging technology is a success for use in the 3D cinema, 3D-capable televisions and broadcasters. The H3D camera used here is built from the 3D Vivant Project (3D Live Immerse Video-Audio Interactive Multimedia) [8] and the purpose is to capture high quality 3D images. The developed camera includes micro-lens array, relay lens, and digital camera sensors. The 3D holoscopic image's spatial sampling is determined by the

number of lenses. It shows that the captured 2D lenslet array views are slight different angle than its neighbour and reconstructed image in relay [8]. The detailed parameters of the camera are shown in Figure. 3.2.

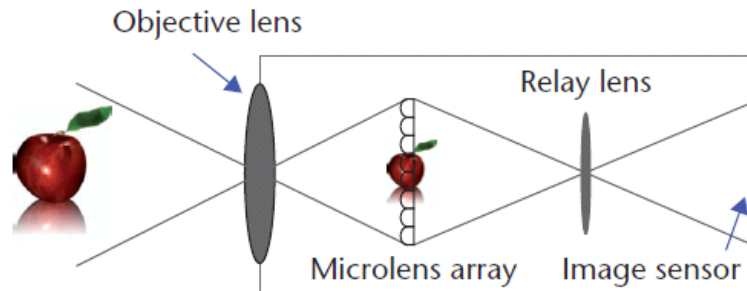


Figure 3.7: Principle of the holoscopic 3D camera. The microlens array is placed between objective and relay lens to produce fly eye style images [8].

The holoscopic 3D camera sensor has unique optical components which support the continuous parallax RGB image system, and contain the depth information viewpoint images. The Figure shows the H3D imaging having the full colour with full parallax. H3D imaging is comprised of the 2D array of micro images.

The H3D sensor is a crucial requirement for the capture of the objects. This database uses the H3D imaging system to support the dynamic and static RGB data. And it's not only can record the continuous motion, but repetitive lens array can extract different angles viewpoint images. The uniqueness to encourage the innovation of gesture capture and recognition.

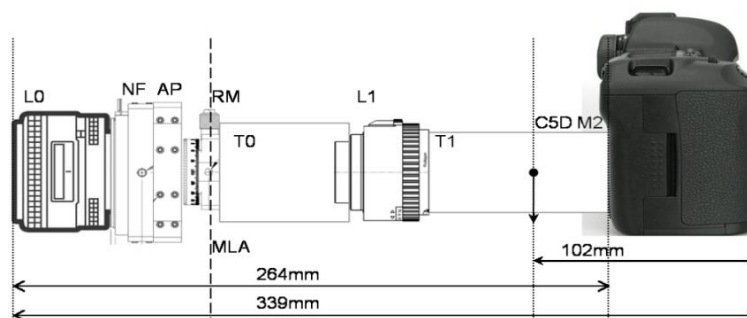


Figure 3.8: Assembled holoscopic 3D camera[13].



Figure 3.9: H3D Data acquisition configuration.

### 3.3.4 Recording Setup

The recording H3D gesture place, as in Figure. 3.9 a green screen room is used for the recording where it can offer clear and professional recording background to reduce noise. Four locations are used to capture the micro-gesture movements. 1 and 2 are used for the left-hand recording while 3 and 4 are used for right-hand recording. The different settings are in consideration of different control behaviours of a left-hander and a right-hander. In order to enrich the data diversity and test the various kinds of angles, there are set up two distances capture viewing angle. Due to the gesture movement on air are not stable. In the closed position, we set a hollow frame to help the participant to find the 3D micro-gesture capture zone shown in Figure. 3.10. Before the recording, the holoscopic 3D camera adapted and surface are set up in advance. Canon 5D camera has been used as sensor. And the camera settings are configured to ISO200, shutter 1/250. Holoscopic 3D camera adaptor is calibrated and the lens is corrected.

Considering the influence of distances, angles, and backgrounds, we prepared 4 positions for participants. In order to enrich the data diversity and test the various kinds of angles, there are set up two distances capture viewing angle. Two positions are the close and far locations where the objective lens set to 45cm and 95cm. Figure. 3.10 shown the details. The other two positions are from the left and right hand side for the convenience of the participants. It is noted that hollow frame has been set to help the participant finding the 3D micro-gesture capture zone in the close position.

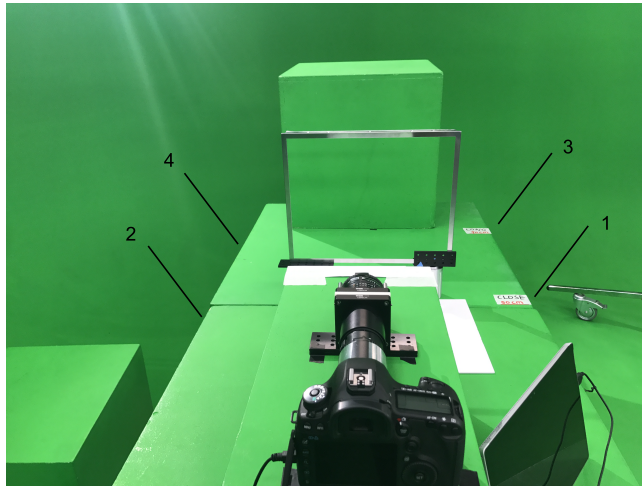


Figure 3.10: H3D Data acquisition positions,(1)(2) are close positions of 50cm for left and right hand. (c)(d) are far positions of 90cm for left and right hand record.



Figure 3.11: Debrief for participants before record data.

Before recording, we have a short introduction and tutorial for each participant, which ensures that all participant understand the aim and content of the gesture data collection. Figure. A.1 show the explanation in progress and demonstration of the three gestures before the commencement of shooting. In order to follow the human natural gesture of the concrete thinking [135]. We remind the participants the name of the gestures while recording their finger movements, which encourage participants to perform the said gesture in their own unique way based on their understanding of the function of the gesture. The participants can perform their micro-gesture at their own speed. The recording duration is around 15 minutes for each participant.

We prepared two different colour backgrounds, two different distances of close and far from the end of the camera lens to gesture area. The recorded imaging resolution is 1086 x 1902 pixels, and the microlens is 27 x 27 pixels. Participants are successively stand each pre-established position to play three gestures around 3-5 seconds. The three gestures are

involved button, dial, and slider.

Holoscopic 3D image is composed of the many regular elements images. And the function of the elements image array is a capture process. The square aperture fitted on the front of the camera lens when achieve over 95%. A standard 3D Holoscopic image satisfies the fill elements over 95%, so there is a square aperture to install the front of the camera lens. It gives rise to a regular structure in the intensity distribution by using the square aperture at the front of the camera lens and the regular structure of the square micro-lenses array in the square grid (recording micro-lens array) as shown in Figure. 3.7 below. Each block pixel pattern is called a micro-image and the planar intensity distribution representing a 3D holoscopic image consists of a 2D array of micro-images.

Table 3.2: Data acquisition details.

<b>Parameters</b>	<b>Detailed Information</b>
Micro-gesture	Button (B), Dial (D), Slider (S)
Participants	Male (33), Female (17)
Hand	Right (R), Left (L)
Distance	Close ( 45cm), Far ( 95cm)
Background	Green (G), White (W)
Camera	Canon 5D
Image resolution	1902 x 1086
Lens array	28 x 28
Shutter speed	1/250
Film speed	ISO200
Frame rate	25
Recording length	Between 2 and 20 Sec.

### 3.3.5 Participants

In total, 40 participants attended the recordings including 17 female participants and 33 male participants, who all read the participant information sheet guidance and sign the research ethics application forms before the recording. There is no any limitation of age and race for the participants and we respect the participants' will. The Figure. 3.3.5 shown the gender and age of participants, which may predict different target users' user behaviours and hobbies. Some participants wear married rings and watches during the recording on finger movements. These increase the data's noise and bring more challenges. The participants' double hands have been record, in order to increase the diversity of the data. The detailed information about the data acquisition is summarised in Table. 3.2. The ethics form is attached to the sub-file of final thesis.

Table 3.3: Participants genders and ages.

<b>Gender/Age</b>	<b>20-30</b>	<b>30-40</b>	<b>40-50</b>	<b>50-60</b>
Female	11	5	1	0
Male	15	15	2	1

### 3.3.6 HoMG Database Structure

For the data collection, the recordings from 40 participants are selected to make the HoMG database. The recordings were done under different conditions. One participant has recorded 24 videos. In total, 960 videos are included in the database.

Table 3.4: The summary recording length of videos.

<b>Partition</b>	<b>under 2 Sec</b>	<b>2 - 5 Sec</b>	<b>5 - 10 Sec</b>	<b>10 - 15 Sec</b>	<b>15 - 20 Sec</b>	<b>over 20 Sec</b>
Number	2	361	358	167	53	19

Table 3.5: The summary recording length of videos.

	<b>Training</b>	<b>Development</b>	<b>Testing</b>
Subjects	480	240	240



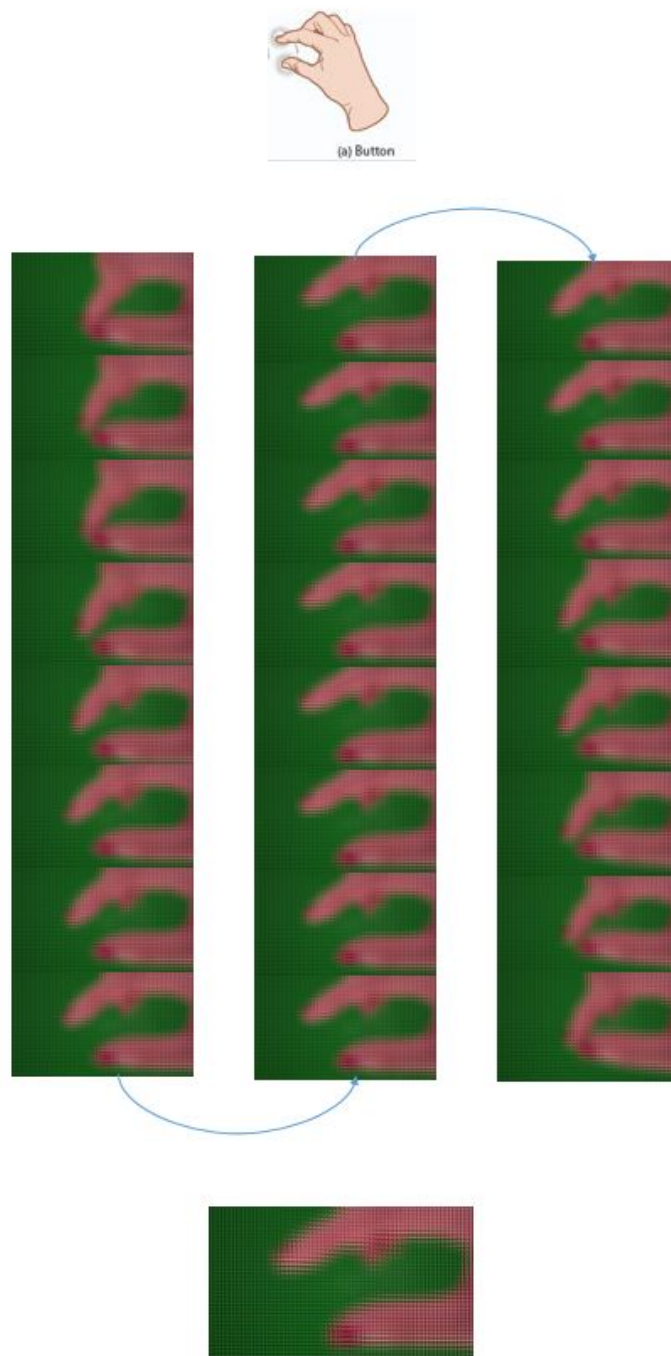


Figure 3.12: Button gesture movement tracking.

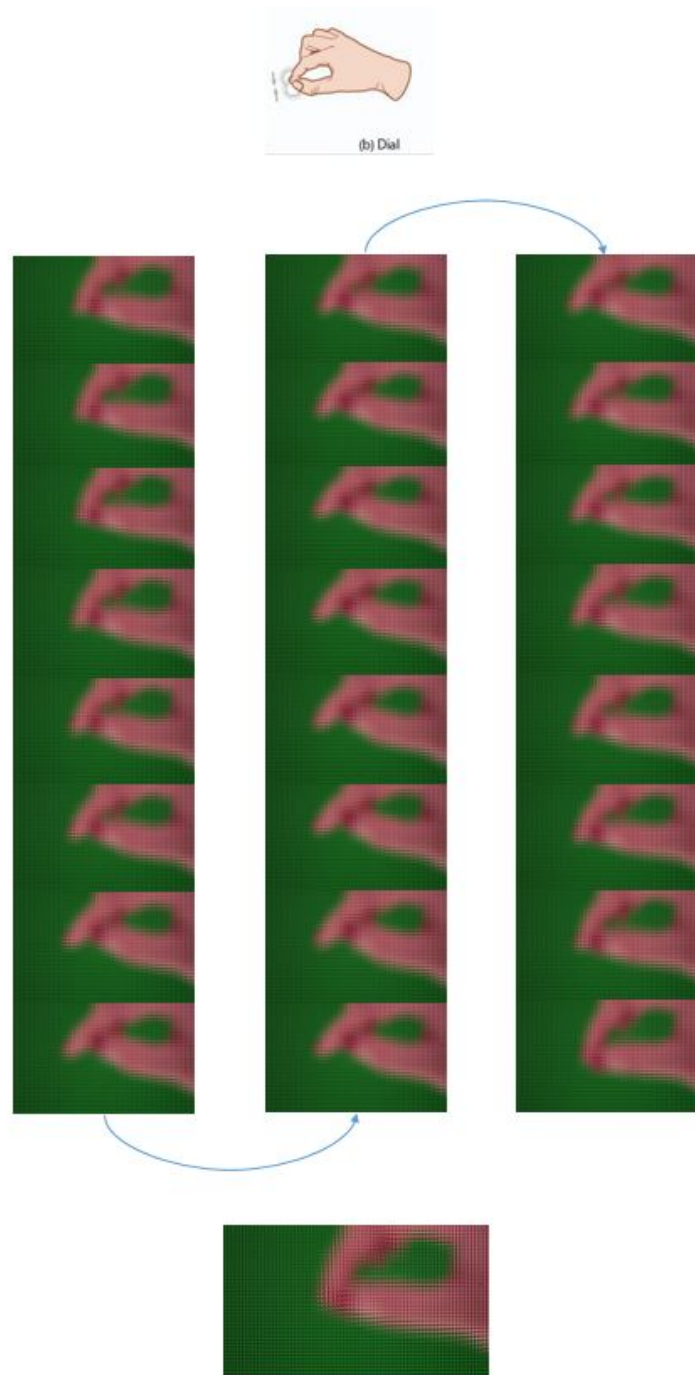


Figure 3.13: Dial gesture movement tracking.

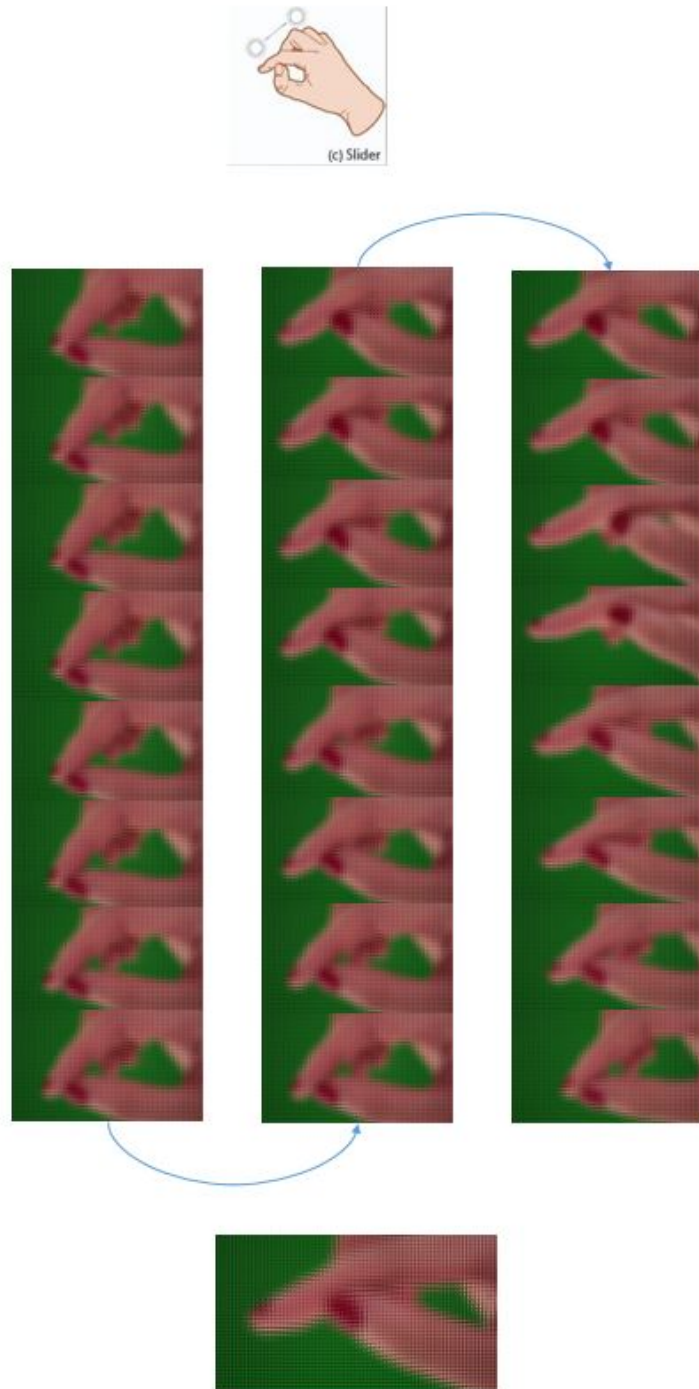


Figure 3.14: Button gesture movement tracking.

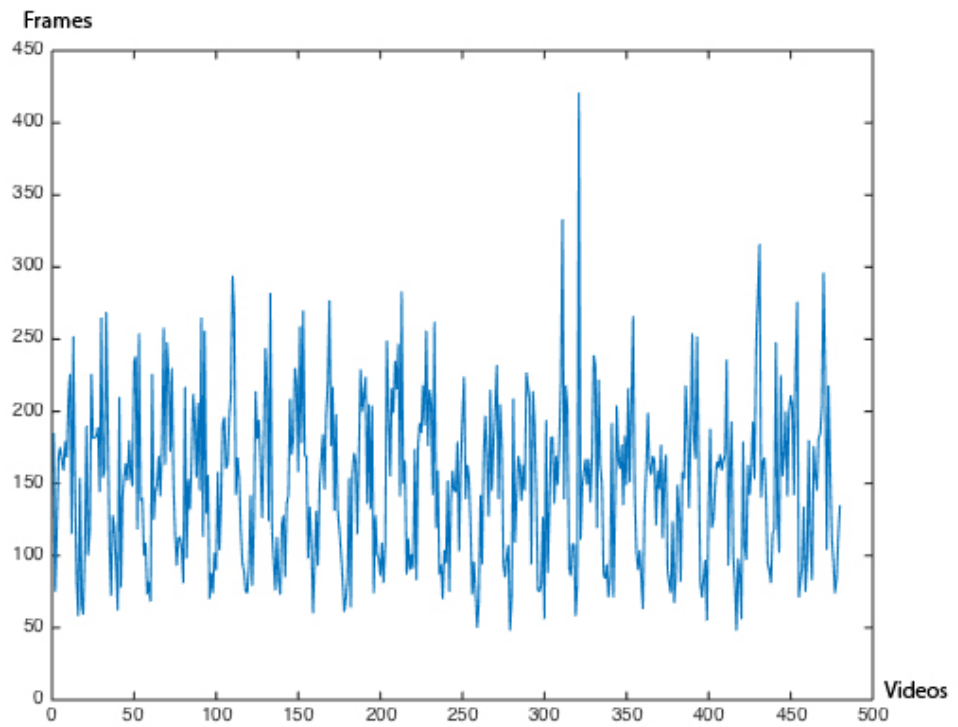


Figure 3.15: Frame numbers of each video in training set.

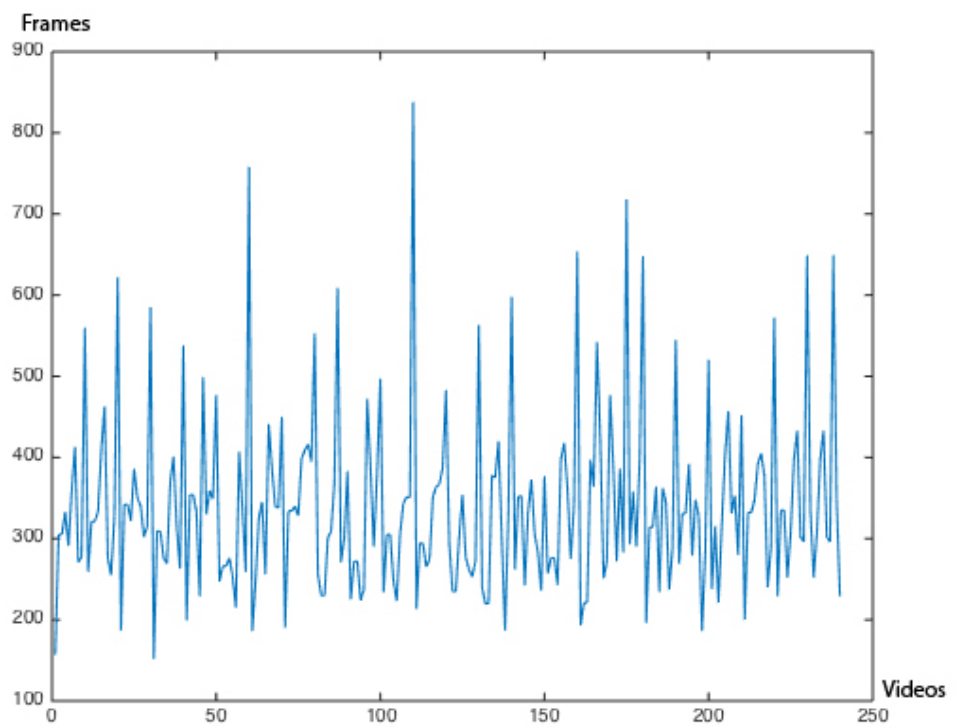


Figure 3.16: Fame numbers of each video in development set.

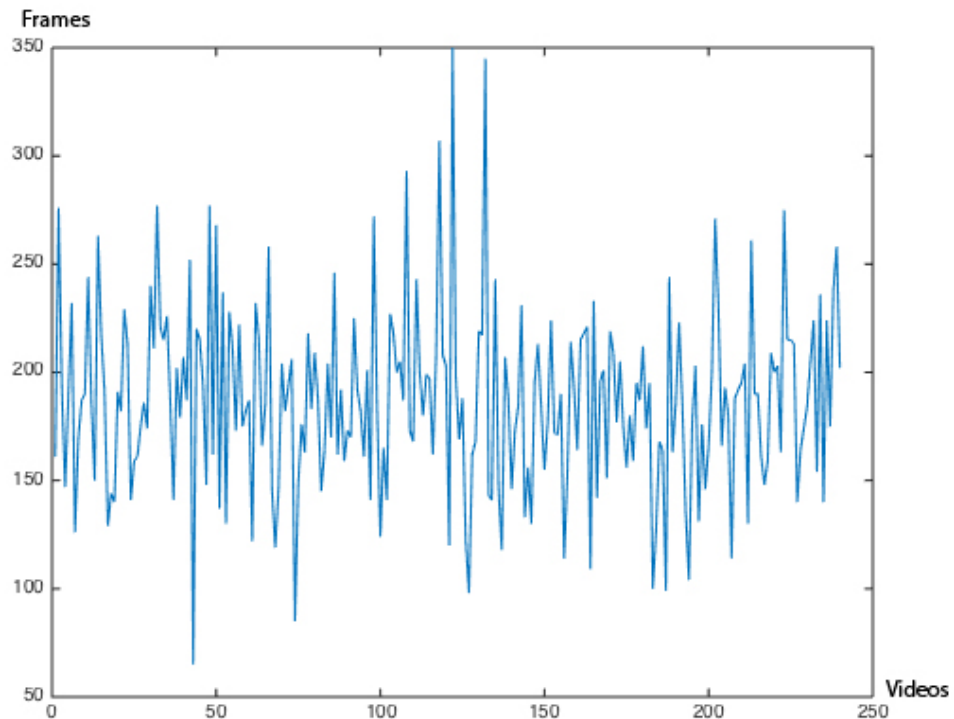


Figure 3.17: Fame numbers of each video in test set.

For micro-gesture recognition, it can be done based on single image or can be done from a short video. So this database was divided into two subsets: image based and video based micro-gesture subsets. Each subset have three classify gestures: Button, Dial, and Slider. The each gesture movements shown in Figure. 3.12, Figure. 3.13, and Figure. 3.14. The action decomposition diagrams of the three gestures fully demonstrate the trajectory of each movement. Although the gesture movements recorded by the H3D camera, and the videos are based on the fly's eye lens to shown. The video that record by close distance has been clear to recognize the finger tracks. Figure. 3.12 demonstrates the two fingers from open to close, which simulate the press the button actions. In the button videos, the user hobbies have been discovered when they want to play the button gestures. There are two types of finger movements: One is to move with two fingers at the same time, and the other is to use one finger as a plane to press down with the other finger. Both of these behaviours are based on the user's previous life experience and intuitive perception of manipulation, especially they feel response without the interface. The second class gesture is dial, and the movement tack can be seen in Figure. 3.13, where each frame shows the high accuracy micro-adjustment using the two fingers friction. Figure. 3.14 shows the slider gesture movements. Due to the nature of the gesture, this gesture takes longer than other gestures to complete. User will use this gesture to choose they want, which means this action is repeated. It can be seen from the three gesture decomposition diagrams that the initial actions of the three gestures are the same, which also means the

importance of data coherence. It is important to note that the two gestures, slider and dial have more coincidence points.

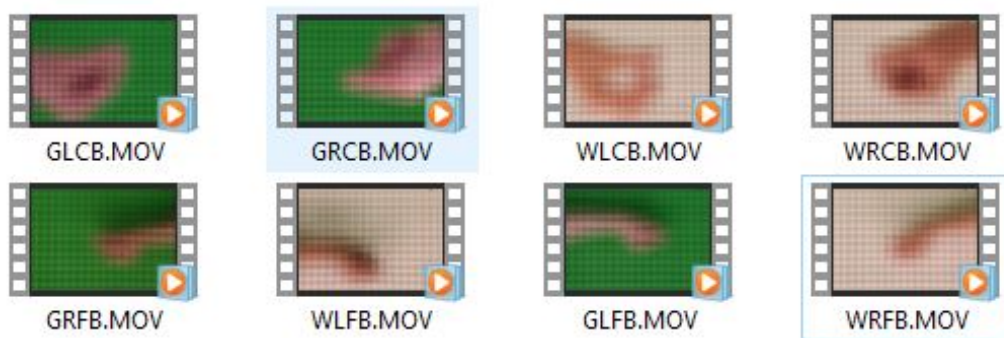


Figure 3.18: Six types of gestures on every class.

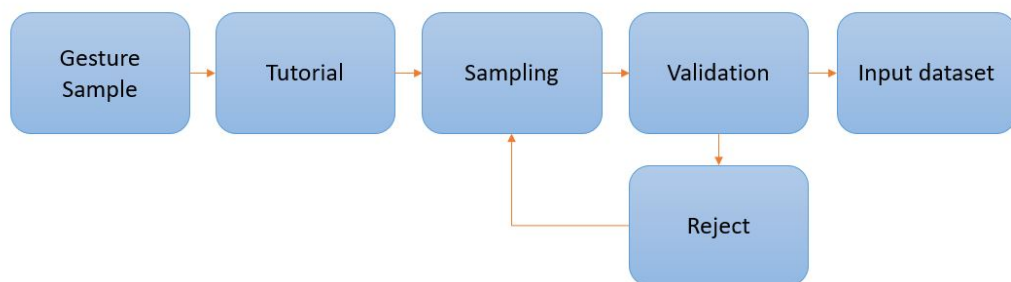


Figure 3.19: Video validation.

### H3D Video Subset

There are 40 subjects and each subject has 24 videos due to the different setting and three gestures. For each video, the frame rate is 25 frames per second and length of videos are from few seconds to 20 seconds and not equally. The data validation process show the Figure. 3.19, and the detail of each video' frame number has shown in Figure. 3.15 Figure. 3.16 Figure. 3.17. Due to all the video record by participant's gesture speed, and every video recorded each movement at least twice to make sure the movement integrity. When we record the videos are based on every participant's order, therefore the original record files are name by location and background colour. There six types of videos for each class. Which show Figure. 3.18 We used the capital of words represent the file content, for example WLFB means white background record left hand play button of gesture type on far distance. Then considering rich the data and made it have multiple categories, the data have been reassignment based on every participant's movements. Firstly, the

whole dataset was divided into 3 parts. 20 subjects for the training set, 10 subjects for development set and another 10 subjects for testing set. In this way, the micro-gesture recognition is person independent. Then, we randomized the video subset of all the subjects, which means the even distribution of different types of videos in the training set, development set, and testing set. Figure. 3.20 show the training set details. Figure. 3.15 shows the each video's frame.

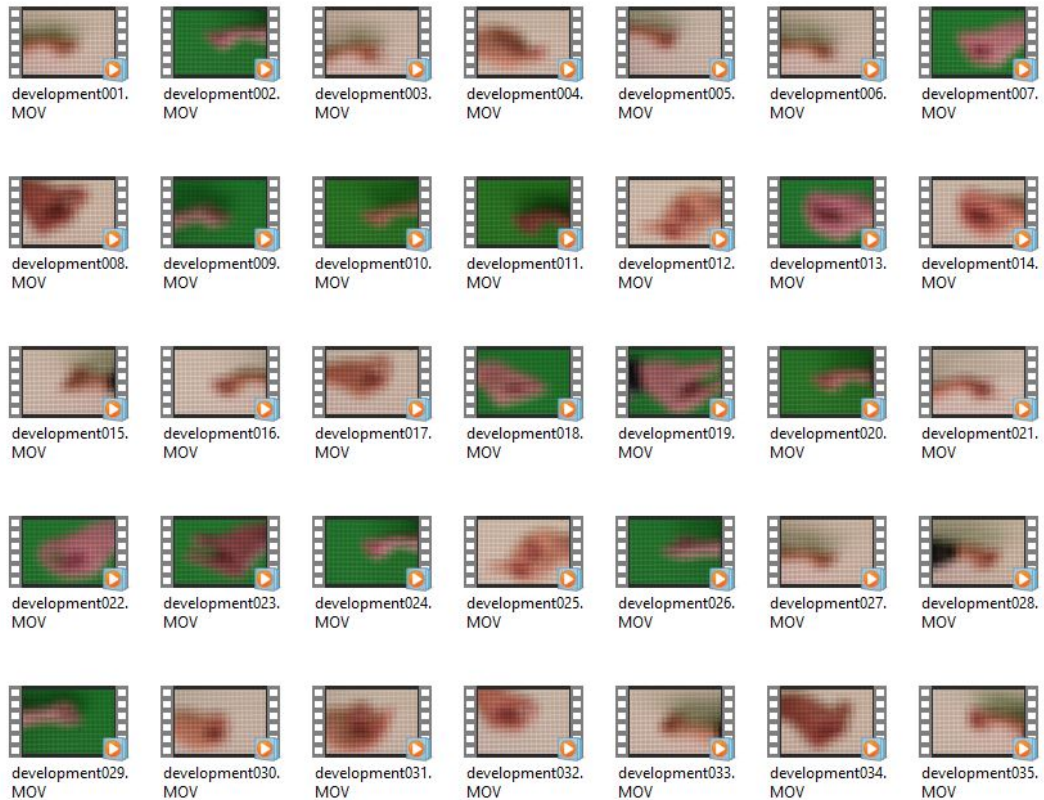


Figure 3.20: Video-based subset data overview.

### H3D Image Subset

Video can capture the motion information of the micro-gesture and it is a good way for micro-gesture recognition. However, it needs more data and takes a long time. It is very interesting to see whether it is possible to recognize the micro-gesture from a single image with high accuracy. Hence, the image-based database has been created, which extract and select from video data. The Figure. 3.21 and Figure. 3.22 shown the image-based subset details. For each gesture movement file is extracted from each gesture videos and we selected every one frame from every 6 frames, which possible keep the movement continuity.

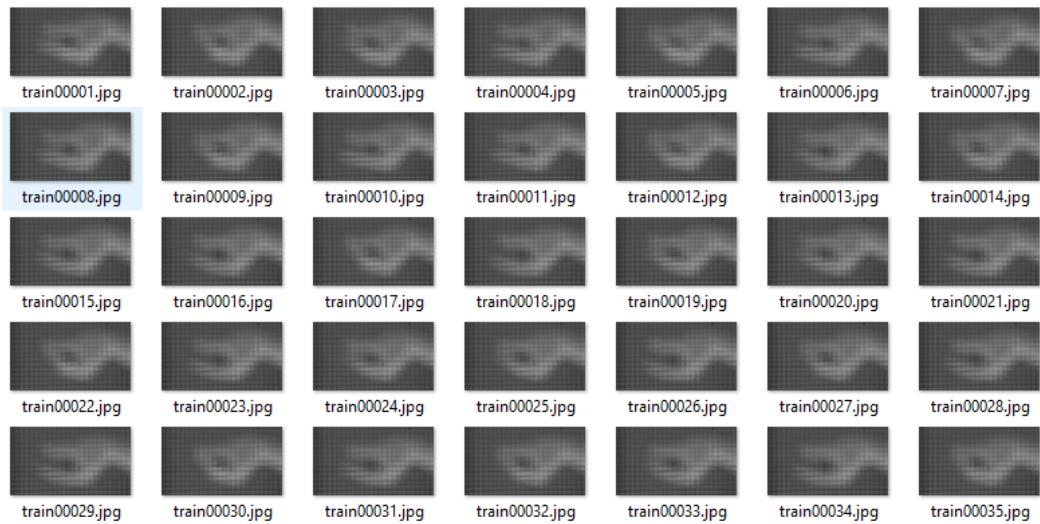


Figure 3.21: Green colour background data of image-based subset overview.

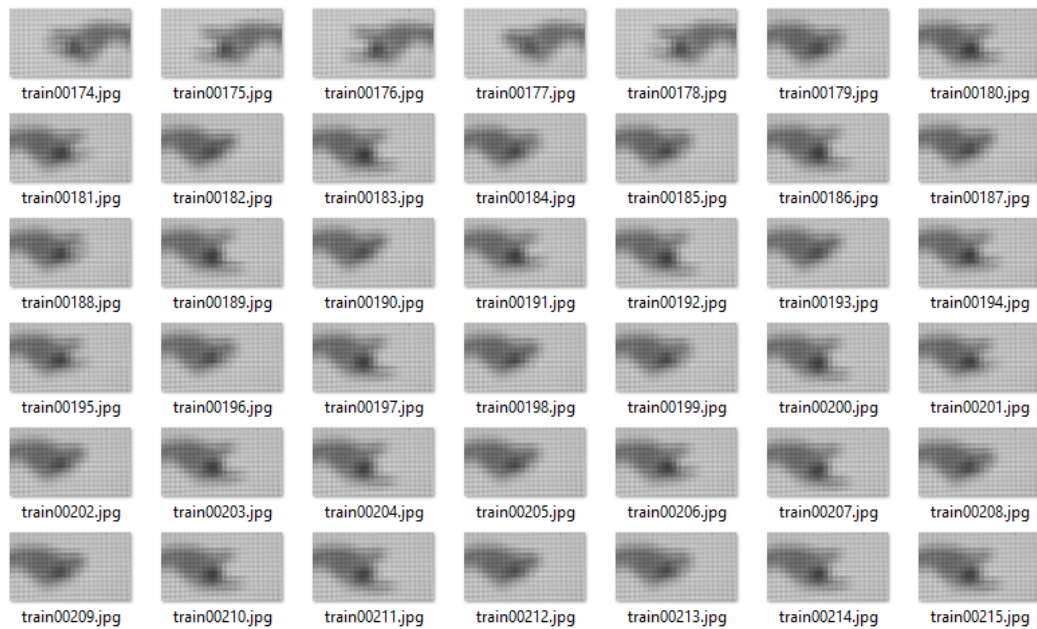


Figure 3.22: White colour background data of image-based subset overview.

From each video recording, the different number of frames were selected as the still micro-gesture images. In total, there are 30635 images selected. The whole dataset was split into three partitions: Training, Development, and Testing partition. There are 15237 images in the training subsets of 20 participants with 8364 in close distance and 6853 in the far distance. There are 6956 images in the development subsets of 10 participants with 3077 in close distance and 3879 in far distance. There are 8442 images in the testing subsets of 10 participants with 3930 in close distance and 4512 in far distance.

The summary of the HoMG database is listed in the Table 3.6.



Table 3.6: The summary of the HoMG database.

<b>Partition</b>	<b>Subjects</b>	<b>Image Set</b>	<b>Video Set</b>
Training	20	16763	480
Development	10	6560	240
Testing	10	7291	240

## **3.4 Micro-Gesture Recognition Validation**

The initial investigation is carried out independently based micro-gesture recognition study from video and image separately. We would like to see how high performance can be achieved from each.

### **3.4.1 H3D Video Subset for Micro-gesture Recognition**

There are many good features that can be extracted from each video to capture the movement of the fingers. Here LBPTOP [136] and LPQTOP [137] are selected. These features can not only calculate the distribution of the local information of each frame, but also the distribution of finger movements along to the time. From each video, the frame size was reduced to 66 x 38 pixels from 1920 x 1080 pixels. Firstly, then a feature vector with the dimension of 768 is extracted using LBPTOP and LPQTOP for the classification. For the classification, it is a three class classification problem. There are lots of classifiers available. Here, most popular ones such as k-NN, Support Vector Machines (SVM) and Naive Bayes classifiers are chosen for comparison purpose. SVM, KNN and Bayes are based on the MATLAB toolbox/library.

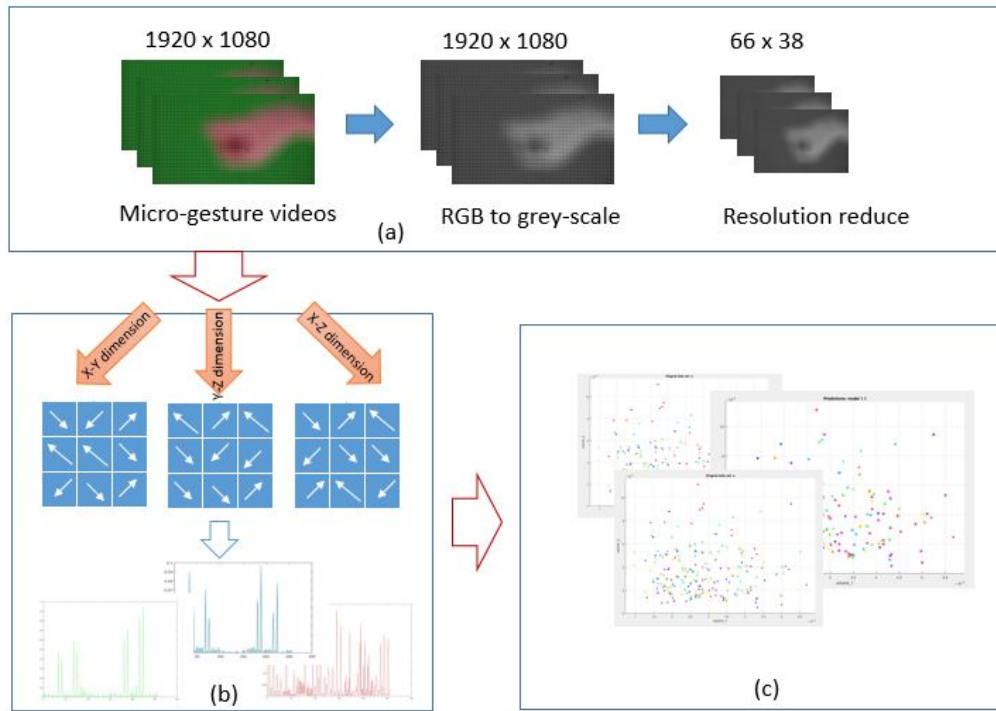


Figure 3.23: Video-based subset micro-gesture recognition process. (a) shows the video data have been resized, (b) use LBP-TOP and LPQ-TOP to extract the feature, (c) use the algorithms of k-NN, Support Vector Machines (SVM) and Naive Bayes classifiers to obtain the final results.

Table 3.7 shows the accuracy using three different classifiers under different distance on video based micro-gesture recognition. From this table, it can be seen that LPQTOP is better than LBPTOP for feature extraction. SVM is better than k-NN and Naive Bayes classifiers in most cases. In general, the accuracy on close distance is better than far distance because the detailed information of the finger movement can be captured. For the testing set, both training and development sets were used for training. Overall, 66.7% accuracy can be achieved even use the feature extraction methods from all videos in the testing set.

Table 3.7: Recognition accuracy (%) of video based micro-gesture recognition on development (Dev.) and testing sets using k-NN, SVM and Naive Bayes classifiers.

Dataset	Distance	Feature	Classifier		
			<i>k-NN</i>	<i>SVM</i>	<i>Bayes</i>
Dev.	Close	LBPTOP	53.3	68.3	52.5
		LPQTOP	56.7	66.7	63.3
	Far	LBPTOP	40.8	53.3	47.5
		LPQTOP	50.8	55.8	49.2
	All	LBPTOP	44.5	52.9	47.9
		LPQTOP	47.9	<b>60.4</b>	51.3
Test	Close	LBPTOP	56.7	53.3	40.8
		LPQTOP	67.5	73.3	65.8
	Far	LBPTOP	55	55	50.8
		LPQTOP	51.7	65.8	58.3
	All	LBPTOP	53.3	59.5	45.4
		LPQTOP	60.4	<b>66.7</b>	57.5

### 3.4.2 H3D Image Subset for Micro-gesture Recognition

For each image, 2D texture features such LBP [138] and LPQ [139] were extracted to represent each image. These two features captured the edge and local information of the 2D image in different ways and form a histogram feature vector with the dimension of 256. Popular classification methods such as k-NN, SVM and Naive Bayes classifiers were used for recognising the three different micro-gestures. The whole process can be seen in Figure. 3.24, (a) Image pre-processing, (b) feature extraction by LBP and LPQ, (c) is classification.

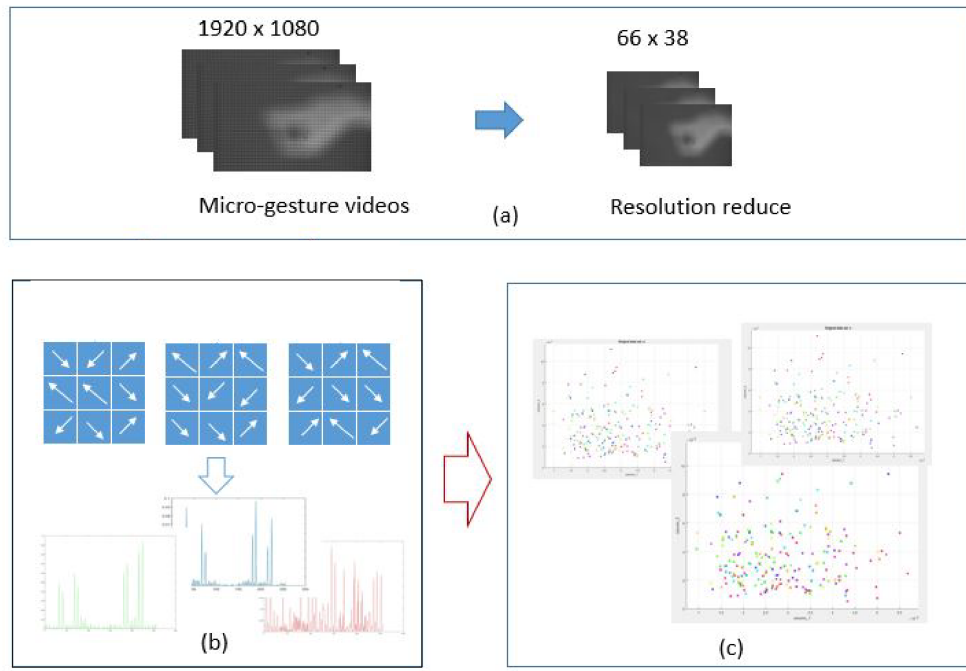


Figure 3.24: Image-based subset micro-gesture recognition process. (a) shows the image data is resized, (b) use LBP and LPQ to extract the feature, (c) use the algorithms of k-NN, SVM and Naive Bayes classifiers to classify

Table. 3.7 should the experimental results on video based micro-gesture recognition by training on the training set and tested on the development, and testing subsets. From this table, it can be seen that for most of the classifications, around 50% accuracy can be achieved.

Table 3.8: Recognition accuracy (%) of image based micro-gesture recognition on development (Dev.) and testing sets using k-NN, SVM and Naive Bayes classifiers under different distance conditions.

Dataset	Distance	Feature	Classifier		
			<i>k-NN</i>	<i>SVM</i>	<i>Bayes</i>
Dev.	Close	LBP	40.9	44.3	46.0
		LPQ	43.4	45.0	42.8
	Far	LBP	35.9	32.1	37.4
		LPQ	36.7	52.6	47.5
	All	LBP	41.0	35.0	39.6
		LPQ	32.9	<b>51.6</b>	50.6
Test	Close	LBP	49.7	33.6	45.4
		LPQ	44.1	46.4	39.7
	Far	LBP	50.9	37.7	47.2
		LPQ	34.6	51.6	50.0
	All	LBP	44.7	48.9	44.7
		LPQ	46.8	<b>50.9</b>	46.8

### 3.5 Conclusion

This chapter introduces an unique holoscopic 3D micro-gesture database (HoMG), which is recorded under different settings and conditions from 40 participants. The data recording uses the similar the H3D system of fly viewing to capture the participants' precise finger movements. The H3D imaging system supports robust 3D depth micro lens array to capture dynamic and static information. The HoMG database has 3 unobtrusive manipulative gestures in two different backgrounds, two different distances, left and right hands. These micro-gestures can be used to control multifarious displays. This database

would speed up the research in this area.

The database is further divided into video and image subsets. Initial investigation of micro-gesture recognition is conducted. For the comparison, video based method achieved better performance as it has dynamic finger movement information in the data. However, this method needs much more data and computing time. Image based method is convenient for the user and might have more applications, especially on the portable devices. Even with the standard 2D feature extraction methods and basic classification methods, 66.7% recognition accuracy can be achieved for micro-gesture videos and over 50.9% accuracy for micro-gesture images. This baseline methods and results will give a foundation for other researchers to explore their methods.

From the initial investigation, it can be seen that the recognition accuracy can reach around 66% even just using the 2D image processing methods. For 3D image processing methods, such as extracting the different viewing point images and extract 3D information of the micro-gesture, high accuracy will be achieved. This will be our future works. In addition, more type of gestures can be added into the dataset for wide applications.

HoMG dataset has been published and it holds an international challenge in 2018. Four teams from all over the world participated in the challenge. They all presented many novel approaches to achieved tasks. Meanwhile, some particular problems appeared and wait to solve.

## Chapter 4

# CNN and Decision Fusion for

# Image-based Micro-gesture Recognition

In Chapter 3, it is designed three ubiquitous 3D micro-gesture for VR and AR interaction of manipulation gesture, then the those 3D micro-gestures have been tested capture by holoscopic camera. In holoscopic micro-gesture (HoMG) database, we recorded 50 participants(40 subjects using for database, 10 subjects using for backup) to perform the three classes of 3D micro-gestures in different angles, distances and backgrounds. HoMG database creation follows the human-centre principle to design. Therefore we have invited many participants of different races, genders, and ages to record micro-gestures. Then the HoMG database has been publicised to our website using for further research.

To inspired more researcher to work this dataset, some classical and potential machine learning methods have been used micro-gesture recognition, and the initial result have been public as the baseline. Meanwhile, we hold a international challenge in order to inspired more researchers and companies join this new research. After that, based on the feedback from the participants and the suggestions and comments on the baseline, the following feedback summary was made: first of all, deep learning is a popular trend of gesture recognition, and it can provide effective algorithms for micro-gesture recognition. Therefore, deep learning become the main algorithm reference for further research. Secondly, as the characteristics of the new database H3D imaging system are not fully utilised. Hence, using the 3D information implied in the data is another focus on the next study. Finally, an effective system for improving the recognition rate of micro-gestures is established. The above summaries are the problems, that will be a focus and tackle in this

chapter.

## 4.1 State-of-the-Art Methods

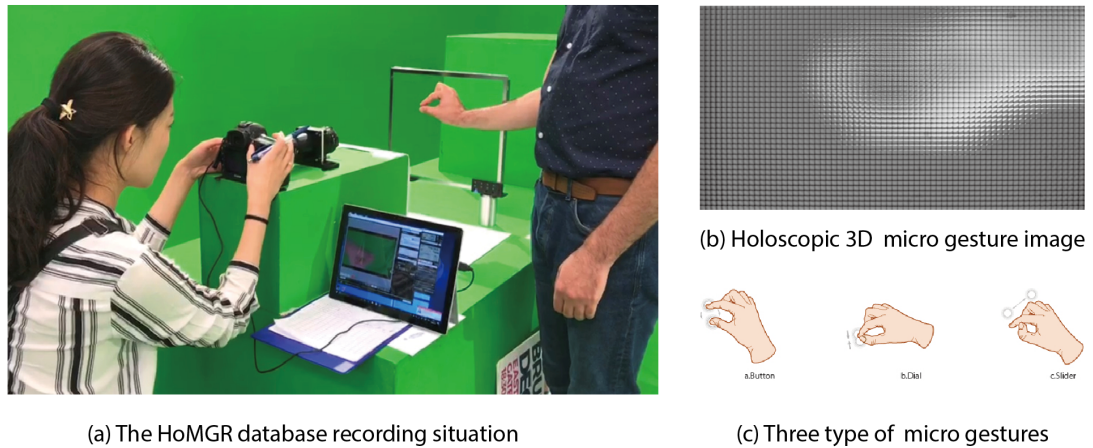


Figure 4.1: Holoscopic 3D micro-gesture image capture. (a) Recording setting, (b) Obtained H3D image, (c) Three types of micro-gestures.

Recently, the first Holoscopic Micro-Gesture (HoMG) database [140] has been created and made it publicly available. Figure 4.1 (a) shows the scene of the HoMG database recording. Four positions were set up to capture the gesture from different distances, and different sides of the hand. The obtained image is shown in Figure 4.1 (b) where 3D information was embedded inside of the H3D image. Figure 4.1 (c) presents the three gesture types: button, dial and slider used in the capture. We also organized the first Holoscopic Micro-Gesture Recognition Challenge and attracted the researchers all over the world to invest their efforts to this challenge [22] [21] [20] [23]. Although significant progress has been made on the performance of micro-gesture recognition based on H3D imaging, there are two key problems that haven't yet been solved. The first one is that the pre-processing of the holoscopic image is insufficient, and attempts to extract detailed 3D information from H3D images were unsuccessful. The second one is that micro-gesture recognition rate is still unsatisfactory for real-world applications due to the limited 3D information. In this research, a new H3D image pre-processing method used to takes the advantages of 3D information and then use deep neural network for pattern recognition. In addition, the several decision fusion approaches has been applied to combine multiple viewpoint recognition decisions.



The research on holoscopic 3D micro-gesture recognition has been accelerated after the first holoscopic 3D micro-gesture database HoMG was created [140]. Additionally, the HoMG database was made publicly available and the first HoMGR challenge competition was held [140]. Created by using the H3D imaging system, the 3D micro-gesture database supported high resolution static image data as well as high quality dynamic video data. It is renowned that the 3D micro-gesture database obtained has high quality micro-gestures and true 3D advantage over traditional 3D capture device. Although there are image and video subsets in the HoMG dataset, we only focus on the image subset, as the data processing methods will be quite different for video subset. The dynamic information extraction is the key in video based micro-gesture recognition systems. For image based micro-gesture recognition, good efforts have been made recently [140] [22] [23] [20] [21]. Traditional 2D image feature extraction and classification methods were used in the baseline paper [140]. Zhang et al. [22] proposed a method on 2D micro-gesture images using CNN models with fine-tuning. The method achieved the best accuracy by averaging the probabilities predicted from different models and different epochs. Lei et al. [20] proposed a bi-directional morphological filter and a fast fuzzy C-Means Clustering (FCM) method [141] to reconstruct 2D images from a H3D image. It is a good method for solving the problem of blur and distortion grids in the H3D image. It was ranked in the second place of the challenge competition in the image subset. The main limitation of this method is that the reconstructed image has a lower resolution and loses some detailed information. Sharama et al. [23] considered that each micro-lens captured the image at its respective angle which was different from the other lenses. They extracted a viewpoint image by selecting a pixel from each micro-lens and used feature fusion technique on both hand-crafted and deep features extracted from the neural network. The experiments show that their proposed method outperforms the baseline by an absolute margin of 26.67%. Peng et al. [21] proposed a deep residual network with attention mechanism. The experiments show that the attention design can highlight the micro-gesture part and reduce the noise introduced from the wrist and background. An accuracy of 82.1% on the image subset was achieved.

From the methods of all the participants in the challenge on this dataset, it can be seen that there are three potential issues that may help to improve the performance of the micro-gesture recognition. Firstly, because the original H3D images contain a lot of noise, the appropriate image processing method is crucial for extracting correct 3D information from H3D images. Although Lei et al. [20] and Garima et al. [23] have made attempts on this issue, the noise such as dark borders and geometric distortions were not eradicated. More importantly, none of the participants took the advantage of the full 3D information for micro-gesture recognition. Secondly, most participants accomplished the task of micro-gesture recognition by using deep learning methods and obtained significantly

improved results than the baseline method. However, most of the deep learning models were applied for H3D images directly or some 2D images with limited view angles. The power of the deep learning models were not fully used. Thirdly, Lei et al. [20] and Zhang et al. [22] used the decision fusion methods to improve the recognition accuracy. This is a good way and should be explored further. In what follows, we will address all these issues one by one and to make a better system for micro-gesture recognition.

## 4.2 System Development

This section introduces the detailed information on our solutions on each issue identified in current methods and then proposes our solution. We present our method below in detail.

### 4.2.1 Micro-gesture Recognition System

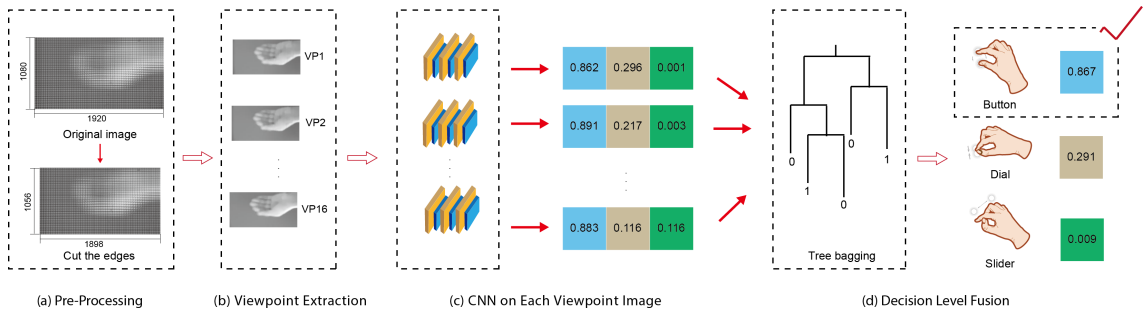


Figure 4.2: Block diagram of the proposed micro-gesture recognition system. There are four stages: (a) Pre-processing; (b) Viewpoint image extraction; (c) Deep learning for prediction on viewpoint images; (d) Decision level fusion.

Figure. 4.2 shows the framework of our proposed method that includes four main stages. Firstly, in the pre-processing stage, the original H3D images are cut on the four boundaries in order to localize the Element Images (EIs). An EI is a local area in the H3D image that was captured by one of the micro-lens array. The captured H3D images might have various offsets depending on the positions of the micro-lens array inside of the camera. This is a preparation step for View Point (VP) image extraction where the VP image is a 2D image of the scene from a particular viewing angle. Secondly, multiple shifted VP images are extracted from one H3D image where each VP image has different shifts from horizontal and vertical positions in the angle of view. Three simple and efficient patch-based rendering approaches were proposed by Georgiev and Lumsdaine [142], which

were used by Yang et al. [143] in holoscopic image coding scheme. However, our proposed method here is simpler and the EIs can be cut out automatically. Thirdly, CNN model with an attention block is used to extract features from each VP image and gives predictions (e.g. the possibilities it belongs to each type of micro-gesture from its fully connected layer). Finally, the decision fusion methods are used to combine the predicted decision values from each VP images together and produce the final prediction of the type of micro-gesture.

### 4.2.2 H3D Pre-processing

In the recording process of the original H3D image, an object is captured through an array of micro-lenses, where each micro-lens captures a perspective 2D elemental image of the object from a specific angle. The final captured image contains the intensity and directional information of the corresponding 3D scene in 2D form. This 2D elemental image is called the EI that is a small grid area in the H3D image. The standard pre-processing process can be found in [144], which includes lens correction, distortion correction, EI extraction, viewpoint extraction, etc. The first step of our pre-processing is to create an automated method to detect the edges of EIs and cut out the EIs from the original H3D image. Because of the lens distortion led to the many incomplete EIs which cannot extract the useful 3D information.

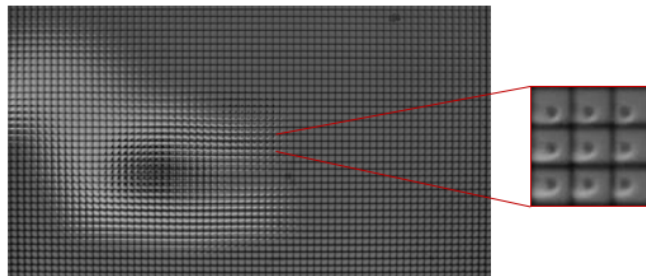


Figure 4.3: H3D micro-gesture image is consist of multiple 2D Element Images (EIs). Here 9 EIs are enlarged from the original H3D image.

Figure. 4.3 shows an example of a holoscopic 3D micro-gesture image that consists of many 2D EIs. Roughly, each EI is a approximately square area with small values (dark colour) on the edge. However, some boundaries are not straight lines due to the distortion of the micro-lens, especially the ones near the H3D image boundary as the associated micro-lens are far from the centre. Barrel distortion is caused by spatial imaging in narrow space, which results in the obvious distortion in the corner and edges. Although the dis-

tortion is not easy to be noticed by human eye, it affects greatly in extraction process [84].

In this work, all the EIs are cut out based on straight lines and the distortion will be dealt with later by the algorithm for VP image extraction. On the boundaries of the H3D image, some EIs are not captured fully, so only completed EIs will be cut out and used later for VP image extraction.

The cutting algorithm is based on the detection of the minimum values of the rows and columns of a H3D image. For a H3D gray-scale image  $H(i, j), i = 1, 2, \dots, I_o, j = 1, 2, \dots, J_o$  with a resolution of  $I_o \times J_o$ , all the values are summarized to  $H_c$  and  $H_r$  according to column and row as shown in the equations 4.1 and 4.2 respectively. Then the local minima of function  $H_c$  and  $H_r$  are picked up as the boundaries of the EIs. For example, if the H3D image has a resolution of  $1080 \times 1920$ , we can get two vectors as shown in Figure. 4.4 and Figure. 4.5. In Figure. 4.4, there are 40 local minima detected where the left and right minimum points were removed as they are the edge of an incomplete EI. Therefore, 38 EI edges were finally chosen. In the same manner, in Figure. 4.5, there are 70 local minima detected where the left and right minimum points were removed, Thus, EI edges were finally chosen. In the end,  $38 \times 68$  EIs were cut out from one H3D image. This method is a fast algorithm that can quickly produce all the EI images.

$$H_c(i) = \sum_{j=1}^{J_o} H(i, j), i = 1, 2, \dots, I_o \quad (4.1)$$

$$H_r(j) = \sum_{i=1}^{I_o} H(i, j), j = 1, 2, \dots, J_o \quad (4.2)$$

In practice, the camera calibration is not executed perfectly as the micro-lens cannot be placed perfectly on the horizontal and vertical lines of the H3D image. There might be a small angle  $\delta$  between the boundary of EIs and the horizontal and vertical lines of the H3D image as shown in Figure. 4.6. A small shift image  $H_\delta$  of the original H3D image can be used in the above methods and then the best  $\hat{\delta}$  can be obtained by minimizing the local minima of the summary of row and column pixels as shown in the following Eq. 4.3:

$$\hat{\delta} = \arg \min_{\delta} \left( \sum_{i=1}^{I_o} H_c^\delta(i) + \sum_{j=1}^{J_o} H_r^\delta(j) \right) \quad (4.3)$$

In this way, the best cut of H3D image  $I_c$  is obtained with  $\hat{\delta}$ .

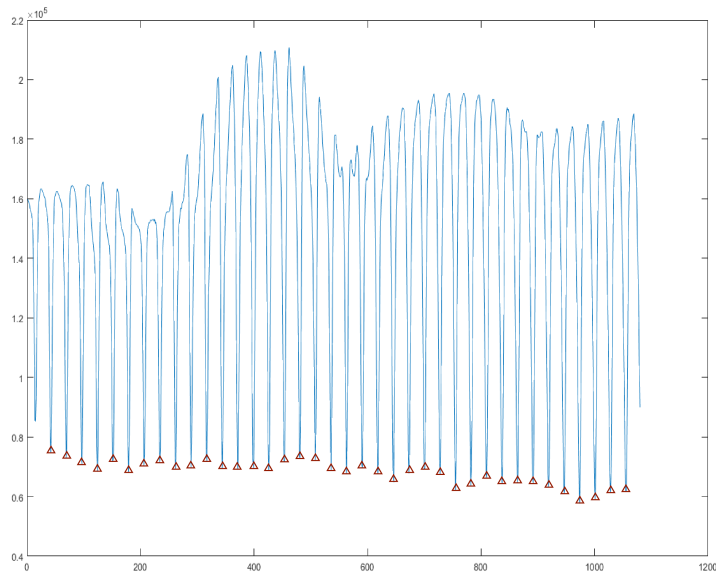


Figure 4.4: The minimal values of the summarized rows of a H3D image. The 38 points marked with small red triangles are the selected boundaries of the EIs.

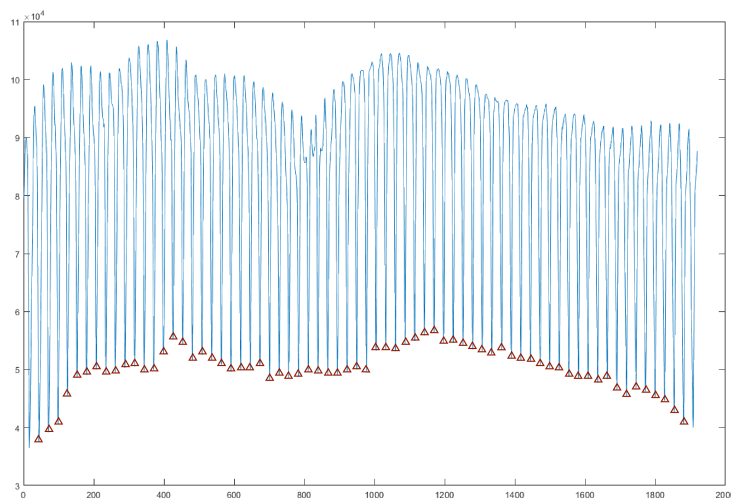


Figure 4.5: The minimal values of the summarized columns of a H3D image. The 68 points marked with small red triangles are the selected boundaries of the EIs.

### 4.2.3 H3D Viewpoint Extraction based on Shifting Method

From all the obtained EIs, VP images can be extracted. The VP image is a low-resolution orthographic projection type of rays from a particular direction. It can be extracted from the pixels of all the EIs. The basic principle of the recording process is the object to image through a micro-lens array, where each micro-lens has the intensity and directional information from the captured specific angle. Figure. 4.7 shows the relationship between 5 EIs and 3 focus layers. The captured H3D image has micro-lens array of 5 lenses with 3 pixels per lens. In this particular example, there are 3 planes per slice with the image

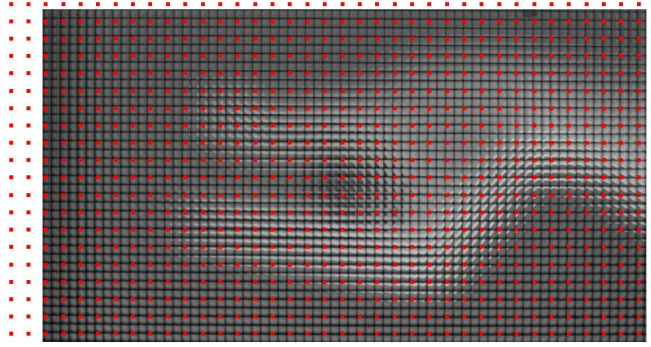


Figure 4.6: The horizontal and vertical lines of the H3D images are adjusted for the EI extraction.

size of 5 pixels, which is the VP image. In the recording stage, each micro-lens of local pixel position is involved direction as shown in Figure 4.7(a). To each VP image, the construction integrate all the pixels from the same location under different micro-lenses. And all the VP images, such as VP1, VP2 and VP3, are orthographic image as shown in Figure. 4.7 (b), and they are reconstructed by all pixels from same location in the 5 EIs. It should be mentioned here that the focus plane of each VP image might be different as shown in Figure. 4.7 (b). Holographic 3D image is changing the focus plane that is all the light ray to converge the ideal virtual depth plane. However, the viewpoint rendering pixel used to refocus at different depth planes.

In holographic 3D imaging principle, each EI, which is captured by a micro-lens contains a pixel from each layer of the 3D scene. In the same way, all EIs contribute to create a single aperture holographic 3D scene in the space.

H3D viewpoint extraction process is to select appropriate pixels at the same location from every EI of the H3D image to reconstruct an orthographic viewpoint image. The principle of the proposed viewpoint extraction method is illustrated in Figure. 4.8 where there are  $3 \times 3$  pixels in each EI, and the 9 EIs constitute an omni-directional H3D image.

In general, for a well-cut H3D image  $I_c$  with  $n \times m$  EIs, each EI can be represented as EI  $(p, q)$  where  $p = 1$  to  $P$  and  $q = 1$  to  $Q$ . The VP image  $VP(p, q)$  will have the dimension of  $n \times m$  as there is one pixel extracted from each EI. The values of  $P$  and  $Q$  are decided by the resolution of the cut H3D image  $I_c$  because  $I_c$  resolution will be  $(m \times P, n \times Q)$ .

The equation for VP extraction can be represented as the following:

$$VP_{p,q}(i, j) = I_c((i-1)P + p, (j-1)Q + q) \quad (4.4)$$

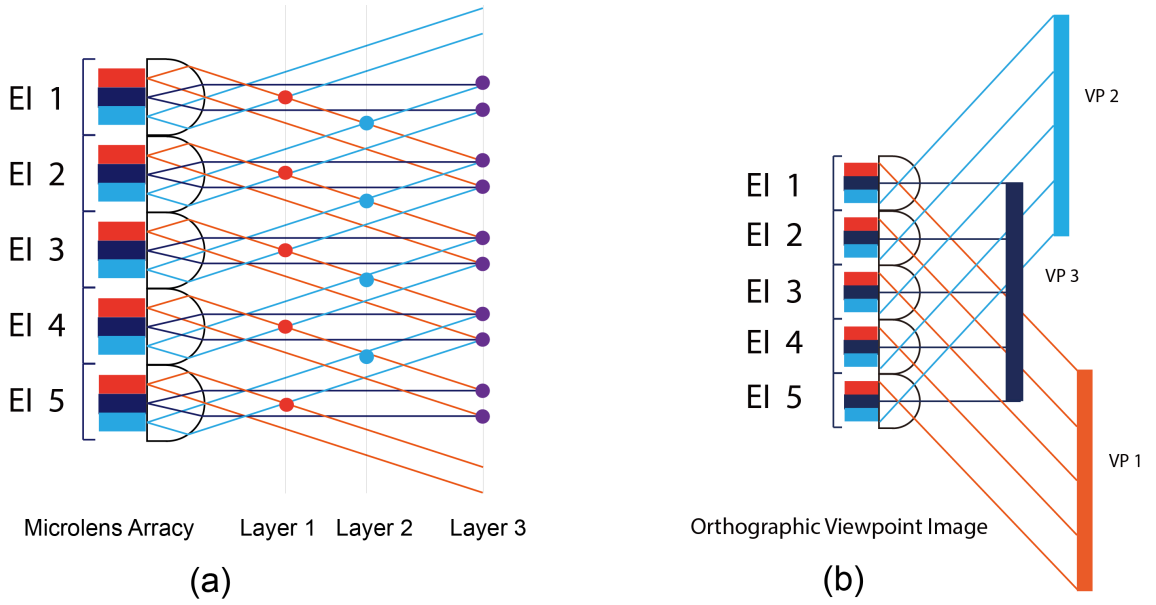


Figure 4.7: The relationship between EIs and focus layers of the holoscopic 3D capturing system. (a) micro-lens array recording system, (b) Orthographic viewpoint images from different perspectives.

where  $i = 1, \dots, m$  and  $j = 1, \dots, n$  are the coordinates of the VP image and  $p$  and  $q$  are the index of the horizontal and vertical positions of VP images respectively. The VP image  $VP_{p,q}(i, j)$  has a resolution of  $n \times m$  pixels.

In principle,  $P \times Q$  VP images can be extracted from one H3D image where every pixels in all EIs can be picked up. However, it is not working as expected in practice due to several issues. Firstly, there is very small difference between two adjacent pixels in one EI and there is no much additional information provided by picking up all the VP images. Secondly, intensity value of one pixel might vary due to the lighting conditional and random noise. It will reduce the quality of the VP images. Thirdly, there are barrel distortion effect on the boundaries of each EI while the distances between the object and the micro-lens are bigger.

In our proposed method, we address all these issues in order to obtain the high quality VP images. Firstly, we only extract small number of VP images that have big difference between each other. Secondly, our VP images are extracted from patches instead of single pixels of each EI. Thirdly, only central area in one EI are selected and used for VP image extraction in order to avoid the distortions. The patches are shifted in horizontal and vertical directions and only small number of viewpoint images are extracted.

Figure. 4.9 shows one EI where the boundary pixels are not used and only the central pixels are selected for VP image extraction. The central pixels are made of 16 patches,

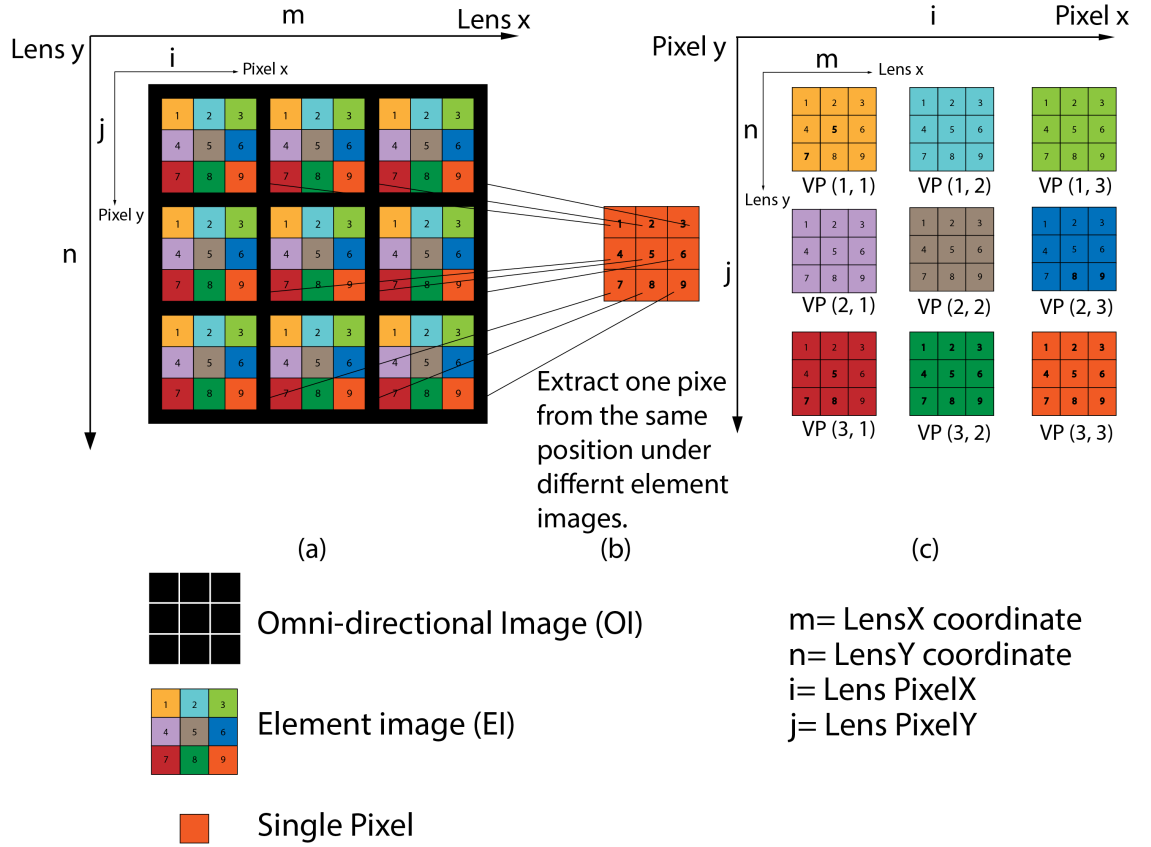


Figure 4.8: Illustration of the principle of H3D image viewpoint image extraction. (a) The  $3 \times 3$  pixels under each micro-lens, (b) One viewpoint image extracted from same position under different micro-lenses, (c) Nine viewpoint images extracted from  $3 \times 3$  EIs.

each of them has  $3 \times 3$  pixels. all the  $3 \times 3$  pixels contribute to one pixel in the VP image. From each patch, one VP image is reconstructed and totally 16 viewpoint images are extracted.

#### 4.2.4 Convolutional Neural Network Model

Convolutional neural network (CNN) is a biologically-inspired model and very successful in image-related recognition tasks [145]. The important component of the CNN is the shared-weight and sub-sampling. Figure. 4.10 shows the general structure of a CNN. The input layer receives images with the same size. After processed by a convolution kernel, each small neighborhood in the input layer will form a value in a feature map (each plane in the layer). The  $i^{\text{th}}$  feature map  $C^i$  can be expressed as:

$$C^i = f(x * W^i + b^i) \quad (4.5)$$

where  $f$  is the activation function,  $x$  is the input VP image,  $W$  and  $b$  are the weight of



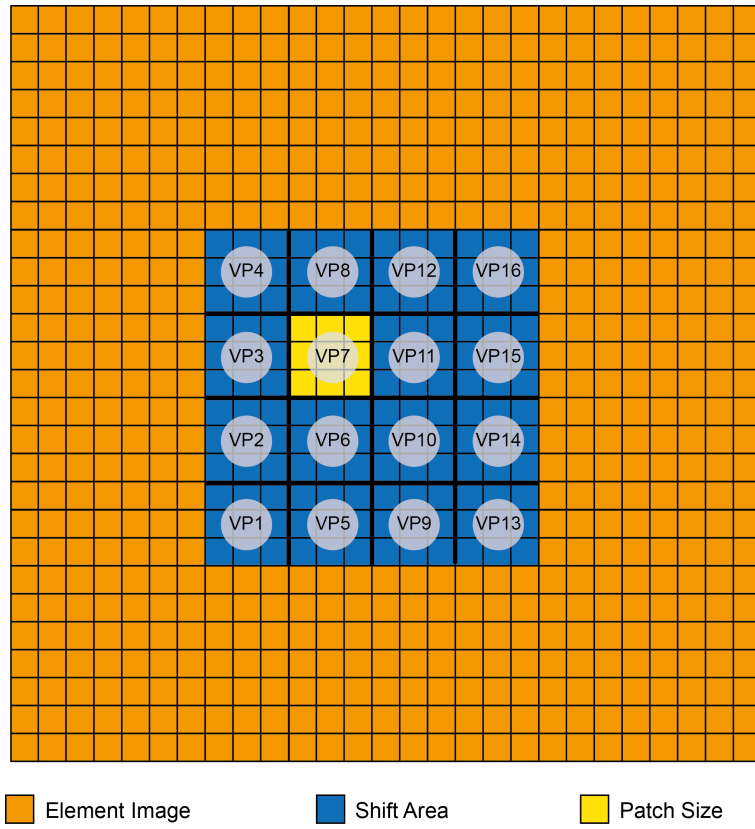


Figure 4.9: This is one EI image with the resolution of  $27 \times 27$ . The boundary pixels are avoided, only the central pixels are selected. Each patch has  $3 \times 3$  pixels. In the end, only 16 patch areas are selected from this EI.

the convolution kernel and bias, respectively. Each feature map shares the same  $W$  and  $b$ . In a convolutional layer, there is normally more than one convolution kernel, so multiple feature maps are calculated. The  $i^{th}$  feature map  $P^i$  in the pooling layer (S1 layer) can be calculated by using:

$$P^i = f(\beta * S(C^i)) + \alpha \quad (4.6)$$

$\beta$  and  $\alpha$  are the coefficient and bias, respectively.  $S(\cdot)$  denotes the sub-sampling operation for a convolutional feature map. It can be written as:

$$S(C^i) = \max_{s,l} C_{s,l}^i \quad \|s\| \leq \frac{N_s}{2}, |l| \leq \frac{N_s}{2}, s, l \in \mathbb{Z}^+ \quad (4.7)$$

where  $N_s$  is the sub-sampling size.

Generally, a deep convolutional neural network formed by stacking multiple convolution layers and sub-sampling layers [146]. The fully-connected layer is a multi-layer percep-

tion feed-forward neural network, the output layer can be written as:

$$p(j|F; \theta) = \frac{e^{\theta_j^T F}}{\sum_{i=1}^J e^{\theta_i^T F}} \quad 1 \leq j \leq J \quad (4.8)$$

where  $p(j|F; \theta)$  denotes the probability that the input feature  $F$  belongs to class  $j$ ,  $\theta$  is the weight vector between the output layer and the previous layer,  $J$  is the number of class.

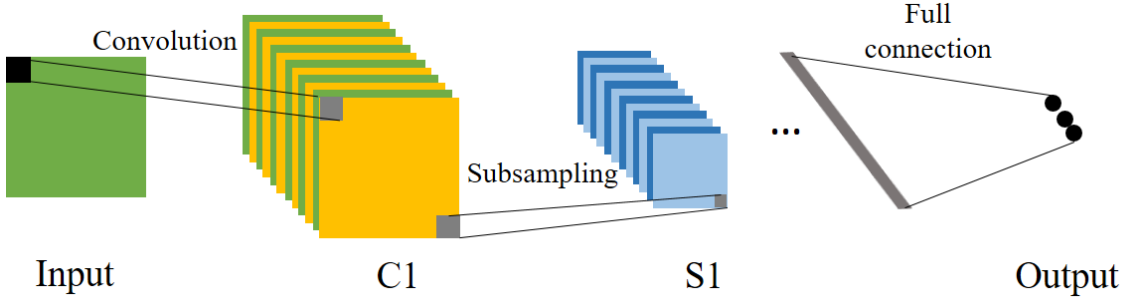


Figure 4.10: The general structure of a CNN. The inputs are the VP images extracted. There are a few pair of convolution layer (C) and Sub-sampling layer (S). Finally, it is fully connected layers with the number of outputs being the number of class.

Many CNN structures, such as GoogLeNet [147] and ResNet [148], have been trained on ImageNet [145] with super performance. Here, pre-trained ResNet model is used and then fine-tuning is carried on the model with our dataset. In addition, we modified the existing CNN model by adding an attention-based residual block.

Figure. 4.11 shows the attention-based residual block, where the dotted-line area is the attention branch that can spotlight the finger micro-gesture and reduce the noise introduced from the wrist and background. For input  $x$ , the overall output is  $O(x)$ .

$$O(x) = F(x) + F(x) \cdot A(x) + x \quad (4.9)$$

$A(x)$  represents the spatial attention mask. This attention branch here has been used in our previous work [21]. It is a bottom-up top-down structure to learn the interesting area in a gesture image as shown in Figure. 4.12. From Figure. 4.12, it can be seen that the attention design puts more attention on the gesture area in higher level layers. We believed that this is special for micro-gesture recognition.

The CNN model adopted this attention design is applied for all the VP images for the micro-gesture recognition. The output probabilities of the CNN was produced. From each sample, the probabilities of three gestures are computed, which are used for decision

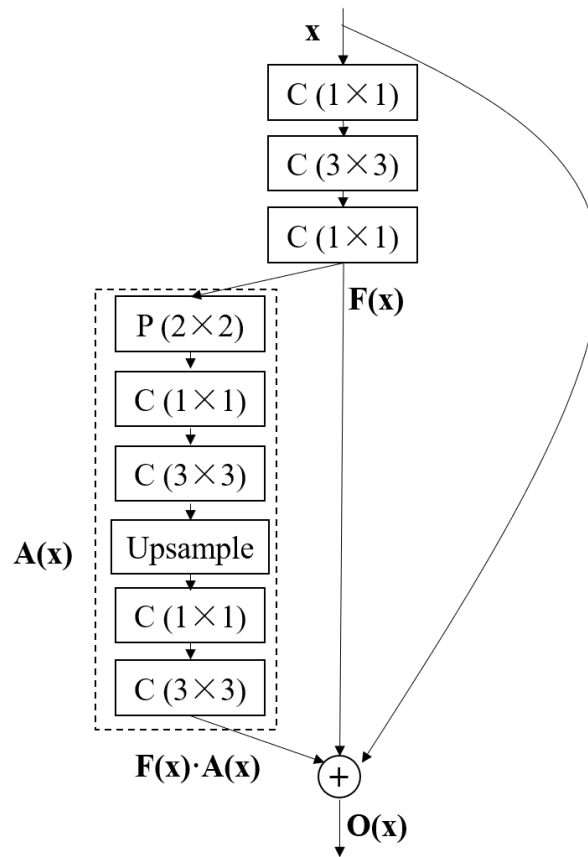


Figure 4.11: Attention-based residual block that was integrated to the CNN architecture.

level fusion.

#### 4.2.5 Decision Level Fusion

The decision fusion also referred as mixture of experts [113] is a method that can contribute to improve the recognition rate by combining all the decisions together. In this work, some ensemble functions based on voting [149] and trainable methods [150] have been explored for combining predictions from multiple VP images efficiently. Specifically, some simple fusion methods such as majority voting, averaging, product of the predictions as well as trainable mixture of experts approach such as bagging learning strategy with REPTree are used on the multiple viewpoint predictions.

Assume there are  $J$  classes for all the H3D images and each H3D image has  $K$  VP images, CNN models will be applied on all VP images separately to produce all the prediction probabilities  $\{p_{j,k}\}$ , where  $\{j = 1, 2, \dots, J\}$  is the index of the class and  $\{k = 1, 2, \dots, K\}$  is the index of the VP images.

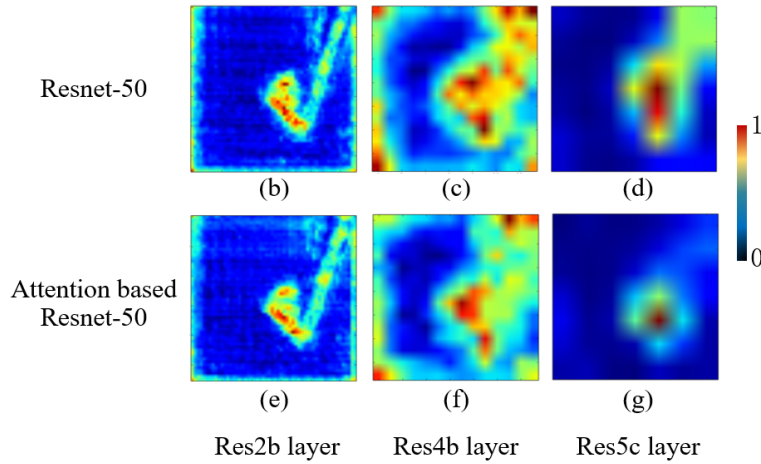


Figure 4.12: First row and second row represent the feature maps learned by ResNet-50 and attention based ResNet-50, respectively at res2b layer (low-level), res4b layer (middle-level) and res5c layer (high level).

**Majority Voting.** For an instance, voting strategies assume that each classifier gives a single class label. For convenience, define the output of the multiple classifier as  $V_k$  as the following,

$$V_k = \arg \max_j (p_{j,k}), \quad k = 1, 2, \dots, K \quad (4.10)$$

The predicted label  $V_M$  for the majority voting can be written as:

$$V_M = \arg \max_j \left( \sum_{k=1}^K I_{\{V_k=j\}} \right) \quad (4.11)$$

where index function  $I_\Delta$  will be one if set  $\Delta$  is non-empty and 0, otherwise.

**Averaging.** Averaging fusion method can be applied to the multiple classifier under the condition that each output of the classifiers is expressed in probabilities. The decision output  $V_A$  for the averaging fusion can be written as:

$$V_A = \arg \max_j \left( \frac{1}{K} \sum_{k=1}^K p_{j,k} \right) \quad (4.12)$$

**Product.** Product probability fusion method is to calculate the product of experts by multiplying individual probabilities. Similarly to the averaging probability fusion method, the product probability fusion output  $V_P$  can be written as:

$$V_P = \arg \max_j \left( \prod_{k=1}^K p_{j,k} \right) \quad (4.13)$$

**Bagging Classification Tree.** Trainable mixture of experts has the ability to learn from individual classifier outputs to form a higher level of experts. In this work, Bagging learning with REPTree has been explored to heighten the multi-viewpoint results. Bagging learning strategy was introduced by Breiman [151] to reduce the variance of a predictor. It is a successful way for improving classification performance. Reduce Error Pruning Tree (REPTree) [152] is a fast decision tree learning method that based on the information gain. The substantial steps of the trainable mixture of experts approach are as follows.

Assume that we have  $N$  instance. For each instance, the number of VP images and classes are  $K$  and  $J$ , respectively. Therefore, the feature dimension of each instance is  $K \times J$ . Firstly, a training set is sampled (with replacement) from the all instances to generate a classifier. Specifically, REPTree algorithms are used as the learning system. And then, same as the first step, the number of  $T$  trails are replicated to form the  $T$  classifiers. Finally, for an instance, the classification result is voted by every classifier for the class with the most votes.

### Mean Probability Voting

For the  $i^{th}$  classifier, suppose the probabilities of each sample to be class  $j$  is  $p_{j,i}$ , the ensemble decision for the mean probability voting can be written as:

$$V_P = \arg \max_j \left( \text{mean} \left( \sum_{i=1}^m p_{j,i} \right) \right) \quad (4.14)$$

### Ranking Decision Fusion

In ranking decision fusion method [153], for the  $i^{th}$  classifier, suppose the probabilities of each sample to be class  $j$  is  $p_{j,i}$ , a new vector  $z_{j,i}$  is generated as follows:

Define a ranking vector  $r=r_j=0; j \in [1 c]$ ;

For  $j=0$  step 1 until  $c$ ,

$$r_j = \arg \max p_{j,i}, i = c - (j - 1) \quad (4.15)$$

Table 4.1: Number of samples in each partition of the HoMG database. "B" stands for Button, D stands for Dial and S stands for Slide.

Subset	Training			Development			Testing		
	B	D	S	B	D	S	B	D	S
Image	5507	5534	5722	2266	2188	2106	2665	2267	2359
Video	160	160	160	80	80	80	80	80	80

## 4.3 Experiments and Evaluation

### 4.3.1 H3D Image Subset of HoMG Database

Holoscopic Micro-Gesture (HoMG) database was recorded for this research that is public available at our website. For the data collection, 40 participants were selected and the recordings were done under 2 different backgrounds, 2 hands (e.g. left and right), 2 distances (e.g. far and close) and 3 micro-gestures (e.g. Button, Dial and Slide). So 24 videos were recorded for one participants. The length of the video is between 2 and 20 seconds with a frame rate of 25 and resolution of  $1902 \times 1086$ . In total, 960 videos are included in the database.

The HoMG database has been made public available [140] for micro-gesture recognition competition where it was divided into two subsets: image based and video based micro-gesture subsets. Also it was divided into training, develop and testing subsets. The detailed information of HoMG is shown in the Table 4.1. In this paper, the work is only done for the image based subset where each micro-gesture is represented by a H3D image.

### 4.3.2 H3D VP Extraction Parameters

For an original H3D image, its resolution is  $1920 \times 1080$  pixels. The resolution of element image from each micro-lens is about  $27 \times 27$  pixels. However, at the edge of the H3D image, there are some EIs that can not have the full resolution due to the in-completion of the micro-lens. So these pixels of the H3D image were cut out. Specifically,  $68 \times 38$  pixels full EI were cut out from one H3D image after rotation of the image and make the straight cutting lines.

After the edge cutting, the H3D image should be estimated in depth and be refocused to extract the VP images. Small patch area of  $3 \times 3$  was chosen from the central area of each EI, and then shift it in horizontal and vertical directions. The shift value also leads to the depth transformation. We obtained 16 points by extracting the viewpoint from

the different refocusing layers, as shown in Figure. 4.9. In the human eyes, the slight movement from the viewing of the continuous VP images can be observed. In the end, for each H3D micro-gesture image, 16 2D VP images were extracted from different depth as shown in Figure. 4.13 .

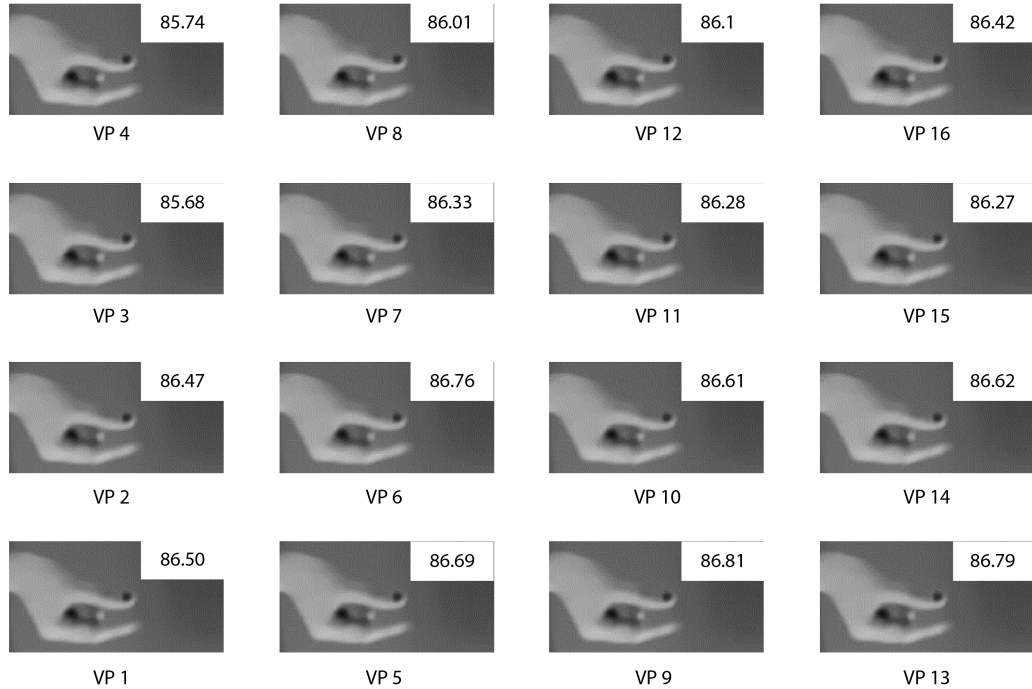


Figure 4.13: The 16 viewpoint images extract from a H3D image and its associated recognition accuracy based on CNN models.

### 4.3.3 Implementation Details of CNN Model and Mixture of Experts

Considering that the micro-gesture is a small-local representation in an image, we adopted the attention mechanism to spotlight the finger micro-gesture area.

Before training, we used the transfer learning strategy to initialize the network with ImageNet database and get a pre-trained model [148]. In the stage of training for micro-gesture recognition, fine-tuning was done for the pre-trained model on the HoMG database. The average values of all pixels of the training set have been subtracted from the input gray-scale image and the input image was further divided by the variance of all pixels of the training set. This is a normalization process. To increase the robustness of the network, each VP image was resized to  $256 \times 256$  and then cropped out the four corners to size of  $224 \times 224$ . Moreover, data augmentation such as color shift (maximum value of 20) and image rotation (maximum degree of 10) are applied to the training set with a

Table 4.2: Classification accuracy(%) of CNN models on each VP images at the testing set of HoMG database.

No.VP	Acc	No.VP	Acc	No.VP	Acc	No.VP	Acc
<b>VP1</b>	86.50	<b>VP5</b>	86.69	<b>VP9</b>	<b>86.81</b>	<b>VP13</b>	86.79
<b>VP2</b>	86.47	<b>VP6</b>	86.76	<b>VP10</b>	86.61	<b>VP14</b>	86.62
<b>VP3</b>	85.68	<b>VP7</b>	86.33	<b>VP11</b>	86.28	<b>VP15</b>	86.27
<b>VP4</b>	85.74	<b>VP8</b>	86.01	<b>VP12</b>	86.1	<b>VP16</b>	86.42

probability of 0.5. The dropout ratio of the last weight layer was set to 0.5. The batch size is set to 9 with momentum of 0.9 and weight decay of 0.0005. The initial learning rate was set to 0.001, which decreases 10 times smaller after every 10 epochs. Our training process was implemented on Caffe framework [154] with an Nvidia Titan X GPU.

After the training, each VP image in the dataset was input to the trained model to get a decision output of the classification layer. Specifically, each value in the output vector represents the probability belonging to the three types of micro-gesture, respectively. Due to each instance in the dataset having 16 VP image, we got  $16 \times 3$  output values as the probabilities of the instance belonging to each type of micro-gesture based on each VP image.

In our mixture of experts procedure, we set the number of bagging trails as 10000 and in each trail, 50% of the instance in the training set has been sampled to generate a classifier. For REPTree training [152], the minimum total weight of the instances in a leaf is set to 2 and the amount of data used for pruning is set to 3.

#### 4.3.4 Experimental Results and Comparison

Table 4.3.4 shows the experimental results using CNN models on separate VP images. Totally, 16 VP images were extracted from one H3D image. Each of them has been applied in the proposed CNN models and the associated classification probabilities have been produced. This accuracy is the percentage of correctly classified micro-gestures in the testing set after the models was trained on the training and development subsets. From this table, it can be seen that similar performance has been achieved for each single VP image. The best results of 86.81% was achieved for viewpoint 9, which is a very good accuracy already.

Table. 4.3.4 shows the results achieved by our proposed method in comparison with other



Table 4.3: Classification accuracy (%) comparison between the proposed method with all the existing methods on the testing subset of HoMG dataset. "A" means attention block. "M.V." means "majority voting".

Author	Methods	Accuracy
Liu et al. [140]	LBP+k-NN	50.90
Liu et al. [140]	LPQ+SVM	52.6
Sharma et al. [23]	CNN+LPQ(Max Vote)	77.57
Peng et al. [21]	A-Resnet	82.10
Lei et al. [20]	FCM+GoogLeNet	84.28
Zhang et al. [22]	ResNet152+DenseNet161+SeResNet50+M.V.	86.70
This work	16VPs+A-ResNet	<b>86.71</b>
This work	16VPs+A-Resnet+M.V.	<b>87.04</b>
This work	16VPs+A-Resnet+(Mean Probability Fusion)	<b>87.04</b>
This work	16VPs+A-Resnet+(Product Probability Fusion)	<b>87.03</b>
This work	16VPs+A-Resnet+(Bagging Classification Tree)	<b>87.15</b>

state-of-the-art methods. Firstly, we combined all the 16 VP images together and applied CNN model and achieved the accuracy of 86.71%. The last four methods of the table denote the the methods combining the CNN outputs based on mixture of experts approach. We can see from the Table. 4.3.4, our proposed pre-processing methods used for H3D image combining CNN with mixture of experts approach obtained a significant performance improvement on micro-gesture recognition. Specially, the proposed methods that combine CNN with Bagging Classification Tree approach gained an improvement about 40% of recognition accuracy than baseline method. It also outperforms the method of Zhang et al. [22] (87.15% vs 86.70%). Besides, compared with the method of Peng et al. [21], the CNN model achieved an accuracy improvement of 5% approximately. It even slightly higher than the method achieved by Zhang et al. [22] although voting method was used in their work (86.71% vs 86.70%). Consequently, the proposed pre-processing methods used for H3D image is validated to be effective. Moreover, the last four methods in the table show that the mixture of experts approach makes a great contribution to the improved recognition rate. Notably, the proposed trainable mixture of experts based on bagging classification tree is more superior than the voting and probability fusion method.

## 4.4 Conclusion

Image based finger micro-gesture recognition is a very challenging problem. In this study, we presented a recorded public micro-gesture dataset using a holoscopic 3D imaging sensor for this particular research area. Then, A new micro-gesture recognition system has

been proposed and evaluated it on the dataset in comparison with the state of the art methods. We proposed a fast and robust pre-processing methods for H3D images, which can extract the element images. In addition, we presented our simplified viewpoint image extraction method. Each viewpoint image is clean and prominent, which highlights the micro-gesture area and creates full representation of the original H3D image. Furthermore, a pre-trained CNN model with the attention mechanics is applied for each VP image for the predicted probabilities of each gesture. Finally, some mixture of experts methods based on voting strategy and trainable model have been explored to achieve better classification results. The achieved recognition accuracy is better than all existing state of the art methods. The accuracy of 87% might be good for some applications already. The main reason is that the attention-based network can learn to focus on the interest area for each viewpoint image, and the decision fusion method ensemble the classification results of each viewpoint efficiently.

In addition, it demonstrated that holoscopic imaging is a potential way for real-world applications. This image sensor can be embedded into a general digital camera with small additional cost.

The proposed system might be improved further in the following ways. Firstly, the number of VP images can be optimized. Better performance might be achieved if larger number of VP images are extracted. However, that means more computing cost and slower speed. Secondly, feature level fusion methods might be considered, and achieve better performance. Thirdly, the whole system might be treated as a learning system, in which the optimized parameters can be learned together for the best performance.

## **Chapter 5**

# **Pseudo Viewpoint and Deep Learning**

# **for Video-based H3D Micro-gesture**

# **Recognition**

The HoMG database has been introduced in chapter 3, and the recognition task focuses on the HoMG image subset in chapter 4. However, The HoMG image subset still have many limiting factors for micro-gesture movements, such as time sequence and data consistency. As the video subset is able to offer rich information, they become a new trend for HCI applications. Apart from the implicit 3D and depth information, the HoMG video subset supports each micro-movement from each time sequence.

This chapter focuses on the HoMG video subset for H3D micro-gesture recognition. In order to further develop the algorithms for the HoMG video subset, relevant recognition algorithms have been investigated, such as macro-gesture recognition, facial expression and lip-reading. Because the H3D micro-gesture video has richer information than the H3D image, the more advanced system has been created to further improve the performance on the HoMG database.

In summary, all the methods mentioned above have many advantages, for example the 3D information still not fully used to recognition, and the accuracies are not satisfactory. Therefore, micro-gesture recognition based on video subset of HoMGR dataset will be further development in this chapter.

## 5.1 Methodology

### 5.1.1 Proposed Video-based H3D Micro-gesture Recognition System

In this section, the PVP based 3D micro-gesture recognition is described in detail.

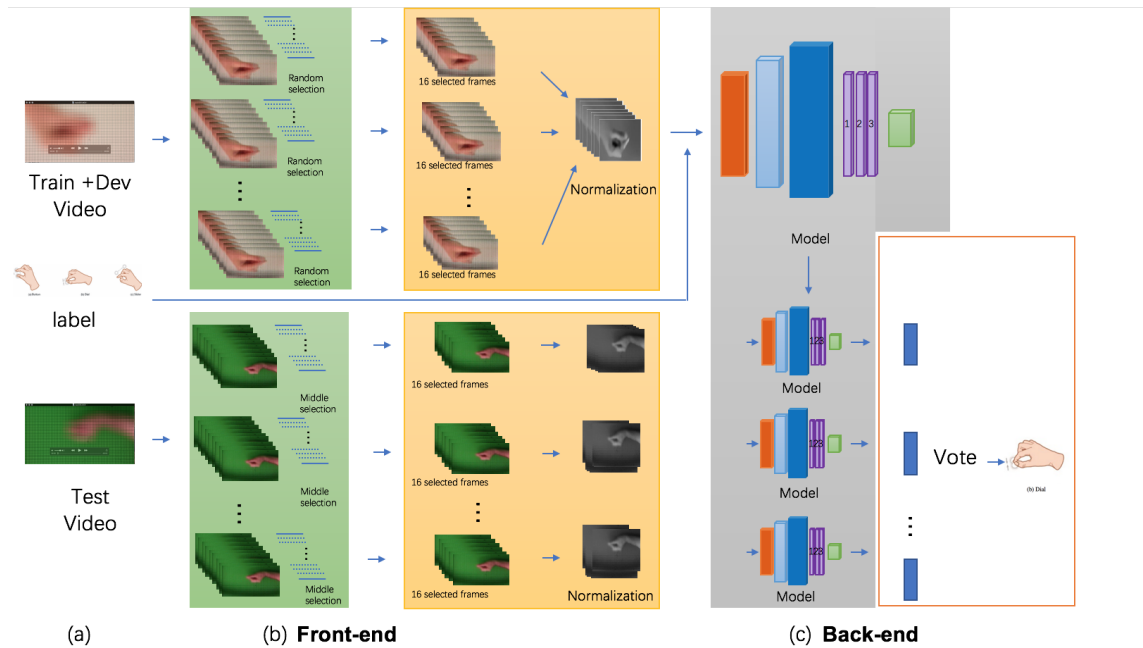


Figure 5.1: The video-based H3D micro-gesture recognition pipeline: (a) Sample video data of HoMG (b) is front-end, which consists the PVP extraction and pre-processing (c) described the back-end, which consist the deep network models and majority voting.

Figure. 5.1 shows the framework of the whole recognition pipeline. Firstly, each H3D frame has been transform in the original video to several PVP images according to the imaging principle of H3D cameras. In this transformation, the 25 frames in each second have all been kept, although the duration of each video is different. The issue is that videos of different durations do not consider pre-processing in the early stage. However, the input method will tackle this problem when the input is fed to deep neural networks. Besides these transformation problems, several videos have been obtained with different amounts of Pseudo View-Points of the single input video. The PVP extraction method is similar to VP extraction principles, which extracts pixels from each EI image at the same location, and then reconstructs the pixels of each patch into PVP images with different depth and multi-view. It is noted that PVP images cannot be displayed and human eye can

hardly recognise the image integrity. A deep neural network extracts features in each PVP frame, therefore obtaining the prediction results of each PVP. Compared to static images, gesture recognition of videos is more challenging, because it is difficult to represent their temporal features and the training process is more time-consuming, especially in real-time applications [155]. In this section, four advanced deep learning architectures have been trained on the data. This system used end-to-end models, which not only reduce the input and output process, also learn spatiotemporal visual features and a sequence model. The layers of deep learning models have also been adjusted for back-end comparison. Majority voting has been performed on all the results to get the final prediction of the input H3D video.

### 5.1.2 The VP based Front-end

For viewpoint extraction, it is a common method to used to obtain different depth images. The principle follows Figure. 5.4, which shows the relationship between EI and the focus layers from H3D capture system. For HoMGR video data, the pre-processing is a necessary step, because the captured spatial imaging is in a narrow space, which lead to the image has distortion. Viewpoint was widely used in display area, the traditional method are complicated. Following the traditional method to cut the straight lines and correct the distortion [156]. In order to cut the incomplete EIs on a batch basis, the automatical cutting pre-processing method has been used. The cut-out is along the straight lines, and the complete EIs are used to extract and reconstruct viewpoint images. For extraction of viewpoint images , distortion of all the EIs have been corrected in order to extract the VPs without noise pixels. This method was commonly used in the H3D imaging system of 3D display, and therefore gesture movements in the reconstructed VPs are easily recognized by human eyes. In this section, as the VP extraction is based on the traditional principle. This process is similar to the method used for the image database in the last chapter, but the VPs resolution is higher than image database because of selected patch size.

Original H3D image is composed of the a set of 2D element images, which is against to direct using for deep learning model even it has many 3D information inner. The resolution of H3D video frames are  $1080 \times 1920$  pixels, which means consuming a mass of the time for data training. Therefore, the viewpoint extraction is necessary. For VP front-end, following the traditional principle extraction. As the 3D scenes of the intensity and directional information are recorded in 2D form. The omni-directional H3D imaging system is include the parallax and 3D information from recorded the direction, which is record by spherical microlens. And the element image (EI) is consist of the single pixels, Figure. 5.4 explain  $3 \times 3$  pixels the single pixels constitutes the element images and the

viewpoint extraction. The H3D images are orthographic, which is the parallel rays of projected at various angles from the object, then forming viewpoint images (VP). And the viewpoints use zigzag manner to select from pixel mapping. Figure. 5.4 shows the example of the  $3 \times 3$  pixels element images of H3D image extract the all-in-focused viewpoint images [156]. (a) proposed  $n \times m$  EIs and each element images have  $3 \times 3$  pixels. (b) is explanations of the sub-sampling pixels from (a). (c) described the shifting of extraction location. The equation of the VP extraction can be represented as below:

$$VP_{p,q}(i,j) = I_c((i-1)P + p, (j-1)Q + q) \quad (5.1)$$

where  $i = 1, \dots, m$  and  $j = 1, \dots, n$  are the coordinates of the VP image and  $p$  and  $q$  are the index of the horizontal and vertical positions of VP images respectively. The VP image  $VP_{p,q}(i,j)$  has a resolution of  $n \times m$  pixels.

Viewpoint extraction follows the traditional method, which is based on the holoscopic 3D imaging system principle. Extraction process extract a set of the patch from the each of the 2D EIs, then re-construct the extracted pixels. Before the extraction stage, the cut off EIs in the edges should be considered. There is obvious barrel distortion due to the spacial imaging in narrow space, and the most boundaries are not straight lines. Additionally, each EI has small values under the dark square. Based on the above of the situations, there are many limited conditions for extract visible level viewpoint.

Therefore, the visible level viewpoint extraction is complicated. First, the cut the H3D images edges from horizontal and vertical in order to obtain the H3D image is consists of the full element images. Then the lens correction should be considered due to the barrel distortion, which is caused by the spatial imaging in narrow space. The distortion has been negative affection on extraction process, therefore the most distortion are not easy noticed by human eye. In this part work, the principle of cutting algorithm is based on the minimum values of H3D image from rows and columns, the incomplete element images have been removed from horizontal and vertical.

For viewpoint extraction and shift, every pixels under all EIs have been used for selection and extraction. Every time has a patch pixels from same location under each EIs have been selected and extracted, then those patch pixels use for each VP reconstruction. The shifting patch has been selected from horizontal and vertical direction respectively. It is noted that only select the central pixels for shifting due to boundaries' pixel not exertion.

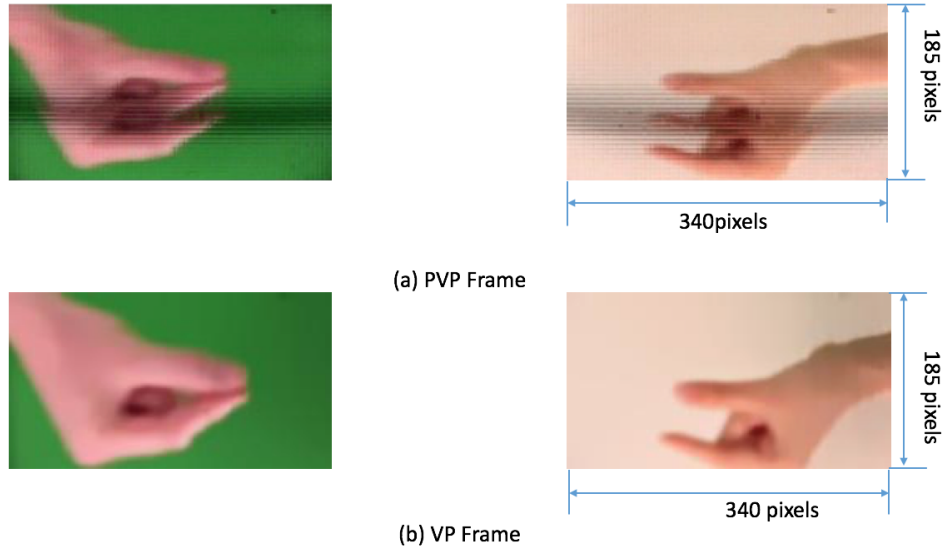


Figure 5.2: Pseudo viewpoint(PVP) frame and viewpoint frame. (a) obtained PVP frame, and the resolution is 340 by 185.(b) VP frames extraction.

### 5.1.3 The PVP based Front-end

For pseudo viewpoint extraction, obtaining the image from several different depth planes according to the principle of all-in-focused. One time-consuming step in the traditional viewpoint based process is the way to tackle the distortion effect and the incomplete element image (EI) in the boundaries. In the previously works, most researches are correct distortion first, then used manual software to remove the incomplete EI images. In this study, this process has been simplified by temporarily, and ignoring the distortions in the corner and the edges, which is follow the principle of integral imaging extraction and reconstruction, the Figure. 5.4 shows the concept of extraction. All the pixels have been kept, even some of which are partly or completely distorted. Hence, the Figure. 5.4(a) and (b) shows the PVP images have many noises, Specifically, a H3D image resolution is  $1920 \times 1080$  pixels, then resize the each frame to  $1836 \times 999$  pixels, which in accord with the VP. As the each EI is  $27 \times 27$  pixels and the number of EI could be approximately  $68 \times 37$ . The PVP images resolution depend on the number of EIs and patch size. A several local patches in each EI have been sampled, then obtain the final PVP frame by putting all the patches from all EIs together in a single frame. To make a fair comparison with the traditional methods, PVP16 has been extracted for training and comparison .

In order to further compare with multi-viewing affect, PVP 25 has been extracted, which means add the one more selection and reconstruction from horizontal and vertical direction respectively, Figure. 5.5 shown the each viewpoint extraction direction. Due to the

---

**Algorithm 1** Algorithm

---

**Input:** Video frames  $f(o), o = 1, \dots, O$   
**Initialisation:**  
**for** Resize frames:  $image = f.resize(1836, 999)$  **do**  
    Array:  $a = asarray(image)$ ,  
    Result:  $result = zeros([result_h, result_w, 3])$ ,  
    Number of Video  $num = 0$   
**Horizontal Sampling Patch:**  
    **for**  $i$  in range (5): **do**  
**Vertical Sampling Patch:**  
        **for**  $j$  in range (5): **do**  
**Horizontal Unit Numbers :**  
            **for**  $height$  in range (37): **do**  
**Vertical Unit Numbers:**  
                **for**  $width$  in range (68): **do**  
                     $result[height*5:(height+1)*5,width*5:$   
                     $(width+1)*5...] = a[(height*27+j*5):(height*27+(j+1)*5),$   
                     $(width*27+i*5):(width*27+(i+1)*5),...]$   
                    **for**  $u=1$  to  $U$  (For 2) **do**  
                        Number of Video  $num+ = 1$   
                    **end for**  
                **end for**  
            **end for**  
        **end for**  
    **end for**  
**end for**  
**Output** result resolution (340 \* 185)

---

Figure 5.3: Pseudocode for Pseudo viewpoint(PVP) extraction algorithm.

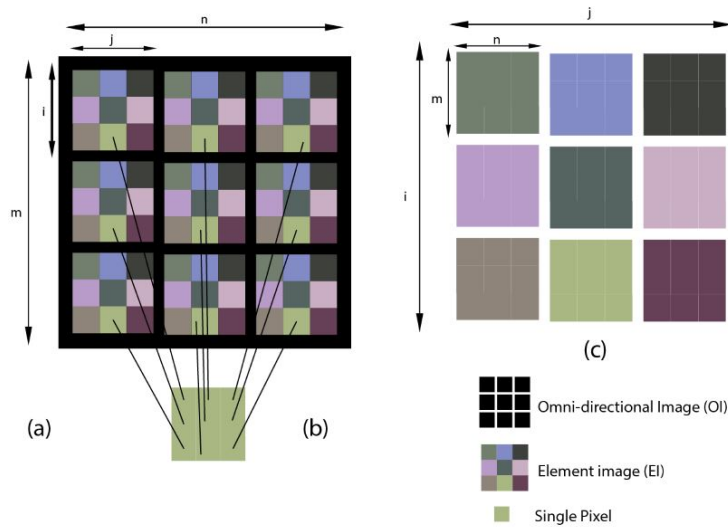


Figure 5.4: The principle of viewpoint extraction



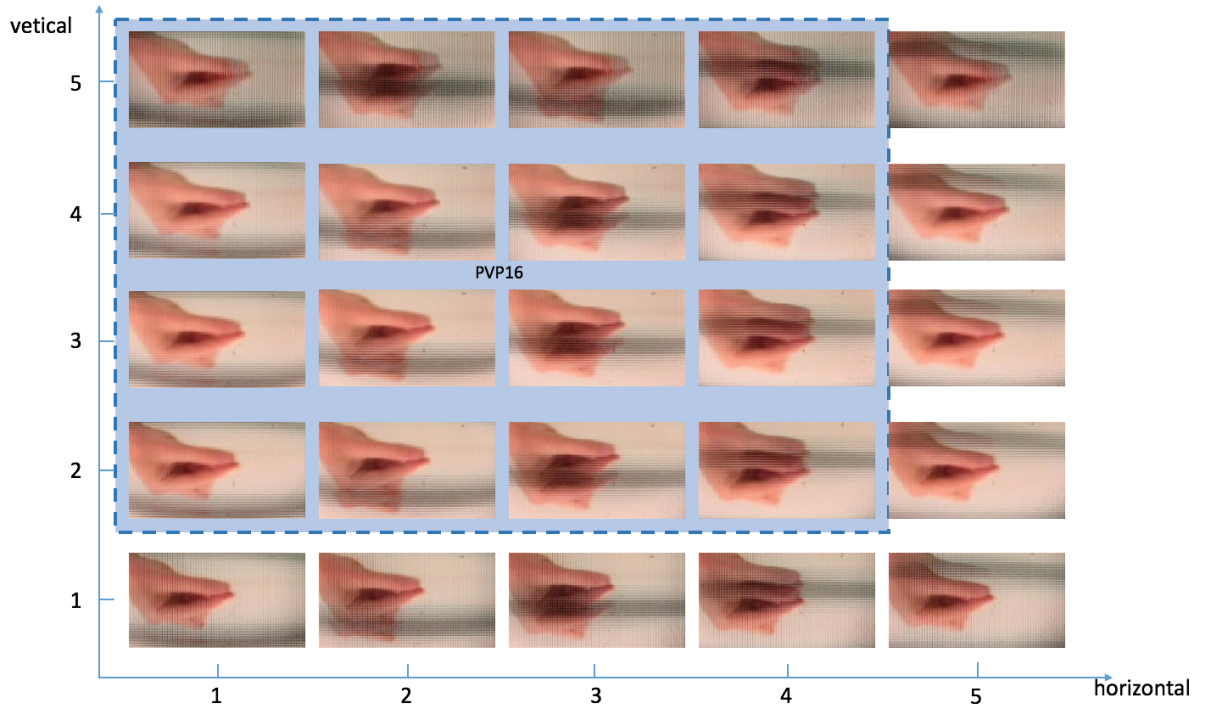


Figure 5.5: PVP16 and PVP 25 extraction from horizontal and vertical.

PVP is not based on the visible level to extract the viewpoint image, and the extraction and reconstruction of PVP image is based on the EI number and sampling patch size. For example, the each EI size is  $27 \times 27$  pixels, if every time shift 1 pixels and every time sampling 1 pixels, which enable to shift 25 times from horizontal and vertical direction respectively. The PVP25 has been proven the high accuracy than PVP 16.

#### 5.1.4 Deep network Architecture Based Back-end

##### Architecture Comparison

In this section, the four deep learning models have been use for training, the architectures shown in the Figure. 5.6. Then the evaluation of each model and result. Therefore, four of the models is apart from these differences, the architectures share the configuration of based convolution neural network [36], and this allows us to directly compare the performance across different input designs. The time sequence has a notable capability that is proposed by peer works. It is based on the spatio-temporal feature description, and most works are using CNN architectures which combination of the RGB and optical flow CNN streams [35].

Based on these four types of models, the aim that provides a to afford a relatively com-

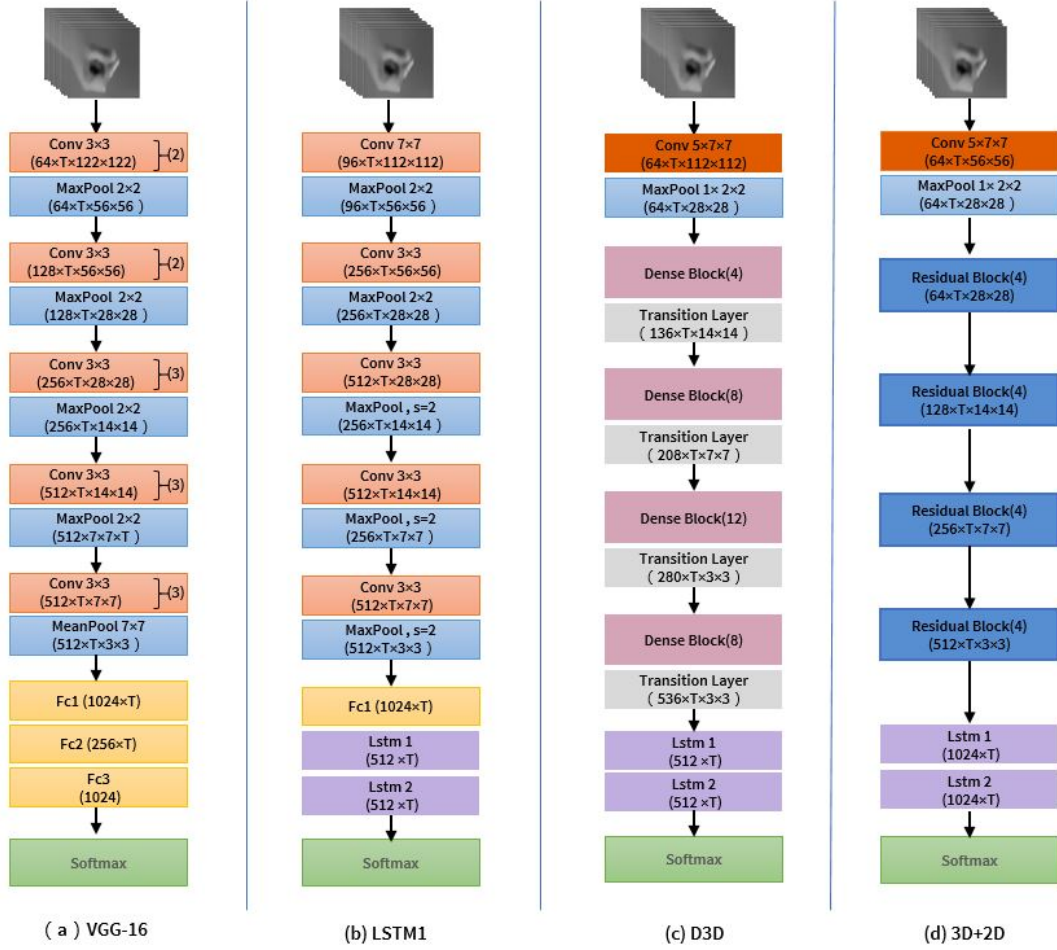


Figure 5.6: The four network architectures. (a) VGG-16,(b) LSTM-1,(C) D3D,(d) 3D+2D

plete analysis and comparison of the current advanced deep learning networks. In this comparison, fully 2D convolutional layers, fully 3D convolutional layers and mixing 3D and 2D convolutions layers have been testing performance on HoMG video data.

Convolutional neural networks (CNN) have four crucial elements: local connections, shared weights, pooling and the use of many layers, which are making major advances for solving problems. The good performance of the feature representation under image based on two dimensional convolution. Based on this performance, many peer work used to CNN to train the gesture and increase hand gesture recognition. One representative work [157] is using CNN to automatically extract the spatial and semantic feature from the gesture. Modified CNN structure has been proved to be effective against popular 2D gesture dataset such as Cambridge Hand Gesture Data set. The signals, sequence, and language usually use the 1D, images or audio spectrograms have a process by 2D, 3D usually has been used for video or volumetric images [98]. With the video dataset is main trend, the spatiotemporal feature learning become a crucial factor of deep deep 3-dimensional convolutional networks (3D ConvNets). However, different modalities use to process the various forms of multiple arrays. Based on 3D convolution front-end network is popular using lipreading recognition, which transforms the raw input video into spatial-temporal features based on three 3D convolutional layers. And its feed them to the following gated recurrent units (GRUs) to generate the final transcription [158]. Sharma et al [23]. proposed use the GRUs and LSTM obtaining these features while the final accuracy has not presented the merit. The 2D spatial convolutional layer is widely use to gesture recognition for extraction and classify feature in the spatial domain. Whereas the 3D temporal convolution layer becomes a new upsurge due to the video data popularization. With the developing of novel models, a combination of 2D convolutional layer and 3D convolution layer to achieve a new network has been using to training data by peer works. This type of architecture enhances the features and demonstrate their strong performances. For example, this model increased the accuracy of lip reading recognition [159] [158]. Enlightened by the good performance of high accuracy micro-movement recognition, therefore architecture enable to use for micro-gesture recognition.

The convolutional layers of VGG-16 and LSTM are completely composed of 2D convolutional layers. VGG-16 has achieved an appealing performance on the ImageNet challenge. LSTM as a model is often used to inference for temporal sequences. And this model is very popular to use for continuous dynamic movement recognition, such as facial emotion recognition and lipreading. “D3D” model transforms the 2D DenseNet into a 3D counterpart, which is the same as “3D+2D” front layers. Three models of “LSTM”, “D3D”, and “3D+2D” have back layers of same structure, which contains a two-layer LSTM and softmax to perform the final recognition. In order to better test the function

and performance of different functional layers, the number of layers remains the same as possible in all network structures. To perform a fair comparison, four models used softmax as the last layer, and LSTM, D3D, and 3D+2D are combined with a back-end network of the same structure which contains a two-layer bi-directional LSTM to perform the final recognition.

Considering that the amount of data in the HoMG database is smaller than that of ImageNet and other databases when selecting a deep network framework, the number of layers above the convolutional layers should be reduced to avoid over-fitting and reduce unnecessary waste.

The performance of the model combining 2D and 3D convolutional layers is better than fully 3D front-end, which was proven on the lip-reading. A probable reason for 2D convolutional layers for extracting fine-grained features in the spatial domain is a necessity, which is beneficial for discriminating similar lip movements. For micro-gesture movements, it has many similar characters to lip-reading movements and the performance may affect as well.

Consider a single image  $X_0$  that is passed through a convolutional network. The network comprises  $L$  layers, each of which implements a non-linear transformation  $H_l(\cdot)$ , where  $l$  indexes the layer.  $H_l(\cdot)$  can be a composite function of operations such as Batch Normalization (BN) [160], rectified linear units (ReLU) [161], Pooling [162], or Convolution (Conv). The output of the  $l^{th}$  layer is denoted as  $X_l$ .

Traditional convolutional feed-forward networks connect the output of the  $l^{th}$  layer as input to the  $(l + 1)^{th}$  layer [145], which gives rise to the following layer transition:  $X_l = H_l(X_{l-1})$ . The algorithm for ResNet can be understood as follows [34] :

$$X_l = H_l(X_{l-1}) + X_{l-1} \quad (5.2)$$

To further improve the information flow between layers, a different connectivity pattern is proposed [34]: direct connections are introduced from any layer to all subsequent layers. Consequently, the  $l$ th layer receives the feature-maps of all preceding layers,  $X_0, \dots, X_{l-1}$ , as input:

$$X_l = X_l([X_0, X_1, \dots, X_{l-1}]) \quad (5.3)$$

where  $([X_0, X_1, \dots, X_{l-1}])$  refers to the concatenation of the feature-maps produced in layers  $X_0, \dots, X_{l-1}$ .

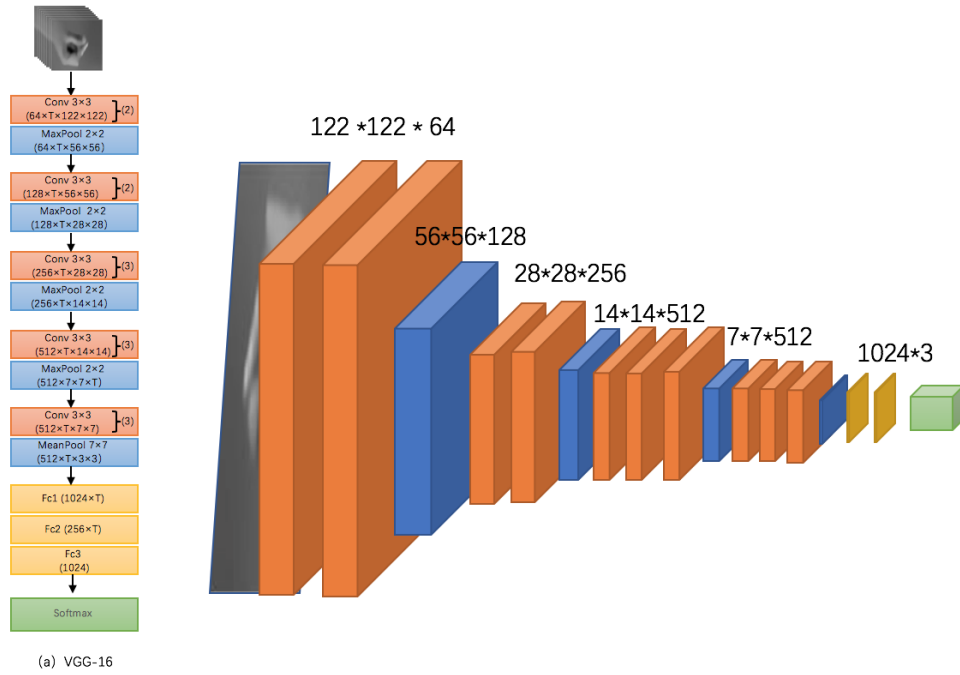


Figure 5.7: VGG-16 is classic deep learning model, which consists of the 13 2D convolutional layers, 5 pooling layers, 3 fully connect layers and softmax layer.

### VGG-16

With the wide development of deep learning, a number of researchers have started to take advantage of them to tackle the micro-gesture recognition problem. In this research, a classic VGG architecture has been used, known as Visual Geometry Group, which is a group from Oxford, UK. In 2014, they published the 16-layer and 19-layer models, which have achieved successful performance on ImageNet [146]. This architecture uses multiple small kernel-sized filters to replace the large kernel-sized filters, which could capture the high-accuracy movements and reduce the information redundancy [146]. VGG-16 and VGG-19 are the networks they trained for visual object recognition amongst 14 million images of over 20,000 categories [163]. This has shown good performance on many of the different object databases. VGG-16 network contains 16 layers in total. It contains 13 convolutional layers, 3 fully connected layers, 5 pooling layers, and 1 softmax layer as output layer. Through filters, the input image generates the output feature map of the same size. Each of the convolutional layers has a kernel size of  $3 \times 3$ , stride of 1 and padding of 1.

This property makes it much appropriate for our task. In the training process, the input PVP videos into several non-overlap clips are divided and a frame is randomly selected from each clip to generate a new short sequence. Therefore, the input are fixed-size  $122 \times$

122 grey scale images. The images are passed through a stack of convolutional layers of filters with very small  $3 \times 3$  receptive field.  $1 \times 1$  convolutional filters have been used for a linear transformation of the input channels, which follows non-linearity. The maximum pooling replaces the pixel values with the largest pixel value in the corresponding area of the filter, which is used for dimension reduction. Here, the filters used by the pooling are  $2 \times 2$  in size, so the image size obtained after pooling is  $1/2$  of the original size. There are five maximum pooling layers, which follow the convolutional layers. It is noted that in this experiment VGG-16 consists of 13 convolutional layers, 3 fully connected layers, 5 pooling layers, and 1 softmax layer as output layer. Each of the convolutional layers has a kernel size of  $3 \times 3$ , stride of 2 and padding of 1. Three fully connected layers follow a stack of convolutional layers. Each neuron in the fully connected layer is connected to each neuron of the upper layer, and the output features of the previous layer are integrated. In VGG-16 architecture proposed by Simonyan et al. [164], the first fully connected layer FC1 has 4096 channels, and the upper pool have 51200 channels. Similarly, the second fully connected layer FC2 has 4096 channels, and the last FC3 has 1000 neurons. In my experiment, three fully connected layers have 1024 channels on each layer. As the end of the architecture, a softmax layer is used for classification and normalisation. And final softmax equation shown equation 5.4.

$$f(z_j) = \frac{e^{z_j}}{\sum_{i=1}^n e^{z_i}} \quad (5.4)$$

The standard exponential function is applied to each element  $z_j$  of the input vector  $z$  and these values are normalized by dividing by the sum of all these exponentials. This normalization ensures that the sum of the components of the output vector is 1.

All these short sequences will be used as individual samples to be fed into the network for training. This can be regarded as a special type of data augmentation, which is helpful for the network's learning process. The network has to learn to discriminate the correct regions of the gestures in each frame where both the distortions of the corner and the noise pixels exist. For the test process, multiple different PVP videos corresponding to the single test video are obtained and majority voting is performed based on the predictions from all the PVP videos to obtain the final prediction.

## LSTM

Long short-term memory(LSTM) is a type of recurrent neural network(RNN). LSTM model was commonly used in emotion recognition, face recognition, and so on. Kuchaiev et al. [148] proposed to combine Bidirectional LSMTs and CNN to recognize the six types

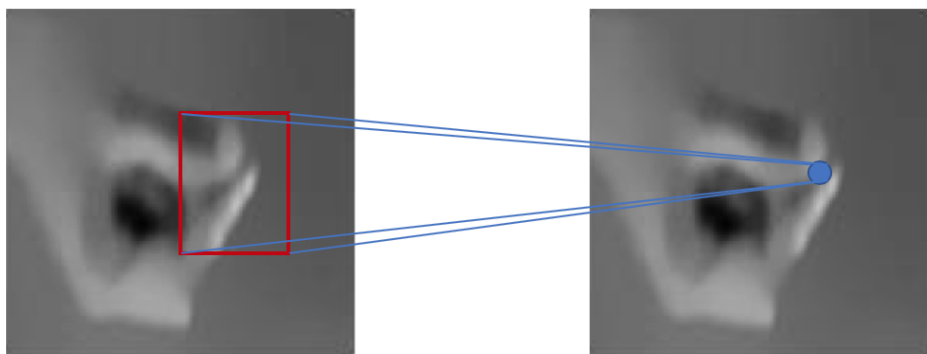


Figure 5.8: 2D convolutional layer generally used on Image data, the kernel slides are along 2 dimensions on the data.

of 3D hand gesture and obtained high accuracy. However, this gesture recognition system captured hand skeletal by a Leap Motion. In this experiment, the same convolutional layers of VGG-16 model has been used to make a fair comparison. LSTM-1 consists of the 2D convolutional layers and LSTM layers. 2D convolutional layers are used to extract image features, while the LSTM layers are used to remember the temporal information through the input gate, output gate, and forget gate. Finally, the softmax is used for classification.

### DenseNet in 3D Version (D3D)

In the previously research works, the different action categories are hard to be captured in the real-world scenarios. There is a significant performance improvement of 3D CNN over 2D CNN, because 3D CNN can represent temporal features. Kurmaji et al. [165] proposed a novel method, which used 3D convolutions to extract feature and classify gestures more accurately. Their experiment indicates that the 3D CNN model has significant advantages compared with frame-based 2D CNN for most training tasks, especially in small positive training samples.

DenseNet have more narrow and fewer parameters than traditional convolutional networks, that would benefit the each layer has direct access to the gradients from the original input signal and loss function.

DenseNet in 3D Version (D3D) model was created by Yang et al. in 2018 [158]. This model transforms the 2D DenseNet into the 3D counterpart, which is fully 3D convolutional architecture. DenseNet is more advanced than ResNet, DenseNet generates the

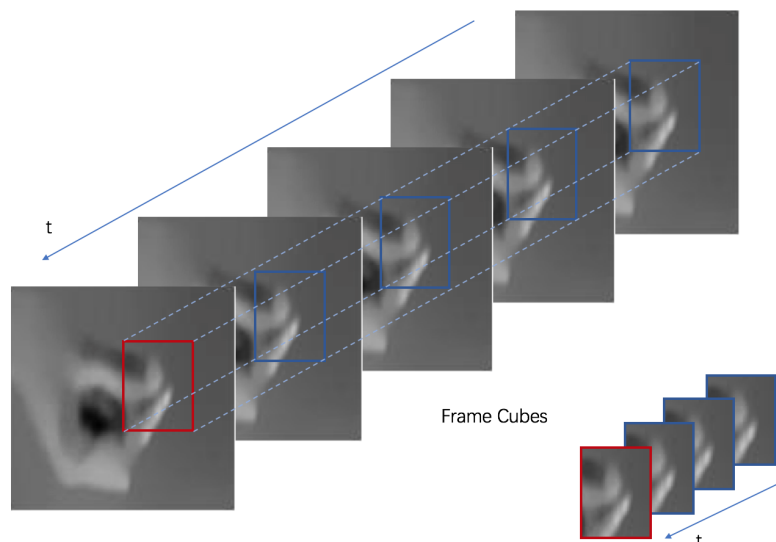


Figure 5.9: The kernels are applied in 3D convolution,  $t$  is temporal.

feature map from different layers, And keeps feature sizes consistent in transition layers. The Figure. 5.10 shows the D3D model details and it is used for training and recognition of lipreading, because the characteristic of the dataset is complex. In this experiment, similar characteristics between micro-gesture movement and lip-movement were found. So we believe that commonalities are able to achieve a good performance in the same architecture.

### 3D + 2D

The 2D convolutional layers extract features from local neighbourhoods on the feature map of the previous layer, and only use 2D spatial-dimension feature maps. 3D convolutional layers apply a 3D convolution kernel to the cube formed by stacking multiple consecutive frames. In this configuration, each feature map in the convolutional layer is connected to multiple adjacent consecutive frames in the previous layer, thus capturing motion information.

Many researchers have investigated the potential of 2D over 3D CNN's for representing temporal features and hand gesture classification in videos. In particular, the sequence of frames of gestures is mapped to a chronological pattern to capture the dynamics of hand motion in a single frame. Therefore, both 2D and 3D convolutional functions are used to achieve better performance. Many research works take advantage of combining 2D and 3D convolutional layers [158] [146] [155]. In this experiment, the first layers of the model are the same as those of D3D. Hence, this architecture consists of 3D convolutional layers, Residual Blocks, and two LSTM layers.



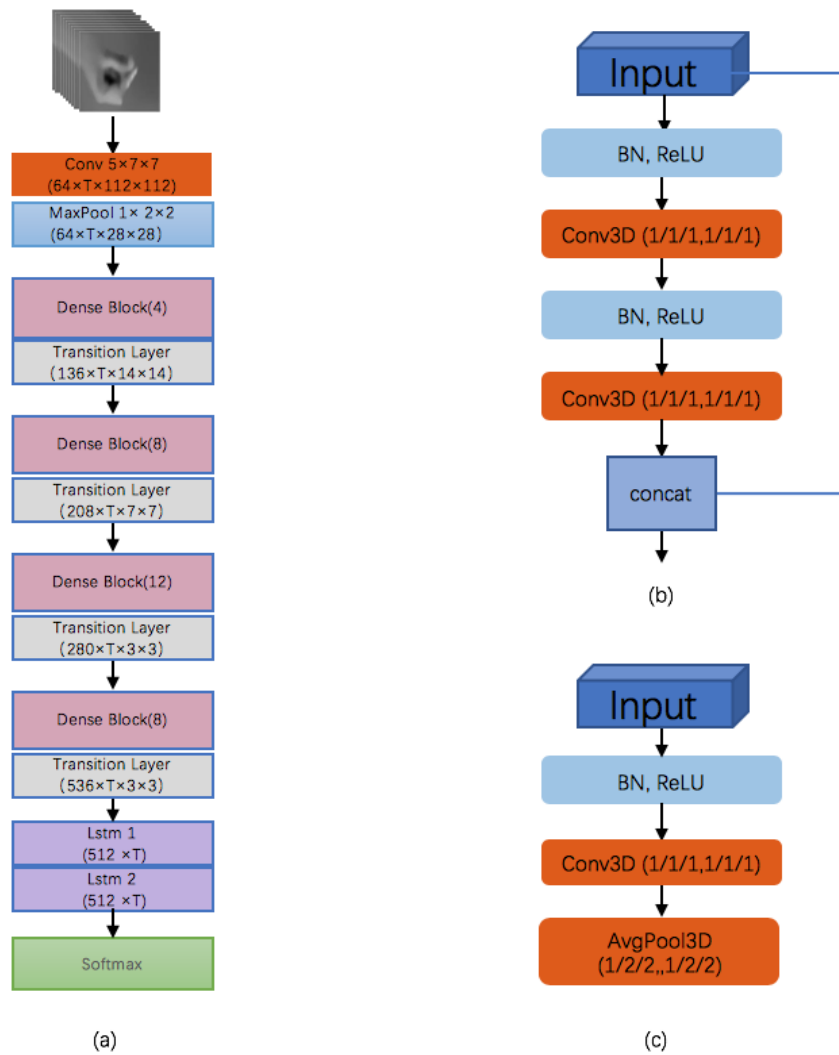


Figure 5.10: The D3D network(DenseNet in 3D version) (a) The D3D model (b) The structure of each Dense Layer (c)The structure of each Trans Block

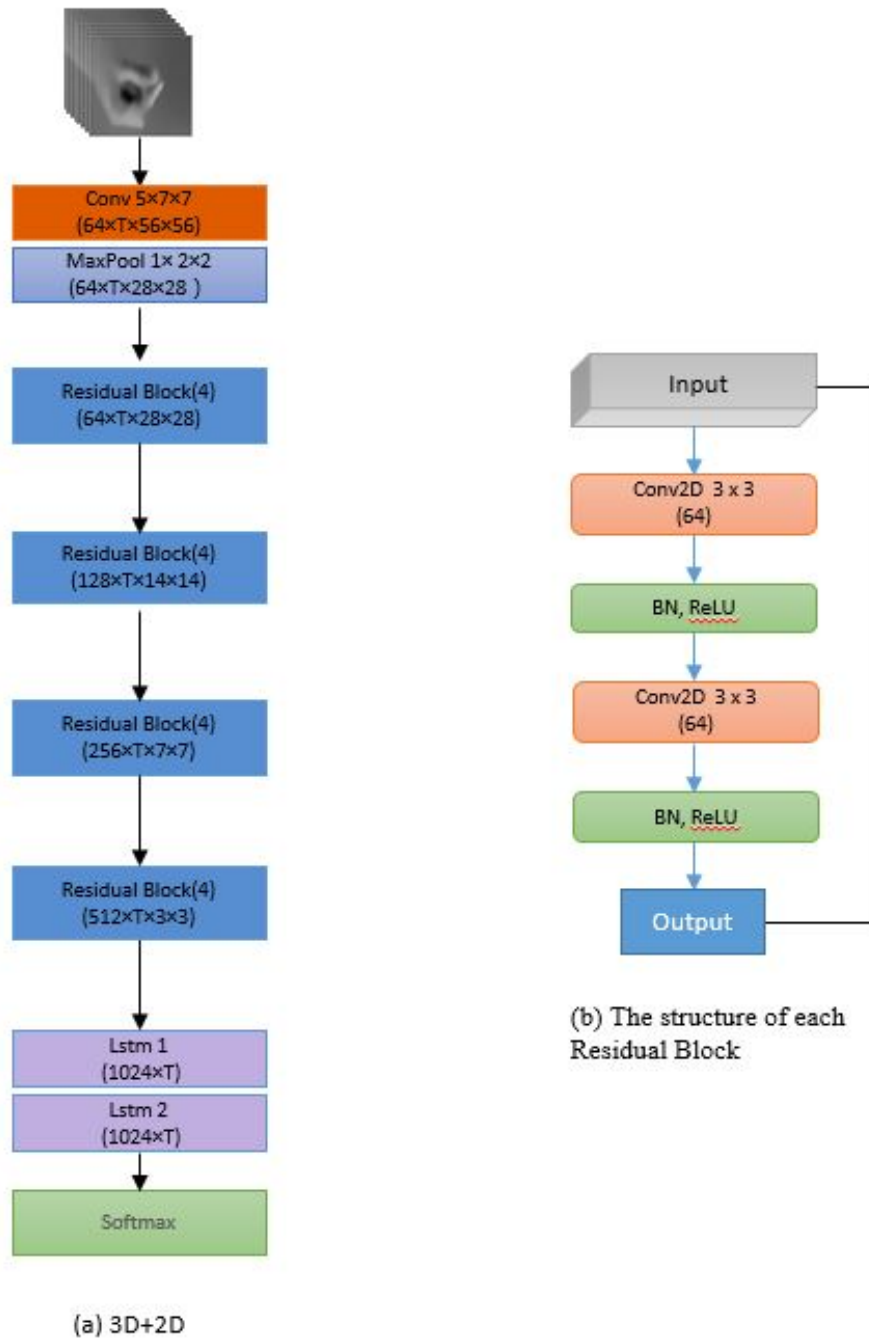


Figure 5.11: The architecture of mixing 2D and 3D convolutional layers (a) The 3D+2D model, (b) The structure of each Residual block Layer

It is noted that the 3D convolutional kernel can only extract one type of feature from the cube, because the weight of the convolutional kernel is the same in the entire cube. That is, the shared weight is the same convolutional kernel (The lines of the same colour in the figure represent the same weight). Multiple convolutional kernels can be used to extract multiple features.

The mixed 2D and 3D convolutional architecture combines the functions of 2D and 3D convolutions. 2D convolutional layers extract the fine-grained features in spatial domain and discriminate the similar movements of micro-gestures. 3D convolutional layers have proved effective in action recognition, because the spatial-temporal convolutional layer is able to capture the temporal dynamics in the sequence. In this experiment, the model of a 2D residual network is referred to two spatial-temporal convolutional layers to extract the relevant features (Stafylakis and Tzimiropoulos [166]). Meanwhile, this model has proved successful in lip movements.

In this experiment, our architecture is based on the model in [166], which is a combination of spatiotemporal convolutional, residual blocks and bidirectional Long Short-Term Memory networks. Each block consists of two convolutional layers, batch normalization (BN) and rectified linear units (ReLU). This architecture was used for ImageNet. Residual blocks gradually reduces the spatial dimension with the max pooling layer until its output becomes a single dimensional tensor at each time step.

## 5.2 Experiments

### 5.2.1 Database

The HoMGR dataset contains 3 gestures, i.e, Button, Slider, and Dial. In this chapter, focus on the video subset, HoMGR video subset is consists of the 40 subjects and each subject is involved 24 videos due to a different setting and three classify gestures. There are 25 frames per second in each video, and the length of the videos are from a few seconds to 20 seconds. The each frame resolution is  $1920 \times 1080$  pixels. Each EI is selected to be size  $27 \times 27$  pixels and the patch to be  $5 \times 5$  pixels in each EI.

### 5.2.2 Parameters Setting

Our implementation is based on the PyTorch and the models are trained on servers with four NVIDIA Titan X GPUs with 12GB memory. Training a single net took 2–3 days de-

pending on the different architectures. We use the Adam optimizer with an initial learning rate of 0.001, with  $\beta = (0.9, 0.99)$ .

### 5.2.3 H3D Pre-processing

#### Viewpoint(VP)

For viewpoint image extraction, the method is similar with the HoMG image subset proposed the method, which means the incomplete EIs on the four edges in the images need remove. The remove approach is based on the minimum value of the H3D row and column in the boundaries and this method was mentioned in the image data. After cutting of the incomplete EI, there produce 68 and 37 EIs on the horizontal and vertical direction respectively. Then, each patch of  $5 \times 5$  pixels is sampled, which may reduce the dark line effects of not straight lines. Finally, a result frame of  $340 \times 185$  pixels is obtained for each viewpoint, as shown in Figure. 5.2 (b). In order to acquire different depth information from data, 16 viewpoints in total are extracted, by shifting 1 pixel along the horizontal and vertical direction 4 times respectively. It noted that the 16 viewpoints are maximum extraction from  $27 \times 27$  EIs, because every EIs have dark boundaries affect, that part cannot achieve VP extraction.

#### Pseudo Viewpoint(PVP)

For pseudo viewpoint image extraction, each patch is selected in the same way but the incomplete EIs have not removed. Therefore, H3D frames are resized for  $1836 \times 999$  pixels, which match the  $68 \times 37$  EIs. Then a similar selection and shift process as detailed above is applied to obtain the final frames. The final resolution of each frame is  $340 \times 185$  pixels, resolution of the PVP and VP are same, that ensure the fairness of competition. One resulting PVP image is shown in Figure. 5.2 (a) and we can easily see the noisy pixels. The noise and interference factors of the extracted viewpoint images are random. Although some viewpoint images contained a lot of noise, their appearance in three gesture frames are relatively balanced, which also ensures the fairness of the three gesture predictions. In order to obtain the more information of the deep and view, PVPs extraction have been shifting 1 pixel along the horizontal and vertical direction 5 times respectively. Figure. 5.5 shows the PVP16 and PVP25 shifting direction.

All the VP and PVP images are converted to grayscale and normalized with their mean and variance. When fed into the models, the frames in each sequence are cropped in the same random position for training and development, then cropped in the centre position for testing. All the frames have been resized to  $122 \times 122$  pixels and then cropped to  $112 \times 112$  pixels for training and testing. In the meanwhile, the frames are randomly flipped

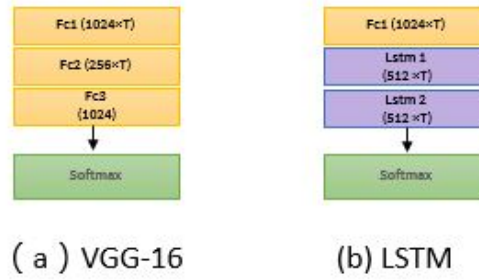


Figure 5.12: Initial testing on multilayer perception and LSTM. (a) is based on the VGG-16 last four layers (b) is based on the LSTM last four layers.

to avoid overfitting. It is noted that multiple tower architectures are used, this model is referred by [146]. For example, VGG input size is  $122 \times 122 \times T$ ,  $T$  is number of Towers. It is no explicit time-domain connectivity between frames before convolutional layers, and  $T=16$  towers with common convolutional layers (with shared weights), each of which takes an input frame. Figure. 4.2 (c) shows the frames have been selected and input in the models. Training frames have been random selected and testing frames only used middle number of each video as the testing data input the model. And the convolutional layers linked an maxpool layer, which has been activated and concatenate along a new dimension. Moreover, 2D convolutions and 3D convolutions are performed in the same manner. The Tower has been activated, which are concatenated channel-wise after first maxpool, and produced an output activation with 1200 channels. The subsequent  $1 \times 1$  convolution is performed to reduce this dimension, to keep the number of parameters at convolutional layer at a manageable level.

## 5.2.4 Experiment Results

For the back-end experiments have three parts: Firstly, as the first aim is test different function layer work on the HoMG video data. The VGG-16 and LSTM are used same 2D convolutional layers to test the performance of the softmax layer and Bidirectional LSTM layers. Therefore, the 2D convolution layer forward input feature layers have been closed, the results of output only shown the last four layers performance in the preliminary test and the results shown Table. 5.1. The importance of the convolutional layer was also verified in this experiment. Subsequently, the convolutional layers of the two networks were opened to retest the performance of the model on the database for comparison with the latter two networks and the results shown in Table. 5.4.

The second experiment is using the VGG-16 network which is through a series of con-

Table 5.1: Classification accuracy (%) initial testing on multiplayer perception and LSTM layers. "M.V." means "majority voting".

Model	VGG-S.M.	M.V.	LSTM-1	M.V.
Original	55.83%	-	63.33%	-
VP16	58.96%	60.42%	67.06	68.08%
PVP16	45.63%	48.33%	58.67%	61.67%
PVP25	53.98%	57.92%	73.07%	77.50%

Table 5.2: Classification accuracy (%) comparison between the proposed methods and Majority Vote on the testing dataset of HoMG dataset. + M.V. means the model results used Majority Vote.

	VGG-16+M.V.	LSTM+M.V.	D3D+M.V.	3D+2D+M.V.
VP16	83.27	83.21	77.92	83.33
PVP16	82.50	85.10	81.25	85.01
PVP25	85.42	85.10	83.75	<b>85.83</b>

volutional layers and pooling layers to out put a data feature, then use three full connection layers, finally, using softmax of regularization to get final result. This structure has achieved appealing performance on the ImageNet Challenge and famous boom. The second network is LSTM-5 network that is proposed by Chung et al. [146]. It is based on multi-tower structure to faster to train deep models. The third network was using for lipreading area which involve front-end of three spatial temporal convolutional layers. The fourth network transformed by DenseNet which apply the fully 3D convolutional layers as front-end. It is noted that our training process does not use pre-trained models, because our task is different from lipreading. It is noted that most PVP accuracies are higher than VP, and the PVP25 accuracies are higher than PVP16. Because of the more PVP have more 3D information as well as give the models are more useful training information.

### Majority Vote

In order to further improve the final accuracies, the majority vote has been applied to fusion the final prediction label on each VP. Then, using the each VP's accuracies to vote the most likely results. The final result shown in the Figure. 5.2. In order to further understand the work process of majority vote Figure. 5.13 shown the accuracies of each models, the red points represent the wrong predictions, the black represent the right predictions. Most results used majority vote able to improve 1-3%, some results failed to greatly increase because some of gestures have many similar movements and hard to recognize.

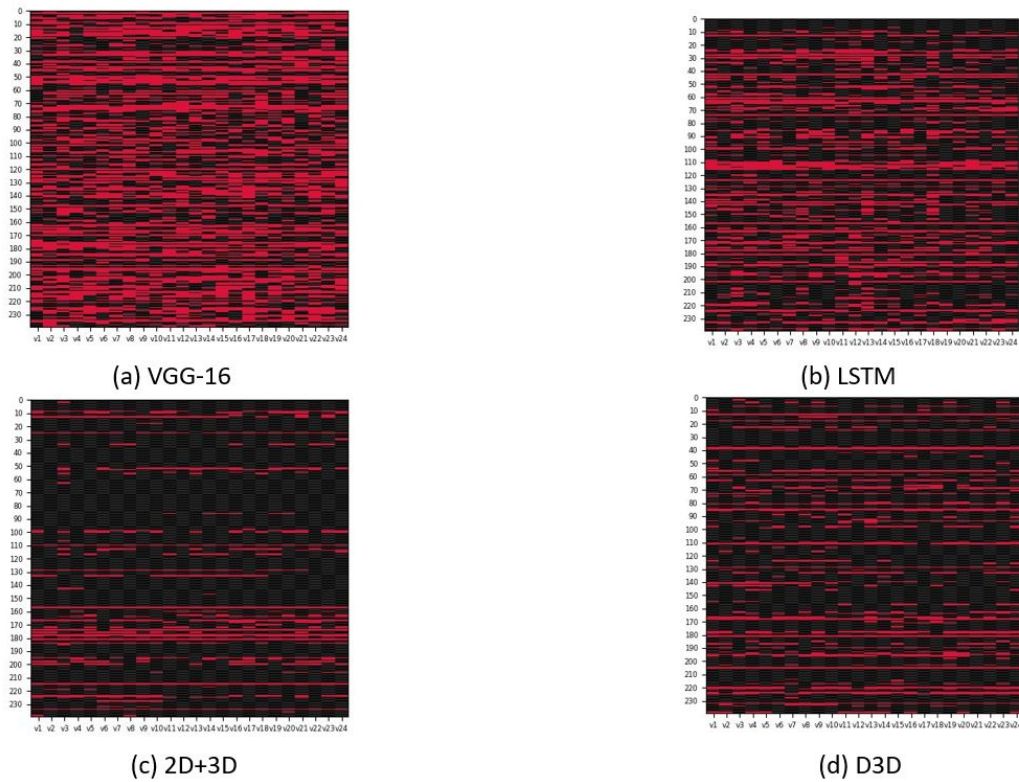


Figure 5.13: Each VP accuracies input the majority vote to fusion. (a)VGG-16 (b)LSTM (c)2D+3D (d)D3D

Table 5.3: Classification accuracy (%) comparison between the proposed methods on the development dataset of HoMG dataset.

	VGG-16	LSTM-1	D3D	2D+3D
Original	48.25	51.66	72.5	79.16
VP16	82.68	61.43	75.67	80.29
PVP16	93.21	85.40	90.13	94.79
PVP25	94.02	94.37	<b>94.70</b>	94.13

Table 5.4: Classification accuracy (%) comparison between the proposed methods on the testing dataset of HoMG dataset.

	VGG-16	LSTM	D3D	3D+2D
Ori	60.83	62.50	68.75	78.75
VP16	82.03	82.31	77.92	83.33
PVP16	82.01	84.27	80.05	83.37
PVP25	84.67	84.48	83.06	<b>85.07</b>

### 5.2.5 Comparison with State-of-Art Method

To avoid the tedious extraction of exact viewpoint images, this section proposes to simply sample pseudo viewpoint images and thus presents an innovative 3D micro-gesture recognition method. This is performed by extracting the features of several depth planes to generate PVP based descriptions and then learning the gestures' pattern by a deep neural network based back-end. Comprehensive experiments are carried out and the results have outperformed all the state-of-the-art methods that evidently prove the robustness and effectiveness of our method. Because the proposed method is able to be expanded for any H3D recognition problems and the deep neural network based back-end testing could be reference for future works.

Table. 5.6 shows the 3D+2D model achieved best accuracies, which use the PVP based front-end and the traditional VP based front-end. The VGG-16 is classic architecture work for image feature extraction. The LSTM architecture, although it relies only on the 2D convolutional layers, the bi-directional LSTM still working on it. The 3D convolution benefit for capturing short-term motion information, which has been proven important information for micro-gesture movements. However, the fully 3D model cannot surpass the model combining 3D and 2D convolutional layers and those four models have not much different in the results. Hence, this result proves the necessity of 2D convolutional layers for extracting fine-grained features in the spatial domain, which is quite useful for micro-gesture recognition. In addition, Table. 5.6 can be easily seen that PVP based front-end performs better than VP based front-end. May the PVP extraction without pre-processing and noise pixels do not interfere with model training. And VP25 performance better than VP16, which maybe due to the VP25 providing more useful information. On the development set shown Figure.5.3, our PVPs based method gives an accuracy of 90.5% and 94%,



Table 5.5: Classification accuracy (%) development dataset of HoMG dataset.

	ResNet152	DenseNet161	SE-ResNet50
Video training set [22]	93.00	93.00	92.00
Video development set [22]	90.00	87.00	90.00
Video test set [22]	82.00	82.00	82.00

Table 5.6: Classification accuracy (%) comparison "M.V." means "majority voting".

	Dev	Dev+M.V.		Test	Test+M.V.
Original	48.3	-	Original	60.8	-
VP16	82.7	83.3	VP16	82.0	83.3
PVP16	90.5	94	PVP16	82.2	83.5
PVP25	92.0	94.1	PVP25	84.7	85.4

which is about 10% higher than the traditional viewpoint based method. On the test set, our method achieves 85.4%. Zhang et al. [22]. achieved 90% on development set, and [22] used three deep learning models gave the accuracies shown 5.5. It is noted that all the methods perform better on the development set.

Table 5.7 shows the results of the proposed approach in comparison with other models. The initial results [19] show that LBPTOP and LPQTOP features with SVM classifier can predict the 3D micro-gestures with an accuracy of 59.5% and 66.7% respectively. In Sharma's paper [23], they used the original frames in the H3D video directly with Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) and reported an accuracy of 69.2%. In 2018, Zhang et al [22] used three advanced deep networks, ResNet152, DenseNet161 and SeResNet50, together with majority voting and achieved an accuracy of 82%. In our approach, only a single VGG network is used and an accuracy of 85.4% is achieved, which surpasses all existing results and clearly demonstrate the effectiveness of our method shown in Figure.5.14.

### 5.3 Conclusion

This section presents a novel effective H3D based micro-gesture recognition system, which is end-to-end model. For front-end network, there are two methods have been used for evaluation. The VP front-end follows the traditional H3D image method to extract and reconstruct the multi-viewpoint images, and the PVP front-end is a novel idea to create a simple front-end using for deep learning. In order to evaluate the performance, there four advanced deep networks have been used for training. Compared with tradi-

Figure 5.14: Classification accuracy (%) comparison between the proposed method with all the existing methods on the testing subset of HoMG dataset.

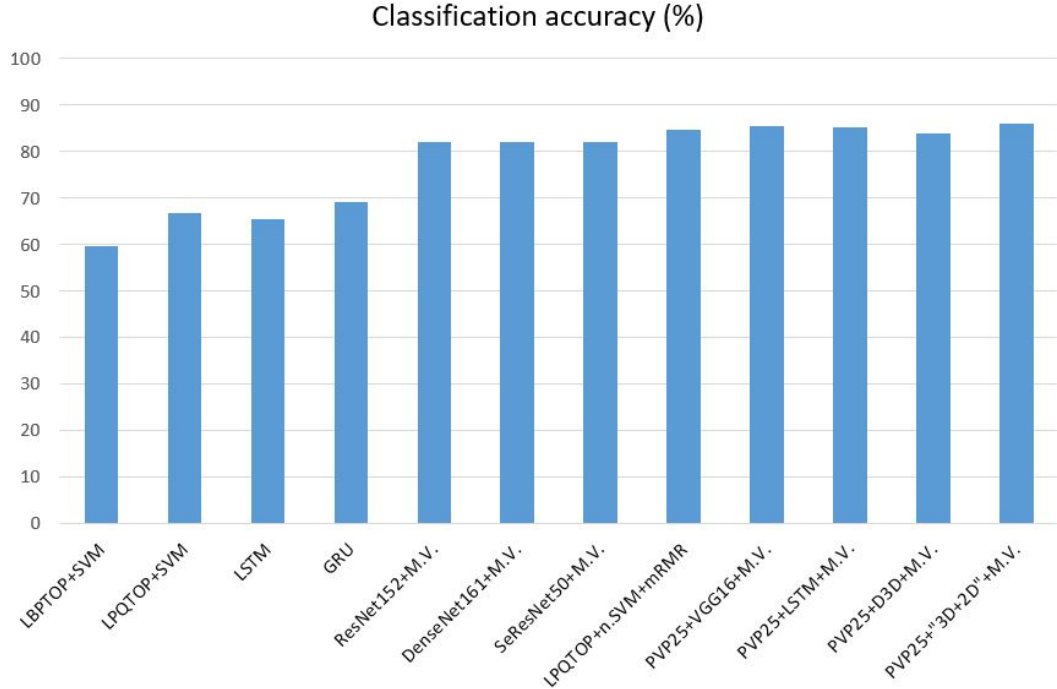


Table 5.7: Classification accuracy (%) comparison between the proposed method with all the existing methods on the testing subset of HoMG dataset. "M.V." means "majority voting" GRU means "Gated Recurrent Unit". "n.SVM" means "non-Linear SVM"

Author	Methods	Acc
Liu et al.[19]	LBPTOP+SVM	59.5
Liu et al.[19]	LPQTOP+SVM	66.7
Sharma et al.[23]	LSTM	65.4
Sharma et al.[23]	GRU	69.2
Zhang et al.[22]	ResNet152+M.V.	82.0
Zhang et al.[22]	DenseNet161+M.V.	82.0
Zhang et al.[22]	SeResNet50+M.V.	82.0
Qin et al.[167]	LPQTOP+n.SVM+mRMR	84.6
The proposed works	PVP25+ VGG16 +M.V.	85.42
	PVP25+ LSTM +M.V.	85.10
	PVP25+ D3D +M.V.	83.75
	PVP25+'3D+2D'+M.V.	<b>85.83</b>

tional VPs front-end, the PVP has remarkable improve. PVP experiment is benefit to the H3D imaging system further development, because the pseudo viewpoint extraction method is different from the past complex handcrafted feature extraction. This purpose of this method is to replace the traditional complex extraction methods and improve the understanding and learning of H3D image systems in deep networks. In order to identify the effectiveness of this method, the method and quantity of the viewpoint/ pseudo viewpoint extraction were verified and the influencing factors were compared in the experiment.

For the four back-end networks, there are four architectures experiments consist of the three parts. It is noted that the input size and frames in the time sequence has been adjusted that in order to keep consistency. The different models given the special function layer to training the data and given the feed backs. The first experiment is testing the single softmax and LSTM function layers. Then, the second experiment used the same front 2D convolutional layers to test the VGG-16 and LSTM of the different back layer. The third experiment is used the same 3D convolutional front layers to compare the ResNet and DenseNet of different back layers. The purpose of these experiments is understanding the H3D data performance on deep learning models. Finally, in order to further improve the accuracy of video-based database, the majority vote has been applied.

## Chapter 6

# Conclusion and Future Work

## 6.1 Conclusions

### 6.1.1 HoMG dataset and baseline

Micro-gesture recognition based on the holoscopic 3D imaging system is a novel topic, which includes the human centre interaction, gesture design, holoscopic 3D imaging system and gesture recognition. The aim of this research create a novel micro-gesture interaction system in order to improve H3D micro-gesture recognition performance. Therefore, the first micro-gesture database (HoMG) based on the holoscopic 3D imaging has been designed and created. This design use a novel advanced gesture interaction method for application, which can use for disabled facilities. HoMG database record by H3D camera, which offer RGB continuous video and image with 3D information. Two types of data are riches the models and encourage different algorithms. In order to evaluate and promote HoMG database development, the video subset baseline used LBPTOP and LPQTOP to extract the feature and the image subset baseline used LBP and LPQ to extract the feature, then used three classical methods of machine learning to classify the three type gestures, the result has been published for encourage the more people to join this research. And this database holds an international challenge in IEEE automatic face and gesture recognition 2018 and four international groups joined this challenge and proposed a serial of novel methods for HoMG further developing.

### **6.1.2 Image-based Micro-gesture recognition**

The novel system has been proposed to tackle the problem of using H3D special information and improve the performance based on image subset, and the results surpass all the existing methods on HoMG image subset. Therefore, the holoscopic 3D imaging of 3D information benefit the CNN model. Firstly, the H3D imaging system has been understanding to improve the 3D micro-gesture system. Then, the automatic cut out algorithm has been proposed for large-scale H3D data pre-processing. The principle of extraction and re-construction method enable to use, the particular extract patch size, re-constructed VP image number and size are based on the after cutting out H3D images. The 16 view-points have been re-constructed, which are from different depth layers. For developing the micro-gesture recognition, the 16 VPs input to CNN model with attention block, which is used to extract features from each VPs, then predict each label possibility. Finally, in order to further improve the accuracy and performance, the five fusion decision methods have been tried for improving the recognition accuracy. The bagging classification tree method has been proven the best performance, and the final accuracy surpassed all the existing methods on the HoMG image-based database.

### **6.1.3 Video-based Micro-gesture recognition**

It propose an efficient and robust H3D micro-gesture recognition end-to-end system, which have innovative Pseudo View Points (PVP) based front-end and efficient deep networks based back-end. PVP front-end is tackled the tedious 3D information extraction problem. While the VP16, PVP16 and PVP25 have been used for comparison, the PVP performance better than traditional VP front-end. For the back-end, the four advanced networks: VGG-16, LSTM, D3D and 3D+2D have been applied for training the HoMG video data, which test data on basic softmax layer, bi-directional LSTM layer, fully 2D convolution network, 3D convolution network, and combing the 3D and 2D convolution network. The performances of four deep networks have been evaluated. In order to further improve the performance, majority voting is utilised to further improve the accuracies. Finally, the proposed method achieved 85.83% which is state-of-the-art performance and outperforms all existing methods on a public database.

## **6.2 Future Work**

For the database design and creation, there are three micro-gesture movements have been recorded in HoMG database, further work should acquire more gesture types. In addition, there are still some issues with the proposed micro-gestures database that can be worked

on to improve the database further, for example, the micro-gesture design only focus on the wearable devices, the three types of micro-gestures are similar with the start movements, and data record backgrounds are too clear. Therefore, further improvements can be to develop more types of gestures, which are not only limited to manipulative micro-gestures and wearable applications. With the development of the Internet of Things (IoT), more scenarios and manipulations should be considered. Besides, the HoMG database used camera and microlens array can be improved, with a specific focus on noise reduction which will also reduce data pre-processing procedure.

For the video-based subset, there are other areas of improvement that possibly make the descriptor more robust and more feature character for training and classification. For future improvements of performance and accuracy. Firstly, it can be extract more 3D information from PVP, and the models can be modified and created based on the data features. Then, it can be use the varied fusion decision methods to improve accuracy. Finally, the recognition accuracy should be improve due to satisfy the requirements of most applications.

# Bibliography

- [1] Google Soli Project. Virtual tool gestures. URL: <https://atap.google.com/soli/> [cited 04.08.2019].
- [2] Leap Motion. Leap motion's software and hardware platform brings your bare hands directly into virtual and augmented reality. URL: <https://www.leapmotion.com/technology/> [cited 01.02.2019].
- [3] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction*, 9(3):171–193, 2002. URL: <http://portal.acm.org/citation.cfm?doid=568513.568514>, doi:10.1145/568513.568514.
- [4] 3D Stereo. MS Windows NT kernel description. URL: [www.3dstereo.com](http://www.3dstereo.com) [cited 25.09.2018].
- [5] Computer Desktop Encyclopedia. Computer desktop encyclopedia. (2010)3d glasses. URL: <https://www.pcmag.com/encyclopedia/term/60650/3d-glasses/> [cited 05.07.2018].
- [6] Obaidullah Abdul Fatah. *Post-production of holoscopic 3D image*. PhD thesis, Brunel University London, 2015.
- [7] ToF. What is time of flight camera and how to use one. URL: [https://www.asus.com/3D-Sensor/Xtion\\_PRO\\_LIVE/](https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/) [cited 04.10.2019].
- [8] A. Aggoun, E. Tseklevs, M. R. Swash, D. Zarpalas, A. Dimou, P. Daras, P. Nunes, and L. D. Soares. Immersive 3D holoscopic video system. *IEEE MultiMedia*, 20(1):28–37, Jan 2013. doi:10.1109/MMUL.2012.42.
- [9] O. A. Fatah, A. Aggoun, M. Nawaz, J. Cosmas, E. Tseklevs, M. R. Swash, and E. Alazawi. Depth mapping of integral images using a hybrid disparity analysis

- algorithm. In *IEEE international Symposium on Broadband Multimedia Systems and Broadcasting*, pages 1–4, June 2012. doi:10.1109/BMSB.2012.6264257.
- [10] Alfredo Canziani, Eugenio Culurciello, and Adam Paszke. Analysis of deep neural network models. pages 1–7. arXiv:arXiv:1605.07678v4.
- [11] Dietrich Kammer, Jan Wojdziak, Mandy Keck, Rainer Groh, and Severin Taranko. Towards a Formalization of Multi-touch Gestures. *ACM Int. Conf. Interact. Tabletops Surfaces*, 3/94:49–58, 2010. URL: <http://portal.acm.org/citation.cfm?id=1936662>, doi:10.1145/1936652.1936662.
- [12] Shahrouz Yousefi, Mhretab Kidane, Yeray Delgado, Julio Chana, and Nico Reski. 3D gesture-based interaction for immersive experience in mobile VR. *Proc. - Int. Conf. Pattern Recognit.*, pages 2121–2126, 2017. doi:10.1109/ICPR.2016.7899949.
- [13] M. R. Swash, O. Abdulfatah, E. Alazawi, T. Kalganova, and J. Cosmas. Adopting multiview pixel mapping for enhancing quality of holoscopic 3D scene in parallax barriers based holoscopic 3D displays. In *2014 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pages 1–4, June 2014. doi:10.1109/BMSB.2014.6873560.
- [14] John M. Carroll. What is human-computer interaction (HCI). URL: <https://www.interaction-design.org/literature/topics/human-computer-interaction1> [cited 10.10.2016].
- [15] J.K. Aggarwal and M.S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, April 2011. URL: <http://doi.acm.org/10.1145/1922649.1922653>, doi:10.1145/1922649.1922653.
- [16] Spencer D Kelly, Sarah M Manning, and Sabrina Rodak. Gesture Gives a Hand to Language and Learning : Perspectives from Cognitive Neuroscience , Developmental Psychology and Education. 4:569–588, 2008.
- [17] Hong Cheng, Zhongjun Dai, and Zicheng Liu. IMAGE-TO-CLASS DYNAMIC TIME WARPING FOR 3D HAND GESTURE RECOGNITION Hong Cheng , Zhongjun Dai , Zicheng Liu University of Electronics Science and Technology of China , Microsoft Research Redmond. *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. doi:10.1109/ICME.2013.6607524.
- [18] Atheerair. Augmented reality for the enterprise. URL: <https://atheerair.com/> [cited 04.08.2017].



- [19] Yi Liu, Hongying Meng, Mohammad Rafiq Swash, Yona Falinie A Gaus, and Rui Qin. Holoscopic 3D Micro-Gesture Database for Wearable Device Interaction. *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 802–807, 2018. doi:10.1109/FG.2018.00129.
- [20] Tao Lei, Xiaohong Jia, Yuxiao Zhang, Yanning Zhang, Xuhui Su, and Shigang Liu. Holoscopic 3D Micro-Gesture Recognition Based on Fast Preprocessing and Deep Learning Techniques. *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 795–801, 2018. URL: <https://ieeexplore.ieee.org/document/8373921/>, doi:10.1109/FG.2018.00128.
- [21] Min Peng, Chongyang Wang, and Tong Chen. Attention Based Residual Network for Micro-Gesture Recognition. *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 790–794, 2018. URL: <https://ieeexplore.ieee.org/document/8373920/>, doi:10.1109/FG.2018.00127.
- [22] Weizhe Zhang, Weidong Zhang, and Jie Shao. Classification of Holoscopic 3D Micro-Gesture Images and Videos. *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 815–818, 2018. URL: <https://ieeexplore.ieee.org/document/8373924/>, doi:10.1109/FG.2018.00131.
- [23] Garima Sharma, Shreyank Jyoti, and Abhinav Dhall. Hybrid Neural Networks Based Approach for Holoscopic Micro-Gesture Recognition in Images and Videos. *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 808–814, 2018. URL: <https://ieeexplore.ieee.org/document/8373923/>, doi:10.1109/FG.2018.00130.
- [24] Daniel K. Schneider. Wuser interaction and user interface design. URL: [http://edutechwiki.unige.ch/en/User\\_interaction\\_and\\_user\\_interface\\_design](http://edutechwiki.unige.ch/en/User_interaction_and_user_interface_design) [cited 09.09.2016].
- [25] J Jankowski and M Hachet. Advances in Interaction with 3D Environments. 34(1):152–190, 2015. doi:10.1111/cgf.12466.
- [26] A Y C Nee, S K Ong, G Chryssolouris, and D Mourtzis. CIRP Annals - Manufacturing Technology Augmented reality applications in design and manufacturing. *CIRP Annals - Manufacturing Technology*, 61(2):657–679, 2012. URL: <http://dx.doi.org/10.1016/j.cirp.2012.05.010>, doi:10.1016/j.cirp.2012.05.010.
- [27] Steve Mann. Veillance and Reciprocal Transparency : Surveillance versus Sousveillance , AR Glass , Lifeglogging , and Wearable Computing. *2013 IEEE*

- International Symposium on Technology and Society (ISTAS): Social Implications of Wearable Computing and Augmented Reality in Everyday Life*, pages 1–12, 2013. doi:[10.1109/ISTAS.2013.6613094](https://doi.org/10.1109/ISTAS.2013.6613094).
- [28] Steve Mann. Wearable Computing: A First Step Toward Personal Imaging. (February):25–32, 1997.
- [29] Jiang Xu, Patrick J. Gannon, Karen Emmorey, Jason F. Smith, and Allen R. Braun. Symbolic gestures and spoken language are processed by a common neural system. *Proceedings of the National Academy of Sciences*, 106(49):20664–20669, 2009. URL: <https://www.pnas.org/content/106/49/20664>, arXiv: <https://www.pnas.org/content/106/49/20664.full.pdf>, doi:[10.1073/pnas.0909197106](https://doi.org/10.1073/pnas.0909197106).
- [30] Ursula Bellugi Edward S. Klima. *The Signs of Language*. Harvard University Press, 1979. doi:[10.1017/CB09781139163910](https://doi.org/10.1017/CB09781139163910).
- [31] Wendy Sandler and Diane Lillo-Martin. *Sign Language and Linguistic Universals*. Cambridge University Press, 2006. doi:[10.1017/CB09781139163910](https://doi.org/10.1017/CB09781139163910).
- [32] G. Lippmann. Épreuves Réversibles Donnant La Sensation Du Relief. *Journal de Physique Théorique et Appliquée*, 7(1):821–825, 1908. doi:[10.1051/jphystap:019080070082100](https://doi.org/10.1051/jphystap:019080070082100).
- [33] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. Deep Residual Learning for Image Recognition. 2017. URL: <http://arxiv.org/abs/1703.10722>, arXiv: [1703.10722](https://arxiv.org/abs/1703.10722), doi:[10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [34] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, 2017-Janua(9):2261–2269, 2017. doi:[10.1109/CVPR.2017.243](https://doi.org/10.1109/CVPR.2017.243).
- [35] Yusuf Tas and Piotr Koniusz. Cnn-based action recognition and supervised domain adaptation on 3d body skeletons via kernel feature maps, 2018. arXiv:[1806.09078](https://arxiv.org/abs/1806.09078).
- [36] Chinmaya R. Naguri and Razvan C. Bunescu. Recognition of dynamic hand gestures from 3D motion data using LSTM and CNN architectures. *Proc. - 16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2017*, 2018-Janua:1130–1133, 2018. doi:[10.1109/ICMLA.2017.00013](https://doi.org/10.1109/ICMLA.2017.00013).

- [37] C. Zhang, X. Yang, and Y. Tian. Histogram of 3D facets: A characteristic descriptor for hand gesture recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, April 2013. doi:10.1109/FG.2013.6553754.
- [38] Maria Karam and m. c. Schraefel. A Taxonomy of Gestures in Human Computer Interactions. *Technical Report, Eletronics and Computer Science.*, pages 1–45, 2005. doi:10.1.1.97.5474.
- [39] Aashni Haria, Archanasri Subramanian, Nivedhitha Asokkumar, Shristi Poddar, and Jyothi S. Nayak. Hand Gesture Recognition for Human Computer Interaction. *Procedia Computer Science*, 115:367–374, 2017. doi:10.1016/j.procs.2017.09.092.
- [40] Maria Karam. *A framework for research and design of gesture-based human computer interactions*. PhD thesis, King’s College London, 2006.
- [41] Ying Wu and Thomas S. Huang. Vision-based gesture recognition: A review. In Annelies Braffort, Rachid Gherbi, Sylvie Gibet, Daniel Teil, and James Richardson, editors, *Gesture-Based Communication in Human-Computer Interaction*, page c, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.
- [42] G. Ren and E. O’Neill. 3D marking menu selection with freehand gestures. In *2012 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 61–68, March 2012. doi:10.1109/3DUI.2012.6184185.
- [43] Johnny Chung Lee. Hacking the nintendo wii remote. *IEEE Pervasive Computing*, 7(3):39–45, July 2008. URL: <http://dx.doi.org/10.1109/MPRV.2008.53>, doi:10.1109/MPRV.2008.53.
- [44] Ken Perlin. Quikwriting: Continuous stylus-based text entry. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, UIST ’98, pages 215–216, New York, NY, USA, 1998. ACM. URL: <http://doi.acm.org/10.1145/288392.288613>, doi:10.1145/288392.288613.
- [45] Hoggan, Williamson, Oulasvirta, Nacenta, P Kristensson, and A Lehtio. Multi-Touch Rotation Gestures: Performance and Ergonomics. *Proceedings of CHI 2013*, pages 3047–3050, 2013. URL: <https://mpi-sb.mpg.de/~joantti/pubs/hoggan-rotations-CHI13.pdf>  
`file:///Users/grahamwilson/Documents/Papers/2013/Hoggan/Hoggan2013ProceedingsofCHI2013.pdf`  
`papers:`

- [//c80d98e4-9a96-4487-8d06-8e1acc780d86/Paper/p10865,](https://doi.org/10.1145/2512349.2512817) doi:  
[10.1145/2512349.2512817.](https://doi.org/10.1145/2512349.2512817)
- [46] Ryosuke Aoki, Masayuki Ihara, Atsuhiko Maeda, Minoru Kobayashi, and Shingo Kagami. Unicursal gesture interface for TV remote with touch screens. *Digest of Technical Papers - IEEE International Conference on Consumer Electronics*, pages 99–100, 2011. doi:[10.1109/ICCE.2011.5722941](https://doi.org/10.1109/ICCE.2011.5722941).
- [47] Andrew Bragdon, Eugene Nelson, Yang Li, and Ken Hinckley. Experimental analysis of touch-screen gesture designs in mobile environments. *Proc. 2011 Annu. Conf. Hum. factors Comput. Syst. - CHI '11*, page 403, 2011. URL: <http://dl.acm.org/citation.cfm?doid=1978942.1979000>, doi:[10.1145/1978942.1979000](https://doi.org/10.1145/1978942.1979000).
- [48] Muhammad Shahzad, Alex X Liu, and Arjmand Samuel. Behavior Based Human Authentication on Touch Screen Devices Using Gestures and Signatures. 16(10):2726–2741, 2017.
- [49] Yang Li, Ken Hinckley, Ken Hinckley, Zhiwei Guan, and James A. Landay. Experimental analysis of mode switching techniques in pen-based user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '05*, pages 461–470, New York, NY, USA, 2005. ACM. URL: <http://doi.acm.org/10.1145/1054972.1055036>, doi:[10.1145/1054972.1055036](https://doi.org/10.1145/1054972.1055036).
- [50] Volker Roth and Thea Turner. Bezel swipe: Conflict-free scrolling and multiple selection on mobile touch screen devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09*, pages 1523–1526, New York, NY, USA, 2009. ACM. URL: <http://doi.acm.org/10.1145/1518701.1518933>, doi:[10.1145/1518701.1518933](https://doi.org/10.1145/1518701.1518933).
- [51] Alvin R. Tilley and Henry Dreyfuss Associates. *The Measure of man and woman: human factors in design*. Phaidon, New York, 1993.
- [52] J. G.Vargas Yeppez, A. Yepes Moreno, and S. Roa Prada. Conceptual design of a robotic assistant system based on projections onto flat surfaces for human-machine interaction applications. *2016 IEEE Colombian Conference on Robotics and Automation, CCRA 2016 - Conference Proceedings*, pages 3–6, 2017. doi:[10.1109/CCRA.2016.7811424](https://doi.org/10.1109/CCRA.2016.7811424).
- [53] A. Argyros X. Zabulis, H. Baltzakis. Hand Gesture Recognition for Human Computer Interaction. *The Universal Access Handbook. Boca Raton, FL, USA: CRC Press*, 115:367–374, 2009. doi:[10.1016/j.procs.2017.09.092](https://doi.org/10.1016/j.procs.2017.09.092).

- [54] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19(2):4–10, Feb 2012. doi:10.1109/MMUL.2012.24.
- [55] Xbox 360. Xbox 360 “project natal” 101, 2009. URL: <http://mb.cision.com/Public/MigratedWpy/93727/664834/9c48f222ba2bbae3.pdf>.
- [56] L.A. Times. E3: Microsoft shows off gesture control technology for xbox 360. URL: <https://latimesblogs.latimes.com/technology/2009/06/microsofte3.html> [cited 14.06.2018].
- [57] Gavin Gear. Tapping into the power of kinect for windows. URL: <https://latimesblogs.latimes.com/technology/2009/06/microsofte3.html> [cited 14.12.2016].
- [58] Tim Stevens. Kinect for windows sdk beta launches, wants pc users to get a move on, 2011. URL: [https://consent.yahoo.com/collectConsent?sessionId=3\\_cc-session\\_41ce4122-c737-46b8-83ae-797dc37c099c&lang=en-US&inline=false](https://consent.yahoo.com/collectConsent?sessionId=3_cc-session_41ce4122-c737-46b8-83ae-797dc37c099c&lang=en-US&inline=false).
- [59] Kinect. Kinect for windows announces new version of sdk coming march 18. URL: <https://msdn.microsoft.com/en-us/> [cited 12.5.2018].
- [60] ASUS. Use xtion pro developer solution to make motion-sensing applications and games. URL: [https://www.asus.com/3D-Sensor/Xtion\\_PRO\\_LIVE/](https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/) [cited 14.01.2017].
- [61] H Cheng, L Yang, and Z Liu. A Survey on 3D Hand Gesture Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1, 2015. doi:10.1109/TCSVT.2015.2469551.
- [62] Thomas B Moeslund and Erik Granum. A Survey of Computer Vision-Based Human Motion Capture. 268:231–268, 2001. doi:10.1006/cviu.2000.0897.
- [63] Ramprasad Polana and Randal Nelson. Low Level Recognition of Human Motion. *Proceedings of the 1994 IEEE Workshop on Motion of Non-Rigid and Articulated Objects, 1994.*, pages 77–82, 1994. doi:10.1109/MNRAO.1994.346251.
- [64] Y. Sang, L. Shi, and Y. Liu. Micro hand gesture recognition system using ultrasonic active sensing. *IEEE Access*, 6:49339–49347, 2018. doi:10.1109/ACCESS.2018.2868268.
- [65] Jun Gong, Yang Zhang, Xia Zhou, and Xing Dong Yang. Pyro: Thumb-tip gesture recognition using pyroelectric infrared sensing. *UIST 2017 - Proceedings of the*

- 30th Annual ACM Symposium on User Interface Software and Technology*, pages 553–563, 2017. doi:[10.1145/3126594.3126615](https://doi.org/10.1145/3126594.3126615).
- [66] Takanori Okoshi. *Three-dimensional imaging techniques*. Academic Press, London;New York (etc.);, 1976.
- [67] I. Sexton and P. Surman. Stereoscopic and autostereoscopic display systems. *IEEE Signal Processing Magazine*, 16(3):85–99, 1999.
- [68] Pierre Boher, Thierry Leroux, Thibault Bignon, and Véronique Collomb-Patton. Multispectral polarization viewing angle analysis of circular polarized stereoscopic 3D displays. In Andrew J. Woods, Nicolas S. Holliman, and Neil A. Dodgson, editors, *Stereoscopic Displays and Applications XXI*, volume 7524, pages 247 – 258. International Society for Optics and Photonics, SPIE, 2010.
- [69] B. Toperverg, O. Nikonov, V. Lauter-Pasyuk, and H. J. Lauter. Towards 3d polarization analysis in neutron reflectometry. *Physica B: Physics of Condensed Matter*, 297(1):169–174, 2001.
- [70] Andreas Kolb, Erhardt Barth, and Reinhard Koch. ToF-sensors: New dimensions for realism and interactivity. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 2008. doi:[10.1109/CVPRW.2008.4563159](https://doi.org/10.1109/CVPRW.2008.4563159).
- [71] David Droeschel, Jörg Stückler, and Sven Behnke. Learning to interpret pointing gestures with a time-of-flight camera. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 481–488. ACM, 2011.
- [72] Marvin Lindner, Andreas Kolb, and Thorsten Ringbeck. New insights into the calibration of ToF-sensors. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, (1), 2008. doi:[10.1109/CVPRW.2008.4563172](https://doi.org/10.1109/CVPRW.2008.4563172).
- [73] Herbert E. Ives. Optical properties of a lippmann lenticulated sheet. *J. Opt. Soc. Am.*, 21(3):171–176, Mar 1931. URL: <http://www.osapublishing.org/abstract.cfm?URI=josa-21-3-171>, doi:[10.1364/JOSA.21.000171](https://doi.org/10.1364/JOSA.21.000171).
- [74] A. P. Pentland. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(4):523–531, July 1987. doi:[10.1109/TPAMI.1987.4767940](https://doi.org/10.1109/TPAMI.1987.4767940).
- [75] Ara Chutjian and Robert J. Collier. Recording and reconstructing three-dimensional images of computer-generated subjects by lippmann ntegral photography. *Applied optics*, 7 1:99–103, 1968.

- [76] Ivars J Vilums. Optical imaging system using lenticular tone-plate elements, November 7 1989. US Patent 4,878,735.
- [77] Mathias Hain, Wolff von Spiegel, Marc Schmiedchen, Theo Tschudi, and Bahram Javidi. 3d integral imaging using diffractive fresnel lens arrays. *Opt. Express*, 13(1):315–326, Jan 2005. URL: <http://www.opticsexpress.org/abstract.cfm?URI=oe-13-1-315>, doi:10.1364/OPEX.13.000315.
- [78] D. Zarpalas, E. Fotiadou, I. Biperis, and P. Daras. Anchoring graph cuts towards accurate depth estimation in integral images. *Journal of Display Technology*, 8(7):405–417, 2012.
- [79] Amar Aggoun. *3D Holographic Imaging Technology for Real-Time Volume Processing and Display*, pages 411–428. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. doi:10.1007/978-3-642-12802-8\_18.
- [80] Alan Bundy and Lincoln Wallen. *Constructive Solid Geometry*, pages 21–22. Springer Berlin Heidelberg, Berlin, Heidelberg, 1984. doi:10.1007/978-3-642-96868-6\_40.
- [81] R. Zaharia, A. Aggoun, and M. McCormick. Adaptive 3d-dct compression algorithm for continuous parallax 3d integral imaging. *Signal Processing: Image Communication*, 17(3):231 – 242, 2002. URL: <http://www.sciencedirect.com/science/article/pii/S0923596501000200>, doi: [https://doi.org/10.1016/S0923-5965\(01\)00020-0](https://doi.org/10.1016/S0923-5965(01)00020-0).
- [82] Silvia Manolache Cirstea, S.Y. Kung, Malcolm McCormick, and Amar Aggoun. 3d-object space reconstruction from planar recorded data of 3d-integral images. *Journal of VLSI signal processing systems for signal, image and video technology*, 35(1):5–18, Aug 2003. doi:10.1023/A:1023386402756.
- [83] A. Aggoun. Pre-Processing of Integral Images for 3-D Displays. *Journal of Display Technology*, 2(4):393–400, Dec 2006. doi:10.1109/JDT.2006.884691.
- [84] M R Swash, A Aggoun, O Abdulfatah, B Li, J C Fernández, E Alazawi, and E Tseklevs. PRE-PROCESSING OF HOLOSCOPIC 3D IMAGE FOR AUTOSTEREOSCOPIC 3D DISPLAYS. pages 3–7, 2013. doi:10.1109/IC3D.2013.6732100.
- [85] N. Davies and M. McCormick. Holographic imaging with true 3-d content in full natural colour. *The Journal of Photographic Science*, 40(2):46–49, 1992. arXiv:<https://doi.org/10.1080/00223638.1992.11737166>, doi:10.1080/00223638.1992.11737166.

- [86] Manuel Martínez-Corral, Bahram Javidi, Raúl Martínez-Cuenca, and Genaro Saavedra. Formation of real, orthoscopic integral images by smart pixel mapping. *Opt. Express*, 13(23):9175–9180, Nov 2005. URL: <http://www.opticsexpress.org/abstract.cfm?URI=oe-13-23-9175>, doi:10.1364/OPEX.13.009175.
- [87] Jun Arai. In memoriam: Fumio Okano, innovator of 3D display. In Bahram Javidi, Jung-Young Son, Osamu Matoba, Manuel Martínez-Corral, and Adrian Stern, editors, *Three-Dimensional Imaging, Visualization, and Display 2014*, volume 9117, pages 109 – 114. International Society for Optics and Photonics, SPIE, 2014. doi:10.1117/12.2058117.
- [88] Bahram Javidi, Raúl Martínez-Cuenca, Genaro Saavedra, and Manuel Martínez-Corral. Orthoscopic long-focal-depth integral imaging by hybrid method. In Bahram Javidi, Fumio Okano, and Jung-Young Son, editors, *Three-Dimensional TV, Video, and Display V*, volume 6392, pages 21 – 28. International Society for Optics and Photonics, SPIE, 2006. doi:10.1117/12.686659.
- [89] John R. Koza, Forrest H. Bennett, David Andre, and Martin A. Keane. *Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming*, pages 151–170. Springer Netherlands, Dordrecht, 1996. doi:10.1007/978-94-009-0279-4\_9.
- [90] Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–27, 2009. doi:10.1561/22000000006.
- [91] Andreas M. Tillmann. On the computational intractability of exact and approximate dictionary learning. *IEEE Signal Processing Letters*, 22(1):45–49, 2015. arXiv:1405.6664, doi:10.1109/LSP.2014.2345761.
- [92] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing over-complete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, Nov 2006. doi:10.1109/TSP.2006.881199.
- [93] Stevan Harnad. The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence (PUBLISHED VERSION BOWDLERIZED). pages 1–28, 2008. URL: <http://eprints.soton.ac.uk/262954/1/turing.html>.
- [94] P. Barrett, R. O. Duda, and D. Nitzan. Use of range and reflectance data to find planar surface regions. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 1(03):259–271, jul 1979. doi:10.1109/TPAMI.1979.4766922.



- [95] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012.
- [96] Vladimir Vapnik and Rauf Izmailov. Knowledge transfer in svm and neural networks. *Annals of Mathematics and Artificial Intelligence*, 81(1):3–19, 2017.
- [97] William W. Hsieh. *Machine learning methods in the environmental sciences: Neural networks and Kernels*. 2009.
- [98] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. 2015. doi: [10.1038/nature14539](https://doi.org/10.1038/nature14539).
- [99] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. arXiv:1206.5538, doi:10.1109/TPAMI.2013.50.
- [100] Jürgen Schmidhuber. Deep learning in neural networks : An overview. *Neural Networks*, 61:85–117, 2015. URL: <http://dx.doi.org/10.1016/j.neunet.2014.09.003>, doi:10.1016/j.neunet.2014.09.003.
- [101] Adam H. Marblestone, Greg Wayne, and Konrad P. Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10:94, 2016. URL: <https://www.frontiersin.org/article/10.3389/fncom.2016.00094>, doi:10.3389/fncom.2016.00094.
- [102] Bruno A Olshausen and David J Field. Code for Natural Images. *Nature*, 381(June):607–609, 1996.
- [103] Joseph J LaViola Jr. An introduction to 3d gestural interfaces. In *ACM SIGGRAPH 2014 Courses*, page 25. ACM, 2014.
- [104] Pedro Trindade, Jorge Lobo, and P Barreto. Hand gesture recognition using color and depth images enhanced with hand angular pose data \*. 2012. doi:10.1109/MFI.2012.6343032.
- [105] Mahmoud Elmezain, Ayoub Al-hamadi, and Bernd Michaelis. Hand Gesture Recognition Based on Combined Features Extraction. 3(12):2389–2394, 2009.
- [106] Dandu Amarnatha Reddy and Samit Ari. Hand Gesture Recognition Using Local Histogram Feature Descriptor. (Icoei):199–203, 2018.

- [107] G R S Murthy. Hand Gesture Recognition using Neural Networks. *2010 IEEE 2nd International Advance Computing Conference (IACC)*, pages 134–138, 2010. doi:[10.1109/IADCC.2010.5423024](https://doi.org/10.1109/IADCC.2010.5423024).
- [108] VLADIMIR I. PAVLOVI C RAJEEV SHARMA, MEMBER and THOMAS S. HUANG. Toward Multimodal Human – Computer Interface. 86(5), 1998.
- [109] Thomas Stiefmeier, Daniel Roggen, and Gerhard Tr. Fusion of String-Matched Templates for Continuous Activity Recognition. 2007.
- [110] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6:21–45, 2006. doi:[10.1109/MCAS.2006.1688199](https://doi.org/10.1109/MCAS.2006.1688199).
- [111] Lisa Osadciw and Kalyan Veeramachaneni. *Fusion, Decision-Level*, pages 593–597. Springer US, Boston, MA, 2009. doi:[10.1007/978-0-387-73003-5\\_160](https://doi.org/10.1007/978-0-387-73003-5_160).
- [112] D. Roggen, G. Tröster, and A. Bulling. 12 - signal processing technologies for activity-aware smart textiles. In Tünde Kirstein, editor, *Multidisciplinary Know-How for Smart-Textiles Developers*, Woodhead Publishing Series in Textiles, pages 329 – 365. Woodhead Publishing, 2013. URL: <http://www.sciencedirect.com/science/article/pii/B9780857093424500122>, doi: <https://doi.org/10.1533/9780857093530.2.329>.
- [113] Seniha Esen Yuksel, Joseph N. Wilson, and Paul D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012. doi:[10.1109/TNNLS.2012.2200299](https://doi.org/10.1109/TNNLS.2012.2200299).
- [114] R. Häuslschmid, B. Menrad, and A. Butz. Freehand vs. micro gestures in the car: Driving performance and user experience. In *2015 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 159–160, March 2015. doi:[10.1109/3DUI.2015.7131749](https://doi.org/10.1109/3DUI.2015.7131749).
- [115] I. Wang, M. B. Fraj, P. Narayana, D. Patil, G. Mulay, R. Bangar, J. R. Beveridge, B. A. Draper, and J. Ruiz. Egnog: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 414–421, May 2017. doi:[10.1109/FG.2017.145](https://doi.org/10.1109/FG.2017.145).
- [116] F. Garcia, D. Aouada, T. Solignac, B. Mirbach, and B. Ottersten. Real-time depth enhancement by fusion for RGB-D cameras. *IET Computer Vision*, 7(5):1–11, October 2013. doi:[10.1049/iet-cvi.2012.0289](https://doi.org/10.1049/iet-cvi.2012.0289).

- [117] Giulio Marin, Fabio Dominio, and Pietro Zanuttigh. Hand gesture recognition with jointly calibrated leap motion and depth sensor. *Multimedia Tools Appl.*, 75(22):14991–15015, November 2016. doi:[10.1007/s11042-015-2451-6](https://doi.org/10.1007/s11042-015-2451-6).
- [118] Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihoud, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Trans. Graph.*, 35(4):142:1–142:19, July 2016. doi:[10.1145/2897824.2925953](https://doi.org/10.1145/2897824.2925953).
- [119] Robert Y Wang. Real-Time Hand-Tracking with a Color Glove. 28(3):1–8, 2009. doi:[10.1145/1531326.1531369](https://doi.org/10.1145/1531326.1531369).
- [120] Jerome R Bellegarda. Spoken Language Understanding for Natural Interaction: The Siri Experience. In Joseph Mariani, Sophie Rosset, Martine Garnier-Rizet, and Laurence Devillers, editors, *Natural Interaction with Robots, Knowbots and Smartphones*, pages 3–14, New York, NY, 2014. Springer New York.
- [121] Pramod Kumar Pisharady and Martin Saerbeck. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141:152–165, 2015. doi:<https://doi.org/10.1016/j.cviu.2015.08.004>.
- [122] Rodrigo Ibañez, Álvaro Soria, Alfredo Teyseyre, and Marcelo Campo. Easy gesture recognition for Kinect. *Advances in Engineering Software*, 76:171–180, 2014. doi:<https://doi.org/10.1016/j.advengsoft.2014.07.005>.
- [123] Leigh Ellen Potter, Jake Araullo, and Lewis Carter. The leap motion controller: A view on sign language. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, OzCHI '13, pages 175–178, New York, NY, USA, 2013. ACM. doi:[10.1145/2541016.2541072](https://doi.org/10.1145/2541016.2541072).
- [124] W. Lu, Z. Tong, and J. Chu. Dynamic hand gesture recognition with leap motion controller. *IEEE Signal Processing Letters*, 23(9):1188–1192, Sept 2016. doi:[10.1109/LSP.2016.2590470](https://doi.org/10.1109/LSP.2016.2590470).
- [125] B. Kaufmann and M. Akil. 3D images compression for multi-view auto-stereoscopic displays. In *International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)*, pages 128–136, July 2006. doi:[10.1109/CGIV.2006.3](https://doi.org/10.1109/CGIV.2006.3).
- [126] Tae-kyun Kim and Roberto Cipolla. Gesture Recognition Under Small Sample Size. pages 335–344, 2007.

- [127] Shanxin Yuan, Guillermo Garcia-hernando Bj, Gyeongsik Moon, Ju Yong, Chang Kyoung, Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Lihao Ge, Junsong Yuan, Xinghao Chen, and Guijin Wang. Depth-Based 3D Hand Pose Estimation : From Current Achievements to Future Goals. pages 2636–2645, 2018. doi: [10.1109/CVPR.2018.00279](https://doi.org/10.1109/CVPR.2018.00279).
- [128] Isabelle Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner. Results and analysis of the ChaLearn gesture challenge 2012. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7854 LNCS:186–204, 2013. doi:[10.1007/978-3-642-40303-3\\_19](https://doi.org/10.1007/978-3-642-40303-3_19).
- [129] Steve Krug. *Don't Make Me Think, Revisited: A Common Sense Approach to Web Usability, Third Edition*. New Riders, 3 edition, 2013.
- [130] Amy L. Parsons. Emotional design: Why we love (or hate) everyday things. new york, ny: Basic books, a member of the perseus books group 2004. 257 pp. \$15.95 (paperback). *Journal of Consumer Marketing*, 23(2):115–116, 2006.
- [131] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey Mckinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A low power, fully event-based gesture recognition system. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:7388–7397, 2017. doi: [10.1109/CVPR.2017.781](https://doi.org/10.1109/CVPR.2017.781).
- [132] Tae-Kyun Kim and Roberto Cipolla. Gesture recognition under small sample size. In *Asian conference on computer vision*, pages 335–344. Springer, 2007.
- [133] Tomás Mantecón, Carlos R. del Blanco, Fernando Jaureguizar, and Narciso García. Hand gesture recognition using infrared imagery provided by leap motion controller. In Jacques Blanc-Talon, Cosimo Distanto, Wilfried Philips, Dan Popescu, and Paul Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, pages 47–57, Cham, 2016. Springer International Publishing.
- [134] Shanxin Yuan, Qi Ye, Siddhant Jain, and Tae-kyun Kim. Big Hand 2 . 2M Benchmark : Hand Pose Data Set and State of the Art Analysis. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2007.

- [135] Michael Studdert-Kennedy. Hand and Mind: What Gestures Reveal About Thought. *Language and Speech*, 37(2):203–209, 1994. doi:[10.1177/002383099403700208](https://doi.org/10.1177/002383099403700208).
- [136] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, June 2007. doi:[10.1109/TPAMI.2007.1110](https://doi.org/10.1109/TPAMI.2007.1110).
- [137] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Transactions on Cybernetics*, 44(2):161–174, Feb 2014. doi:[10.1109/TCYB.2013.2249063](https://doi.org/10.1109/TCYB.2013.2249063).
- [138] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996. doi:[https://doi.org/10.1016/0031-3203\(95\)00067-4](https://doi.org/10.1016/0031-3203(95)00067-4).
- [139] Ville Ojansivu and Janne Heikkilä. *Blur Insensitive Texture Classification Using Local Phase Quantization*, pages 236–243. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. doi:[10.1007/978-3-540-69905-7\\_27](https://doi.org/10.1007/978-3-540-69905-7_27).
- [140] Yi Liu, Hongying Meng, Mohammad Rafiq Swash, Yona Falinie A. Gaus, and Rui Qin. Holoscopic 3D micro-gesture database for wearable device interaction. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 802–807, 2018. arXiv:[1712.05570](https://arxiv.org/abs/1712.05570), doi:[10.1109/FG.2018.00129](https://doi.org/10.1109/FG.2018.00129).
- [141] T. Lei, X. Jia, Y. Zhang, L. He, H. Meng, and A. K. Nandi. Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering. *IEEE Transactions on Fuzzy Systems*, 26(5):3027–3041, Oct 2018. doi:[10.1109/TFUZZ.2018.2796074](https://doi.org/10.1109/TFUZZ.2018.2796074).
- [142] Andrew Lumsdaine Todor G. Georgiev. Focused plenoptic camera and rendering. *Journal of Electronic Imaging*, 19(2):1 – 11, 2010. arXiv:[021106](https://arxiv.org/abs/021106), doi:[10.1117/1.3442712](https://doi.org/10.1117/1.3442712).
- [143] L. Yang, P. An, D. Liu, R. Ma, and L. Shen. Three-dimensional holoscopic image-coding scheme using a sparse viewpoint image array and disparities. *Journal of Electronic Imaging*, 27(033030):1 – 15, 2018. doi:[10.1117/1.JEI.27.3.033030](https://doi.org/10.1117/1.JEI.27.3.033030).

- [144] A. Aggoun, E. Tsekleves, M. R. Swash, D. Zarpalas, A. Dimou, P. Daras, P. Nunes, and L. D. Soares. Immersive 3d holoscopic video system. *IEEE MultiMedia*, 20(1):28–37, 2013.
- [145] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012. [arXiv:1102.0183](https://arxiv.org/abs/1102.0183), [doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007](https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007).
- [146] Karen Simonyan & Andrew Zisserman. Very deep convolutional network for large-scale image recognition. *International Conference on Learning Representations*, 75(6):398–406, 2015. [arXiv:1409.1556v6](https://arxiv.org/abs/1409.1556v6), [doi:10.2146/ajhp170251](https://doi.org/10.2146/ajhp170251).
- [147] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015. [doi:10.1109/CVPR.2015.7298594](https://doi.org/10.1109/CVPR.2015.7298594).
- [148] Jian Sun Kaiming He, Xiangyu Zhang, Shaoqing Ren. Deep Residual Learning for Image Recognition Kaiming. 2017. URL: <http://arxiv.org/abs/1703.10722>, [arXiv:1703.10722](https://arxiv.org/abs/1703.10722), [doi:10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [149] Abhijit Sinha, Huimin Chen, D. G. Danu, Thia Kirubarajan, and M. Farooq. Estimation and decision fusion: A survey. *Neurocomputing*, 71(13-15):2650–2656, 2008. [doi:10.1016/j.neucom.2007.06.016](https://doi.org/10.1016/j.neucom.2007.06.016).
- [150] Mike Gashler, Christophe Giraud-Carrier, and Tony Martinez. Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous. *2008 Seventh International Conference on Machine Learning and Applications*, pages 900–905, 2008. [doi:10.1109/ICMLA.2008.154](https://doi.org/10.1109/ICMLA.2008.154).
- [151] L Breiman. Bagging predictors - Springer. *Machine learning*, 140:123–140, 1996. [doi:10.1007/BF00058655](https://doi.org/10.1007/BF00058655).
- [152] Sushilkumar Kalmegh. Analysis of WEKA Data Mining Algorithm REPTree , Simple Cart and RandomTree for Classification of Indian News. *International Journal of Innovative Science, Engineering & Technology*, 2(2):438–446, 2015.
- [153] Utthara Gosa Mangai, Suranjana Samanta, Sukhendu Das, and Pinaki Roy Chowdhury. A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)*, 27(4):293–307, 2010. [doi:10.4103/0256-4602.64604](https://doi.org/10.4103/0256-4602.64604).

- [154] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe. *Proceedings of the ACM International Conference on Multimedia - MM '14*, pages 675–678, 2014. [arXiv:1408.5093](https://arxiv.org/abs/1408.5093), [doi:10.1145/2647868.2654889](https://doi.org/10.1145/2647868.2654889).
- [155] Meghdad Kurmanji. A Comparison of 2D and 3D Convolutional Neural Networks for Hand Gesture Recognition from RGB-D Data. pages 2022–2027, 2019.
- [156] Obaidullah Abdul Fatah. *Post-production of holoscopic 3D image*. PhD thesis, Brunel University London, 2015.
- [157] Xing Yingxin, Li Jinghua, Wang Lichun, and Kong Dehui. A Robust Hand Gesture Recognition Method via Convolutional Neural Network. *Proc. - 2016 Int. Conf. Digit. Home, ICDH 2016*, pages 64–67, 2017. [doi:10.1109/ICDH.2016.023](https://doi.org/10.1109/ICDH.2016.023).
- [158] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. 2018. URL: <http://arxiv.org/abs/1810.06990>, [arXiv:1810.06990](https://arxiv.org/abs/1810.06990).
- [159] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. *Proc. IEEE Int. Conf. Comput. Vis.*, 2015 Inter:4489–4497, 2015. [arXiv:arXiv:1412.0767v4](https://arxiv.org/abs/1412.0767v4), [doi:10.1109/ICCV.2015.510](https://doi.org/10.1109/ICCV.2015.510).
- [160] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, 1:448–456, 2015. [arXiv:1502.03167](https://arxiv.org/abs/1502.03167).
- [161] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. *Journal of Machine Learning Research*, 15:315–323, 2011.
- [162] LeCun Y, Bottou L, Bengio Y, and Haffner P. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [163] ImageNet. MS Windows NT kernel description, 2019. URL: <http://www.image-net.org/>.
- [164] Karen Simonyan & Andrew Zisserman. VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION. *ICLR 2015 : International Conference on Learning Representations 2015*, 75(6):398–406, 2015. [arXiv:1409.1556v6](https://arxiv.org/abs/1409.1556v6), [doi:10.2146/ajhp170251](https://doi.org/10.2146/ajhp170251).

- [165] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D Convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. doi:[10.1109/TPAMI.2012.59](https://doi.org/10.1109/TPAMI.2012.59).
- [166] Themis Stafylakis and Georgios Tzimiropoulos. Combining Residual Networks with LSTMs for Lipreading. pages 4–8. [arXiv:arXiv:1703.04105v4](https://arxiv.org/abs/1703.04105v4).
- [167] R. Qin, Y. Liu, M.R. Swash, H Meng, T. Lei, and T. Chen. A fast automatic holoscopic 3d micro-gesture recognition system for immersive applications. In *The 15th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 2019.



# **Appendices**

# Appendix A

## Participant Information Form

This document is provided to all participants to inform them of the purpose of the experiment and the content of the participation. Before taking part in the experiment, they can ask questions and train the three micro-gestures.

Then, after the participants have sufficiently understand the role of each micro-gesture action and gesture function, they proceed to the shooting area and is introduced to the experimental equipment and individual shooting points. Figure.teach shows the principal investigator explaining each micro-gestures to the participant.

After the recording, each participants' videos is recorded with a serial number to represent every position and gesture performed. Figure [A.2](#). shows the video record process.

Figure A.1: Explanation of the micro-gesture movements.



Figure A.2: Data recording and classification.

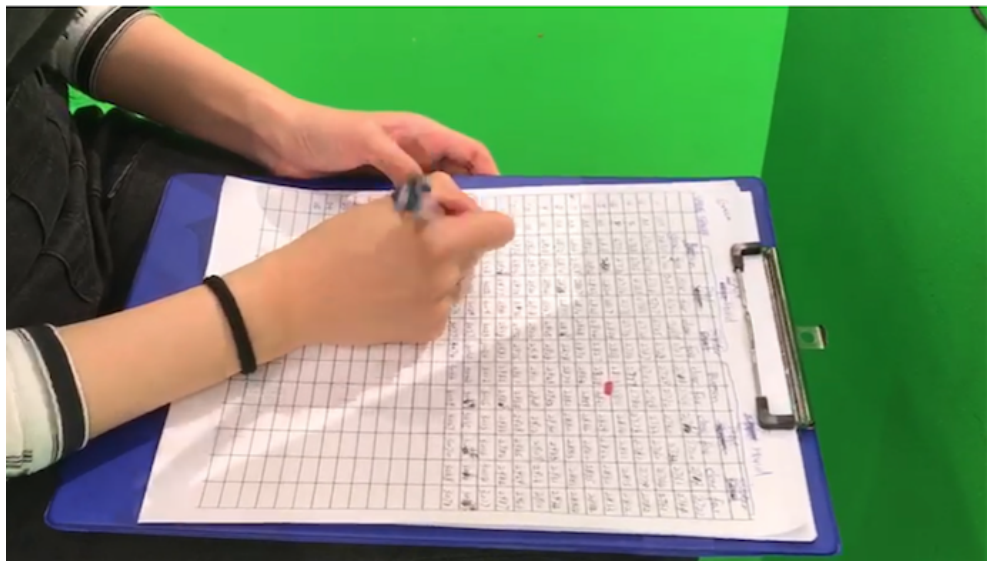
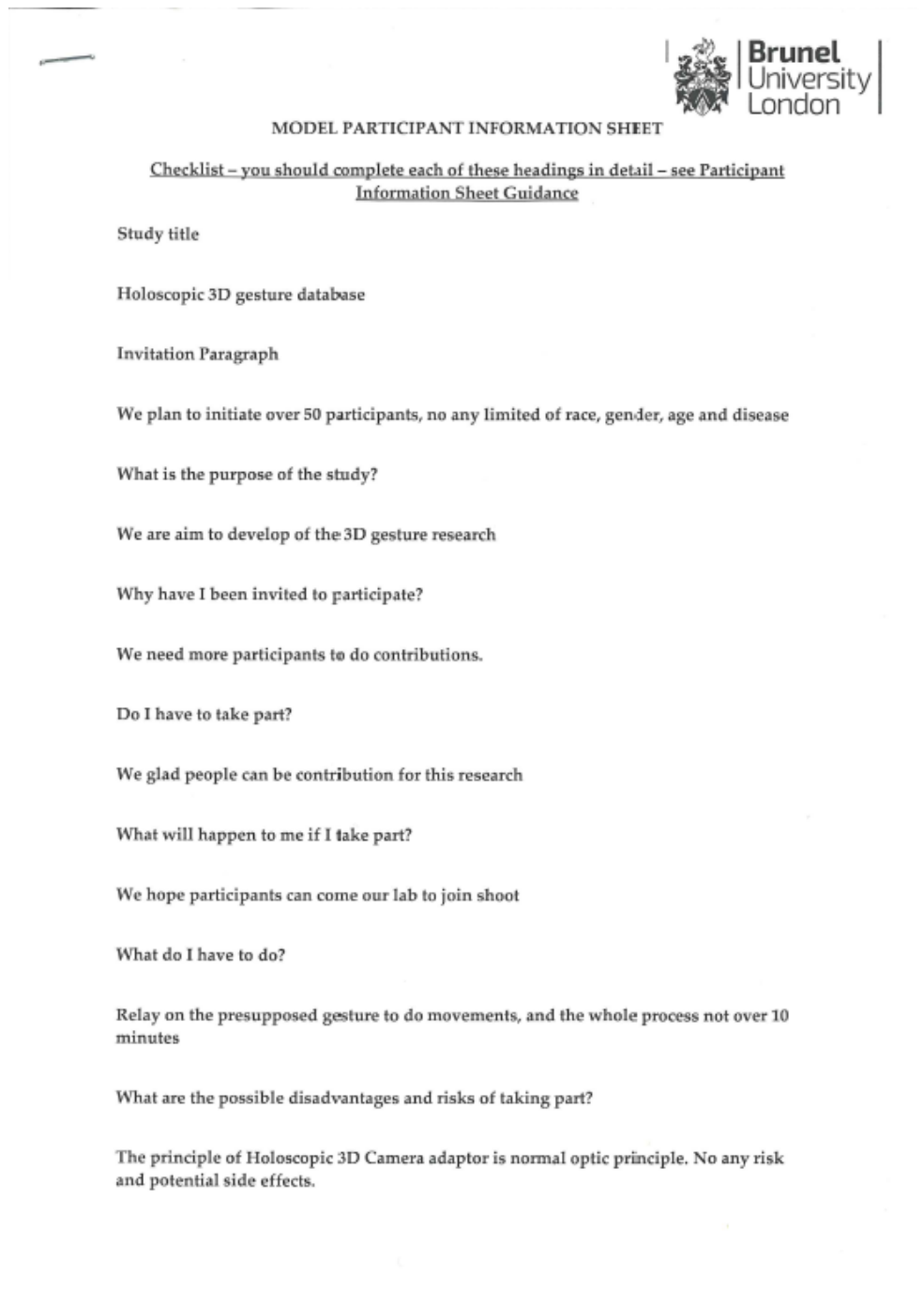



Figure A.3: Participant information sheet



The image shows a participant information sheet template from Brunel University London. It includes a title, a checklist of sections to complete, and several example text blocks. The text is as follows:



**MODEL PARTICIPANT INFORMATION SHEET**

Checklist – you should complete each of these headings in detail – see Participant Information Sheet Guidance

Study title

Holoscopic 3D gesture database

Invitation Paragraph

We plan to initiate over 50 participants, no any limited of race, gender, age and disease

What is the purpose of the study?

We are aim to develop of the 3D gesture research

Why have I been invited to participate?

We need more participants to do contributions.

Do I have to take part?

We glad people can be contribution for this research

What will happen to me if I take part?

We hope participants can come our lab to join shoot

What do I have to do?

Relay on the presupposed gesture to do movements, and the whole process not over 10 minutes

What are the possible disadvantages and risks of taking part?

The principle of Holoscopic 3D Camera adaptor is normal optic principle. No any risk and potential side effects.

What if something goes wrong?

If you are harmed by taking part in this research project, there are no special compensation arrangements. If you are harmed due to someone's negligence, then you may have grounds for a legal action but you may have to pay for it.

Will my taking part in this study be kept confidential?

All information which is collected about you during the course of the research will be kept strictly confidential. Any information about you which leaves the University/hospital/surgery/local authority premises, etc., will have your name and address removed so that you cannot be identified from it.

Who is organising and funding the research?

This is research from H3D group

What are the indemnity arrangements?

Participants should be informed if participation in a study might affect health-related insurance.

Who has reviewed the study?

Department of H3D researcher

Include a passage on the University's commitment to the UK Concordat on Research Integrity

Brunel University is committed to compliance with the Universities UK Research Integrity Concordat. You are entitled to expect the highest level of integrity from our researchers during the course of their research

Contact for further information and complaints

## **Appendix B**

# **End User License Agreement**

Figure B.1: End User License Agreement

End User License Agreement HoMGR Database

By signing this document the user, he or she who will make use of the database or the database interface, agrees to the following terms.

With provider, we denote both the actual data as well as the interface to the database.

With provider, we denote HoMGR Database group.  
(Introduce the people of this subject)

1. Commercial use

The user may not use the database for any non-academic purpose. Non-academic purposes include, but are not limited to:

- Proving the efficiency of commercial systems
- Training or testing of commercial systems
- Using screenshots of subjects from dataset in advisements
- Creating military applications
- Developing governmental systems used in public spaces

2. Responsibility

This document must be signed by a person with a permanent position at an academic institute (the signee). Up to five other researchers affiliated with the same institute for whom the signee is responsible may be named at the end of this document which will allow them to work with this dataset.

3. Distribution

The user may not distribute the database or portions thereof in any way, with the exception of using small portions of data for the exclusive purpose of clarifying academic publications or presentations.

Only data from subjects who gave consent to have their data used in publications and presentations may be used for this purpose. Note that publications will have to comply with the term stated in article 5.

4. Access

The user may only use the database after this End User License Agreement (EULA) has been signed and returned to the provider. The signed EULA should be returned in digital format by uploading it to the website when requesting an account at:

<http://3dvie.co.uk/>

Only if the user is not capable of requesting an account in this manner, accounts may be requested by sending the signed EULA via traditional mail to:

HoMGR2018@gmail.com

Dr. Rafiq Swash  
Department of Electronic and computer Engineering  
Brunel University London  
Kingston Line, London, Uxbridge, UB8 3PH, U.K.

The user may not grant anyone access to the database by giving out their user name and password.

5. Publications

Publications include not only paper, but also presentation for conference or educational purposes.

The user may only use data of subjects in publications if that particular subject has explicitly granted permission for this. This is specified with every database element.

All documents and papers that report on research that use any parts of the HoMGR Database will acknowledge this. The following paper has to be cited:

Y. Liu, H. Meng, M. R. Swash, Y. F. A. Gaus, and R. Qin, "Holoscopic 3D Micro-Gesture Database for Wearable Device Interaction," *2018 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG 2018)*, pp. 802-807, 2018.

The user will send a copy of any document or papers that reports on research that uses the HoMGR Database to Dr. Rafiq Swash and Dr. Hongying Meng.

6. Academic research

The user may only use the database for academic research.

7. Warranty

The database comes without any warranty. The provider cannot be held accountable for any damage (physical, financial or otherwise) caused by the use of the database. The provider will try to prevent any damage by keeping the database virus free.

8. Misuse

If at any point, the administrators of HoMGR database and/or the provider have a reasonable doubt that the user does not act in accordance to this EULA, he/she will be notified of this and will immediately be denied the access to the database.

User: \_\_\_\_\_  
User's Affiliation: \_\_\_\_\_  
User's address: \_\_\_\_\_  
User's email: \_\_\_\_\_  
Additional Researcher1 \_\_\_\_\_  
Additional Researcher2 \_\_\_\_\_  
Additional Researcher3 \_\_\_\_\_  
Additional Researcher4 \_\_\_\_\_  
Additional Researcher5 \_\_\_\_\_

Signature:

Date/place: