Hand Gesture Recognition using Deep Learning Neural Networks

By

Norah Meshari Alnaim

A thesis submitted for the degree of

Doctor of Philosophy

Department of Electronic & Computer Engineering
School of Engineering and Design and Physical Sciences
Brunel University London
December 2019

Abstract

Human Computer Interaction (HCI) is a broad field involving different types of interactions including gestures. Gesture recognition concerns non-verbal motions used as a means of communication in HCI. A system may be utilised to identify human gestures to convey information for device control. This represents a significant field within HCI involving device interfaces and users. The aim of gesture recognition is to record gestures that are formed in a certain way and then detected by a device such as a camera. Hand gestures can be used as a form of communication for many different applications. It may be used by people who possess different disabilities, including those with hearing-impairments, speech impairments and stroke patients, to communicate and fulfil their basic needs.

Various studies have previously been conducted relating to hand gestures. Some studies proposed different techniques to implement the hand gesture experiments. For image processing there are multiple tools to extract features of images, as well as Artificial Intelligence which has varied classifiers to classify different types of data. 2D and 3D hand gestures request an effective algorithm to extract images and classify various mini gestures and movements. This research discusses this issue using different algorithms. To detect 2D or 3D hand gestures, this research proposed image processing tools such as Wavelet Transforms and Empirical Mode Decomposition to extract image features. The Artificial Neural Network (ANN) classifier which used to train and classify data besides Convolutional Neural Networks (CNN). These methods were examined in terms of multiple parameters such as execution time, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood, negative likelihood, receiver operating characteristic, area under ROC curve and root mean square. This research discusses four original contributions in the field of hand gestures. The first contribution is an implementation of two experiments using 2D hand gesture video where ten different gestures are detected in short and long distances using an iPhone 6 Plus with 4K resolution. The experiments are performed using WT and EMD for feature extraction while ANN and CNN for classification. The second contribution comprises 3D hand gesture video experiments where twelve gestures are recorded using holoscopic imaging system camera. The third contribution pertains experimental work carried out to detect seven common hand gestures. Finally, disparity experiments were performed using the left and the right 3D hand gesture videos to discover disparities. The results of comparison show the accuracy results of CNN being 100% compared to other techniques. CNN is clearly the most appropriate method to be used in a hand gesture system.

Statement of Originality

The whole work covered in this thesis is completely that of the author unless otherwise stated. Except what is acknowledged, none of the work presented here has been published or distributed by anyone other than the author.

Norah Meshari Alnaim
December 2019, London

Acknowledgements

This accomplishment would not have been possible without support of my supervisors, Dr Maysam Abbod, and Dr Mohammad Rafiq Swash. I would like to express my sincere appreciation thankfulness to my supervisors for their support, advice and help during the period of my PhD research. I am literally thankful to my parents who were besides me from beginning of PhD journey till the end.

I would love to thank all my friends in KSA and UK who were beside me during my research and supporting me when I challenge the difficulties of life.

I am heartily thankful to my colleagues who have been supporting me for first day of my study till this moment. I am indebted to show my gratitude to Imam Abdulrahman bin Faisal University for providing me full scholarship to continue my study. Last but not least, my deep appreciation goes to everyone supported me through my study.

Table of Contents

Table of Figures

Table of tables

List of Acronyms

Acronym	Stands for
2D	Two-Dimensional
3D	Three-Dimensional
3D	3D pixels per inch in space
3DTV	Three-Dimensional Television
ADCNN	Adapted Deep Convolutional Neural Network
AI	Artificial Intelligence
API	Application Programming Interface
ANN	Artificial Neural Network
ANPR	Automatic Number Plate Recognition
ASL	American Sign Language
CGI	Computer-Generated Imagery
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
CT	Computed Tomography
CWT	Continuous Wavelet Transform
DBN	Daubechies Wavelets
DOF	Six Degrees of Freedom
DSC	Dice Similarity Coefficient
DTW	Dynamic Time Warping
EMD	Empirical Mode Decomposition
ES	Evolutionary Strategy
FPGA	Field-Programmable Gate Array
HD	High Definition

HDTV	High-Definition Television
HEVC	High Efficiency Video Coding
HMM	Hidden Markov Model
ICAP	Internet Configuration Access Port
IK	Inverse Kinematics
IMF	Intrinsic Mode Function
IQ	Intelligence Quotient
ICWT	Inverse Continuous Wavelet Transform
IT	Information Technology
IVPP	Image and Video Processing Platform
KCF	Kernelized Correlation Filters
MLA	Micro-lens Array
MOCAP	Motion Capture
MR	Magnetic Resonance
MRF	Markov Random Field
MRI	Magnetic Resonance Imaging
NN	Neural Network
NURBS	Non-uniform rational basis spline
OCR	Optical Character Recognition
PCA	Principle Component Analysis
PE	Permutation Entropy
RDF	Random Decision Forest
ReLU	Rectified Linear Unit
SD	Secure- Digital
SD	Standard Deviation
SLR	Single-lens Reflex

SPECT	Single-Photon Emission Computed Tomographic
SS	Self-Similarity
SVM	Support Vector Machine
TDNN	Time Delay Neural Network
ToF	Time of Flight
VIP	Video and Image Processing
URL	Uniform Resource Locator
WT	Wavelet Transforms

Chapter 1
Introduction

1.1 Preface

A Gesture is defined as the physical movement of the hands, fingers, arms and other parts of the human body through which the human can convey meaning and information for interaction with each other [1]. There are two different approaches for human–computer interactions, the data gloves approach and the vision-based approach. The vision-based approach was investigated in the following experiments including, the detection and classification of hand gestures. A Hand gesture is one of the logical ways to generate a convenient and high adaptability interface between devices and users. Applications such as, virtual object manipulation, gaming and gesture recognition be used in HCI systems. Hand tracking, as a theory aspect, deals with three fundamental elements of computer vision: hand segmentation, hand part detection, and hand tracking. The best communicative technique and the common concept used in a gesture recognition system is hand gestures. Hand gestures can be detected by one of these following techniques: posture is a static hand shape ratio without hand movements, or a gesture is dynamic hand motion with or without hand movements. Using any type of camera will detect any type of hand gesture; keeping in mind that different cameras will yield different resolution qualities. Two-dimensional cameras have the ability to detect most finger motions in a constant surface called 2D [2].

Sign language is one of the common examples for a hand gesture system. It is defined as a linguistic system based on hand motions besides other motions. For instance, most hearing-impaired people around the world use universal sign language. Sign language contains three fundamental parts: word level sign vocabulary, non-manual features and finger spelling [3]. One of best methods to communicate with hearing-impaired people is sign language.

Recently, sign language may be achieved by some types of robotics using some appropriate sensors used on the body of a patient [3]. Another example is stroke rehabilitation. People who have experienced stroke can have paraplegia which prevents them moving their lower limbs. Stroke rehabilitation can play a significant role to solve this type of issue. Additionally, some people who have stroke cannot communicate adequately with other people.

Researchers presented different studies in hand gestures, such as object detection and object motions. Gaming takes a keen interest in the area of Three-Dimensional (3D) hand tracking. At the outset of the 2010s, recent movie releases, such as Avatar, revolutionised cinema by combining content production and 3D technology with real actors, leading to the creation of the new type [4]. After the success of 3D cinema, the different electronic companies focused on production of Three-Dimensional Television (3DTV) technology. The researchers proposed the dome auto stereoscopic display that is used to view the position that is still limited [4]. The two different technologies such as stereo and multi-view rely on the brain to fuse the two images to create the effect of 3D [4].

Most studies have similar phases to implement their experiments. The first phase used for most studies is pre-processing, which is simply preparing the image to be appropriate to enter the second phase. Next, image processing prepares to receive the whole image so that it may be tracked using image processing tools like Wavelet Transform (WT) and Empirical Mode Decomposition (EMD). Artificial intelligence releases many classifiers such as Neural Network (NN) and Convolutional Neural Network (CNN); each one with the ability to classify data, that rely on its configuration and capabilities. WT and EMD techniques are the most capable tools to extract the image feature. For classification, the type of ANN used for some experimental works is Feedforward. It is the most efficient classifier type beside CNN for gesture recognition. The next section presents the research aim and objectives of the thesis.

1.2 Research Aim and Objectives

The aim of the research is to develop a system for 2D, and 3D hand gesture recognition using any type of camera, background, illuminations or position of hand, by finding the most appropriate algorithms to implement the system and test the validation of system. This system helps individuals with special needs and people who have experienced stroke to communicate accurately. Using WT and EMD algorithms for feature extraction and AI for classification provides different results while CNN provides an accurate result.

The objectives of the research include investigations, experimentation and development of appropriate algorithms for hand gesture recognition. The main objectives are highlighted below:

1- Study the concept of gesture recognition, detection phases and algorithms used in gesture recognition detection such as WT and EMD for feature extraction and AI and CNN for classification.

2- Determine different articles related to hand gesture recognition field including the holoscopic 3D imaging system, depth, different techniques and applications. All determined articles will be used for the literature review.

3- Determine the type of gestures, record them using different cameras, such as mobile cameras and a holoscopic imaging system camera, as well as apply them into a pre-processing phase before analysing them.

4- Analyse gestures with different image extraction tools, such as WT and EMD.

5- Develop a classification system using ANN and CNN classifiers.

6- Implement the system.

1.3 Research Original Contributions

This original section presents the summaries of contributions in the field of hand gesture recognition using deep learning methods:

1- 2D video gesture recognition

Ten hand gestures, in short and long distances, are recorded using an iPhone 6 Plus camera with 4k resolution, with a plain background and specific illumination.

The system is used ten times to obtain the mean of ten hand motions. Three algorithms (WT, EMD and CCN) are considered by different methods. The data is extracted using WT and EMD and fed into AI to prepare it for training then classifying. CNN is used as the classifier besides other algorithms. This comparison proposed an accurate method for the next experiments.

2- 3D video gesture recognition

This experimental work included twelve gestures in 2D and 3D, recorded using a holoscopic imaging system camera with size 135 x 75. The data is extracted into three frames, left, centre and right. The data is composed into single LCR frames and combined. Single, combines and all frames are fed, as input, into CNN algorithm.

3- Stroke Patients Gesture Recognition

Most studies proposed different systems in stroke rehabilitation. There is no study which discusses the issue of communication between people who have experienced stroke and other able-bodied persons. The proposed system supports people who have experienced stroke to communicate effectively using different common sign gestures which are meaningful.

Therefore, the experimental work is novel in that field. Twenty hand gestures, for different ages, display seven 2D and 3D universally common hand gestures size 227 x 227, using different mobile cameras, backgrounds, illumination, and positioning of the hand and shape of hand. A hundred and forty videos are read, to create 24,698 image frames which are fed as inputs into the CNN algorithm.

4- Disparity

This experiment discusses the disparity between left and right frames in 3D video over short distance. The size of left and right frames is 550 x 310. Four directories are generated for three different people. The first disparity image size is 59 and the window size 20, whereas the second and third disparity images are 49 and the window size is 31. The stereo match function is applied to estimate the disparity of the left and the right images.

1.4 Thesis Outline and Chapters' Summary

The thesis covers eight main chapters; the introduction of the thesis, the research literature review for the recent studies, the introduction of gesture recognition, image processing and recognition theory, 2D video gesture recognition, 3D video gesture recognition, stroke patients' gesture recognition, and conclusion and future work.

Chapter 1 - Introduction

Chapter one is an initial chapter, which presents the background of the research, the aim and objectives, the research original contributions, and the thesis outlines and chapters' summary.

Chapter 2 - Literature Review

Chapter two shows the literature review of the evolution of holoscopic 3D imaging systems, depth, finger movement measurement, classification and image processing. It presents each technique used in each study including the application. Lastly, it also discusses the hand tracking studies. These areas particularly, elaborate the state-of-the-art works on hand gesture recognition.

Chapter 3 - Gesture Recognition

Chapter three presents the HCI background and the history of gesture recognition. It explains briefly the main types of gesture recognition. The hand gesture recognition and its types are

discussed in this chapter. It also proposes the types of cameras used for 2D and 3D images. The holoscopic 3D imaging system camera technique is also explained in chapter three.

Chapter 4 - Image Processing and Recognition

Chapter four includes a brief theory of image processing, video processing, computer vision, empirical mode decomposition, wavelet transforms, artificial intelligence and convolutional neural network, and discusses the functionality of each technique and its uses.

Chapter 5 - 2D Video Gesture Recognition

Chapter five proposes a system created for hand motion recognition using WT and EMD for features extraction. ANN and CNN were used for classification. (10) 2D and 3D motion images in short and long distances were used in this experimental work. The experimental works were performed to compare the implementation of various methods using several measures. The results suggested that the CNN classifier is better than the ANN classifier. CNN was able to classify the hand gestures without using image processing tools for extraction.

Chapter 6 - 3D Video Gesture Recognition

Chapter five presents a system generated for 3D hand gesture recognition using CNN. Twelve 2D and 3D gesture images, in short and long distances, for three subjects used in this experimental work. The experimental works performed to compare the implementation of training and testing using a number of factors.

Chapter 7 - Stroke Patients Gesture Recognition

Chapter seven shows a model generated for hand motion recognition using CNN technique. Seven different 2D and 3D motions for twenty subjects used within short distances. The performance of training and testing in CNN algorithm will be compared. As a result, the experiments show that accuracy of testing is almost 100%.

Chapter 8 - Conclusion and Future Work

Chapter eight provides an overview of all research works as proposed in this thesis.

Appendix A presents the seven different common motions for seventeen subjects.

1.5 Author's Publications

This section shows the list of conference papers and journals published in international conferences.

1- N. Alnaim and M. Abbod, "Gesture Recognition and Classification using Intelligent Systems," *Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik*, vol. 60, no. OpenAccess Series in Informatics (OASIcs), pp. 8–1, 2018.

2- N. Alnaim and M. Abbod, "Mini gesture detection using neural networks algorithms," *Eleventh International Conference on Machine Vision (ICMV 2018)*, vol. 11041, no. Proc. SPIE, pp. 1–8, Mar. 2019.

3- N. Alnaim and M. Abbod, "Hand Gesture Detection Using Neural Networks Algorithms," *International Journal of Machine Learning and Computing*, vol. 9, no. 6, pp. 782–787, Dec. 2019.

4- N. Alnaim, M. Abbod, and A. Albar, "Hand Gesture Recognition Using Convolutional Neural Network for People Who Have Experienced A Stroke," *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1–6, Dec. 2019.

5- N. Alnaim, M. Abbod, and R. Swash, "Recognition of Holoscopic 3D Video Hand Gesture Using Convolutional Neural Networks," *Technologies*, vol. 8, no. 2, p. 19, Apr. 2020.

Chapter 2
Literature Review

2.1 Introduction

This chapter presents some current studies regarding different techniques used for gesture recognition includes the holoscopic 3D imaging system. It provides a list of journal and conference papers which proposed the problems faced by researchers and the proposed solutions.

Professor Gabriel M. Lippmann proposed the use of microlenses array at the image surface [4]. He presented this concept to the French Academy of Sciences as La Photography Integral [4] [5]. The spatial image with full parallax in all directions recorded completely by Professor Lippmann and is known as a fly's-eye lens array. Essentially, the display system was a screen containing a number of small lenses [4][5]. During the 1920s, some scientists, such as Herbert Ives, started to think about simplifying Lippmann's concept by joining a lenticular lens sheet, which contains a signaller array of spherical lenses known as lenticules. It was designed to view different angles of images to provide a pixel from each micro picture [4][5]. The lens sheet is transparent, and the back face creates the focal plane which is flat. One example is the lenses used in lenticular production where the technology is used to provide an illusion of depth, i.e; creating images which appear to move or change as the image is seen from many different angles. This innovative technology may also be used for generating 3D images on a flat sheet display. So, if the motion of the pictures is considered, this would result in a 3D Holoscopic video [4][5] [6].

Lately, Film industry releases, such as Avatar have transformed the cinema by joining 3D technology and component production with actors which lead to the innovation of a new category. The accomplishment of 3D film compelled some major consumer electronics industries to launch 3D TVs and broadcasters to present 3D content. Newsday's 3D TV technology uses a stereo vision system presenting left and right eye images via spatial and temporal multiplexing, to viewers wearing a pair of glasses. The next stage in 3D TV systems will probably be a multi-view autostereoscopic imaging system. This system will record and present many video gestures on a display. This will not compel viewers to wear glasses [4][5] [6].

Adedoyin et al. [7] used evolutionary strategy (ES) for joint motion and disparity estimation in order to compress 3D integral video sequences. The integral video sequences were decomposed to viewpoint video sequences and jointly exploit motion and disparity redundancies to maximise the compression using a self-adapted evolutionary strategy. Finding valid arrangements of movement vectors just for a solitary perspective and using this connection to limit the general coding unpredictability could be extremely useful. In order to further improve the quality of video, the half-pixel algorithm was than applied. They revealed that ES and the half-pixel algorithm joint motion and disparity estimation can be up to 1.5 dB. This can be achieved without any additional computational cost. It reduces the computation cost by up to 90%. Digital advances in holographic printing, holographic projection and holographic television have been presented in a survey paper in [8]. Between 1970 and 1980 holograms became a part of everyday life. Holography applications extended to areas including microscopy, metrology, non-ruinous testing, and the formation of diffractive optical elements. 3D high-dimensional holographic presents system based on dense ray sampling and integral imaging capture who presented in [9]. Amplitude modulation of the lenses and blind deconvolution technique were combined during capture stage for the dense ray sampling to present the depth of field extension technique. The strategy can be executed by setting a twofold cover before the focal point.

In another work, Wang et al. [10] combined integral imaging and augmented reality and named it 3D augmented reality micro integral imaging display system. It had greatly enhanced the viewer's perception of reality. There exist two advantages of this system; First one is it facilitates 3D augmented reality display, and second it has a compact design. Aggoun et al. [4] developed 3D holoscopic video systems for 3D TV applications. They used a field lens and a square aperture. Self-Similarity (SS) coding approach outperforms the standard High Efficiency Video Coding (HEVC) scheme and they also explore that search and retrieval effectiveness relies on the depth map's quality. A combined state build-up estimation was extended to perceive a larger number of motions dependent on their momentary directions. O. A. Fatah et al. [11] proposed a technique for motion picture rendering on holoscopic content to generate content for stereoscopic systems. The rendering technique used in the proposed method relied upon the sampling, shift and integrating of different views. For their experiment, the authors used a single aperture camera and holoscopic imaging to produce a high-resolution stereoscopic image. Their experimental results revealed that using this method improved the resolution of images. The stereo content generated with this rendering method who played back

on polarised stereoscopic system. Holoscopic 3D camera adaptors are usually designed for large-scale Single-lens Reflex (SLR) cameras. Albar and Swash [12] proposed a technique for prototyping a holoscopic 3D camera adaptor for a credit card sized board computer, called Raspberry Pi. The two prototypes are relatively cheap to assemble. This innovative holoscopic 3D adaptor design can be used in a variety of applications such as surveillance, medical, entertainment and also in equipment where 3D depth sensing and measurement are the main concerns. The principle behind this invention was that one can record numerous basic pictures of a 3D scene in a 2D network sensor with each natural picture putting away an alternate point of view of the 3D scene. This is accomplished by integrating a Microlens Array (MLA) before a sensor or utilize a camera cluster to catch the 3D object from different perspectives.

Fatah et al. [13] targeted the digital refocusing where they proposed a method. This method uses Michelson contrast formula to extract all-in-focus images. The highest contrast values at different points in space can return the focused points where the objects are initially positioned, which make it possible to obtain all-in-focus image. . It requires a new system which effectively separates the equivalent situated pixels under each Exposure Index. Oliveira [14] presented in their thesis a 2D image extraction technique for holoscopic 3D images. They named it Disparity Assisted Patch Blending which outperforms existing methods. The second contribution in their research work is the identification of potential non-reference image quality assessment metrics. These metrics are able to measure 2D image extraction to extractions to compare with human perception.

A 3D finger motion measurement system which is based on soft sensors are proposed by Park et al. [15]. These sensors are made out of Ecoflex (soft material), having embedded micro channels filled with conductive liquid metal Eutectic gallium-indium (EGaln). These sensors have the capability to embed in such environments where other traditional sensors cannot be embedded. Joint of thumbs are modelled to specify the location of the sensors. An algorithm is proposed to decouple the signals and extract the motions such as flexion, extension, abduction etc. They compare their technique with the camera-based motion capture system.

2.2 Image Depth

Depth estimation or extraction concerns the assembly of technology and algorithms to represent the spatial structure of a scene, in other words, to calculate the depth of each point in a scene accurately [16]. The basis of depth is characterised by two methods, active and passive. Active methods work by emitting energy into a scene before passively processing the energy reflected. These methods were proposed before passive methods as micro processing had not

yet been invented [16]. The main disadvantage presented by active methods is the energy required to operate. Nevertheless, it found their reliability to be much higher, and some of them are used to obtain ground-based evidence. Light-based calculation of depth for distance measurement is the first type of energy. An example of this can be seen in experiments with incandescent light. However, many light sources can be used and therefore many different algorithms, configurations and hardware are available as well. Ultrasound based methods utilise the Time of Flight (ToF) principle to measure distances. A real-world example of this technique is when imaging a foetus in a mother's womb [16]. The main point of this proposition is the high accuracy and processing rates (up to 100 fps) on account of CMOS and LED based light.

Passive Methods operate using natural light in the atmosphere where the optical data of the captured image can be used to estimate depth. Such techniques capture images with image sensors, which fixes the computer problem. The advantage of this method is the low amount of operations needed to process a single image, instead of two or more. There are two former classes in this algorithm family, monocular and Multiview solutions [17].Monocular depth estimation is the task of estimating a depth from an RGB image [17]. To obtain the depth map, a single image or a video sequence may be used in this approach. The advantage of using this approach is that a small number of operations is needed to process each image [17]. Depth-on-defocus is an approach that uses monocular information to provide an absolute measurement of distance based on the focus properties of the image. This approach estimates the distance of each point in an image following the human visual focusing system by calculating the defocusing level of such points. This defocusing calculation is done primarily with Laplacian operators, which measure the second spatial derivative in each direction for each point in an N pixel neighbourhood [16].

Multiview depth estimation approaches have different algorithms dealing with two or more images to calculate the depth map. Stereo vision is a case of this array, using two objects. It can use stereo vision and multiview for more than two images for clarification purposes when two images are involved. Absolute measurements in some conditions may be required and the depth-on-focus offers a precise measure of depth in a very narrow field. [16]. In a three-dimensional space, the set of objects is called a "3D scene." Furthermore, the scene is always seen in a particular area. The blurred picture seen at this stage is the scene's so-called projection. This projection consists of a series of rays that cross a small aperture to the so-called projection plane [16].

Some of the depth estimation applications include smoothing blurred image parts, improved 3D scenery rendering, self-driving cars, robotic grasping, robot-assisted surgery, automatic 2D-to-3D film conversion and 3D computer graphics shadow mapping. There are various studies about depth estimations. A hand gesture recognition technique based on depth data is proposed by Dominio et al [18]. To detect a more complete 3D hand gesture, depth information was used to measure the posture by sectioning the hand depth map into its various parts, adjusting the AI approach for the full body that was performed using a Kinect. At first, the hand is extracted from depth maps acquired, along with colour information from associated views. The next step is segmentation of the palm and finger region from the hand. Two feature descriptors are extracted, the first one based on distances of the fingertips from the hand centre and the second one on the hand contour curvature. To recognise the gestures, a multiclass Support Vector Machine (SVM) classifier is employed. A high accuracy is achieved on depth data.

According to Liu [19], a deep CNN model is used to tackle depth estimation from single monocular image problems. It also aims to explore the capacity of deep CNN and continuous Conditional Random Field (CRF). The proposed scheme learns the unary and pairwise potentials of continuous CRF. Moreover, a model based on fully convolutional networks and a novel super-pixel pooling method is proposed which is about ten times faster. This efficient model is a better performing CNN design. Experiments on indoor and outdoor scene data shows that the proposed method outperforms the state-of-the-art depth estimation approach.

2.3 Finger Movement Measurement

Rash et al. [20] were among the first to investigate 3D hand motions. They performed 3D video motion analysis in order to measure hand motions. The goal was to illustrate the validity of this technique by contrasting it with a two-dimensional movement, considered the 'highest quality level'. Their investigation was performed to decide if (1) markers put on the dorsal part of the hand and fingers precisely measure joint edges, and (2) the 3-D method for evaluating finger movements is accurate by utilizing a standard movement investigation framework.

Hue et al. [21] then tracked fingertip positions using two stereo cameras in order to minimise the error between mapped and measured fingertip positions, they constrained an Inverse Kinematics (IK) solver. The user wears the data glove and moves fingers openly while the wrist is fixed on the table. The vision framework records a progression of finger movements and measures the specific fingertip places of each finger. Simultaneously, the uncalibrated crude

sensor estimations of the data glove are additionally recorded. Teleoperation tests show that the human hand model offers adequate precision for teleoperation task. Their experimental results show a degree of error of less than 5 mm. Sridhar et al. [22] used a different approach to recover poses from depth sensor data. First, they extracted the hands by filtering the data and then applying principal component analysis to resolve the orientation. Signaller SVM classifier was used to classify the fingertip locations. After all those steps, a pose estimation algorithm was used. The algorithm was able to match fingertip locations to poses in a database with similar fingertip locations.

Dynamic Time Warping (DTW) based technique for capturing detailed finger and body motions was proposed by Majkowska et al. [23]. Capturing body and finger motion is a challenging task due to the size of motion differences and markers. They suggested two sessions for capturing hand and body motions, where the finger motion was recorded in a smaller area where the subject remained standing or seated. Four markers were placed at the hand, wrist and forearms. The position of the markers allows for later alignment of the hand and body motions. A three-step algorithm was proposed. In the first step, stroke, hold and retraction was matched using acceleration, a velocity profile based DTW. In the second, step frames were aligned to the frames of the full body motion. In the last step, a smoothing of the resulting motions for seamlessly fitting together was applied.

Van den Noort et al. [24] proposed a system they named PowerGlove. This new system has multiple miniature inertial sensors with an opt-electronic marker system. This 3D measurement system can measure the finger motion while performing finger tasks. The subjects perform different finger tasks such as flexion, fast flexion, tapping, hand open/close and circular pointing. The median root and mean square difference for all the finger task presented above was then calculated. It revealed that fast and circular pointing tasks have the largest differences while the smallest differences were observed in flexion tasks. This system measured the 3D hand and finger kinematics and their position in an ambulatory setting. These results may help in hand function and quantifying hand motor symptoms in clinical practice.

Krupicka et al. [25] also developed a measurement tool for objective measurement of the finger tapping test. A contactless 3D capture system using two cameras and wireless reflexive markers were used in the measurement system. An algorithm for extracting, matching and tracking markers was proposed. They compared the performance of their system with OptiTrack, a commercial motion capture system.

2.4 Image Classification

Damasio and Musse. [26] proposed a system using data glove and an artificial neural network system, for recognising hand postures. This system utilizes uniquely designed gloves with flexible sensors to acquire and transmit data to a computer thereby enabling interaction between real and virtual humans. Salomon and Weissmann [27] investigated the mapping angular measurements received from gloves (with sensors) to predefined hand gestures. This is another example of using a classification approach rather than motion reconstruction. They also used neural network classifier, using training sets comprised of 200 hand poses. They made a comparison between the propagation neural network and the radial basis functional neural network. They concluded that simply trained back propagation neural network classifies the poses better than radial basis function neural networks.

Plancak and Luzanin et al. [28] in their experiment, used 5DT data glove 5 ultra, which is cheap glove. Probabilistic neural network was trained on the data in order to classify gestures of fully open and closed hands. The training dataset size is reduced, using some popular clustering algorithms, allowing for faster execution time with slight loss in training quality. Expectation Maximization (EM) clustering algorithm was selected to represent the core of the clustering ensemble.

2.5 Image Processing

To extract some useful information or enhancing the image, some operations are performed on images; the methods of performing these operations on images are called image processing. Just like signal processing, the input is some image, while the output may be some enhanced image, or some attribute related to the input image. Two types of techniques are used for the two types of image. Analogue image processing targets analogue images such as photographs, while digital image processing focuses on digital images. Since the focus of this document is digital image processing so, we will restrain this discussion to digital image processing.

Digital image processing can be defined as processing an array of real numbers presented by a number of bits. The field of image processing plays a vital role and contributes towards the solution of many problems including security, remote sensing applications, industries, medicine, etc. [29]. There exist several steps involved in processing a digital image ranging from image pre-processing, image segmentation, image feature extraction and image classification. The following discussion includes techniques and applications in the area of 2D–3D image and video processing. A research paper related to these techniques is presented for each technique and application. The first part consists of techniques such as Field-

Programmable Gate Arrays (FPGAs) while the second part contains the image processing research papers and applications.

In this section is shown various techniques used in some current studies. Each technique has a capability to be used in a specific field.

2.5.1 Field Programmable Gate Arrays (FPGAs)

Desktop personal computers are used for most of the image processing system [30], which are not capable of fully utilising the power of image processing and so they are more generic. These small systems find it difficult to meet the requirements for image processing when it comes to real-time processing. FPGAs are used for recent image processing systems and the following discussion presents some research work related to FPGAs for image processing.

Desmouliers et al. [31] proposed an image and video processing platform named Image and Video Processing Platform (IVPP), which is based on FPGAs. This platform can be used to process complex algorithms for image and video processing applications. IVPP can be used for a variety of image and video processing tasks, such as recognition and detection, surveillance and encoding/decoding purposes. They developed IVPP using high-level synthesis design flow based on C-language. The Altera Video and Image Processing (VIP) Suite is an assortment of IP centre (MegaCore) capacities that can be utilized to encourage the improvement of custom video and picture preparing plans. In an earlier work Liu et al [32] used FPGA to develop a laser image detector, which helps in detecting flying planes using an algorithm based on calculating Zerike moment of plane. This article presents a real-time Multi-Object-Tracker implemented on a Field Programmable Gate Array (FPGA). [33] presents a system that is capable of tracking for three objects at the same time using diverse algorithms to obtain the best results. It is a user choice to decide algorithm from different algorithms for each object. The algorithm can be exchanged during the run-time by interrupting the object tracking. They used Xilinx FPGA which has a dynamic and partial reconfiguration capability. Internet Configuration Access Port (ICAP) is used to achieve an automatic reconfigurable system, which needed a minimum time to exchange the algorithms. They showed that their design achieved a theoretical maximum throughput of the ICAP of 400 MB per second.

Alali et al. [34] proposed a hardware design based on FPGA. This hardware design is used for various algorithms of image processing, enhancement and filtering. FPGAs can exploit spatial-temporal parallelism, so they are mostly used in real-time image processing. They used the windowing operator (WO) technique for traversing image pixels and later apply filtering

techniques on them. They performed their experiments on 585 x 450 image size and can be used on larger images as far as the memory of FPGA can hold image processing algorithms.

2.5.2 Image Segmentation

Image segmentation is one of the important parts of image processing techniques [35] [36]. Several researchers propose their work on image segmentation from various domains. Since there is no single solution to the image segmentation problem, domain knowledge is combined with these techniques for the solution belongs to different domains. Segmentation is the process of dividing the image into several parts depending on the nature of the problem [35] [36].

2.5.2.1 2D/3D Images

Several researchers proposed different techniques for 2D image segmentation. Xu et al. [37] in their earlier work on 2D image segmentation, proposed an algorithm for grey-level image segmentation. The image is partitioned into connected homogeneous regions. They tried to reduce the time complexity of the region partitioning problem via the minimum spanning tree partitioning problem. Their approach suggests that the partitioning results are satisfactory, as well as it is also insensitive to noise in the image. The experimental implementation using penalty functions provide better segmentation results. These penalty functions also reduced the memory consumption.

Another work by [38] proposed a non-rigid approach for 2D image segmentation. This approach also helps in the 3D–2D pose estimation. This approach fulfils both the tasks while previous approaches which perform both tasks of pose estimation and image segmentation require exact knowledge of 3D. They also show their algorithm robustness to noise, deformation, occlusion and sharp recovery. They generate two 3D training sets comprised of a number "4" and teacups.

Vargas et al. [39], present a hybrid model that is a combination of Imesh image-based segmentation with Markov Random Field (MRF) models. Their model improves the traditional mesh segmentation model which follows a set of well-defined rules, not have statistical information of mesh triangle. They compare their results with Imesh segmentation results and recorded that Imesh+MRF outperform Imesh. They also recorded significant improvements in the computation time of segmentation, as compared to MRF segmentation alone. In the most recent work, Hemalatha et al. [40] exploited the segmentation technique for 2D and 3D images in a medical field. They compared segmented intima-media of 2D and 3D carotid artery images

and implemented in unified technology learning platform. Their technique enhances the processing performance of an image to 120 nanosecond (ns).

Modern medical machines produce huge amounts of high dimensional and high-resolution image data, so there is a need for highly efficient tools for segmentation of these medical images. The latest tools are equipped with state-of-the-art algorithms but most of them are limited to 2D interactive methods like mouse/keyboard. Nyström et al. [41] tackle the issue of 3D image segmentation problems in their research work. With emphasis on the 3D interaction with stereo graphics and haptic feedback, they implemented several segmentation algorithms such as fuzzy connectedness, deformable models, fast marching and live wire which allow the user to interact the 3D image data in an effective way. They performed live segmentation in CT images and recorded high accuracy and precision.

Edwards et al. [42] presented a three-dimensional interactive tool. They proposed an algorithm named Live Mesh, which provided user interaction of efficient 3D image segmentation. their algorithm is based on the concept of Live Wire and 3D implementation of Dijkstra's algorithm. It makes it possible for the user to drag mesh patches over the 3D object. The tool, named SimpleSeg, is an implementation of their proposed algorithm (Live Mesh). A predictive model for object boundary which can integrate information from any source has been presented by [43]. It focuses on the accurate presentation of foreground object, rather than background objects, attributes like in MRF, graph cut and CRF methods. It is also an interactive model and it allows for user interaction in a 3D environment. They used a noisy 3D medical image dataset in their experiments. A similar method for Interactive learning strategies for 3D image segmentation is presented by [44]. They describe the 3D image segmentation problem as a classification problem, and by incorporating active learning, provids the user with ease of automatic iterative input. Based on boundary, regional, smoothness and entropy terms, the given segmentation is evaluated by construction an uncertain field over the image domain. Maximum uncertainty in batch query step is calculated. The user can provide assistance on labelling the data on query plane, providing additional training data. They compare their method with random plan selection and recorded Dice Similarity Coefficient (DSC) improvement of up to 10% in the first five plan selection. They claim that the user can save 64% of their time due to active learning.

2.5.2.2 2D/3D videos

Guimarães et al. [45] proposed a method for video segmentation problems by transforming it into a 2D image segmentation. Video segmentation problems relate to the identification of boundaries between consecutive images, known as transition.. They proposed methods for cut, flash and fade detection. An operator of mathematical morphology and digital topology is used for solving the video segmentation problems. They used pattern detection for an event, where each event is transformed into a 2D image which they named visual rhythm. They performed a comparative analysis of their method for cut detection with other methods and they suggested that their method outperforms other methods. By histogram they distinguished various types of adjustment, primarily blurs. The visual cadence by sub-testing is a rearrangement of the video content spoke to by a 2D picture. The relationship between histogram esteems and dark scale esteems is resolved by standardizing every histogram esteem autonomously. This technique diminished the odds of truncation of data.

Video segmentation has also been discussed by [46], who give an inside analysis of different within and between-frame affinities for video segmentation. In order to obtain good video segmentation, they propose a frame-based super-pixel segmentation combined with appearance and motion-based affinities.

Spatial and temporal video segmentation has been addressed by [47] in their research paper entitled "*Spatial-temporal video segmentation of static scenes and its application*". A segmentation map for each frame is initialised and then link the correspondences among various frames. This is an iterative process where at each step the process is refined and where statistical data is collected until a set of spatial-temporally consistent volume segments is achieved. The application areas are 3D reconstruction, video editing and semantic segmentation, where they demonstrated their method on these applications using some complex videos examples.

2.5.3 Feature Extraction

Feature detection is an important part of 2D and 3D image processing [48]. Before applying any feature extraction technique, the image data is pre-processed and different pre-processing techniques are applied to images, such as binarization, thresholding, and normalisation etc. Features are than extracted and used for classification purposes. Features of an image present the behaviour of an image. A good feature set contains attributes that contain high information gain and can provide good classification of images into different classes. Zhang et al. [49]

present a method that uses symmetric properties from the visual data to detect spare and stable image features. A qualitative symmetry operator with quantitative symmetry range information is used to form the regional features. This method successfully proved that scale from-symmetry is successfully applicable as a modular inset for feature detection.

Extraction and classification of local image structure is discussed by Gevers et al. [50] in their research paper. For the most of image processing and computer vision tasks, such as object recognition, stereo vision and 3D reconstruction, extraction and classification of local image structure are very important. Using the geometric and photometric information, they proposed a method which classified the physical nature of local image structure. This strategy included various types of picture handling systems which are used to standardize pictures. The Gaussian and MoG were considered the most appropriate for feature discovery based on the efficiency and accuracy of the results.

Yalla et al. [51] developed a system that includes a 3D feature detection module and a 3D recognition module. They used a biometric object, where the 3D feature detection module processes the 3D surface map of that object and determines whether there is any type of 3D feature on the 3D surface map. If there does exist any 3D feature, they extracted it along with its type. The next step is that 3D recognition model matches the 3D feature with the biometric dataset to identify the person.

2.6 Image Processing Applications

In the following discussion, light is shed on daily-life applications where image processing is widely used. These covers the medical, the fingerprint detection, the face recognition, the object tracking, and the motion detection fields.

2.6.1 Medical Image Applications

With advancement in medical images such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scan and ultrasound technologies, the need to process the image data in order to extract valuable information increases. This need attracts experts from many fields such as statistics, applied mathematics, biology, physics, engineering, computer science and medicine. In this section, it discusses some the research work in the domain of medical image processing. Mahmoudi et al. [52] presented a detail review of different web-based interactive software tools for 2D/3D medical image processing. In an earlier work [53] proposed an image processing and visualisation algorithm that works interactively for diagnosing moving organs such as the heart during the cardiac cycle. The system is tested on Magnetic Resonance (MR)

and Single-Photon Emission Computed Tomographic (SPECT) images. These MR and SPECT images allow the user to change classification parameters and to zoom or rotate images on the screen.

One of the major problems in the medical field is breast cancers in women, and in men also. It is estimated that approximately 20% of cancer patients relate to breast cancer in developed countries. Sameti et al. [54] suggest a method of feature extraction and classification in order to diagnose breast cancer in its early stage. First, they collect 62 texture and photometric image features and after a stepwise discriminate analysis, six of them can be used to detect the affected and non-affected areas of the breast; 72% average classification results are recorded. This system can be used by radiologists to analyse any pattern in mammograms. The regions identified by the system can have 72% of chances of developing a malignant mass. This could help in earlier diagnose of breast cancer. The goal of this system is to flag the suspicious area.

Segmentation and visualisation of medical images has been discussed by Nystrm et al. [41] where 3D interaction is achieved using stereo graphic and haptic feedback. Creating 3D models of human organs from the 2D images is a hot topic in medical fields these days. These 3D models of CT and MRI scans provide better perspective than 2D images. In order to construct 3D images, sequences based on CT and MRI, surface and volume rendering techniques are mostly used. Geometry of surface and rendering are two-phase modelling techniques for constructing 3D model for complex biological structure. One of such work is by Patel and Mehta [55] in their research work, where they present their Marching Cubes algorithm for modelling. Image segmentation is the necessary pre-processing step used to detect the boundary of regions of interest. This segmentation is especially important in the clinical field as it increases the accuracy of the results. This procedure incorporates volume rendering and surface rendering. The surface rendering is utilized increasingly because of its quick speed and low stockpiling utilization.

2.6.2 Motion Detection

In an earlier work Irani and Anandan [56] provide a unified model for detecting moving objects both in 2D and 3D scenes. The method used in this study is based on seperating the object moving detection problem into different categories based on complexity and using a set of techniques to solve these problems which correspondingly increase in complexity. Examples from real image arrays were then used to illustrate these techniques.

Zhang [57] presents a data mining approach for motion detection in huge surveillance video databases collected by military surveillance cameras. They follow completely qualitative approach, based on signaller system consistency analysis, called QLS. This approach focuses on what is necessary to compute the solution hence reducing the computational cost and increasing the efficiency. Yaun et al. [58] proposed a technique for detection motion regions in video sequences. This technique classifies image pixels into motion regions by applying 2D planar homographs, popular and geometric consistency constraints. Their main contribution is geometric consistency constraints derived from the camera poses from three successive frames. It is implemented within the Plan + Parallax framework.

In 2007, Verbeke et al. [59] presented a Principle Component Analysis (PCA)-based approach to detect motion in surveillance videos. Ten frames are considered, where each of the ten frames are associated with one dimension of feature space. Then they apply PCA to map data in lower dimensional space. These ten frames are than split into blocks. To detect motion within the blocks, inertia ellipsoids of the projected block are used. They recorded very few false positives and satisfying number of connected components as compared to other same purpose algorithms. Automatic detection of motion in human bodies has been discussed by Fablet and Black [60]. Using a low-dimensional spatial-temporal model they develop a presentation model that learned using motion capture data of humans.

2.7 Hand Tracking

Human beings communicate with each other through voice communication or speaking certain languages. Beside voice communication, hands are another means of communication between hearing impaired people. Hand gesture recognition is an interesting topic among researchers from different fields, such as computer vision, human computer interaction and image processing. Researchers from the computer science domains, such as gesture recognition, virtual object manipulation and gaming take a keen interest in the field of 3D hand tracking. Pradipa and Kavitha [61] presented a survey paper on the technique used in gesture recognition. According to them, the main aim of gesture recognition is to develop a system which can detect human actions and use them to extract meaningful information for device control. In the case of human–computer interactions, hand gestures can play a key role where people with verbal disabilities can take full advantages of computer systems. This will also help in reducing the use of hardware devices involved in operating the computer system and as a result lead to less greenhouse gas emitting.

Poudel [62] focuses on techniques to improve the results of accurate 3D hand tracking. He proposed three techniques in order help the HCI society. The first technique is a region-based skin colour technique. A clustering technique based on spatial distance and skin colour to extract the region from the colour. These regions are named as super-pixels. CRF is then applied to improve the results. The proposed region-based technique is used on one of the popular Compaq datasets containing over 14,000 web images. The results reveal that this region-based skin colour technique achieves 97.17% true positive rate over the Compaq dataset. The second technique is a model-based technique and was improved by using proposed depth–foreground–background features, palm deformation module and context cue. One of the major problems associated with model-based techniques is that they expensive in terms of computational cost. In order to overcome this drawback, the third technique, called discriminative techniques was also proposed.

Colour and shapes in 2D images are the main source of information for hand tracking. When it comes to variable light conditions, visual sensor-based hand tracking methods are very sensitive, and when it comes to 3D image, hand tracking techniques used in 2D images do not give satisfying results. Park et al. [63] suggest a new technique for real-time 3D hand tracking using 3D depth sensor and Kalman filter in depth space. They identify hand candidates using motion cluster and predefined wave motion. Hand locations are tracked using a Kalman filter. They compare the performance of their technique with traditional visual based methods and revealed that 3D hand tracking using Kalman filter in depth space outperforms traditional visual based methods.

Manresa et al. [64] proposed an algorithm for tracking and recognising hand gestures to interact with video games. This real-time algorithm includes three steps. The first step is hand segmentation for which they use colour values of human skin due to its variant properties and computational simplicity. The second step is to overcome the errors due to the segmentation process and named it tracking. A constant velocity model and pixel labelling approach is used to perform tracking. The third step is gesture recognition. Several different hand features are extracted during the second steps and finite state classifier is used to identify the hand configuration. There are four motion classes and the classifier based on the hand feature, classify the hand in one of the gesture class. They demonstrate the usability of the proposed algorithm in a controlled video game environment.

Krejov [65] addresses in his thesis, the challenge of real-time hand pose estimation. In this work, the author proposes three approaches. In the first approach, the task of detecting the

fingertips with using sensor is approached. The extrema of the hand are located using graph approach. After that the gestures of the hand are identified. The second contribution is to identify the pose of the hand and region of hand; Random Decision Forest (RDF) is used to the feature that sample depth. In the final approach, machine learning and model-based approaches are jointly used to overcome the drawbacks of both when used in isolation.

2.8 Summary

This chapter presents comprehensive literature review of 2D/3D image and video-related techniques and applications. Starting with the evaluation in 3D cameras research papers related to our topic of interest were discussed. In this section, the scholarly work related to evaluation of the holoscopic 3D camera is also presented. Research papers related to image depth are described in the next section. The third section discusses the figure movement measurement techniques.

After finger movement measurement techniques, the use of classification algorithm for gesture detection were discussed, followed by image processing techniques and applications for 2D/3D videos and images. This presented a detailed analyse of image processing technique such as FPGA, segmentation, feature extraction, and image processing application like medical and motion detection.

Literature work related to hand tracking algorithms and techniques were presented in the final section. Existing and state-of-the-art algorithms were presented in this section, including the current methods proposed by various researchers in the field of hand tracking. The next chapter defines the history of gesture recognition including its types. It describes the different hand gestures and the types of cameras used to detect different gestures.

Chapter 3

Gesture Recognition

3.1 Background

This chapter presents the background of HCI and the history of gesture recognition with its fundamental types. An overview of hand gesture recognition and its types is discussed in this chapter, involving the types of cameras used for 2D and 3D images. Lastly, the chapter will focus specifically on the fundamental concept of holoscopic 3D imaging system camera.

Users interact with computers through the provided interfaces, motions or vocal. These different interactions need to be such that information retrieval is easier and Human Computer Interaction (HCI) is concerned with the way humans interact with technology. It deals with how humans work with computers and how computer systems can be designed to best facilitate the users in achieving their goals. With the advent of third and fourth generation languages, the user interfaces have improved quite dramatically. In future days, Human Computer Interaction HCI will become a field with a variety of sectors that need to characterize it. Users will be able to use any type of interaction which is a potential part of HCI, Interaction can be body movements, facial features and vocals [62][66].

A Human Computer Interaction (HCI) has several types of interaction and one of those is called gestures. One simple definition of a gesture is a non-verbal method of communication utilised in HCI interfaces. The high target of gesture is to design a specific system that can identify human gestures a designedly and use these gestures to convey information for device control.

Recently, HCI has increased in relevance as its usage increases across different applications including human motion acquisition. Initially, it must define the idea of human motion acquisition which records the movements of a human or an object and convey them as 2D or 3D image data. To provide life to 3D digital models or analyse the motions need to study the converted 3D data deeply. Producing a 3D digital object needs special applications and specific tools which are considered exclusive to certain companies [62] [66]. For instance, Disney is one of most famous companies which uses human motion capture as a modern technology in the cartoon world production. Avatar characters, as the first 3D cartoon characters, inspire all production companies to develop their methods in film production. Currently, adults and

children have enjoyed watching 3D movies without realising the way these types of films are produced.

According to Kitagawa and Windsor [67] there were several attempts to apply motion capture technology (MOCAP) by some photographers and producers in the nineteenth and twentieth centuries. Everyone placed his or her thumbprint on this developing technology. Because of that, there are certain major steps needing to be considered. First, preproduction is a part of the procedure that allows the designer to divide everything into parts and organise them before doing anything else. A project pipe signal is the second step which begins with preproduction and ends with post-production, which means the character in a game or animation is eventually where you would prefer the Motion Capture (MOCAP) data to go to. Cleaning and editing data are an important step which needs to be more concentrated and done with high proficiency alongside the last point, which is skeleton editing. However, there are other steps which may be applied depending on the character shape and motion [67].

According to Ye et.al [68], Most of the old approaches depended on the 2D data such as pictures. Recently, the direction of development of the Time of Flight (ToF) cameras and other types of depth sensors became improved by creating opportunities to support this area. The survey presented the overview of traditional approaches which achieve human motion analysis involving depth and skeleton based activity recognition such as facial expression detection, facial performance capture, head pose estimation, hand pose estimation and hand gesture recognition.

One of the primary factors in this research is the matching between the system and the real world which ensures that the system should use the users' movements; following real-world conventions, making information appear in a natural and logical order. The next section defines the subject of gesture recognition in detail in order to cover it accurately.

3.2 Definition of Gesture Recognition

In the present day, HCI is assuming greater significance in our daily lives. Gesture recognition can be named as a method a long this path [69][70][71]. Therefore, what is gesture recognition? In the previous section Gesture Recognition were defined as non-verbal motions used as a method of communication in HCI interfaces [69][70][71]. Gestures are one of the significant aspects of HCI in both interpersonally and in the device interfaces [69][70][71]. Another definition of gestures is physical movements or positions of a human’s fingers, hands, arms or full body used to convert information. Gestures, in a virtual reality system can be used to

navigate, control or interact with a computer [69][70][71]. The process by which gestures are formed in certain ways by a person, are made known to a system, is the main principle of gesture recognition. Signs can be expressed in a multitude of ways by gestures, for example, sign language used by hearing impaired people. Other examples of gestures developed outside the computer field can be seen in use by traffic police, construction labours, and airport ground controllers. Gestures can be static, which means that the user adopts a pose, or dynamic where the motion is a gesture by itself [69][70][71]. Attached devices such as gloves, data suits, Six Degrees of Freedom (6 DOF) trackers generally provide information along all the 3D geometries. For instance, hand and body gestures are used to by pilots to direct aircraft operations aboard aircraft carriers [69][70][71].

Mathematical models based on hidden Markov chains, or methods based on soft computing can handle gesture recognition [69][70][71]. The major advantage of using the hidden Markov model is the ability to recognise a variety of information for gesture recognition [69][70][71]. Any applied implementation of gesture recognition needs the use of diverse imaging and tracking devices or tools such as data gloves, body suits, and marker based optical tracking [69][70][71]. Pens, 2D keyboards, mice and oriented graphical user interfaces are frequently not appropriate to work in virtual systems unlike devices used to sense any part of body orientation and position, facial expression, sound and speech, skin response and other human behaviours or states which may be utilised to present communication between humans and the environment [69][70][71]. Gestures may be static or dynamic or both in certain cases such as sign language. The automatic recognition of gestures needs their temporal segmentation, which usually requires specifying the start and end points of the gesture in terms of the frames of movement, in both time and space. Additionally, the preceding context also affects gestures alongside other gestures [69][70][71].

There are many aspects that have been successfully used for many gesture recognition systems such as computer vision and pattern recognition techniques, including feature extraction, clustering, classification and object recognition. Analysis and detection of texture, shape, motion, colour, image enhancement, optical flow, contour modelling and segmentation are image processing techniques that have been found to be effective [69][70][71]. Gesture recognition uses connectionist methods, including multilayer perceptron, time delay of neural network and radial basis function network [69][70][71].

Static gesture recognition may be achieved by neural networks template matching, and standard pattern recognition. While the dynamic gesture recognition issue includes the use of certain

techniques such as Time Delay Neural Network (TDNN), dynamic time warping, Hidden Markov Models (HMMs), and time-compressing templates [69][70][71].

The last paragraph discussed the principles and background of some of common tools used in gesture recognition [69][70][71]. Essentially, human gestures generate a significant volume of motion uttered by the body, face, and hands [69][70][71]. Recently, gestures have been classified into multi categories; one of these is gesticulation, which is a spontaneous motion of the hands and arms with speech. This means it is combined into a spoken pronouncement, replacing a spoken word or phrase. Another category is a pantomime; gestures that represent objects or actions with or without associated speech. Emblem is also another gesture category which is a familiar gesture such as the V sign which means victory, thumbs up means ok, and various other gestures. The sign language is a linguistic system for example, Universal Sign Language, which is defined very well. It can be defined as a visual language contains three main elements are finger spelling, word level sign vocabulary and non-manual features. A spelling word letter by following the letter is a method used by finger spelling. The second element is utilised for common of communication. Whereas non-manual features are composed of body, facial expression, mouth and position of tongue. One of the protentional fields in gesture recognition is sign language, which is totally useful for hearing impaired people [69][70][71].

For example, robotics can be used to apply the sign language recognition using some suitable sensors used on the body of a patient. By analysing the received values from those sensors, robots can help in-patient therapy. Another example may be stroke rehabilitation. Speech recognition has an ability to record speech and convey it as a text. There are certain types of gesture recognition tools which can record the symbols represented via the sign language into a text [69][70][71].

The operation inside each computer firstly is a gesture which represented as a region in some feature space [69][70][71]. In this approach, the features will be X, Y and Z coordinates of different emphases on the human's body with also the orientations of a few appendages. In an image-based approach, the sensor measures the intensity of a 2D grid of pixels. Generally, the picture is pre-processed, firstly to improve differentiates and to decrease the percentage of noise. The following feature extraction process localises the points around the picture, such as edges. Then, links all these processed features to form the illustration of full limits in the picture. Usually, the illustration limit is the source for a segmentation process which does the separation from each other the regions matching to other parts of human’s body. After that is

done, all positions and orientations of the parts in the picture will be measured and the space of all positions and orientation would be the nominee for the feature space into which the regions of the gesture are well defined [69][70][71].

Sensing machines take each measurement, the computer attempts to recognise the gesture by locating the area in feature space inside which estimation falls and this process may be completed in a pattern recognition or a neural network classifier [67]. These algorithms usually apply a simple model for the gesture areas. Gestures will be defined by storing a limited number of prototypes using algorithms. The provided prototype has a gesture region which will be well-defined as a set of points which are closer to the prototype than other stored prototypes. The prototype similar to the input vector is consistently detected by the recognition process. Algorithms are well-known by the name of nearest neighbour search algorithms. For instance, neural network architecture is a superior network for recognition tasks. By using neural network, the loads of the connections of the network will have stored prototypes. The recognition of a pattern is completed by an algorithm that joins to one of the prototypes [67].

3.3 Types of Gesture Recognition

Gesture recognition has been introduced briefly in the previous sections [69][70][71]. The gestures are made by the user then are recognised by the receiver. It is for the meaningful body motions including movements of the fingers, hands, arms, head, face, or body to convey meaningful information. In this section, some essential types of gesture recognition are addressed briefly.

Hand gesture recognition is one of the understandable ways to generate a convenient, high adaptability interface between devices and users [69][70][71]. Using a series of finger and hand movements through the operation of complex machines is allowed by hand gesture recognition. This technique will eliminate the need for physical interaction between user and device [69]. The next section is introduced facial gesture recognition extensively.

As it is known, the face is a significant feature of a human. A human face is a non-rigid object with a variability in shape, size, texture, and colour. People can recognise and detect features in a scene easily with little or no effort at all. Since the substantial characteristic changes in the visual encouragement because of viewing conditions such as dissimilarity in facial expression, luminance, gender, aging, interruptions or occlusions such as, hair, glasses, hat or other camouflages [69][70][71].

Facial expressions include removing sensitive features from facial landmarks such as areas around the nose, mouth, and eyes of an image [69][70][71]. Frequently, dynamic image frames of these areas are tracked to create appropriate features. Additionally, the dynamics, location and intensity of the facial actions are significant for identifying an expression and the concentration measurement of natural facial expressions is most often harder than that of posed facial expressions [69][70][71].

Facial gesture recognition is an additional technique for generating non-contact interface effectively between users and machines. The main target of facial gesture recognition for machines is to recognise emotions and other communication signs within humans, despite the countless physical variances between users [69][70][71].

The goal of face detection is similar to facial gesture recognition which is identifying and detecting human faces with efficiency despite their scales, positions, poses, locations and illuminations [69][70][71]. Low-bandwidth transmission of facial data, criminal identification, missing children recovery, surveillance, credit card verification, office security, telecommunication, video document retrieval, High-Definition Television (HDTV), human computer interfaces, medicine and multimedia facial queries are examples which require an automatic system for facial gesture recognition [69][70][71].

Automatic facial recognition has two main approaches, firstly analytics which is a flexible mathematical model developed to integrate illumination changes and facial deformation. Discrete local features such as irises and nostrils can be extracted for retrieving and recognising faces [69][70][71]. It may also be implemented on these measurements using statistical pattern recognition techniques such as HMMs [69][70][71]. There are other approaches used in facial recognition involve Wavelet Transform, active contour models and knowledge or rule-based techniques like facial action coding system. The second approach is holistic and involveing grey-level template matching by using worldwide recognition. To represent the entire face template requires using feature vector [69][70][71]. Signaller discriminants, ANNs, PCA, optical flow, singular value and decomposition using eigenfaces are included in the holistic approach [69][70][71].

Many aspects need to be understandable such as overall muscle tension, hand tension, pupil dilation and locations of self-contact [69][70][71]. To specify all these principles, the human body position, configure ration like angles, rotations and movement such as speeds need to be detected. All aspects may be completed through sensing devices attached to the user. The

sensing devices may be magnetic field trackers, data gloves or body suits. Otherwise, using computer vision techniques and cameras can also be called sensing devices. An individual of sensing technology differs laterally, some including accuracy, dimensions, latency, resolution, user comfort, cost and range of motion. The user is required to wear the device and carry cables connecting the device to a computer by using glove-based gestural interfaces [69][70][71].

Normally, there are many meanings of one or more gestures which are unclear for some people. For instance, the raising a hand and the waving of both hands over the head both indicate the concept of 'stop'. It is not just using gestures that have different meanings between different people, speech and handwriting also have these differences. Furthermore, gestures are obtained frequently from language and have a cultural impact [69][70][71]. They may be of these following types: hand and arm gestures are a recognition of a hand pose; sign language and applications used for entertaining, such as kids allowed to play and interact in a virtual reality environment; head and facial gestures, such as shaking or nodding of head, opening the mouth to talk or looks of happiness, anger and fear etc; and lastly, body gestures are full body movements such as understanding the movements of a dancer to create a match between music and graphics and tracking the motions of two humans interacting outside and more [69][70][71].

3.4 Overview of Hand Gesture Recognition

The hand is often well known as the most natural and instinctive interaction for humans' interaction. In the HCI world, an appropriate hand tracking is the first phase to develop instinctive HCI systems that may be used in applications such as, virtual object manipulation, gaming and gesture recognition. Moreover, hand tracking is an interesting principle point which deals with three main parts of computer vision which are segmentation of hand, detection of hand parts, and tracking of the hand. Hand gestures are frequently the most expressive way and the most used in gesture recognition system involving a posture is a static finger shape ration without hand motion and a gesture which is dynamic hand motion with or without finger movements [69][70][71].

A hand gesture requires tracking of 27 degrees of freedom of hand including two major categories, A hand posture is a static hand pose without any movements; While hand motion is any movement of the hand, either the full hand or fingers. A hand movement consists of three major types are data-glove based, vision based and electrical field sensing. Measuring the human body or body parts requires electrical field sensing, and this device is used officially to measure the distance of human hand or other body part from a device. Currently, most of the

significant types almost all researchers are interested in studying, are data-glove-based and vision-based technologies. The data glove based is simply a glove that has multi variety of sensors used to detect hand and finger motions. There are many styles of data glove and each one has its uses, such as MIT Data Glove, CyberGlove III, CyberGlove II, Fifth Dimension Sensor Glove Ultra, X-IST Data Glove and P5 Glove. Each one of these types is will be introduced in detail in Section 3.5 [70].

Recently, vision based is one of the concepts requiring essential development. Simply, it is defined as to detect hand motion using a device such as cameras. Vision based mainly has two approaches which are used in gesture recognition system; model-based and image-based techniques. The main definition of model based is attempting to generate a 3D model of the human hand and use this model for recognition while image based is detecting a gesture by capturing pictures of the user's movement through the sequence of a gesture. Model based, also called spatial gesture models, has two various categories which are 3D model based and appearance based both have diverse types. For example, the 3D model has skeletal and volumetric algorithms whereas appearance based has deformable 2D templates and image sequences. Under volumetric algorithm, there are three other types of algorithms: Non-uniform rational basis spline (NURBS), Super quadrics and Primitives. The next section will provide more information regarding the vision-based concept. Recently, real-time in vision-based is more possible for an HCI system through the assistance of the latest developments in computer vision and pattern recognition field [69][70][71][72].

As usual, there is no perfection in the hand gesture world. It means there are serious issues faced by researchers like, self-occlusion, hand deformation, irregular motion and appearance similarity making 3D hand tracking a challenging mission [62]. The proposed 3D hand tracking technique in this thesis can be used to extract accurate hand gesture features and enable the complex human machine interaction such as gaming and virtual object manipulation [63].

3.5 Types of Hand Gesture Recognition (Data Glove, Vision Based)

Hand gesture is very natural and useful for human machine interaction [71]. This section will briefly discuss the types of data glove and vision-based technique. Figure 3.1 shows the types of hand gesture with a brief definition of each type.

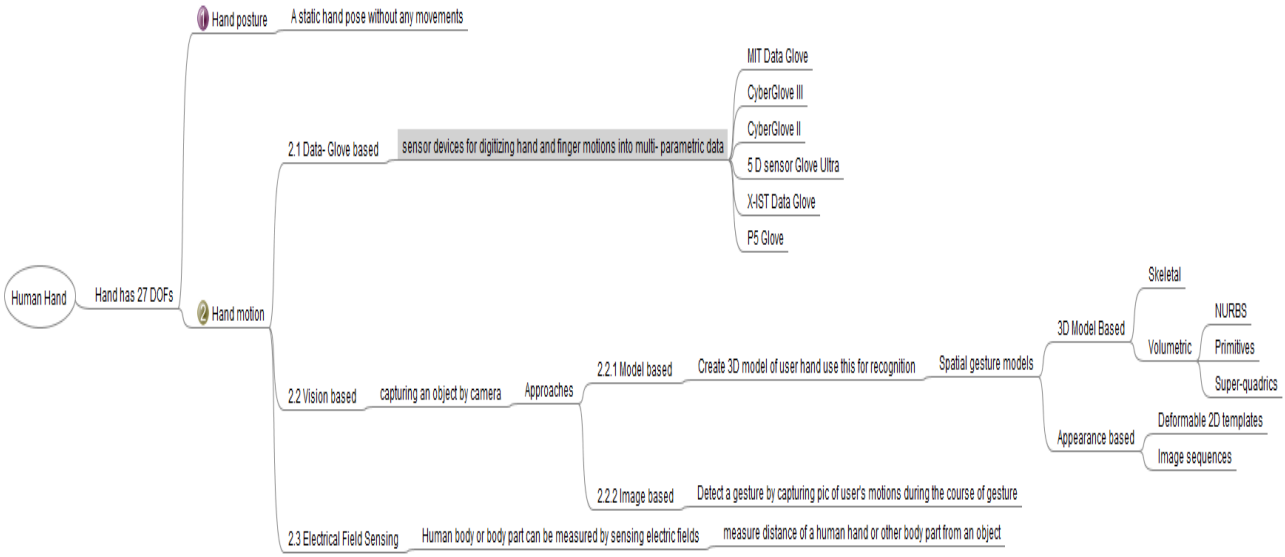


Figure 3.1: Hand Gesture Recognition Map

3.5.1 Data Glove

The history of hand gesture recognition began with the invention of data gloves. Some researchers understood that sign language inspires gestures and that it may be used to suggest simple instructions for a computer [70] [71]. A data glove is a special wired glove with tactile switches or sensors which attach to the fingers or joints of the glove and is worn by a human. Optical goniometers and the tactile switches or resistance sensors estimate the twisting of dissimilar joints when present with basic measurements and that determines if a hand is opened or closed or some finger joints are straight or twisted. A computer is provided results which are mapped to exceptional gestures and interpreted. The benefit of a simple device is there was no need for any type of pre-processing. With very limited processing power on computer back in the 1990s, these types of systems showed some promise regardless of the limitation of manoeuvrability because of cables that were used to connect the data glove to the computer [70].

Lately, since developments in technology, there are now gloves which use wireless technology and may be worn easily unlike the old gloves [70]. This section will address the background of some of these devices and their performances in hand gestures. Data gloves have two different categories which have appeared over many years – active data gloves and passive data gloves [70]. An active data glove consists of many sensors on a glove to measure the movements of

joints or accelerating and have a specific communication track to the host machine via wireless or wired technology. These types of gloves are recognised to confine the user of innovative ability. While a passive data glove contains only pointers or colours for finger detection by devices such us cameras. The glove does not have any sensors on board [70].

Gloves are diverse types, invented from 1977 till recent days, each glove has specific capabilities and functionality. One of the first developed gloves is the Sayre Glove, which was developed in 1977 [70]. It used tubes with a light source at one end and a photocell at the other were riding along each finger of the glove. In 1983, another glove using multiple sensors is called Digital Data Entry glove and it was developed by Gary Grimes. This glove used diverse types of sensors riding on a material. The first commercially available data glove, launched in 1987, was an enhanced version of the first data glove developed by Zimmerman in 1982 which is shown in Figure 3.2 [70] [72][73]. The technology of this glove was similar of the one used in the Sayre Glove [70] [71][72][73].

Figure 3.2: The ZTM Glove [73][74].

3.5.1.1 MIT Data Glove

The MIT Data Glove was a dramatic development presenting diverse capabilities as compared to dissimilar models. The glove was developed by the MIT spinoff company AnthroTronix. The AcceleGlove as shown below in Figure 3.3 [73][74], is a user programmable glove registering hand and finger motions in 3D. This AcceleGlove is used in sports training video games, or body rehabilitation [70].

An accelerometer rests underneath each fingertip and on the back of the hand as shown in Figure 3.3 [73]. The accelerometers may be detecting the 3D positioning of the fingers and palm respect to the importance of when any movement may be made by the hand or the fingers. The precision of these measurements is within a few degrees to let programs differentiate slight changes in the hand location. The glove would allow the user to write or type while wearing the glove [71].

Figure 3.3: MIT Acceleglove with multiple sensors [73].

3.5.1.2 Cyber Glove III

The CyberGlove III is also called a MoCap Glove developed by CyberGlove Systems. The aim of this device is to record gestures precisely for motion capturing used in films and graphic animation industries, as shown in Figure 3.4 [73][74]. Also, the glove consists of Wi-Fi which is used for data communication with a transmission range of 30 m. The single unit has 22 sensors and may run for two to three hours with a rechargeable battery. Likewise, the Secure-Digital (SD) memory card may offer movement recording choices for motion capture animation goals. However, the glove is not targeted at a computer or other peripheral controls [71].

Figure 3.4: CyberGlove III [73][74].

3.5.1.3 Cyber Glove II

The CyberGlove was developed to carry data inputs because of the different flexing motions of joints in the hand. As shown in Figure 3.5 [73] [74], the glove has eighteen sensors which feature two twist sensors on each finger, four capture sensors, and sensors used to measure thumb limit, palm arch, wrist flexion, and wrist capture. Another version of this device includes 22 sensors and has three flexion sensors per finger, four capture sensors, a palm arch sensor, and sensors used to measure wrist flexion and capture. One version of the glove proposes open fingertips which let the user write and type and hold simple objects. The system of CyberGlove motion capture has been used continuously in many applications such as virtual reality, biomechanics, animation and digital prototype evaluation [70].

Figure 3.5: CyberGlove II [73][74].

3.5.1.4 Fifth Dimension Sensor Glove Ultra

The Fifth Dimension Sensor Glove Ultra is a type of glove-based gesture recognition device with extremely high accuracy flexor resolution. The glove contains arrays of sensors which provide 10-bit flexor resolution expected to give highly natural motion capture for film industries. The glove is well known to make high data quality with low cross correlation among dissimilar sensor metrics for real time animations using Bluetooth. Figure 3.6 [73][74] shows the early and present version of Fifth Dimension (5D) Sensor Glove Ultra [70].

Figure 3.6 :5DT Motion Capture Glove and Sensor Glove Ultra. Left: current version, Right: Old version. [73][74].

3.5.1.5 X-IST Data Glove

X-IST Data Glove offers a motion capture result with fingertip touch sensors which may be used for musical applications. The user is not at relaxation while the unit is wired to the computer interface. Each finger joint bend is measured with the movement of the hand. Figure 3.3 [73] shows a glove with a cable connecting the user to the computer [70].

Figure 3.7: X-IST Data Glove [73].

3.5.1.6 P5 Glove

Mind Flux has developed the P5 Glove to offer a cheaper option in the market and which may be used for gaming. As shown in Figure 3.8 [73][74], the P5 Glove combines a twist sensor and remote tracking technologies offering users instinctive interaction with virtual environments and 3D applications like educational software, games and websites. It is one of the rare technologies presently reaching the user as a controlling machine using peripherals rather than the mouse, the keyboard or the joystick. Some of the gloves are used for interacting with a computer for gaming and communication. Some of gloves are used for 3D movie animation and others are used for healthcare applications like controlling of energetic signs, to rehabilitation on injured or healing hands [70].

Figure 3.8: P5 Glove [73][74].

3.5.2 Overview of Vision Based Systems

Gestures recognition is one of the most natural communicative methods between human and computers in virtual environments [70]. Camera techniques are used to identify hand gestures. It started laterally with the early development of the first data gloves. The first computer vision gesture recognition system was reported in the 1980s. Moreover, vision-based recognition is normally natural and comfortable. As shown in Figure 3.9, a flow diagram of a normal gesture recognition plan [71].

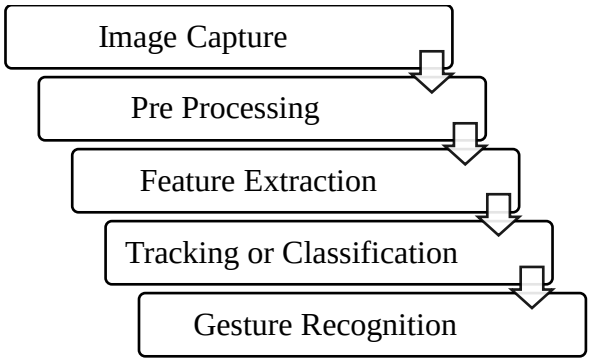


Figure 3.9: Typical computer vision-based gesture recognition approach

Using vision-based techniques require contending with other issues related to occlusion of user's body parts. Although tracking devices had the ability to detect movements of hands quickly while the human's body moving. Vision-based devices could grasp properties such as colour and texture for analysing a gesture, while typical tracking devices may not handle these [4].

Vision-based techniques may also differ between themselves in the number of cameras used, their speed and latency, the structure of the environment, like speed of movement and lighting, user requirements –each user must wear something unique– the low-level features used, such as region, edges, histogram, silhouette and moment; and nor 2D or 3D representation is used and either time is represented [4].

3.6 Types of Cameras

Currently, gestures are detected by different devices while cameras became the first device to detect most gestures. This section will introduce most of the current cameras used in the gesture recognition world. The type of cameras used in gesture recognition, with brief information for each type, is shown in Figure 3.4.

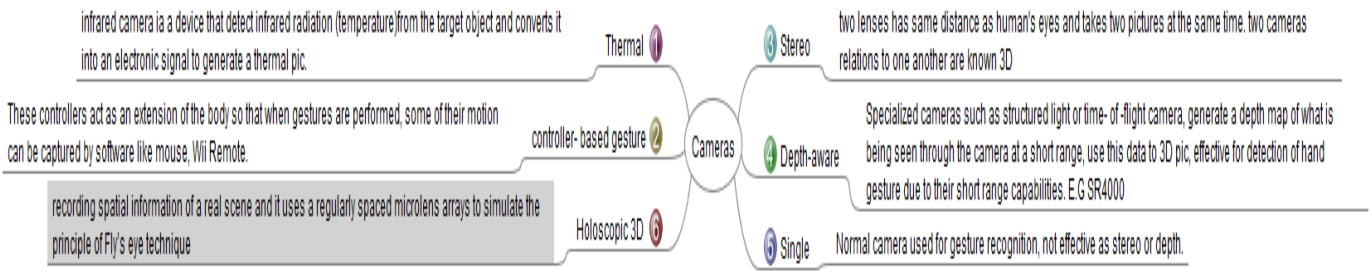


Figure 3.5: Types of Cameras used in gesture recognition

As shown in Figure 3.6 [76], a stereo camera is a camera which has two lenses with almost the same distance separating them, such as human eyes. It takes two pictures at the same time. This copies the method that we humans use to see by and consequently generates the 3D effect when viewed. By using two cameras whose relations to one another are recognised, a 3D representation may be approached by the output of the cameras [75].

Figure 3.11: Stereo Camera [76].

Figure 3.7 [77] shows the depth-aware cameras use cameras such as time-of-flight or structured light cameras. One could create a depth map of what is being seen by the camera at a short range. This data used to estimate a 3D representation of what is being viewed. These cameras may detect hand gestures effectively because of their short-range skills [75].

Figure 3.12: Depth- aware camera [77].

A thermal camera is an infrared camera that detects infrared radiation such as temperature from an object as shown in Figure 3.13 [78]. It converts the temperature of the object into an electronic signal to create a thermal picture on a screen; Or, to make temperature calculations on it. Infrared cameras can capture the temperature and can measure or quantify precisely. However, it is not efficient in detecting hand gestures like other cameras and is negatively affected by the weather. Therefore, the thermal behaviour may be observed but also the relative scale of temperature related issues may be known and distinguished as shown in Figure 3.13 [75].

Figure 3.13: Thermal camera [78].

Controller-based gestures simulate a part of the body. Then, once gestures are made, some of their movements may be captured conveniently by a software as shown in Figure 3.14 [79]. For example, the motion of a mouse device is connected to a sign which is being drawn by a person's hand. Another example is the Wii Remote which may learn the changes in acceleration over time to represent gestures [75].

Figure 3.14: Controller- based gesture [79].

Figure 3.15 [80] shows the single camera which is defined as a normal camera that may be used for gesture recognition where the environment or resources would not be suitable for alternative forms of image-based recognition. A single camera may not be as effective as depth aware or stereo cameras despite a challenge to this concept by Flutter. This is an application that has been released which can be downloaded to Windows or Mac computers with a webcam [75].

Figure 3.15: Single Camera [80].

Figure 3.16 [81] shows the holoscopic 3D camera proposals, the easiest method to accomplish recording and replaying the light field 3D scene. The concept of this technique was proposed by Gabriel M. Lippmann in 1908. The innovative technology contains a microlens array architecture that proposed to double the spatial resolution of a holoscopic 3D camera

horizontally by trading horizontal and vertical resolutions [82]. As shown in Figure 3.17 [83], The holoscopic camera can be in the form of a planar strength distribution, by using MLA [83] [82]. In spite of using the same features of holographic technique, it records the 3D image in 2D form and views it in complete 3D through an optical component, without the required bright light source and restrain dark fine. Moreover, it enables post-production processing such as refocusing [84].

Figure 3.16: Holoscopic 3D camera prototype by 3DVJVANT project at Brunel University. [81].

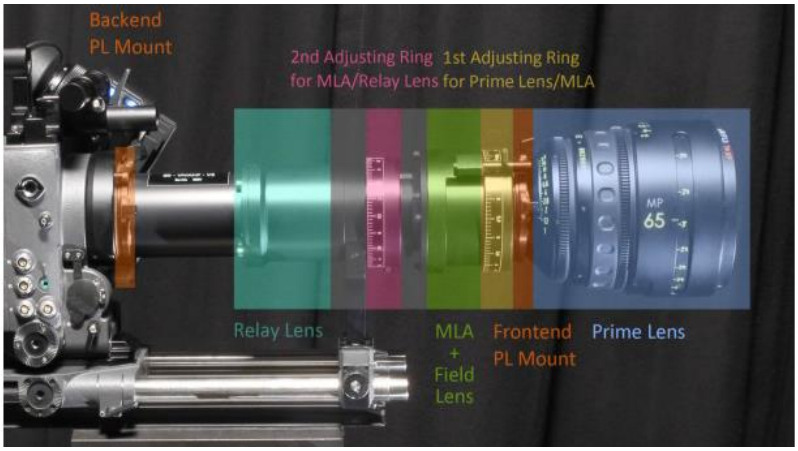


Figure 3.17: 3D integral Imaging camera PL: Prime lens, MLA: Microlens array, RL: Relay lens [83].

Figures 3.17 [83] and 18 [82] show the description of the structure of Holoscopic 3D camera which are L0 = Nikon 35 mm F2 wide-angle lens, NF = Nikon F-mount, AP = adaptor plate, ER = 6 mm diameter extension rods, RM = <5arcminute accuracy rotation mount, MLA = plane of MLA, which is slanted in the process method, T0-T2 = extension tubes, L1 = Rodagon 50 mm F2.8 relay lens ×1.89, C5D M2 = Canon 5D Mark2 DSLR. Arrow displays the position of centre of gravity, SA = Square aperture mouthed to the L0.

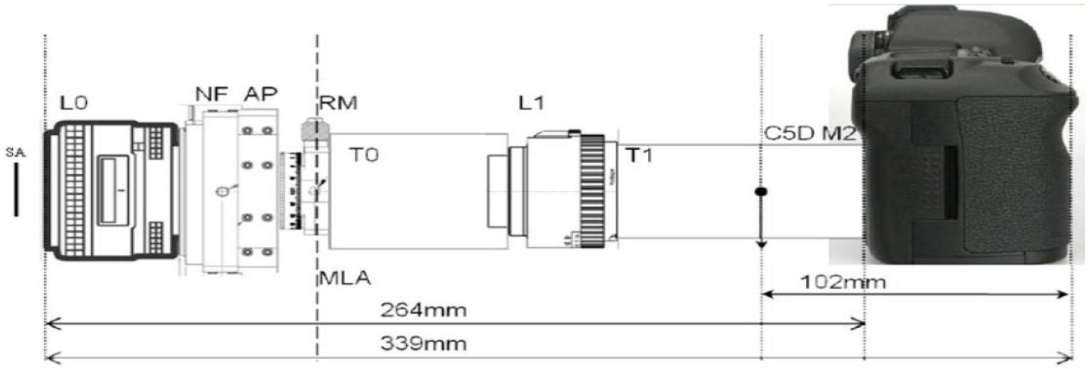


Figure 3.18: Square Aperture Type 2 camera integration with canon 5.6k sensor [82].

3.7 Summary

This chapter introduced the background of HCI as a fundamental field of gesture recognition. HCI is how humans can interact with devices, and how computers respond to humans' requests. There are different types of HCI. One of them, called gestures represent a non-verbal way of interaction applied in user interfaces. Gestures are also defined as a physical movement of any part of the human body such as the finger, hand or full body. For instance, sign language is a model of gestures used by hearing-impaired people. An overview of hand gesture recognition is discussed in this chapter along with its types. Hand gestures are defined as a type of gesture recognition, which is how to move a hand randomly and detect it by certain devices. It has two types, data gloves and computer vision. Data glove is a wired glove with linked sensors attached to fingers or joints of the glove which is worn by a human. Many researchers invented different data gloves. Each data glove has a different purpose. For examples, MIT data glove is used in video games, body rehabilitation or sport training. However, the computer vision is the natural interaction way between humans and devices in a virtual environment. In the computer vision section, it also presents different cameras as types of devices used for detection, such as stereo cameras, depth cameras, thermal cameras, single cameras and a holoscopic imaging system camera were introduced. The next chapter presents the theory of image and video processing tools such as WT and EMD. It also provides a background into AI including ANN, deep learning and CNN.

Chapter 4
Image Processing and Recognition

4.1 Image and Signal Processing

4.1.1 Introduction to Image Processing

Image processing is a way to transform a photo into a digital form and perform other operations on it in order to produce an improved image or gain useful information from it. The input is an image, such as a video frame and the output can also be an object or picture. This is a type of signal propagation [85]. The system of image processing contains different tools such as image acquisition, image enhancement, image restoration, colour image processing, wavelet and multiresolution processing, segmentation and object recognition.

Each tool will be defined as follows: image acquisition is capturing an image and digitising it and then analysing the problem domain then follow the steps according to the problem. Image enhancement is implemented through time and frequency to enhance the image according to the requirements. Image restoration is storing specific parts of an image using a Point Spread Function. The colour image processing tool is used when the image is black and white. Wavelets and multiresolution processing are used if the images are to be rendered in different degrees or wavelets of resolution. Compression reduces the size of images using specific function. Morphological processing is an external structure of the image using dilation and erosion. Segmentation is implemented by splitting an image into different parts. Object recognition used to recognise and save the image description [85][86].

Image processing is faster and more cost-effective. It takes less time to process as well as less film and other equipment to photograph. The Processing of images is more environmentally friendly. No chemicals need to be processed or fixed to capture and process digital images. Printing inks are essential element when digital images are printed. The Microsoft computer vision Application Programming Interface (API) cloud-based tool allows developers to access advanced image processing and data algorithms, transfer pictures or specify image Uniform Resource Locators (URLs), analyse visual content in numerous ways that support inputs and user selections [87]. Amazon Rekognition is a cloud software which is used to incorporate

photo and video analysis to users' applications. It can recognise objects, persons, text, scenes and events of an image or video, or any content that is inappropriate [88]. SimpleCV is an open source computer vision platform that allows users to access various high-powered computer vision libraries such as OpenCV without thinking about bit depths, file formats, colour spaces, buffer management, individual values or matrix versus bitmap storage. Photoshop is a software used to edit digital images [89].

4.1.2 Video Processing

Video processing is based on video data analysis of an allocated time within a video to complete a desired process. Video processing is also a key part of signal processing that transforms the video signal's characteristics, enhances or degrades the video quality, converts the video to a communication channel and storage media etc [90].

Compressed, encoded, and modulated video signal is received from the channel of communication or from the storage device. The signal is extracted, separated, amplified and shifted in the frequency domain if necessary, from other channel or storage signals. It is then demodulated and converted into a virtual stream. There may be one or more compressed video streams, audio channels and data channels in the optical stream [90].

The virtual signal must be demultiplexed in order to extract the specific video bitstream, the corresponding audio streams and the corresponding side data information that could be shown together with video on the screen. Digital video is decoded in the video decoder and stored in backend video processing before appearing on the screen. Side data can be incorporated into video signal side graphic information via the graphic engine [90].

The processing of video starts from a video source in the opposite direction. The source is a video camera in most cases. Devices similar to backend storage, compressed by a video encoder, and multiplexed with audio or other streams will pre-process the signal. The video can be stored or modulated in the storage media at the output and transmitted to the channel of communication [90].

Video processing has various functionalities including object detection, video compression, video filtering and video segmentation. Each functionality is defined as follows: object detection is detecting object instances in a picture in a given class. The object detection objective is to identify all instances of known class objects, including men, cars, or image eyes. Video compression is a vital technology which has developed in all media and communication

fields, including digital television and digital cinema and visual communication in the last four decades, allowing a complete transition from analogy to digital video. It has revolutionized our visual media consumption and communication by turning the internet into the world's first media and visual information exchange environment [90]. Video filtering is multi-frame filters can improve performance of one-frame (image) noise and restoration methods while conversion of video format and reconstruction of a super resolution are inherently multi-frame filtration problems. Lastly, video segmentation is a process where video is partitioned into temporal, spatial and spatial-temporal sections which are homogenous in feature space [90]. Effective video segmentation includes the appropriate choice of attributes and the distance measurement [90].

Some examples of using video processing techniques are traffic applications to detect objects like number plates. Space and target detections beside automatic speech processing are also utilizing video processing tools [90].

4.1.2.1 Empirical Mode Decomposition

The Empirical Mode Decomposition (EMD) provides benefit to the adaptive techniques of data analysis to analyse the non-stationary and non-linear data. The EMD also provides its customers with higher performance and increases their profitability and data analysis techniques [91][92][93]. It provides higher effective data analysis process for improving success in refreshment achievement. The EMD is related to the second generation of the EMD SE which was discussed the progress of the EMD and the specialties of the Intrinsic Mode Functions (IMFs) and PE. The main feature of the EMD is monitoring the activities of the data analysis mode as well as representing the detail about the technology of users and technology burnt during the running. This device is specially designed for the PE users to know keep the data analysis [91][92][93].

The EMD includes different features that support the detection of the frequency variation as well as connectivity with the computer through the technology port. The main feature is related to the improved time series that enhances the durability of this mode. The frequency changes in generation of the technology is based on the new operating systems along with the new applications that are known as permutation entropy filter [91][92].

The innovative technology used in both non-stationary and non-linear data [92]. The functionality of this method is based on decomposing a signal into IMF with respect to the time

domain [92] [93]. The EMD method could be compared to other analysis techniques such as Wavelet Transforms and Fourier Transforms [92][93]. EMD technology might be applied to data related applications such as seismic readings, results of neuroscience experiments, electrocardiograms, gastroelectrograms, and sea-surface height readings [92] [93]. The EMD is defined as follows:

$$x(t) = \sum_{n=1}^{N} c_n(t) + r_n(t) \quad (4.1)$$

where r_n is the mean trend of x (t), the value of cn are the amplitude and frequency modulated output set. The frequency decreases as the value of cn increases.

4.1.2.2 Wavelet Transform

The Wavelet toolbox software is chosen because it provides effective information about the innovative strategies of continuous and discrete analysis that has introduced least asymmetric wavelets reportedly to focus on the software for analysis. It is also defined as one of the image processing algorithms performing the signal analysis where signal frequency differs at the end of time [94]. The technology provides information about the digital services provided by the wavelet toolbox. The technology also has detailed information about this innovative data analysis technique that has reported the time and frequency analysis that is located in antisymmetric of wavelet.

Furthermore, the Wavelet Toolbox software was built through the cooperation of a team and members have used hardware to increase the efficiency of this data analysis technique. In addition to that, the technology has covered the B-spline biorthogonal wavelets. The technology has presented the information about the relationships of team and image and signal characteristics to invent a new technological software about sharper frequency resolution. Furthermore, the technology describes the plans of the frequency resolution, which are looking forward to introducing data analysis technique with effective features for the functionality tools [94].

The technology is based on the changed focus of Continuous Wavelet Transform (CWT) and Inverse Continuous Wavelet Transform (ICWT). The Wavelet toolbox has declared that the wave information will be focusing on software rather than hardware. The wavelet family has experienced much short-term research regarding hardware and now it wants to experience software development. The Wavelet Toolbox has provided the discrete wavelet analysis that is now committed to wearable hardware [94]. The management of Daubechies Wavelets (dbN)

wavelets are the Daubechies' which will be progressing in wearable technology. The Daubechies orthogonal is known as dbN wavelets where N is the number of fading moments. The application is commonly used for audio, speech, image, video, and bio-medical imaging. The Daubechies wavelets are defined as follows:

$$\int x^n \psi(x) dx = 0, \qquad n = 0,1, \dots, K \tag{4.2}$$

The equation has a combination of scaling functions that are used to represent numerical approximations on a secured scale. The value of K is directly proportional to the orthogonality condition.

4.2 Computer Vision Systems

Computer vision is a field that aims to enable computers to interpret, recognize and process objects in the same manner as human vision. It is similar to giving intelligence and instincts to a human computer. In fact, it is a difficult task to recognize computer images of different objects. Computer vision is closely associated with artificial intelligence because machines need to understand what they see and then interpret or act appropriately [95].

Computer vision architecture involves processing digital images via different stages successively. The first stage is image acquisition which captures an image and digitalizes it and then analyses it according to the problem domain. Image Processing is the second stage which is a method to transform an object into a digital form and perform certain operations on it in order to produce an enhanced photo or obtain useful information from it. Image processing is also a form of signal dispensing where the input is an image, such as a video frame or image, and the output can be an object or image-related features. The third stage is image analysis, which extracts a piece of information, and data processing. This method is typically necessary to ensure that certain assumptions suggested by the system are satisfied before a computer vision approach can be applied to image data. Feature Extraction is the fourth stage in computer vision systems which extract features of the object and are derived from the image data at different levels of complexity. The fifth stage is detection/segmentation where a decision is made at some point in the processing whether points or regions of the image require further processing. High-level processing is the sixth stage where the input is usually a small set of data at this level such as a set of points or an image area that should contain a specific object. Lastly, decision making consist of releasing the final decision needed for the application [95].

A sparse 3D point model of a large complex scene can be reconstructed from hundreds of partially overlapping photographs [95]. Stereo matching algorithms can create a detailed 3D model of a building facade consisting of hundreds of photographs taken from the internet. Object tracking algorithms can track a person walking in a street. Face detection algorithms combined with colour-based clothing and hair detection algorithms are able to identify individuals in an image. Combining Computer-Generated Imagery (CGI) with live action videos by monitoring the source video feature points measure 3D camera motion and scene form. Automatic authentication in the form of fingerprint recognition and face detection is also the domain of Computer Vision [95].

The applications of computer vision include Optical Character Recognition (OCR)-interpreting handwritten letter codes and Automatic Number Plate Recognition (ANPR) [95]. An example of a computer vision application is a machine inspection where the quick quality inspection of aircraft wings or auto body parts or X-ray vision defects in steel casting using stereo vision with special lighting [95]. It is also used in retail to classify items for automated checkout lanes. It is used in 3D model creation (photogrammetry) where completely automated 3D photographic aerial models are used in applications like Bing Maps [95]. Moreover, the field of medical imaging utilises computer vision currently and is applied in several ways including capturing preoperative and intraoperative images as well as to perform long-term brain morphology studies in individuals as they age [95].

4.3 Artificial Intelligence

AI is the ability of a machine to perform cognitive tasks and act intelligently. The field of AI tries to understand intelligent entities [1]. AI is a new discipline that began in 1956. With a help of AI, it is possible for machines to learn from their own experience, adapt to new inputs and perform human-like tasks. AI is widely used in finance, education, healthcare, transportation fields and in other industries such as computer vision, medical diagnosis, robotics and remote sensing [96].

The father of computer science and AI is Alan Turing who proposed a ‘Turing test’ in 1950 which was designed to provide an operational definition of intelligence. If a machine passes this Turing test, it is said to be intelligent. But no machines have completely passed this test as of yet. There are other indicators of intelligence such as Intelligence Quotient (IQ) tests and brain size, but none of them convey intelligence in machines. According to Daniel Gilbert,

there is one fundamental element in which our minds differ from the minds of animals and computers; it can experience something that has not yet happened [96].

Intelligence is not defined by behaviour but rather by prediction. Humans can read at a high speed by predicting the future of a sentence at a high rate. It is only when your brain predicts badly that you suddenly feel blocked. Humans are not the best decision makers, and this is what AI needs to make better decisions for users. Factors affecting human decision making are loss aversion, sunk-cost effects, farming effort and omission bias [96].

Computers and robots can exceed the human ability at some tasks that are considered to be 'intelligent' using techniques such as data mining and pattern recognition etc. Lower cognitive tasks that are natural for humans can be extremely complex for machines. For example, a vision system and object recognition, partially concealed objects, same object, different shape, colour, texture and size consistency [96].

Weak AI are machines that are able to act as if they are intelligent, but their thinking is simulated thinking and not real. Strong AI are machines that act as if they are intelligent and they are thinking. Unfortunately, we still only have Weak AI. If a machine passes the Turing test, it is considered as a Strong AI [96].

4.3.1 Artificial Neural Network

ANN is defined as an interconnected assembly of nodes like the neural structure of the human brain and can solve different types of problems in an easy manner. The brain works by learning from experiences [97] [98]. ANN is a system that processes information in a similar manner to the biological nervous system. The key aspect of this system is the unique structure of the information processing system [97] [98]. The system is composed of a large number of unified processing elements working together to solve certain issues [97] [98]. It is specifically configured for data classification or pattern recognition applications via learning processes. The architecture of an ANN is composed of three main layers including an input layer, the hidden layer (one layer or more) and the output layer.

ANN can be trained using a supervised or unsupervised approach. In a supervised approach, ANN is simply trained by matched input and output while the unsupervised approach is an attempt to obtain the ANN to realize the structure of input data. [97] [98]. There are several benefits associated with using ANN such as self-learning and large data handling. The advantage of using an ANN is ANN has the ability to learn and train data models for non-linear

and complicated relationships. Different applications may be used by an ANN such as image processing, object detection and forecasting [97] [98].

4.3.2 Deep Learning

Deep learning is a machine learning based model that instructs systems to perform the task the humans likely to do [99][100]. For instance, deep learning is the basic technology behind the automated cars, helping them to sense the traffic signals and pedestrians. It is also the main idea behind the recognition of audio and voices in different devices such cell phones and tablets. Deep learning becoming famous because it is doing the tasks which could not be performed earlier. A deep learning model is based on the layers of the data which could be pictures, text, or audio, into different and small classification layers. Artificial Intelligence could provide 100% accurate results with close to the human level accuracy and even exceeding the human pace. These models are trained by using large data sets and machine learning techniques such as CNN or ANN which contains many classification layers [99][100].

In machine learning techniques the system would guide how to use the model accurately on the graphics, audios, and text. Deep learning models give precision based accurate results even exceeding the human level. These models are framed according to the data given and transforming that data into artificial neural based systems containing large layers of classified data. Deep learning attains more precision and accuracy ever than before which help it to meet the users' expectations. It is used in useful applications such as automated cars. The advances in the past years have shown that artificial intelligence can even surpass the humans in classifying images [99][100].

Deep learning needs large amount of classified data. For instance, developing automated cars hundreds of thousands of images and videos. Deep learning requires an excessive amount of power. Elite GPUs have an equal design that is proficient for deep learning. Cloud computing and clusters when combined takes less time as compared before when it took weeks [99][100].

As deep learning is consisting of neural networks, so it is also known as deep neural networks. The expression "deep" typically mentions to the quantity of concealed layers in the neural system. Usually neural networks just contain 2-3 concealed layers, while deep systems can have upto of 150 layers. To implement the deep learning models, they must train them. For training these models they need large number of labelled data and neural networks. This will help them to learn the features directly from data without any kind of human interaction.

The CNN is one of the most famous deep neural networks algorithms. It stands for Conventional Neural Network. It involves classified layers of input data and uses 2D convolutional layers to process 2D data [99][100].

Contrastive Divergence (CD) algorithm is different training method to approximate Maximum-Likelihood (ML) learning algorithm which represents the relationship between weights and its error, and it called the gradient. This method implemented to learn the weight of the Restricted Boltzmann Machines (RBMs) with gradient ascent. The formula of this method is shown as follows:

$$\Delta w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\partial \log(p(v))}{\partial w_{ij}} \tag{4.3}$$

The $P(v)$ is the probability of the visible vector which is given by $p(v) = \frac{\sum_k \exp(-E(v,h))}{z}$. The $E(v,h)$ is the energy function allocated to the state of the network which is given by $E(v,h) = -v^T W h$. $\frac{\partial \log(p(v))}{\partial w_{ij}}$ has the simple form of $\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}$. $\langle \dots \rangle_p$ signifies an average with respect to distribution p. Sampling $\langle v_i h_j \rangle_{model}$ requires alternating Gibbs sampling for a long time. The CD changed this phase by running alternating Gibbs sampling for n phases. After n phases, the data are sampled, and that sampled data is utilised in place of $\langle v_i h_j \rangle_{model}$.

Deep learning applications are utilized in projects from computerized heading to clinical devices [99][100]. Automobiles companies are using machine learning models identify traffic signs etc. Due to use of Deep learning accidents of walking people has significantly decreased. In aerospace and defence deep learning is utilized to recognize objects from satellites that find special regions and distinguish guarded or unguarded areas for troops. Cancer analysts are trying to identify malignant growth cells using deep learning models and artificial intelligence. UCLA teams have manufactured a microscope that includes a high-dimensional informational sets used to train a deep learning application to precisely recognize cancer cells. Use of Deep learning models helps workers in their field area where there is heavy machinery by identifying people and things in the safe and unsafe zones. Deep learning is being used in the recognizing the audio and voice, such as the devices that detect your speech and give the results according to it. These all functions are done by deep learning [99][100].

4.3.3 Convolutional Neural Network

A Convolutional neural network (CNN) is a type of artificial neural network specifically designed for image recognition [101][102][103]. A neural network following the activity of human brain neurons is a patterned hardware and/or software system. CNN is also defined as a different type of multi-layer neural network and each layer of a CNN converts one amount of activations to another through a function. CNN is a special architecture used for deep learning [101][102][103]. CNN is frequently used in recognizing scenes and objects, and to carry out image detection, extraction and segmentation.

CNN can be categorised in two phases, namely Training and Inference. To build a CNN- based architecture, it applies three key types of layers: Convolutional Layer, Pooling Layer and the Fully Connected Layer. The first layer is a convolutional layer which is the main block of CNN. It takes many filters that are applied to the given image and creates different activation features in the picture. The second layer is pooling which is used to downsample. It will obtain input from non-linear activation and the output will depend on the window size. The last layer is fully connected where a target is identified to determine the category of final output. Due to the three layers, which removes the necessity for feature extraction by using image processing tools, the image data is learned directly by CNN. CNN causes the recognition results to be unique and it might be retrained easily for new recognition missions while it is allowed to build on the pre-existing network. All the following factors have made the usage of CNN significant in the last few years [101][102][103].

If a correct filter is applied to the temporal and spatial dependency in an image, it can be effectively captured by CNN. The number of parameters (weights) will increase rapidly in a neural network with fully connected neurons as the size of the input increases [101][102][103]. A convolutional neural network reduces the number of parameters with fewer connections, mutual weights and down sampling [101][102][103]. Weight sharing is another major feature of CNNs. CNNs are an efficient extractor for a completely new task or for problems in photo performance, text, audio, video recognition and classification functions. A Convolutional neural network also reduces the number of parameters with fewer connections, mutual weights, and down sampling. Besides that, CNNs remove the need for manual processing of features then discovers the features directly [101][102][103].

CNN Algorithm contains convolutional layers that are represented by an input called map I, many filters K and biases b. In the images case , It may have as input which is an image with height H, width W and $C = 3$ channels which is red, blue and green such that $I \in$

$R^{H\times W\times C}$ Consequently for many D filters will have $K \in R^{k1\times k2\times C\times D}$ and biases $b \in R^{D}$, one for each filter. The output from this convolution process is shown as follows:

$$(I * K)_{ij} = \sum_{m=0}^{k_1-1}\sum_{n=0}^{k_2-1}\sum_{c=1}^{C} K_{m,n,c} \cdot I_{i+m,j+n,c} + b \tag{4.4}$$

The convolution procedure implemented previously is the same as the cross-correlation, exclude that the kernel is flipped horizontally and vertically. For simplicity purposes, It should utilize the argument where the input image is grayscale such as single channel $C = 1$ .The Equation (4.3) will be transformed as follows:

$$(I * K)_{ij} = \sum_{m=0}^{k_1-1}\sum_{n=0}^{k_2-1} K_{m,n} \cdot I_{i+m,j+n} + b \tag{4.5}$$

Search engines, recommender systems and social media are the primary fields to use a CNN in identification and classification of objects. Social media, identification procedures and surveillance are using face recognition which is worth mentioning separately [101][102][103]. This image recognition section involves more complex images such as pictures that could have human or other living beings, including animals, fish and insects. A banking insurance using optical character recognition has been designed to process symbols that are written and printed [101][102][103]. The medical image involves a whole lot of additional data analysis that will spur the initial recognition of the image. A CNN medical image classification detects microorganisms with higher accuracy than the human eye on the X-ray or MRI images. Drug discovery is another important area of health care that uses CNNs extensively. CNN is one of the most innovative implementations used in various fields [101][102][103].

4.4 Summary

The theory of image processing is explained in this chapter. Image processing is a method of applying some techniques to digital images. The applications used in image processing are a remote sensing, entertainment and geological processes. The second theory mentioned in this chapter is video processing which is an analytical technique implemented on video data that is allocated for the time and operation to achieve essential processes. Image processing algorithms such as empirical mode decomposition and Wavelet Transforms are discussed in this chapter. EMD is a method to analyse the non-stationary and non-linear data while WT applies signal analysis where signal frequency changes at the end of time. For classification,

this chapter presents one of the common classifiers, Artificial Neural Network. The main definition of ANN is an electronic model like the structure of human brain neurons. The functionality of ANN is the first node (input) will feed data to a set of hidden nodes to train, then will be classified in the end by producing one node or more (output). The last theory is one of deep learning method namely convolutional neural network. CNN is a multi-layer neural network and each layer of a CNN sends amounts of activations to another layer via function. For feature extractions, Convolution2DLayer, Rectified Linear Unit (ReLU) Layer and MaxPooling2DLayer are applied to data before transforming it to classification layers which are fully-connected layer, Softmax layer and classification output layer. The following chapter presents the first and second experiments which involve 2D video gesture recognition using WT and EMD for extraction as well as ANN and CNN for classification.

Chapter 5
2D Video Gesture Recognition

5.1 Introduction

The hand is often well-known as the most natural and instinctive interaction for humans. People often tend to communicate signals and messages non- verbally using hand gestures. Sign languages have been the only way to communicate with hearing-impaired people for a long time. In Human-Computer Interaction (HCI) world, an appropriate hand tracking is a tracking phase which helps to develop instinctive HCI system that can be used in applications akin to virtual object manipulation, gaming and gesture recognition. Moreover, hand tracking is an interesting principle point which deals with three main parts of computer vision that are hand segmentation, detection, and tracking. The hand gesture is one of the expressive ways used in healthcare, education and the entertainment industry too that could be used by special needs people and elders who become partly incapable of movements.

Hand motion may be detected using any type of camera that offers reasonable image quality. 2D cameras can easily be used in detecting most hand motions on a constant surface such as Microsoft that designed a depth camera with motion sensing named Kinect. Alternatively, Intel produced a small Interactive gesture camera that has reasonable specifications. Apple is also one of the leaders in gesture recognition worldwide by designing the latest versions of the iPhone with a small high-quality camera. The video recording is up to 4K at 24, 30 and 60 frames per second (fps).

It is known that the video contains a large number of images that are connected together to form a clip. An image will down sample into a number of frames thus each frame gets into several stages for image processing. The functionality of all image processing techniques is similar with few differences in each phase. Image processing phases consist of five phases, which are data input, pre-processing, image segmentation, feature extraction, and classification.

There are many detection methods available for hand gestures. In this study, a system is created for hand gesture recognition using the following image processing tools, namely WT, EMD methods, ANN and CNN for gesture classification. These methods are evaluated based on execution time, accuracy, sensitivity, specificity, positive predictive value, negative predictive

value, positive likelihood, negative likelihood, receiver operating characteristic, area under ROC curve, root mean square including Standard Deviation (SD). WT and EMD methods will be employed in two dimensions. For the classification, ANN will be used to classify the gestures using the features extracted as inputs to the ANN. Multiple training sessions would be done where filters would be applied. Each output of one stage will be the input of the next one. All previous stages will be shortened by using CNN deep learning tool. The objective of the study is to identify the best method available to extract features when the classification accuracy is being compared using WT, EMD, and CNN as deep learning techniques.

The remaining of the sections are structured as follows: The details of the proposed system's implementation in Section 5.2. Section 5.2.7 focuses on the discussion and presentation of the results obtained. The conclusion is presented in Section 5.4.

5.2 Short Distance Gesture System Implementations

5.2.1 Hand Gestures Input

In this study, hand gestures are the input to different gesture detection algorithms. Figure 5.1 shows 10 various hand gestures which are recorded in short distance with a plain background that is used in this experiment. Some motions are 2D while others are 3D.

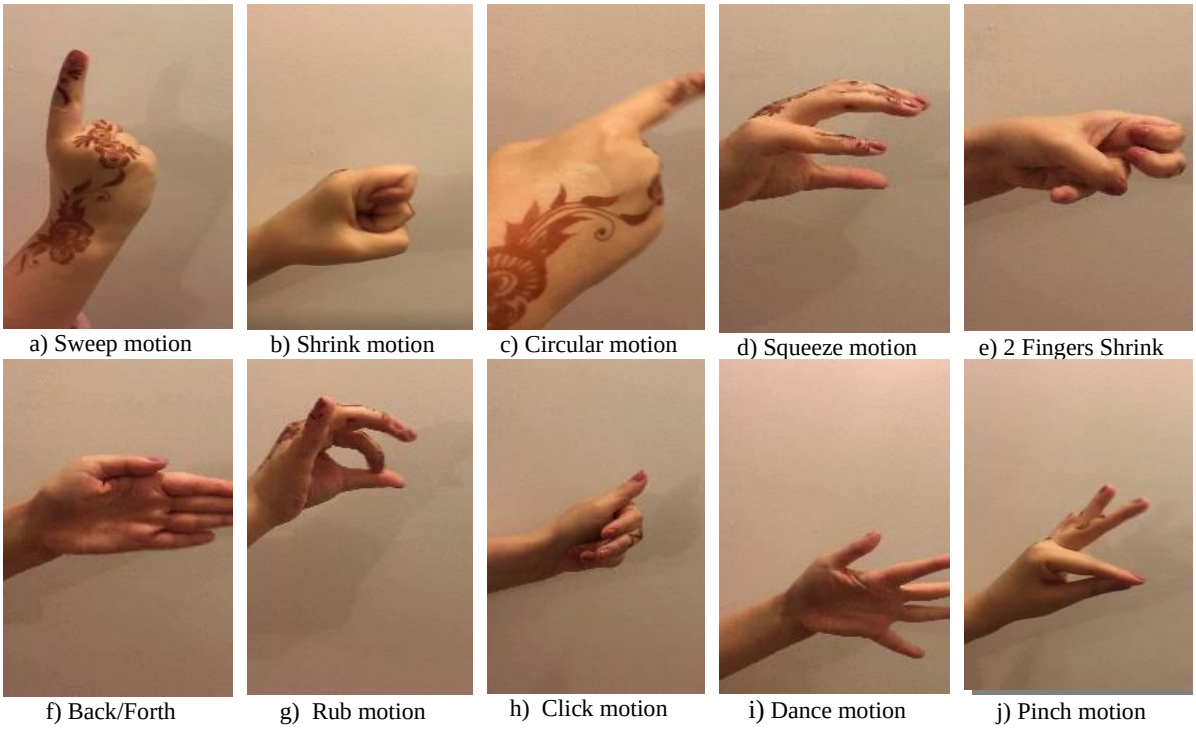


Figure 5.1: Different hand gestures.

Figure 5.2 shows the framework of implementation, explained in the next steps, used to extract and classify hand motions precisely. By using the iPhone 6 Plus camera resolution 4k at 30 fps,

ten different hand motions in short distance are recorded with a plain background. The dataset is the author's hand and it is uploaded directly to the PC to prepare them for multi-processes. Each recording lasts 10 seconds and the resolution of the recorded video is 3840 × 2160. The first system is created using optical flow object by estimating and displaying the optical flow of objects in the video. The length of videos is between 15 to 65 frames. Each video has a different number of frames, which depends on the first section of motion.

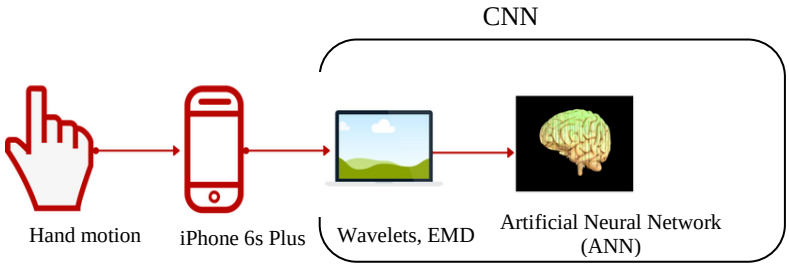


Figure 5.2: Illustrated framework of system implementation.

5.2.2 Computing Platform Specification

The experiment was performed using a Dell laptop XPS 15 9550 with 6th processor Generation Intel Quad Core i7, memory type DDR4 16 GB speed 2133 MHz, 512 GB storage hard drive, 15.6-inch Ultra-HD 15.6" IPS 1920×1080 RGB Optional 3840×2160 IGZO IPS display w/Adobe RGB colour space and touch. The operating system platform is incorporated within windows 10 of 64 bits. The system is implemented using MATLAB R2017bV language.

5.2.3 Feature Detection using Wavelet Transforms Algorithm

The system is implemented using the db8 WT tool into the following stages:

1. Read each video using video reader function
2. Create optical flow object that spreads the object velocities in an image.
3. Estimate and display the optical flow of objects in the video.
4. Divide a video into certain frames; each frame contains 8 IMFs.
5. Apply appcoef2 function which is used to compute an approximation coefficient of 2D signals.
6. Extract each level using *wrcoef* function to reconstruct the coefficients of each level in the video.
7. The execution time of WT is estimated only once.

8. The image data is trained and tested using a Neural Network system. The NN has 20 hidden neurons in a single hidden layer to train data and it stops when the error is reached in 20 epochs. When we add more hidden layers and increase the depth of the neural network, the neural network model becomes a deep learning model. Thus, one single layer is selected for this experiment.
9. The execution time of image data training and testing are also calculated.

5.2.4 Empirical Mode Decomposition Algorithm

The implementation of EMD is similar to WT with some variances. The same 4 steps of WT are used, but with different functions: reshape function (returns the M-by-N matrix) whose elements takes column-wise from X, *ceemd* function (a noise improved data analysis algorithm) complementary collaborates with EMD.

5.2.5 Implement Convolutional Neural Network (CNN)

Deep learning has an intelligent method such as CNN, which is used to train data without requiring any image processing tool. In our experiment, we made a new directory for each video. 10 images are generated to transfer the image frame RGB to grey and resize it to 48×27. All videos have 70 frames. The image's data is split into training and testing datasets. The CNN topology is created in 7 layers; each layer has the following functionality and size: ImageInputLayer Input size [48,27,1], Convolution2DLayer Filter size [5,5], ReLULayer (Rectified Linear Unit), MaxPooling2DLayer Pool size [2,2], FullyConnectedLayer Input size [auto] and Output size [12], SoftmaxLayer and ClassificationOutputLayer Output size [auto]. The hyperparameters of the CNN is generated inside training options function. The value of max epochs parameter is set to 200 epochs.

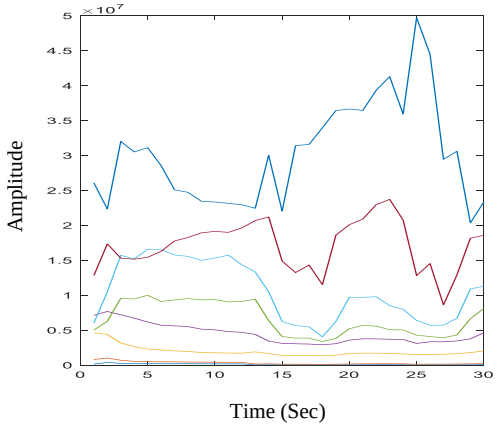
5.2.6 Parameters Selection

In this study, the WT, EMD and CCN algorithms will be compared based upon the following explained parameters. Execution time is the processing amount of duration taken by the software to process the given task. Sensitivity simply measures the percentage of positives which are properly identified. Specificity is a measure of the false positive rate. The PPV and NPV are the percentages of positive and negative results in diagnostic and statistics tests which also described the true positive and true negative results. The LR+ and LR− are one of known measures in diagnostic accuracy. Area under ROC curve (AUC) is the typical technique to measure the accuracy of predictive models.

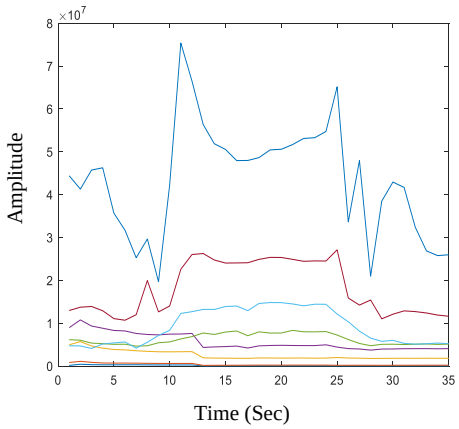
5.2.7 Short Distance Results and Discussion

The system is implemented ten times to obtain the mean of ten-hand motions. Standard deviation is a measure of how extensively values are different from the average value of the group. Two different results in training, testing represented and compared by finding the best tool to detect micro gestures. The training accuracy is achieved by implementing a model on the training data and obtaining accuracy of the algorithm, whereas the testing accuracy is an accuracy for the testing data. The three provided algorithms (WT, EMD and CCN) in this chapter are characterised by different measures. The total time execution for WT in four stages (Data input, Pre-processing, Image segmentation, Feature extraction) is less than the total time execution of EMD.

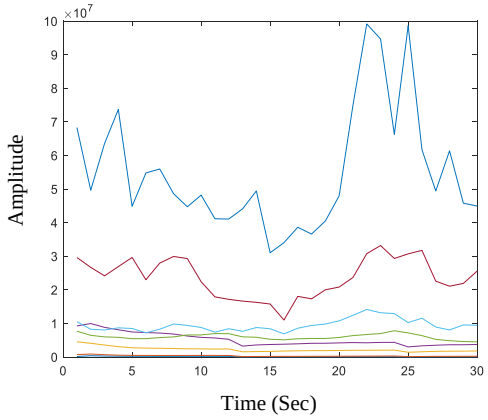
Figures 5.3 and 5.4 show the signal extracted features using Intrinsic Mode Function (IMF) method for 10 different gestures in WT and EMD techniques respectively. The IMF function is used or applied under two conditions: First condition is for the entire data, the number of extrema and the number of zero crossings should also equal or vary at most by one. The second condition is the mean value of the envelope explained by the local maxima or the envelope clarified by the local minima is zero [104]. The extracted features are each assigned a class and fed to ANN for training.



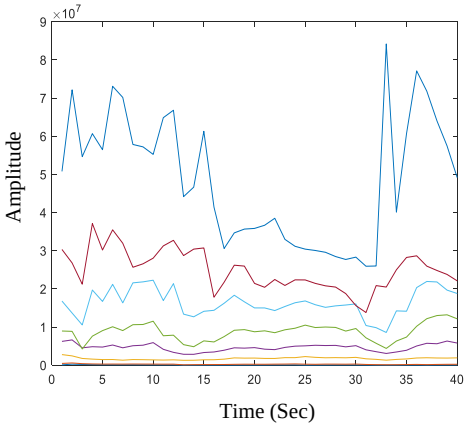
a) Sweep motion



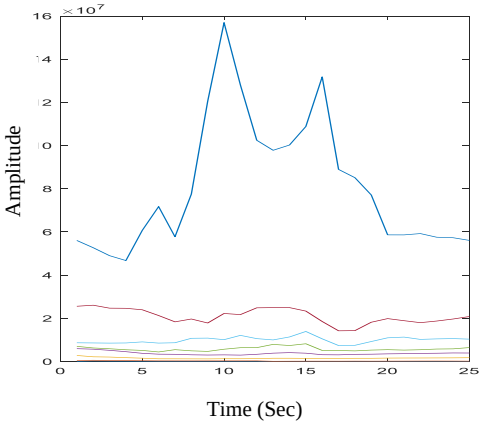
b) Shrink motion



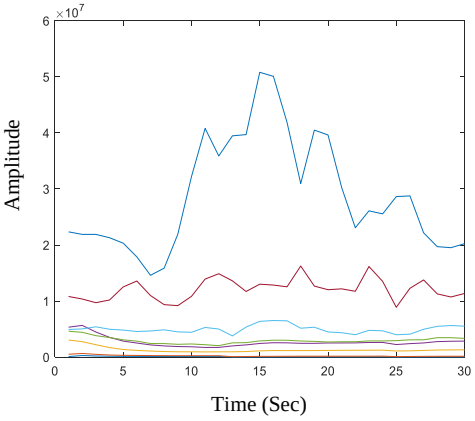
c) Circular motion



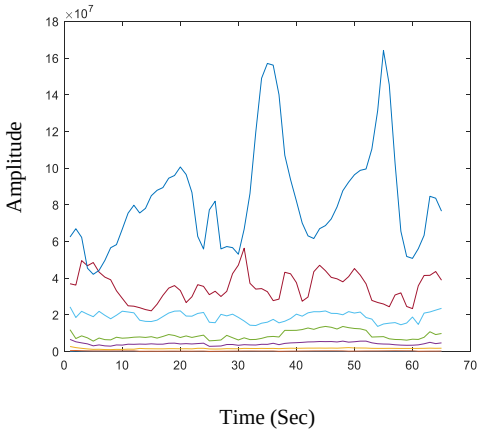
d)Squeeze motion



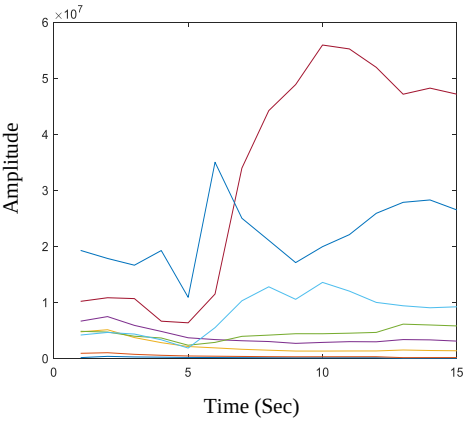
e) 2 Fingers Shrink



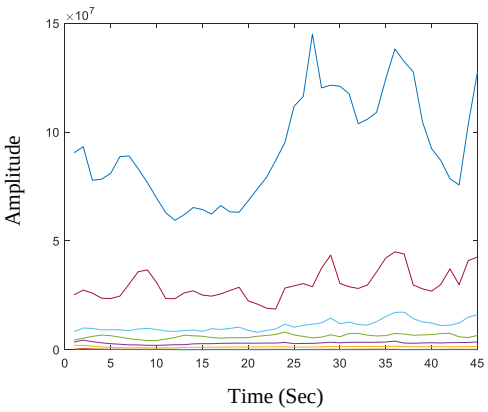
f) Back/Forth



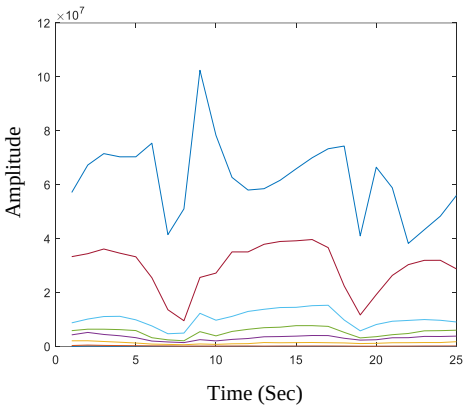
g) Rub motion



h) Click motion



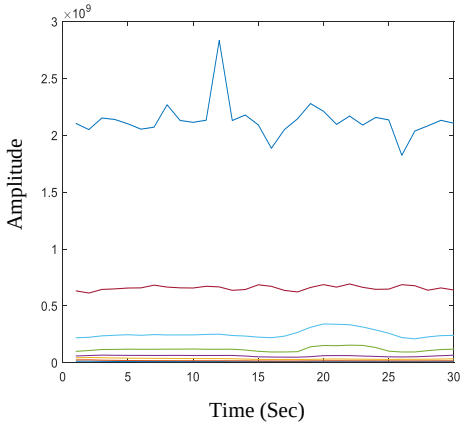
i)Dance motion



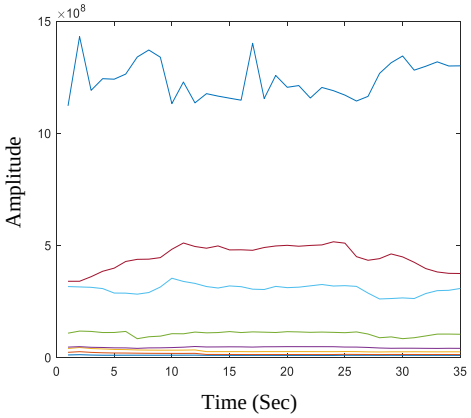
j) Pinch motion

Figure 5.3: IMF for 10 different motions using WT.

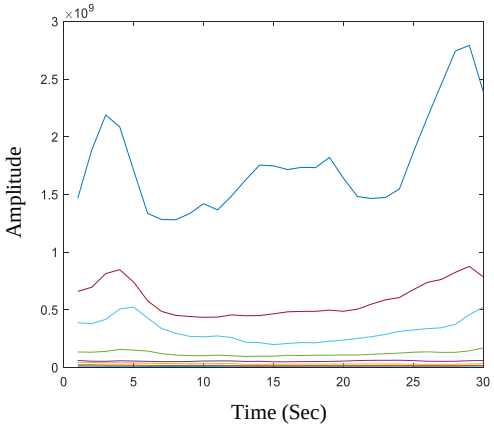
The ten graphs above describe the X-axis in microsecond, Y-axis in frequency with 8 signals of IMF (levels). The speed of motion starts from 0 microsecond till the end of time with a stable signal rate. As shown in Figures 5.3 and 5.4, the blue signal has a slight change whereas other signals are steadier. All figures show a dramatic increase in blue, red and light blue signals while other signals are steady.



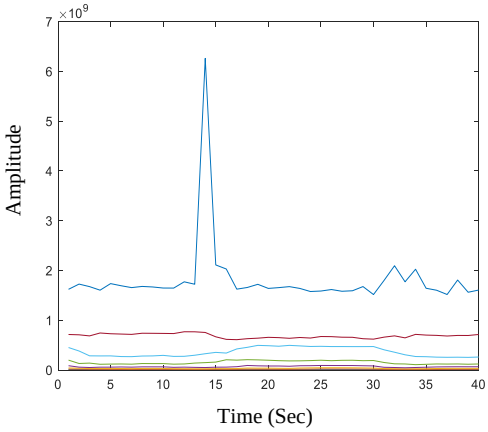
a) Sweep motion



b) Shrink motion



c) Circular motion



d) Squeeze motion

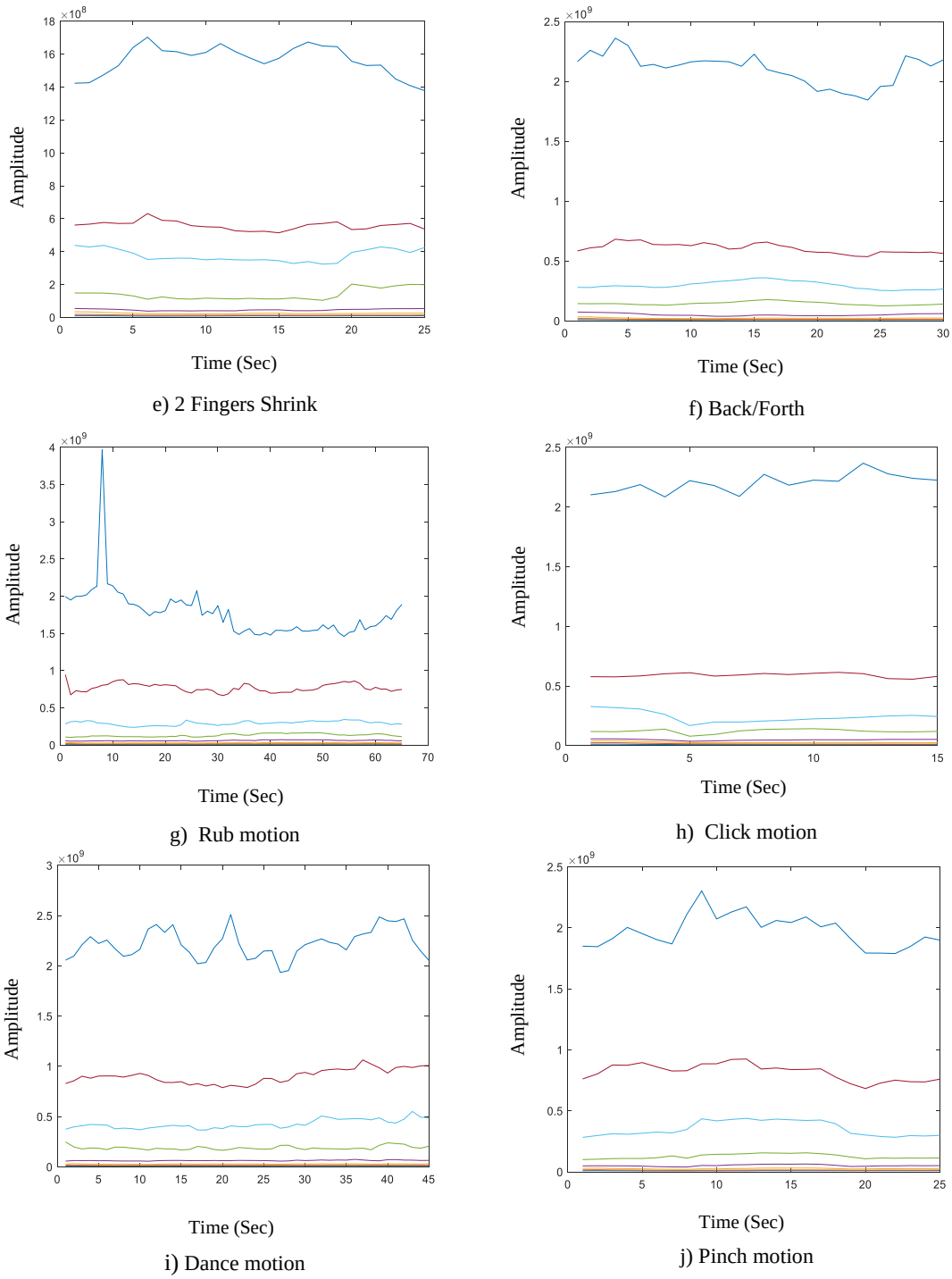


Figure 5.4: IMF for 10 different motions using EMD.

Table 5.1 shows the summary of the values acquired against the parameters when being in training mode. The total time execution for WT, EMD and CNN in training. The execution time of WT is less than the total time execution of EMD and CNN. The accuracy results of CNN exceeded WT and EMD by acquiring the highest value. The value of sensitivity in CNN is larger than that of WT and EMD. Specificity in WT reached the top while EMD is less than

WT and CNN. The PPV and NPV of WT is lower than EMD and CNN. CNN had the best value in LR+ and LR− unlike WT and EMD. The AUC is 0.97 for WT, 0.99 for EMD, and 1 for CNN respectively. The last factor is RMS, the value of EMD and CNN slightly decreased while WT has a high value in Root Mean Square (RMS).

Table 5.1: Comparison between WT, EMD and CNN for Training

	WT+ANN	**EMD+ANN**	**CNN**
Exe Time ± SD (sec)	5.79 ± 0.89	9.86 ± 1.77	713.69 ± 122.64
Accuracy ± SD	0.40 ± 0.07	0.61± 0.12	1 ± 0
Sensitivity ± SD	0.92 ± 0.04	0.98 ± 0.01	1 ± 0
Specificity ± SD	7.75 ± 8.07	0.73 ± 0.23	1 ± 0
Positive Predictive Value (PPV)	0.55 ± 0.13	0.77 ± 0.07	1 ± 0
Negative Predictive Value (NPV)	0.93 ± 0.01	0.96 ± 0.01	1 ± 0
Positive Likelihood (LR+)	18.87 ± 22.71	54.62 ± 64.92	1 ± 0
Negative Likelihood (LR−)	0.71 ± 0.11	0.39 ± 0.12	1 ± 0
RMS ± SD	2.42 ± 1.45	0.85 ± 0.12	1 ± 0
AUC ± SD	0.97 ± 0.02	0.99 ± 0.01	1 ± 0

The parameter values of CNN are constant for all categories. Its execution time is approximate 714 second which is significant and not preferred in experiments. In this study, only ten pictures of different hand movement were involved to train the system. If the input would consist of more images, the execution time would increase proportionally which would be not ideal for any experiment in terms of time constraint. Positive Likelihood (LR+) of EMD is higher as it refers to being more specific towards accuracy as compared to WT and CNN algorithms. Overall, CNN has the highest value in all parameters testing when training was done except for execution time and Positive Likelihood (LR+). Thus, we can compromise on few factors to get best results when training of the system is being done.

Table 5.2 compares the three algorithms performance while they were being tested for this study. CNN had the total execution time, higher than WT and EMD. In accuracy factor, WT achieved a value which is lower than EMD and CNN. The accuracy results of CNN surpassed WT and EMD with a high value. The value of sensitivity in CNN is superior more than that of WT and EMD. Specificity in WT is higher than EMD and CNN. EMD and CNN had higher

values in PPV and NPV comparing with WT which had lower value. In LR+ and LR−, the value of CNN is higher than WT and EMD. The value of RMS for EMD algorithm significantly declined while WT has a higher value.

Table 5.1: Comparison between WT, EMD and CNN for Testing

	WT+ANN	**EMD+ANN**	**CNN**
Exe Time ± SD (sec)	0.20± 0.03	0.19 ± 0.06	713.69 ± 122.64
Accuracy ± SD	0.39 ± 0.07	0.62 ± 0.13	0.97 ± 0.01
Sensitivity ± SD	0.33 ± 0.22	0.55 ± 0.25	1 ± 0
Specificity ± SD	0.94 ± 0.04	0.73 ± 0.37	1 ± 0
Positive Predictive Value (PPV)	0.67 ± 0.42	0.76 ± 0.22	1 ± 0
Negative Predictive Value (NPV)	0.93 ± 0.02	0.97 ± 0.02	1 ± 0
Positive Likelihood (LR+)	9.10 ± 8.78	22.42 ± 24.92	1 ± 0
Negative Likelihood (LR−)	0.68 ± 0.15	0.39 ± 0.19	1 ± 0
RMS ± SD	1.98 ± 0.90	0.84 ± 0.20	1 ± 0
AUC ± SD	0.98 ± 0.03	0.99 ± 0.02	1 ± 0

Again, in the testing phase, CNN algorithm took similar time i.e. 714 second for executing the testing task. This is not a feasible outcome as testing 10 images takes such long time; if mire images would be needed to be tested for experiments, the members of the experiment would have to wait to get results the whole time. CNN has 1 value in all parameters i.e. sensitivity, specificity, PPV, NPV, negative likelihood (LR−), RMS and AUC. In testing phase, the Positive Likelihood (LR+) has 1 value too which is least as compared to WT and EMD that recorded 9.1 and 22.4 respectively. Again, this proves that for testing too, CNN offers best results, but compromise on execution time needs to be ignored.

Figures 5.5 and 5.6 show the Receiver Operating Characteristic ROC curve which is applied in binary classification to learn the output of a classifier. Currently, there are two strategies of ROC to be drawn for multiclass curve, one vs. one and one vs. multi and the experiment used one vs. multi method. According to the WT and EMD graphs, the 10 classes had 10 ROC

curves reached the upper left corner which are %100 True Positive Rate (Sensitivity) and %100 False Positive Rate (1−Specificity). The ROC curve of EMD is slightly near to the upper left corner more than WT.

The functionality of CNN is way better than WT and EMD, according to the results of ten-hand motions. WT and CNN used a low amount of memory while EMD used a large amount of memory as it also takes a long time to run the code.

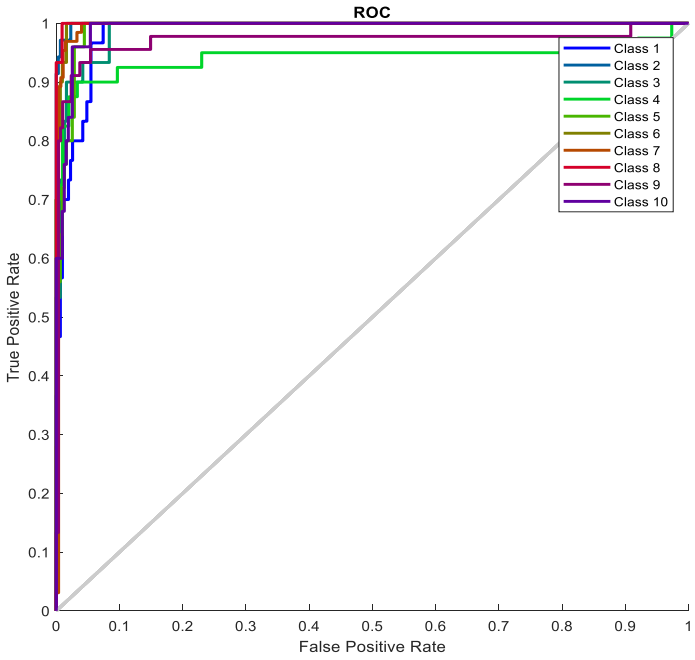


Figure 5.5: ROC for 10 different classes in WT.

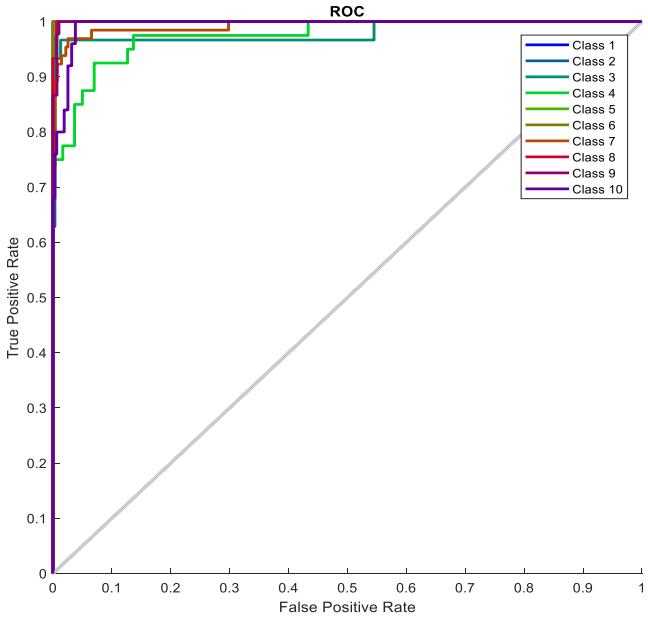


Figure 5.6: ROC for 10 different classes in EMD.

5.3 Long Distance Gestures

Currently, direct contact is the dominant form of interaction between the user and the machine. The interacting channel is based on devices such as the mouse, the keyboard, the remote control, touch screen, and other direct contact methods. Human to human interaction is achieved through more natural and intuitive noncontact methods, such as sound and physical movements. The flexibility and efficiency of noncontact interaction methods has led many researchers to consider exploiting them to support the human computer interaction. Gesture is one of the most important noncontact human interaction methods and forms a substantial part of the human language. Historically, wearable data gloves were usually employed to obtain the angles and positions of each joint in the user's gesture. The inconvenience and cost of a wearable sensor have limited the widespread use of such method. Gesture recognition methods based on noncontact visual inspection are currently popular due to their low cost and convenience to the user. Hand gesture is an expressive interaction method used in healthcare, education and the entertainment industry, in addition to supporting users with special needs and the elderly. Hand tracking is essential to hand gesture recognition and involves undertaking various computer vision operations including hand segmentation, detection, and tracking.

Several gesture-based techniques have been developed to support human computer interaction. According to Pradipa and Kavitha [62], the main aim of gesture recognition is developing a

system that can detect human actions to be used for extracting meaningful information for device control.

Hand motion can be detected using any type of camera supporting reasonable image quality. 2D cameras such as Microsoft's Kinect, Intel's RealSense Technology and Apple's iPhone high quality camera can easily be used in detecting most hand motions on a constant surface. Video content (composed of several images) is processed in several phases including data input, pre-processing, image segmentation, feature extraction, and classification.

The objective of this study is to investigate the best method available to extract features. Deep learning techniques are evaluated including WT, EMD, and CNN comparing their classification accuracy. A hand gesture recognition system was developed based on various image processing methods. The performance of hand gesture recognition methods was evaluated using various metrics including execution time, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood, negative likelihood, receiver operating characteristic, area under roc curve and root mean square. ANN was used to classify the gestures using the features extracted as inputs. Multiple training sessions were performed by applying filters. A CNN deep learning tool was used to minimize the previous stages.

5.3.1 System Implementations

Hand gestures represent the input to different gesture detection methods evaluated in this study. Figure 5.7 illustrates ten 2D and 3D hand gestures with plain backgrounds. They are recorded within long distances and used in the study's experimental work.

The implementation framework illustrating the extraction and the classification steps is shown in Figure 5.8. Using an iPhone 6 Plus camera with resolution 4K at 30 fps, the hand motions shown in Figure 5.8 are recorded. Each recording lasts 10 seconds and the resolution of the recorded video is 3840 × 2160. The first system is created using optical flow object by estimating and displaying the optical flow of objects in the video. The length of videos is between 15 to 65 frames. Each video has a different number of frames, which depends on the first section of motion.

5.3.2 Feature Detection using Wavelet Transforms Algorithm

The system is implemented using the db8 WT tool following the steps outlined below:

1. Read each video using a video reader function.
2. Create an optical flow object that spreads the object velocities in an image.

3. Estimate and display the optical flow of objects in the video.
4. Divide a video into certain frames; each frame contains 8 IMFs.
5. Apply the *appcoef2* function which is used to compute an approximation coefficient of 2D signals.
6. Extract each level using the *wrcoef* function to reconstruct the coefficients of each level in the video.
7. The execution time of WT is estimated only once.
8. The image data is trained and tested using a Neural Network system. The NN has 20 hidden neurons in a single hidden layer to train data and it stops when the error is reached in 20 epochs. When more hidden layers are added the depth of the neural network is increased, the neural network model becomes a deep learning model. Thus, A single layer is selected for this experiment.
9. The execution times of image data training and testing are also calculated.

a) Sweep motion b) Shrink motion

c) Circular motion d) Squeeze motion

e) 2 Fingers Shrink f) Back/Forth

g) Rub motion h) Click motion

i) Dance motion j) Pinch motion

Figure 5.7: Hand gestures used in the study.

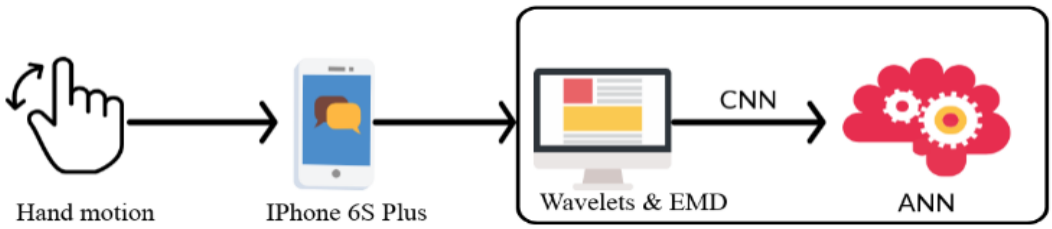


Figure 5.8: The implementation framework.

5.3.3 Feature detection using Empirical Mode Decomposition algorithm (EMD)

The implementation of EMD is similar to WT with the addition of *reshape* function that returns the M-by-N matrix whose elements takes column-wise from X. The function used is *ceemd* representing a noise improved data analysis algorithm.

5.3.4 Implementation of the Convolutional Neural Network (CNN)

CNN forms an integral part of deep learning, as it is used to train data without using any image processing tool. In our experiment, a new directory is created for each video. Ten images are generated to transfer the image frame RGB to grey and resize it to 48×27 from the original image size. All videos have 70 frames. The image's data are split into training and testing datasets. The CNN topology is created in seven layers with each layer having the following functionality and size: *ImageInputLayer* Input size [48,27,1], *Convolution2DLayer* Filter size [5,5], *ReLULayer* (Rectified Linear Unit), *MaxPooling2DLayer* Pool size [2,2], *FullyConnectedLayer* Input size [auto] and Output size [10], *SoftmaxLayer* and *ClassificationOutputLayer* Output size [auto]. The hyperparameters of the CNN are generated inside the training options function. The value of max epochs parameter is set to 200 epochs.

5.3.5 Parameters Comparison

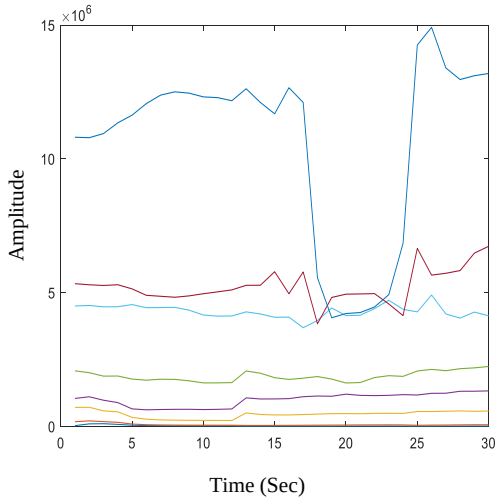
The performance of WT, EMD and CCN algorithms was compared using number of parameters. This includes execution time, that is the duration taken by the software to process the given task. Sensitivity measures the percentage of positives which are properly identified. Specificity is a measure of the false positive rate. The PPV and NPV are the percentages of positive and negative results in diagnostic and statistics tests which also describe the true positive and true negative results. The LR+ and LR− are known measures in diagnostic accuracy. Area under ROC curve (AUC) is the typical technique to measure the accuracy of predictive models.

5.3.6 Comparison between WT, EMD and CNN

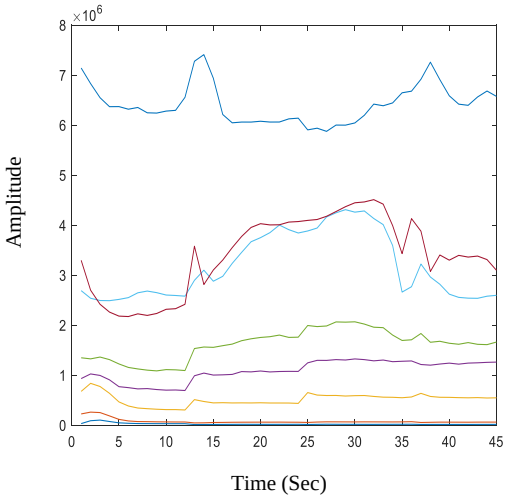
The experiments were executed ten times to obtain the mean of ten-hand motions. Two different training and testing were presented and compared to find the best mini gestures detection tool. Training accuracy is achieved by implementing a model on the training data and determining the accuracy of the algorithm.

Figures 5.9 and 5.10 show the signal extracted features using IMF method for 10 different gestures in WT and EMD techniques respectively. The IMF function is applied under two conditions. The first condition involves the entire data, where the number of extrema and the number of zero crossings are equal or vary at most by one. The second condition is that the mean of the envelope explained by the local maxima or the envelope clarified by the local minima has a value of zero [102]. The extracted features are each assigned a class and fed to ANN for training.

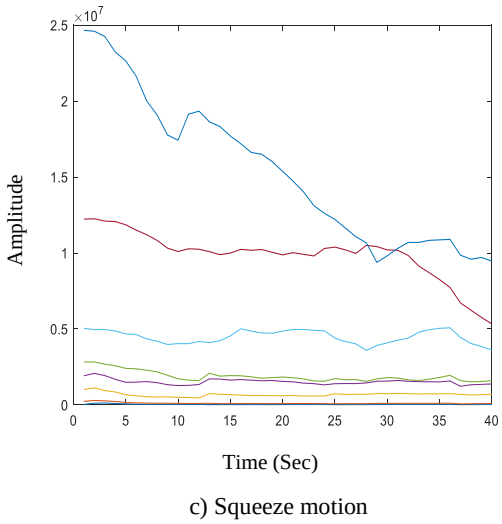
In the IMF graphs shown in Figures 5.9 and 5.10, the X-axis represents time (in microseconds) and the Y-axis represents frequency with 8 signals of IMF (levels). The speed of motion starts from 0 microsecond till the end of time with a stable signal rate. Examining the IMFs for WT shown in Figure 5.9, there is a significant fluctuation in the blueline in the IMF graphs for different hand gestures. In addition, there is notable variation in the red and light blue signals in all graphs. Other signals exhibit steadier behaviours.



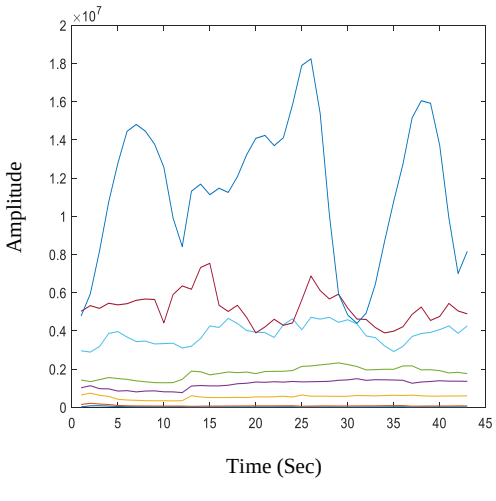
a) Sweep motion



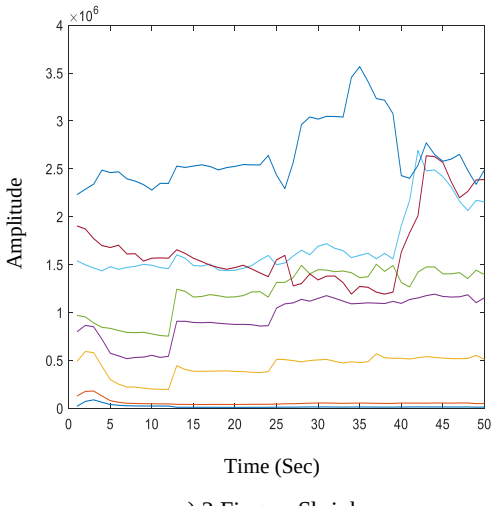
b) Shrink motion



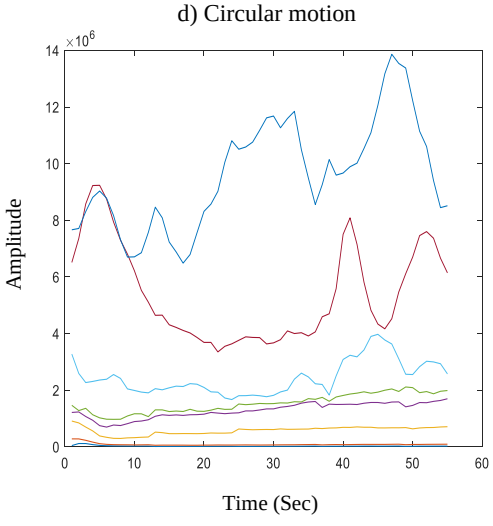
c) Squeeze motion



d) Circular motion



e) 2 Fingers Shrink



f) Back and Forth

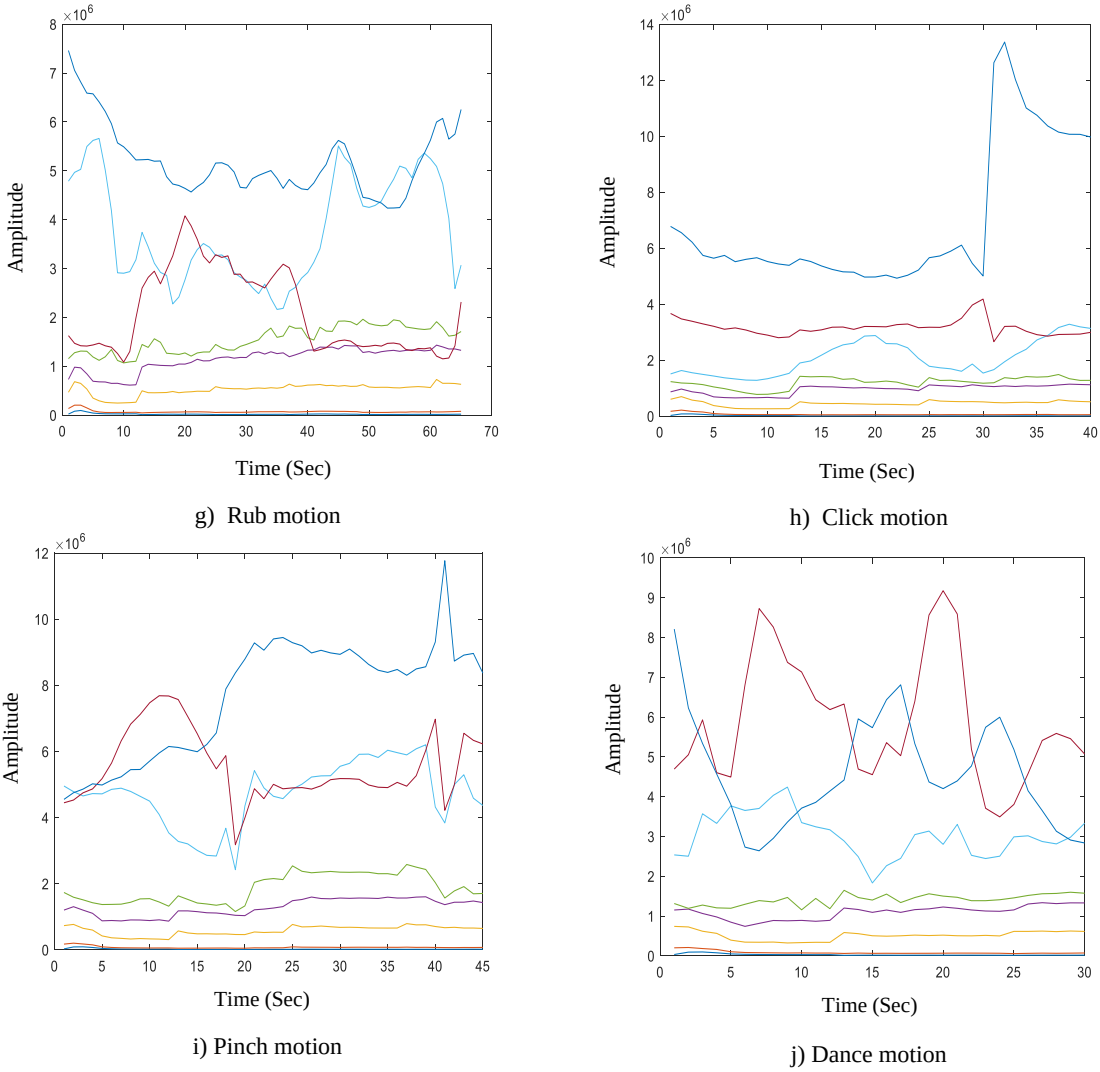
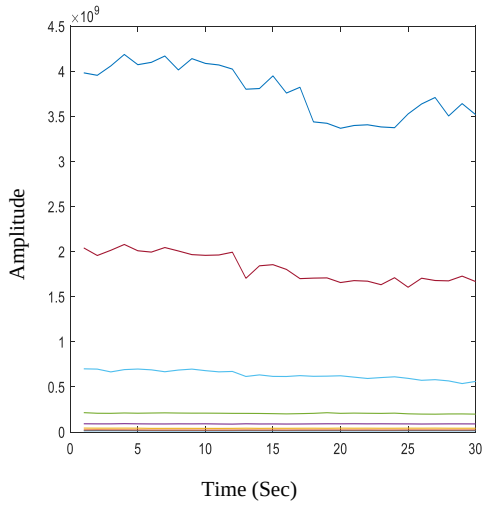
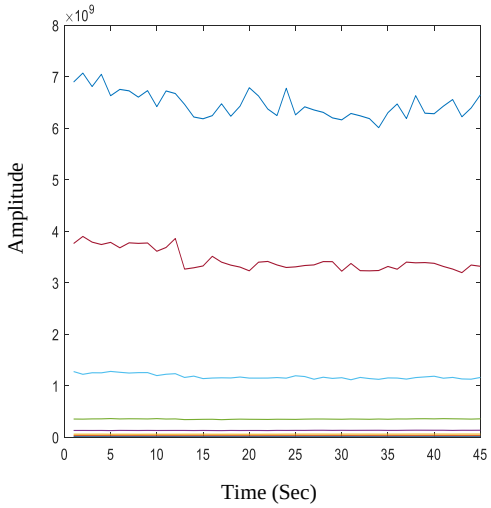


Figure 5.9: IMF for 10 different motions using WT.

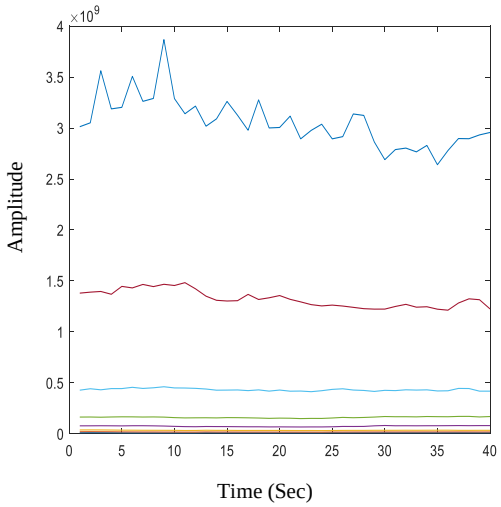
For the IMF graphs for EMD shown in Figure 5.10, signals for different gestures are generally steadier compared to WT IMF graphs shown in Figure 5.10. Slight fluctuations can be noticed in the blue signal, especially for the back and forth hand gesture. Minimum variation can be seen in the path of the red signal for all hand gestures. All other signals show steady lines for different gestures.



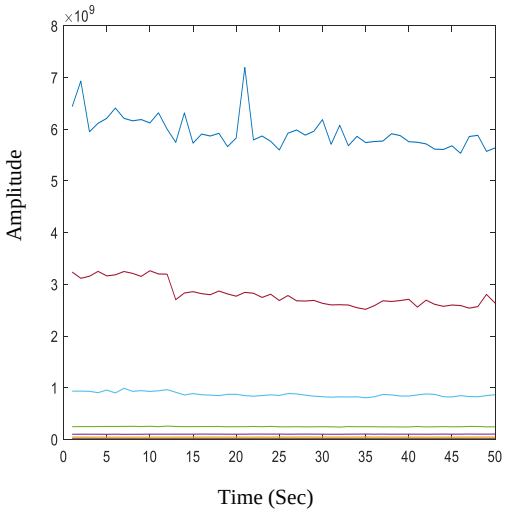
a) Sweep motion



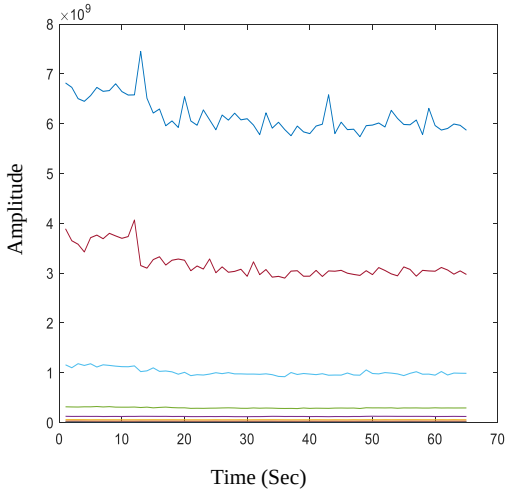
b) Shrink motion



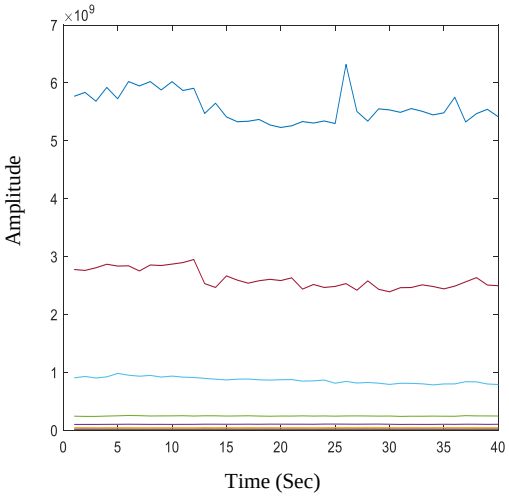
c) Squeeze motion



d) Circular motion



e) 2 Fingers Shrink



f) Back and Forth

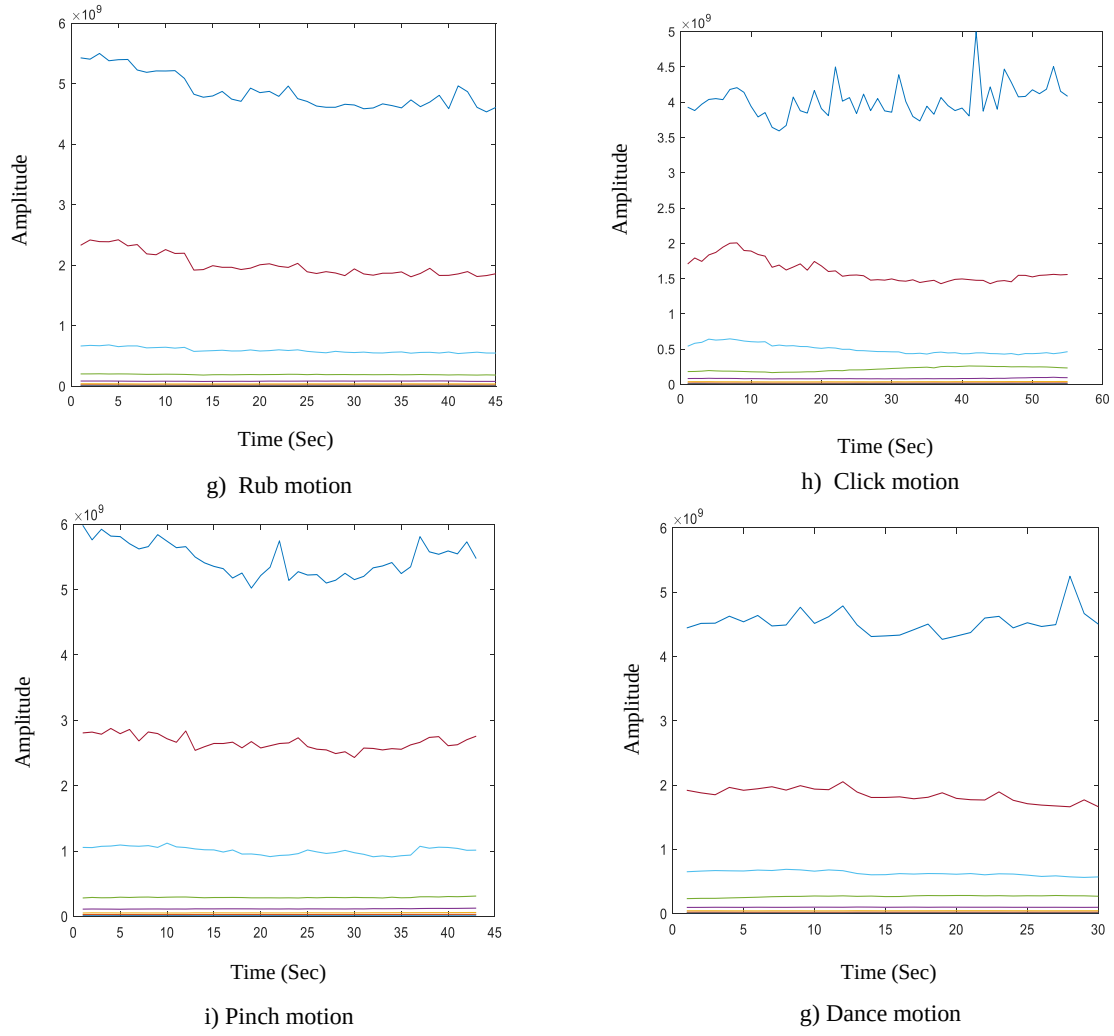


Figure 5.10: IMF for 10 different motions using EMD.

A summary of the values acquired for various parameters in training mode is listed in Table 5.3. It can be noticed that the execution time of WT is less than that of EMD and CNN. The accuracy result of CNN is better than WT and EMD. The value of sensitivity in CNN is higher than WT and EMD. Specificity in CNN is the highest followed by EMD and the lowest result was recorded for WT. The PPV and NPV of WT is lower than EMD and CNN. The best value for LR+ and LR− are recorded for CNN. For RMS, the value of EMD and CNN are lower than WT. Finally, The AUC is 0.90 for WT, 99 for EMD, and 1 for CNN

The parameter values of CNN are constant for all categories. Its execution time is approximate 636 seconds, a substantially long and unacceptable duration to train the system using only ten hand movement pictures (which are used in the experiment). Positive Likelihood (LR+) of EMD is higher indicating more accuracy compared to WT and CNN. Overall, CNN has the best values in most parameters when training was performed except for execution time.

Comparative performance values for the three methods are listed in Table 5.4. CNN's execution time is higher than WT and EMD. For accuracy, WT achieved a lower value compared to EMD and CNN. Accuracy results of CNN outperformed WT and EMD. Similarly, CNN has a higher sensitivity value compared to WT and EMD. Specificity in WT is lower than EMD and CNN. EMD and CNN have higher PPV and NPV values than WT. For LR+ and LR−, CNN values are higher than WT and EMD. The value of RMS for EMD is the lowest while WT has the highest value. Lastly, The AUC is 0.93 for WT, 1 for EMD, and 1 for CNN.

As in the training phase, the duration of CNN execution took similar time i.e. 636 seconds, an impractical timing given that only ten images were tested. It is notable that CNN has a significantly low value of 1 for the Positive Likelihood (LR+) compared to WT (19.29) and EMD (20.81).

Table 5.2: Comparison Between WT, EMD and CNN In Training Mode

	WL+ANN	**EMD+ANN**	**CNN**
Exe Time ± SD (sec)	6.07 ± 1.24	9.17 ± 2.329	636.43 ± 113.92
Accuracy ± SD	0.44 ± 0.07	0.91 ± 0.051	1 ± 0
Sensitivity ± SD	0.43 ± 0.17	0.80 ± 0.151	1 ± 0
Specificity ± SD	0.97 ± 0.01	0.99 ± 0.004	1 ± 0
Positive Predictive Value (PPV)	0.49 ± 0.15	0.97 ± 0.041	1 ± 0
Negative Predictive Value (NPV)	0.96 ± 0.01	0.98 ± 0.010	1 ± 0
Positive Likelihood (LR+)	16.78 ± 13.31	129.41 ±158.59	1 ± 0
Negative Likelihood (LR−)	0.51 ± 0.20	0.24 ± 0.16	1 ± 0
RMS ± SD	1.22 ± 0.13	0.37± 0.05	1 ± 0
AUC ± SD	0.901 ± 0.038	0.99 ± 0.00	1 ± 0

Table 5.3:Comparison Between WT, EMD and CNN In Testing Mode

	WL+NN	**EMD+NN**	**CNN**
Exe Time ± SD (sec)	0.158 ± 0.027	0.195 ± 0.053	636.433 ± 113.922
Accuracy ± SD	0.437 ± 0.079	0.915 ± 0.049	1± 0
Sensitivity ± SD	0.389 ± 0.262	0.769 ± 0.211	1± 0
Specificity ± SD	0.970 ± 0.0162	0.995 ± 0.007	1± 0
Positive Predictive Value (PPV)	0.496 ± 0.254	0.979 ± 0.044	1± 0
Negative Predictive Value (NPV)	0.955 ± 0.012	0.981 ± 0.014	1± 0
Positive Likelihood (LR+)	19.292 ± 18.316	20.81 ± 44.378	1± 0
Negative Likelihood (LR−)	0.622 ± 0.197	0.254 ± 0.215	1± 0
RMS ± SD	1.204 ± 0.122	0.354 ± 0.073	1± 0
AUC ± SD	0.937 ± 0.030	1.00 ± 0	1± 0

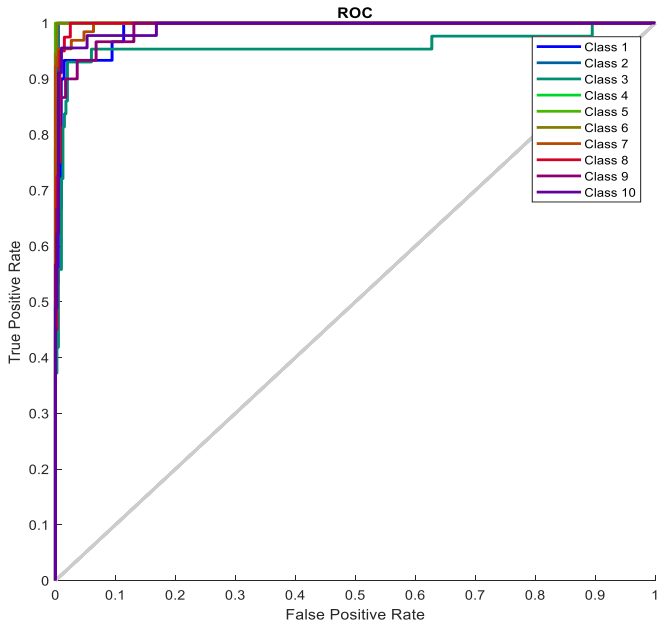


Figure 5.11: ROC for 10 different classes in WT.

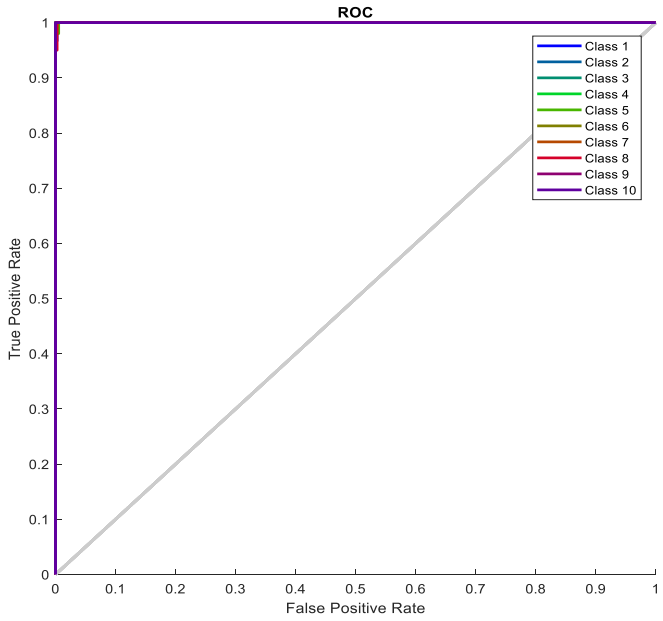


Figure 5.12: ROC for 10 different classes in EMD.

Figures 5.11 and 5.12 show the Receiver Operating Characteristic ROC curve which is applied in binary classification to learn the output of a classifier. There are two strategies of ROC to be drawn for multiclass curve: One VS. One and One VS. Multi, with the latter being used in this study. According to the WT and EMD graphs, the 10 classes had 10 ROC curves reached the upper left corner which are 100% True Positive Rate (Sensitivity) and 100% False Positive Rate (1−Specificity). The ROC curve of EMD is extremely near to the upper left corner compared to WT. CNN provides a better accuracy when compared with WT and EMD. However, CNN's duration of execution is substantially high. WT and CNN memory usage are lower than that of the EMD.

5.4 Summary

Gesture tracking in HCI is the first stage to generate natural HCI system. It can be applied in applications such as gesture recognition. Segmentation of the hand, detection of the hand parts, and the racking of the hands are the most significant parts to deal with in gesture recognition studies. The most frequently used in gesture recognition system is hand gestures. Most of the cameras can detect hand motions in the 2D surface. There are three main phases which must be applied to hand gesture systems. Firstly, an image detects using a camera. In the second phase, the system will receive the image to track it by using any image processing tools such as WT or EMD. In the final stage, the data will be classified using ANN and CNN. In this

study, a system has been created for hand motion detection using WT and EMD to extract the features while the classification is done using ANN and CNN. The results of this experiment show the advantages and disadvantages of each method. CNN is more accurate than WT and EMD. In future work, we will extend the number of motions using 3D horoscopic imaging system.

Hand gesture recognition is essential to support a natural HCI experience. The most important aspects of gesture recognition are segmentation, detection, and tracking. In this study, a system has been created for hand motion detection using WT and EMD for features extraction. Classification is supported using ANN and CNN. Ten 2D and 3D motion images with plain backgrounds and recorded within short and long distances were used. Experiments were performed to compare the performance of various methods using number of measures. Results showed that CNN provides a better accuracy as compared with WT and EMD. However, its computational requirements are relatively high. The memory usage of WT and CNN was lower than that of the EMD. In future work, the number of motions will be extended using a 3D holoscopic imaging system. The next chapter details the 3D video gesture recognition experiments that were performed using a CNN algorithm.

Chapter 6
3D Video Gesture Recognition

6.1 Introduction

The predominant form of interaction between the user and the machine is direct contact. Devices like a mouse, keyboard, touch screen, remote control, and other direct contact methods act as a communication channel. Communication among human to human is achieved through more intuitive and natural non-contact methods, e.g. physical movements and sound. The efficiency and flexibility of these non-contact interaction methods have led several researchers to consider using them to support human–computer communication. The gesture forms a substantial part of the human language and is an important non-contact human interaction method. Historically, to capture the positions and angles of every joint in the user's gesture, wearable data gloves were often employed. The cost and difficulty of a wearable sensor have limited the widespread use of this method. The ability of a computer to understand the gestures and execute certain commands based on those gestures is called Gesture recognition. The primary goal of such gesture recognition is to develop a system that can recognise and understand specific gestures and communicate information.

Currently, the gestures-based recognition methods based on the non-contact visual inspection are popular. The reason for such popularity is their low cost and convenience to the user. A hand gesture is an expressive communication method widely used in entertainment, healthcare and education industry. Additionally, hand gestures can also be used to assist users with special needs and the elderly. Hand tracking is important to perform hand gesture recognition and involves performing several computer vision operations including hand segmentation, detection and tracking.

The use of a microlenses array at the image surface was proposed by Professor Gabriel M. Lippmann. He showed this concept to the French Academy of Sciences as La Photography Integral. Spatial image with full parallax in all directions was recorded by Professor Gabriel M. Lippmann and it’s considered a fly’s eye lens array [5]. The display system was a screen holding several small lenses. Herbert Ives in the 1920s, started to think about how to simplify Lippmann’s idea by joining a lenticular lens sheet which contained a signaller array of spherical lenses called lenticules. A signaller array of magnifying lenses is designed to see from various

angles and images are exaggerated consistently to provide a pixel from each micro picture. The lens sheet is transparent and the back face which creates the focal plane is flat. An example is the lenses used in lenticular production where the technology is used to show an illusion of depth creating moving or changing images as the image is seen from different angles. This innovative technology could also be utilised for producing 3D images on a flat sheet display. Hence, if the motion of the pictures is taken into consideration, this results in 3D holoscopic video [5] [6].

The rest of this paper is structured as follows: some studies of hand gesture recognition techniques and methods used are shown in Section 6.2; Section 6.2.1 proposes the details of the system's implementation; Section 6.2.3 is a presentation and discussion of the results achieved and the conclusion is presented in Section 6.3.

6.2 3D Short Distance Gesture Recognition Systems

Ge et al [105] proposed a 3D CNN method to estimate real-time hand poses from single depth images. The features extracted from images using 2D CNN are not suitable for estimation of 3D hand pose as they lack spatial information. The proposed method takes input as a 3D volumetric representation of the hand depth image and captures 3D spatial structure and accurately regress full 3D hand pose in a single pass. 3D data augmentation is performed to make the CNN method robust to global orientations and hand size variations. The results of the experiment show that the proposed 3D CNN outperforms the state-of-the-art methods on two challenging hand pose datasets. The implementation runs at over 215 fps on a standard computer with a single GPU which is proven to be very effective.

According to Ge et al [106], the method proposed is to increase the accuracy of hand pose estimation. The method involves first projecting the query depth image onto three orthogonal planes and then use the Multi-view projections to regress for two-dimensional heat-maps which can estimate the joint positions on each plane. The generated multi-view projection heatmaps are fused to generate a final estimation of the 3D hand pose. The results of the experiment show that the proposed method outperforms the current state of the art. The generalisation of the model is also proven to be good.

A technique using a depth camera in a smart device for hand gesture recognition is proposed by Keun and Choong [107]. The recognition is made through the recognition of a hand or detection of fingers. For detecting the fingers, the hand skeleton is detected via Distance Transform and fingers are detected by using Convex Hull algorithm. To recognise a hand, a

newly generated gesture is compared with already learned data using the Support Vector Machine algorithm. The hand's centre, finger length, axis of fingers, hand axis and arm centre are reviewed for this. An actual smart device was implemented for the evaluation of this experiment.

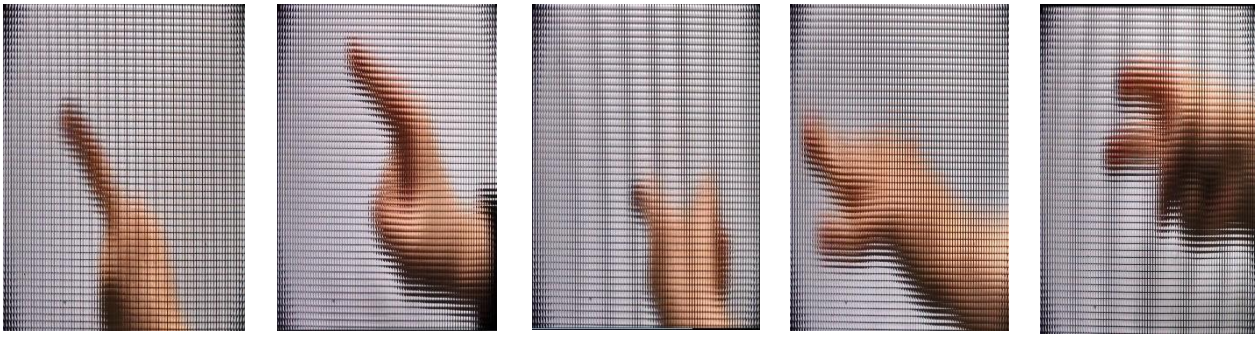
6.2.1 System Implementations

A. Hand Gestures Input

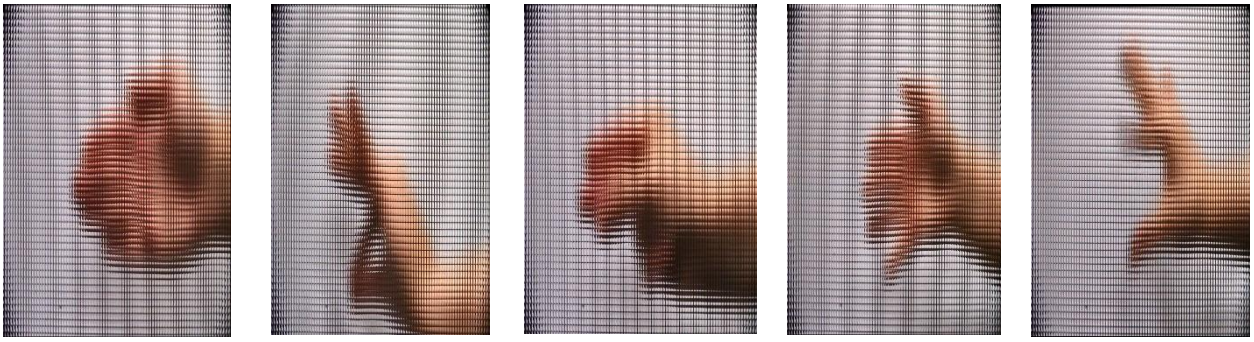
In this experiment, hand gestures are fed as input into CNN. Figure 6.1, Figure 6.4 and Figure 6.7 show twelve random hand gestures recorded in short distance with a plain background using a holoscopic imaging camera system. Some motions are 2D while others are 3D. The images are pre-processed before extracting videos in terms of some steps:

1- For Figure 6.1, the resolution of the camera used is full High Definition (HD) while for Figure 6.4 and Figure 6.7 the resolution is 4K.
2- The camera used in this experiment is a holoscopic imaging camera system with multi lenses. The number of lenses shown in Figure 6.1 is 47 for x-axis whereas for y-axis it is 84. For Figure 6.4 and Figure 6.7, the number of lenses is decreased to 31 on the x-axis is and 55 on the y-axis.
3- The generated images are converted from RGB to grey and images need to be resized to 135 × 75.
4- In figure 6.2, the image is rotated 0.30 degrees to adjust the image position while for figure 5 and figure 8 are rotated 180.20 degrees.
5- Divide lens into seven segments i.e. 7 × 7, the X segment is a constant of 4 while Y is changeable to 2, 4 and 6.
6- Create twelve separate directories for three different left, centre and right images and convert them from RGB to grey colour. Lastly, resize these grey images to size 135 × 75.
7- Combine each left, centre and right images for three people in one directory.
8- Combine the three images i.e. left, centre and right to get one image with a size 405 × 75 in Figure 6.3, Figure 6.6 and Figure 6.9.

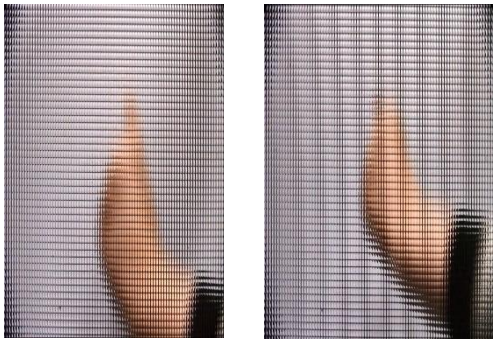
1- <u>Pre- extraction first person's hand motions in short distance</u>



a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion e) 2 Fingers Shrink



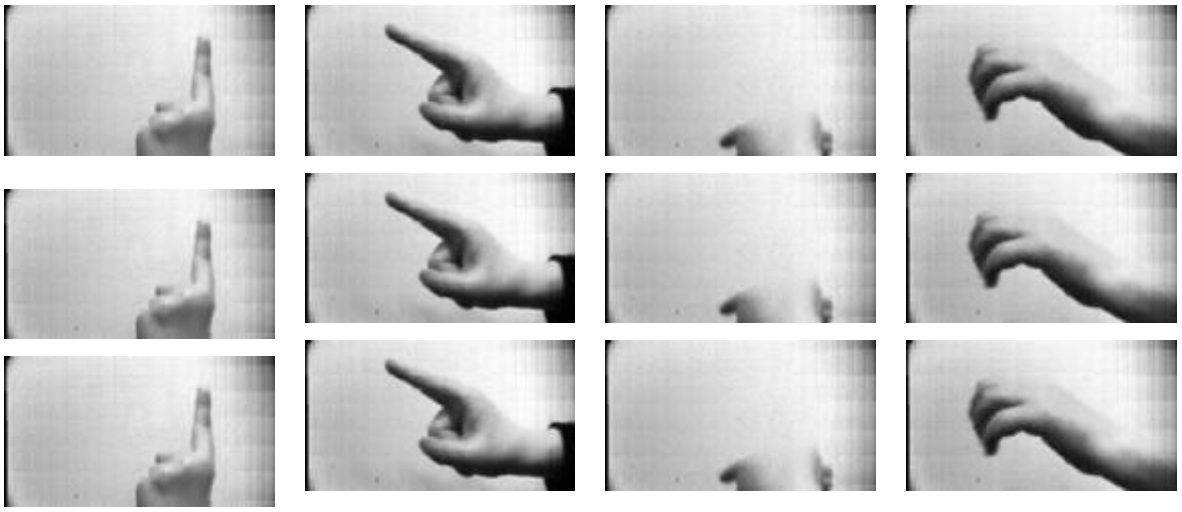
f) Back/Forth g) Rub motion h) Click motion i) Dance motion j) Pinch motion



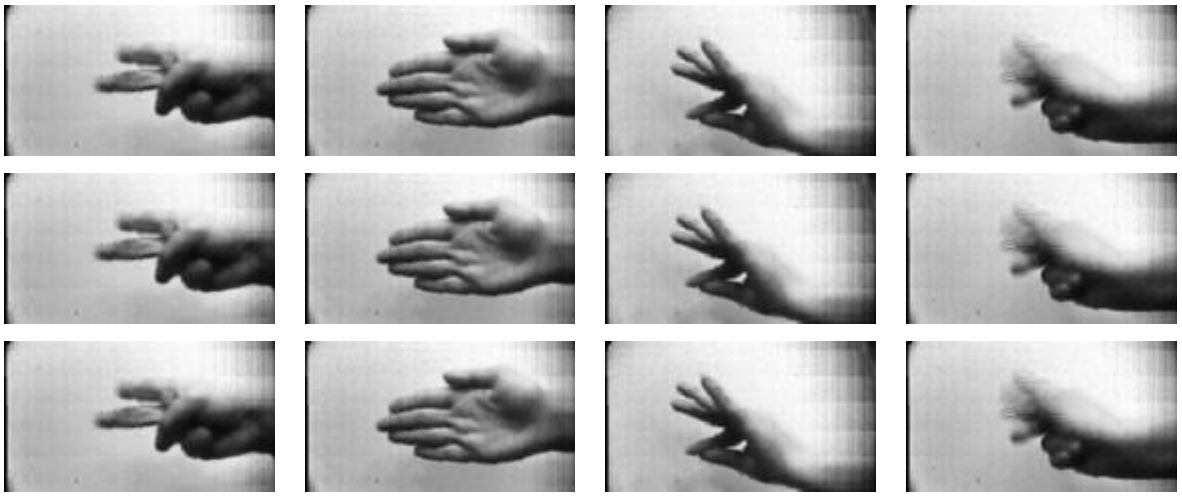
k) write motion l) Click motion 2

Figure 6.1: Pre- extraction first person's hand motions in short distance

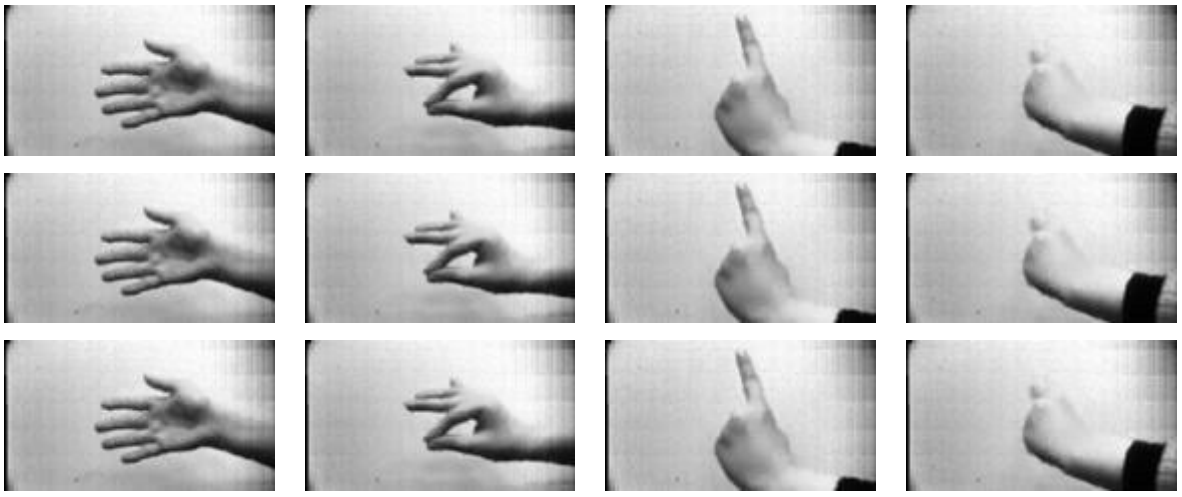
2- Post-extraction first person's hand motions in short distance single (LCR)



a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion



e) 2 Fingers Shrink f) Back/Forth g) Rub motion h) Click motion



i) Dance motion j) Pinch motion k) write motion l) Click motion 2

Figure 6.2: Post- extraction first person's hand motions in short distance

3- Post-extraction first person's hand motions in short distance combined (LCR)

a) Sweep motion

b) Shrink motion

c) Circular motion

d) Squeeze motion

e) 2 Fingers Shrink

f) Back/Forth

g) Rub motion

h) Click motion

i) Dance motion

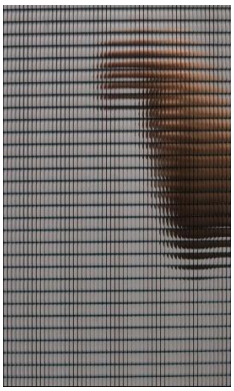
j) Pinch motion

write motion

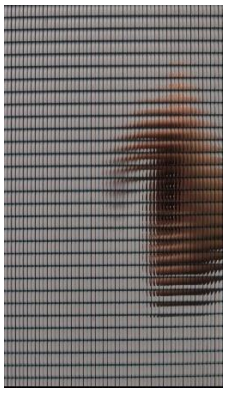
l) Click motion 2

Figure 6.3: Post- extraction first person's hand motions in short distance

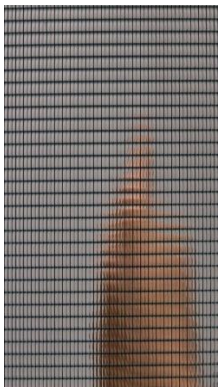
4- Pre- extraction second person's hand motion in short distance



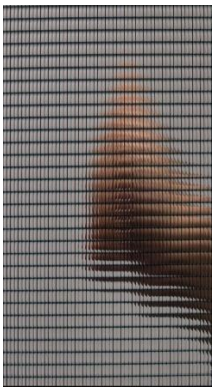
a) Sweep motion



b) Shrink motion

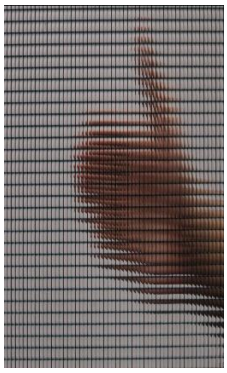


c) Circular motion

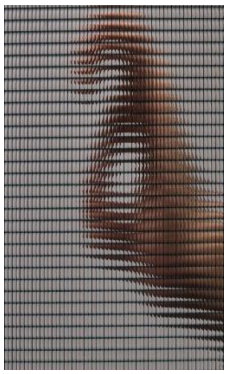


d) Squeeze motion

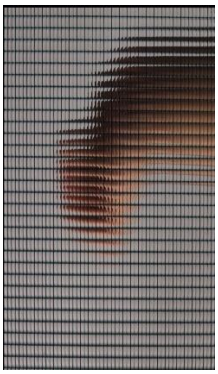
e) 2 Fingers Shrink



f) Back/Forth

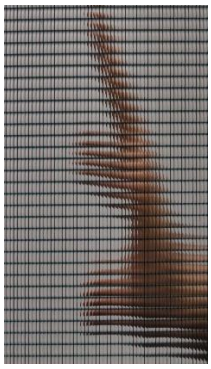


g) Rub motion

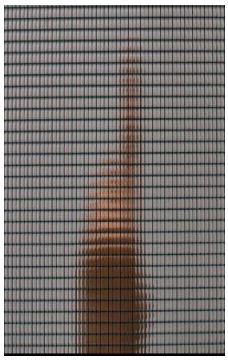


h) Click motion

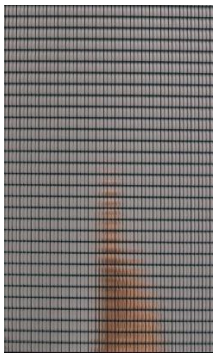
i) Dance motion



j) Pinch motion



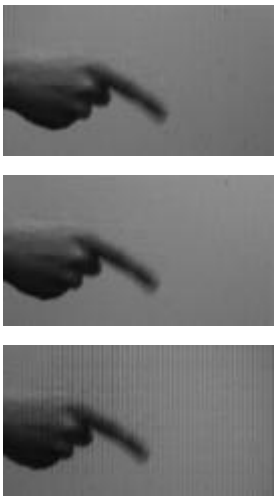
k) write motion



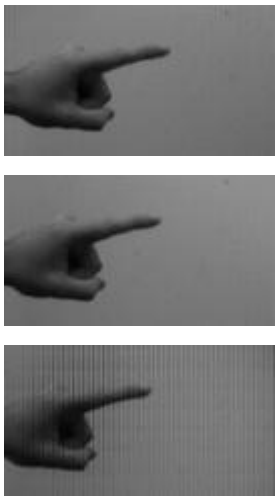
l) Click motion 2

Figure 6.4: Pre- extraction second person's hand motion in short distance

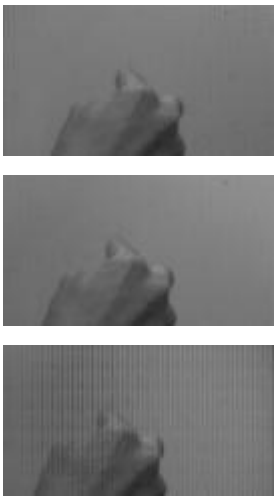
1- Post- extraction second person's hand motion in short distance single (LCR)



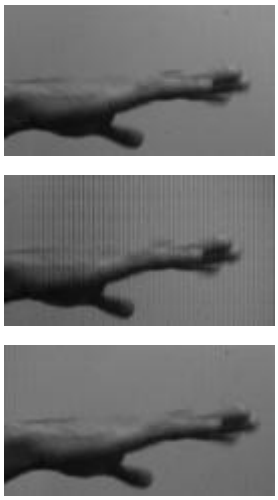
a) Sweep motion



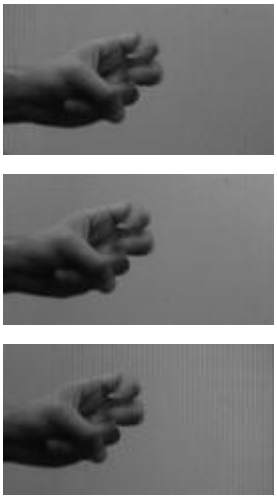
b) Shrink motion



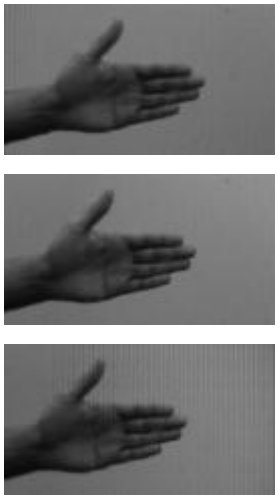
c) Circular motion



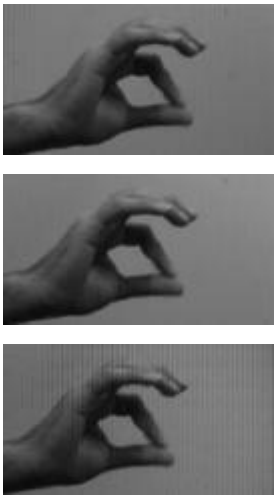
d) Squeeze motion



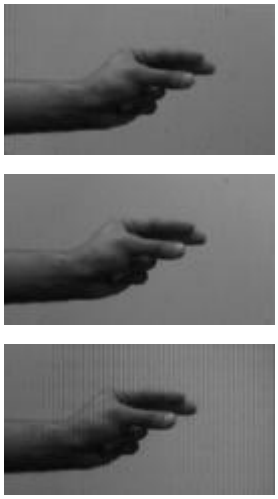
e) 2 Fingers Shrink



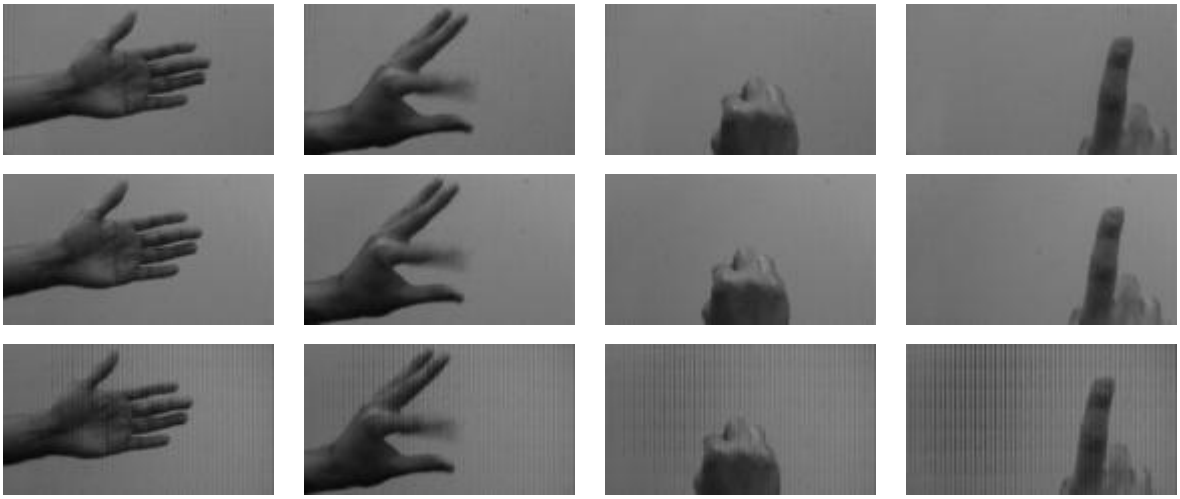
f) Back/Forth



g) Rub motion



h) Click motion



i) Dance motion j) Pinch motion k) write motion l) Click motion 2

Figure 6.5: Post- extraction second person's hand motion in short distance single (LCR)

1- Post- extraction second person's hand motion in short distance combined (LCR)

a) Sweep motion

b) Shrink motion

c) Circular motion

d) Squeeze motion

e) 2 Fingers Shrink

f) Back/Forth

g) Rub motion

h) Click motion

i) Dance motion

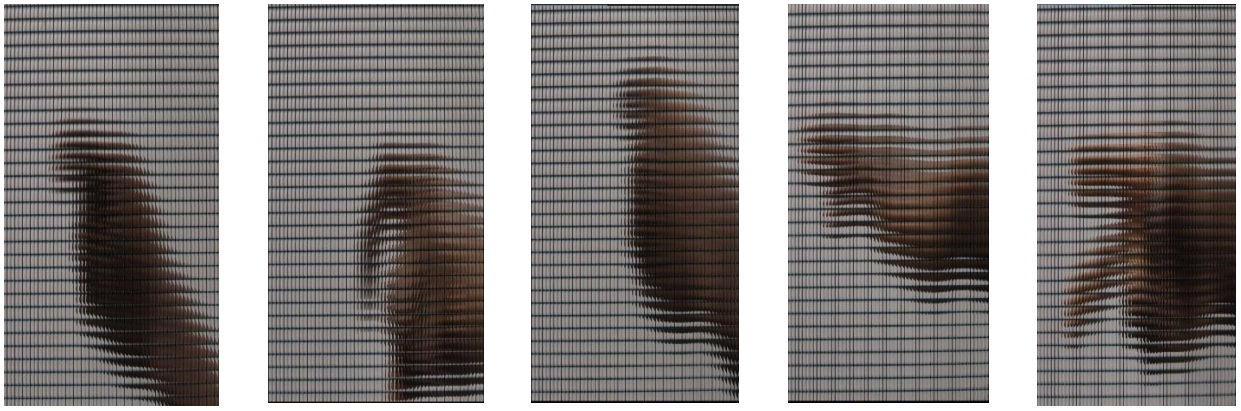
j) Pinch motion

write motion

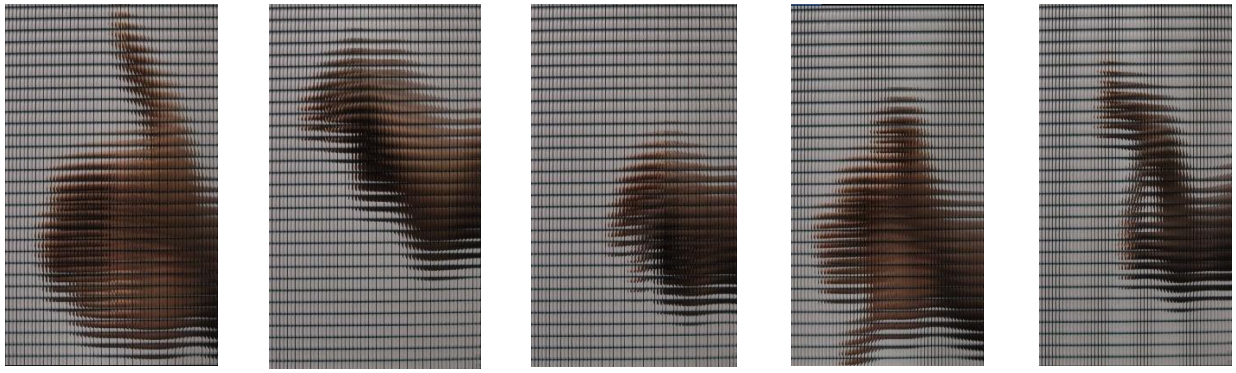
l) Click motion 2

Figure 6.6: Post- extraction second person's hand motion in short distance combined (LCR)

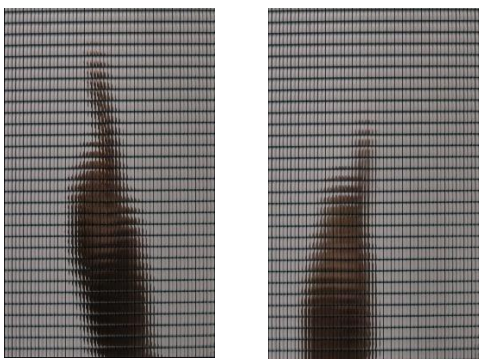
1- Pre- extraction third person's hand motion in short distance



a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion e) 2 Fingers Shrink



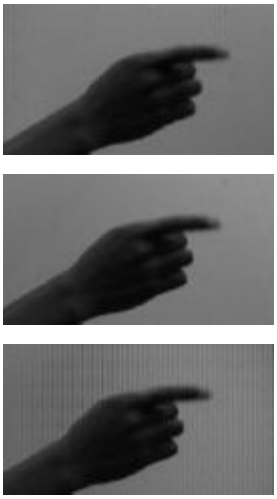
f) Back/Forth g) Rub motion h) Click motion i) Dance motion j) Pinch motion



k) write motion l) Click motion 2

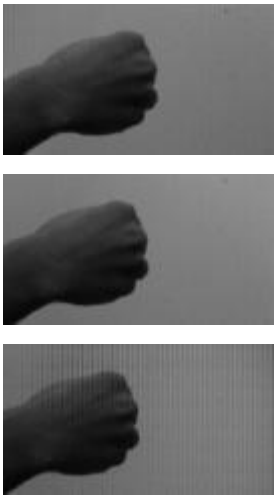
Figure 6.7: Pre- extraction third person's hand motion in short distance

1- Post- extraction third person's hand motion in short distance short distance single (LCR)

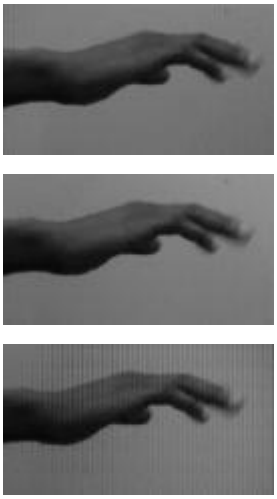


a) Sweep motion

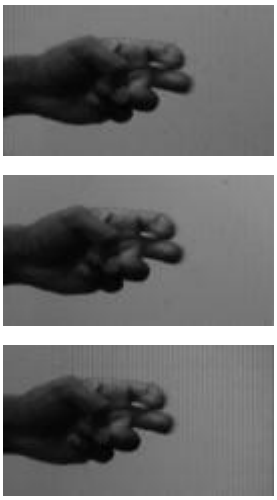
b) Shrink motion



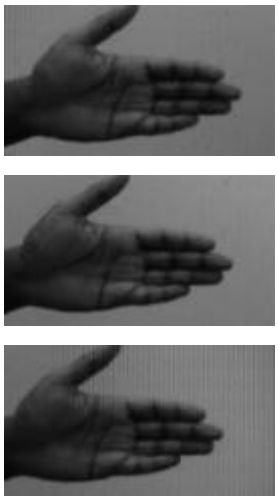
c) Circular motion



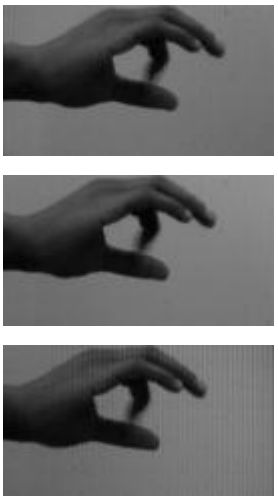
d) Squeeze motion



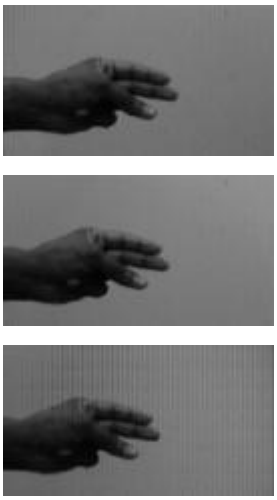
e) 2 Fingers Shrink



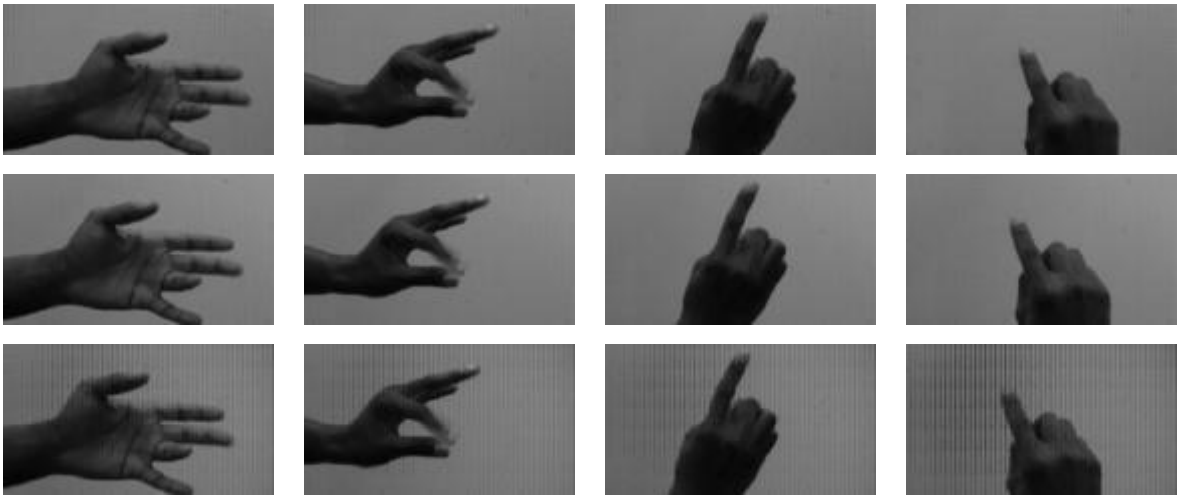
f) Back/Forth



g) Rub motion



h) Click motion



i) Dance motion j) Pinch motion k) write motion l) Click motion 2

Figure 6.8: Post- extraction third person's hand motion in short distance short distance single (LCR)

1- Post- extraction third person's hand motion in short distance short distance combined (LCR)

a) Sweep motion

b) Shrink motion

c) Circular motion

d) Squeeze motion

e) 2 Fingers Shrink

f) Back/Forth

g) Rub motion

h) Click motion

i) Dance motion

j) Pinch motion

write motion

l) Click motion 2

Figure 6.9: Post- extraction third person's hand motion in short distance short distance combined (LCR)

6.2.2 Result

A. Convolutional Neural Network Implementation

Convolutional neural network is an integral part of deep learning since it is used to train data without applying any image processing methods. In this experiment, a new separate directory is created for each video of three people. The topology of CNN is shown in Figure 6.10. The length of each video is 10 seconds for the separated and combined images. Each video will be read to generate 900 images, i.e. 300 for left, 300 for right and 300 for the centre, while combined is 300 images in one directory. The images are divided for training and testing models. The quantity of training frames for separate images is 390 whereas 210 for combined images which are %70. The CNN's topology is produced in seven layers with each layer having the following functionality and size: ImageInputLayer size [135, 75, 1] for separate images whereas [405, 75,1] for combined, Convolution2DLayer with Filter size [5,20], Rectified Linear Unit (ReLULayer), MaxPooling2DLayer Pool size [2,2], FullyConnectedLayer size [auto] and Output size[7], SoftmaxLayer and ClassificationOutputLayer Output size [auto]. The CNN hyperparameters are created inside the training options function. The epochs' parameter value is set to 100 epochs.

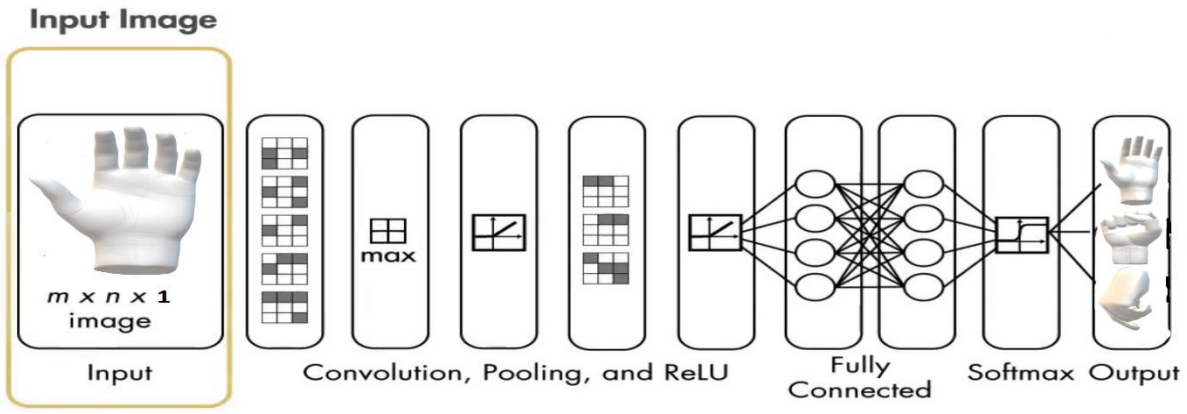


Figure 6.10: CNN topology

B. Parameters Comparison

CNN algorithm’s performance can be compared using several parameters including execution time (H:M:S), i.e. the duration taken by the software to implement the task, training accuracy is calculated by applying model on the training data and finding the accuracy of the algorithm. Testing accuracy is the accuracy for testing data. Sensitivity measures the appropriate count of the identified percentage of positive, specificity measures of the false positive rate, PPV and

NPV percentages of positive and negative results in diagnostic and statistics tests that describe the true positive and true negative results. The LR+ and LR− are identified measures in diagnostic accuracy.

Table 6.1 presents the comparison between three people to find the best results obtained. Single, combined and all three combined results displayed in terms of execution time, training, testing, sensitivity, specificity, PPV, NPV, LR+ and LR−.

In the single images experiment, the execution time of the first person is quite higher than the second and third persons. The result of training for second person is lower than first and third persons. First person has the best testing result of 100%. Sensitivity result for the third person is slightly higher than for the first and second results while all results for the three persons are equal in specificity. The PPV results in this experimental work are equalled whereas the result of NPV for the third person is decreased slightly. LR+ has the best values for three people while LR- result for third person is 0.0425.

In combined images, the result of the first-person experiment is the highest in execution time. The training result for the second person is slightly lower than for the first and third persons. The second person has the best testing result at 99%. The result for the first person is decreased in sensitivity more than the second- and third-persons' results, whereas the result of the third person is slightly lower than first and second persons in specificity. The PPV result for the third person is less than the first and second persons, whereas the result of NPV for the first person is the lowest. LR+ has the highest value for the third person while LR- result for the first person is 0.1333.

The ALL combined experiment shows the performance of all three people's images. The execution time of ALL three persons is the highest. The result of training for all three persons is slightly lower than for first and third persons. ALL has the lowest result in testing comparing to other results. Sensitivity and specific results for ALL is lowest. The results shown in PPV and NPV for ALL are also lower than other results. LR+ value is less than the combined result for third person whereas the result of LR- for ALL is the highest.

The summary of the comparison is that the first person has the best values in all categories in single experiment compared to other persons' results; except the execution time which is the highest. The results of second person in combined are better than first and second's results. The values of ALL experiment in categories is slightly lower than other experiments. Except

the value of training is slightly better than the single of the second person result. Overall, the single of the first person has the best values in most parameters.

Table 6.1: Comparison Between first person, second person and third person in CNN

	First person		**Second person**		**Third person**		**ALL**
	Single (LCR)	**Combined**	**Single (LCR)**	**Combined**	**Single (LCR)**	**Combined**	**Combined**
Execution Time (H:M:S)	02:33:47	02:36:16	00:49:02	00:24:45	00:51:51	00:53:08	02:50:16
Training	1	1	0.99	0.99	1	1	0.99
Testing	1	0.97	0.99	0.99	0.97	0.93	0.92
Sensitivity	1	0.86	1	1	0.95	1	0.79
Specificity	1	1	1	1	1	0.99	0.99
Positive Predictive Value (PPV)	1	1	1	1	1	0.97	0.94
Negative Predictive Value (NPV)	1	0.98	1	1	0.99	1	0.98
Positive Likelihood (LR+)	0	0	0	0	0	506	212.58
Negative Likelihood (LR−)	0	0.13	0	0	0.04	0	0.20

6.2.3 Summary

Hand gesture detection is elementary to provide a natural HCI skill. The most essential aspects in gesture recognition are segmentation, detection and tracking. This experiment is conducted for hand gestures recognition using features extraction and classification using CNN technique. In this experimental work, twelve 3D motions are recorded within short distance for three different people. Experiments were conducted to compare performance of CNN method in terms of multi factors like execution time, training, testing, sensitivity, specificity, PPV, NPV, LR+ and LR−. The results showed that single experiment for the first person provided a better result in all categories.

6.3 3D Long Distance Gesture Recognition Systems

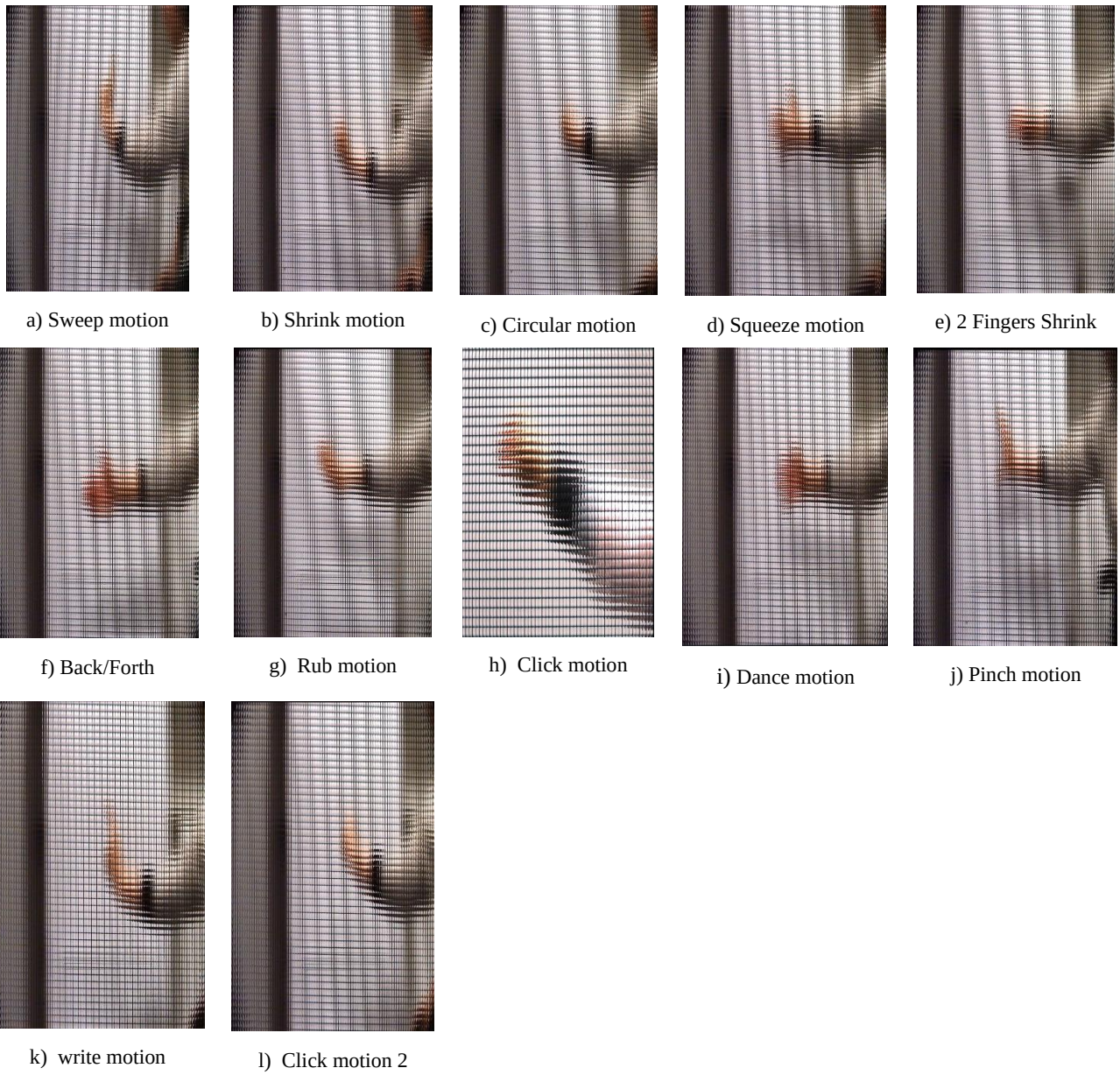
The predominant form of interaction between the user and the machine direct contact. Devices like a mouse, keyboard, touch screen, remote control, and other direct contact methods act as a communication channel. Communication among Human to human is achieved through more intuitive and natural non-contact methods, e.g., physical movements and sound. The efficiency and flexibility of these non-contact interaction methods have led several researchers to consider using them to support human-computer communication. The gesture forms a substantial part of the human language and is an important non-contact human interaction method. Historically, to capture the positions and angles of every joint in the user's gesture, wearable data gloves were often employed. The cost and difficulty of a wearable sensor have limited the widespread use of this method. The ability of a computer to understand the gestures and execute certain commands based on those gestures is called gesture recognition. The primary goal of such gesture recognition is to develop a system that can recognise and understand specific gestures and communicate information.

Currently, the gestures recognition methods based on the non-contact visual inspection are popular. The reason for such popularity is their low cost and convenience to the user. A hand gesture is an expressive communication method widely used in the entertainment, healthcare and education industries. Additionally, hand gestures can also be used to assist users with special needs and the elderly. Hand tracking is important to perform hand gesture recognition and involves performing several computer vision operations including hand segmentation, detection and tracking.

6.3.1 System Implementations

In this experimental work, hand gestures are fed as input to several gesture detection algorithms. Figures 6.11, 6.14 and 6.17 show twelve hand gestures recorded in long distance with a plain background using a holoscopic imaging camera system. It can be seen that some motions are 2D while others are 3D. The images are pre-processed before extracting videos in terms of some steps. The steps implemented in the rest of figures are similar to these of short distance hand gesture.

1- <u>Pre- extraction first person's hand motions in long distance</u>



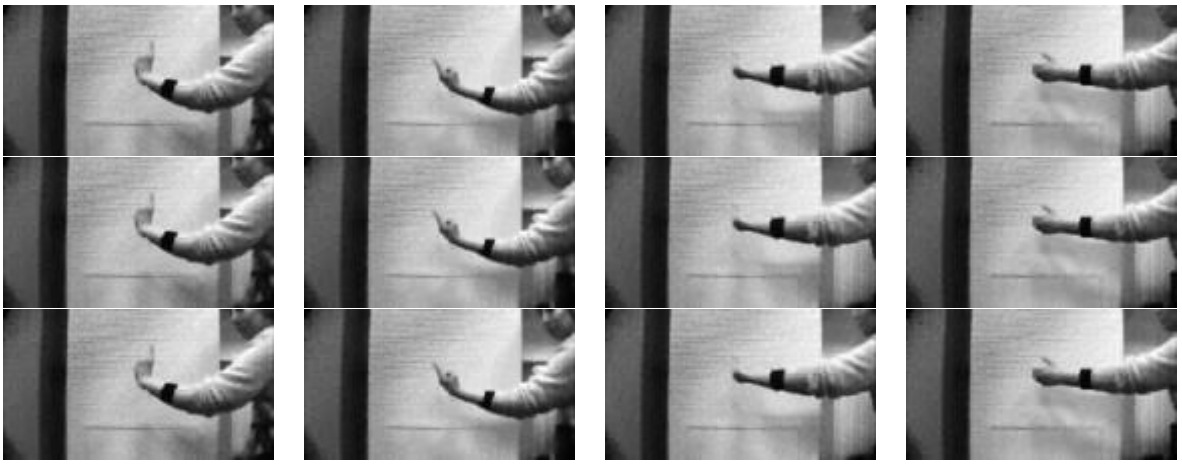
a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion e) 2 Fingers Shrink

f) Back/Forth g) Rub motion h) Click motion i) Dance motion j) Pinch motion

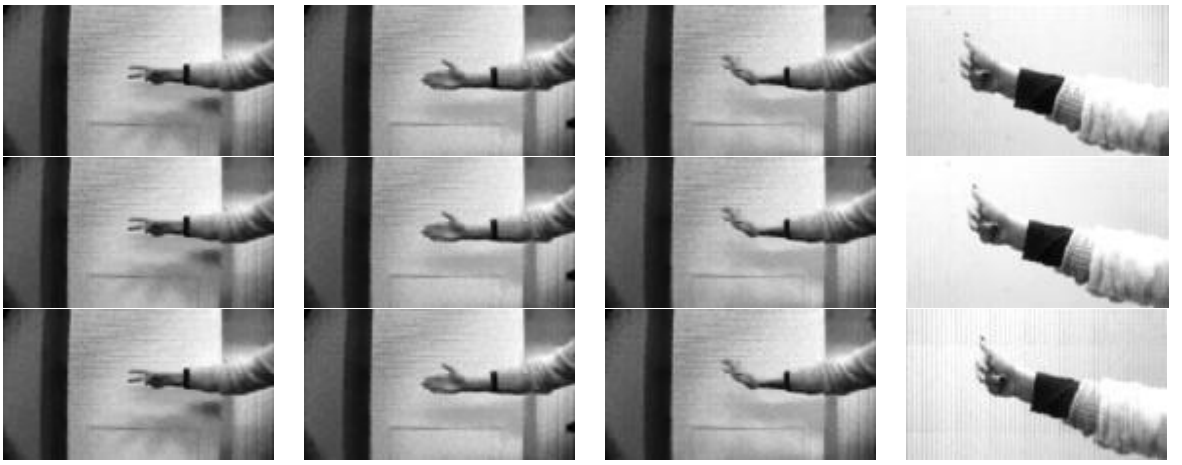
k) write motion l) Click motion 2

Figure 6.11: Pre- extraction first person's hand motions in long distance

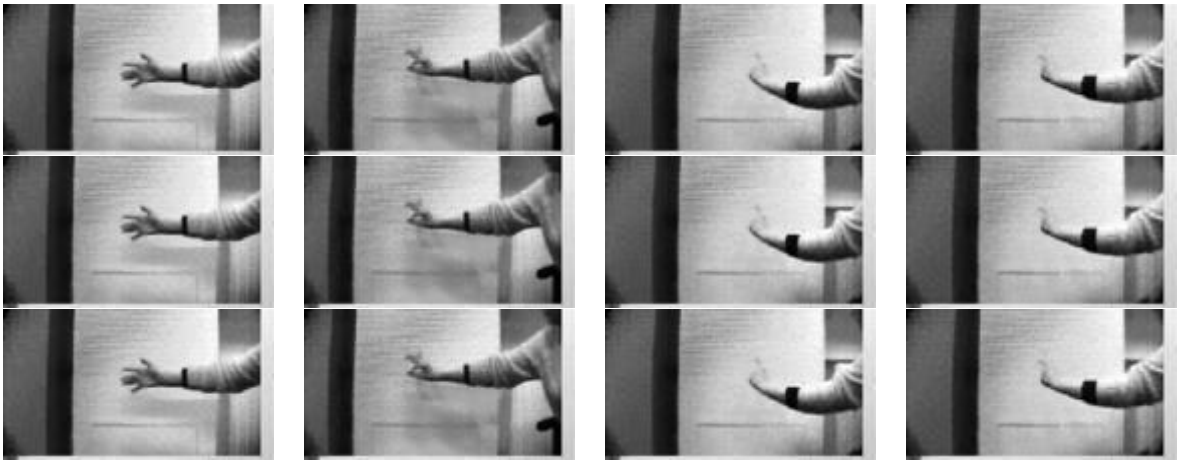
2- Post- extraction first person's hand motions in long distance single (LCR)



a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion



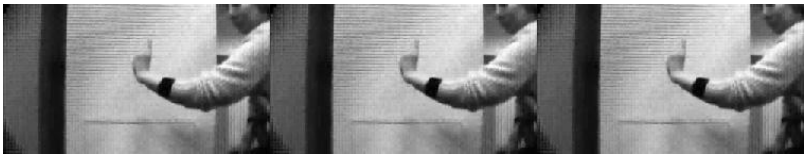
e) 2 Fingers Shrink f) Back/Forth g) Rub motion h) Click motion



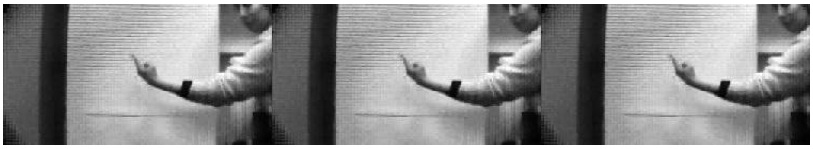
i) Dance motion j) Pinch motion k) write motion l) Click motion 2

Figure 6.12: Post- extraction first person's hand motions in long distance single (LCR)

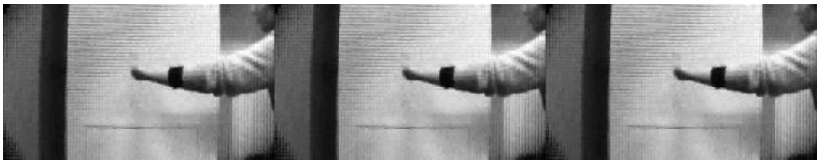
3- <u>Post- extraction first person's hand motion in short distance short distance combined (LCR)</u>



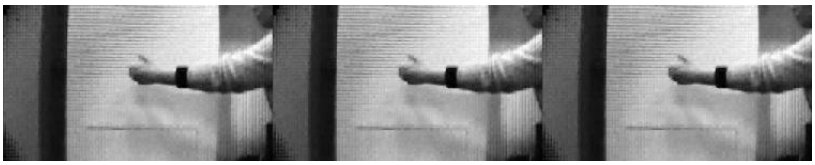
a) Sweep motion



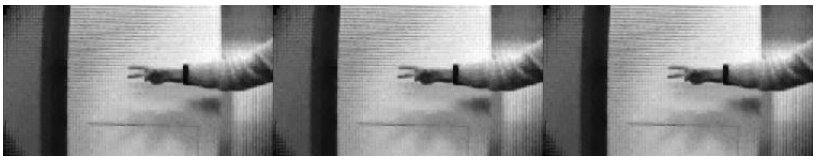
b) Shrink motion



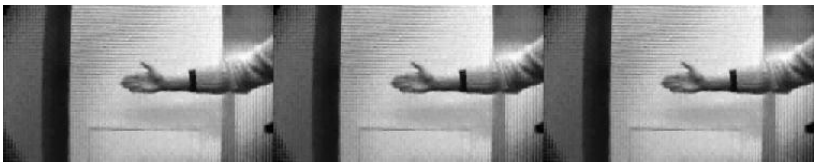
c) Circular motion



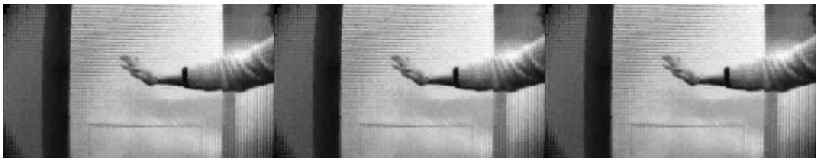
d) Squeeze motion



e) 2 Fingers Shrink



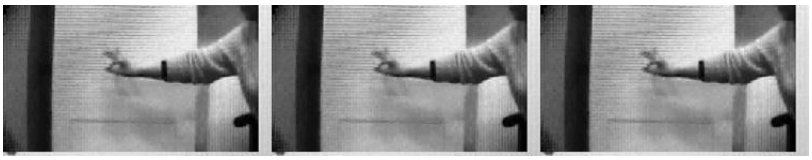
f) Back/Forth



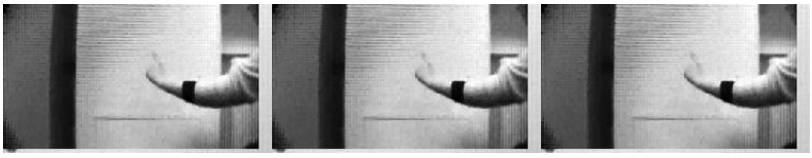
g) Rub motion

h) Click motion

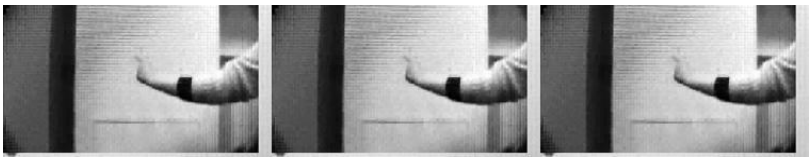
i) Dance motion



j) Pinch motion



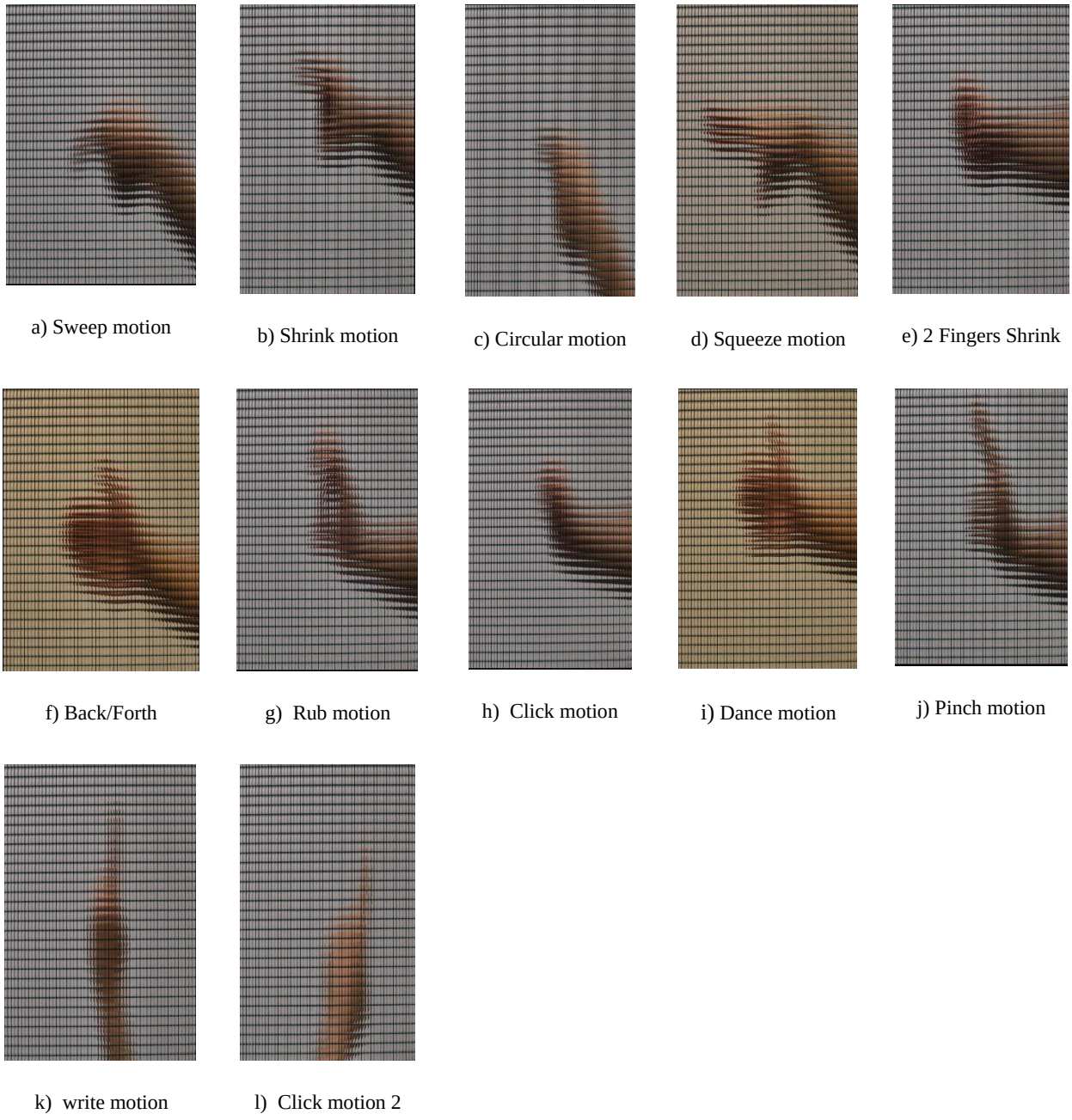
write motion



l) Click motion 2

Figure 6.13: Post- extraction first person's hand motion in short distance short distance combined (LCR)

4- Pre- extraction second person's hand motions in long distance



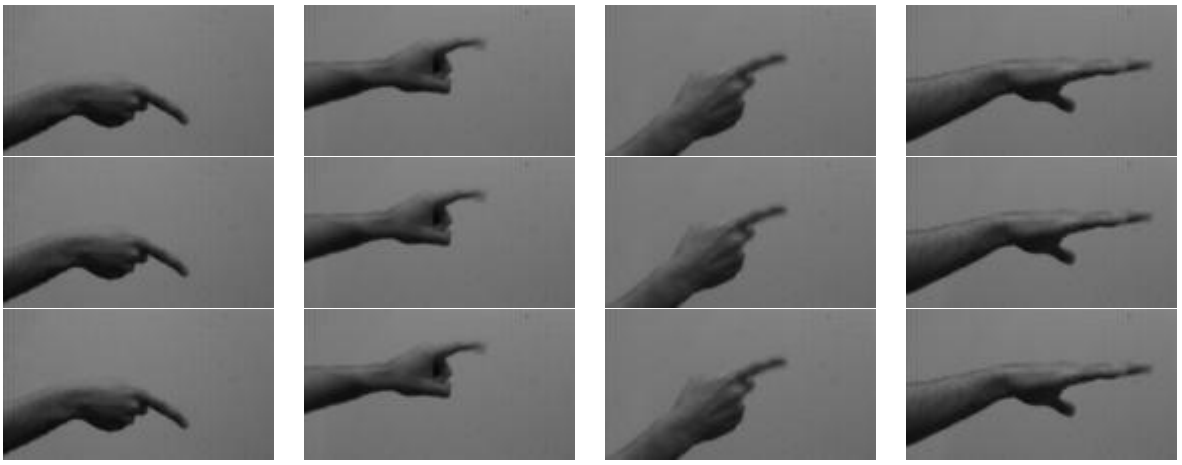
a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion e) 2 Fingers Shrink

f) Back/Forth g) Rub motion h) Click motion i) Dance motion j) Pinch motion

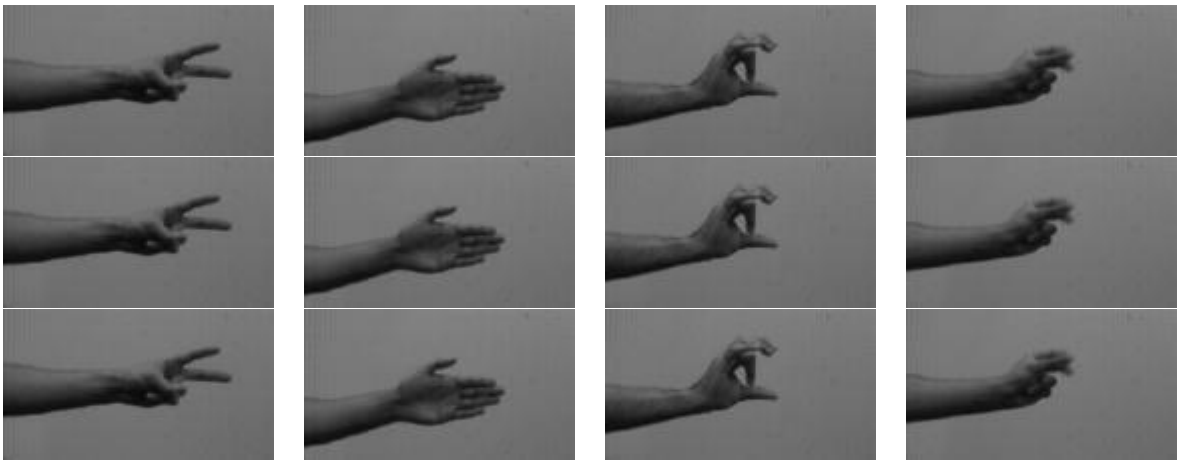
k) write motion l) Click motion 2

Figure 6.14: Pre- extraction second person's hand motions in long distance

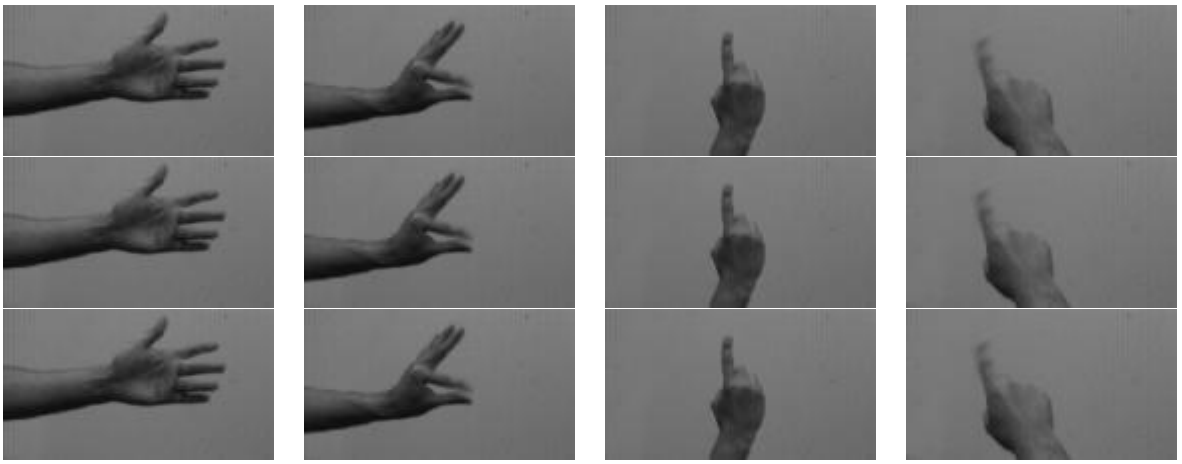
5- Post- extraction second person's hand motions in long distance single (LCR)



a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion



e) 2 Fingers Shrink f) Back/Forth g) Rub motion h) Click motion



i) Dance motion j) Pinch motion k) write motion l) Click motion 2

Figure 6.15: Post- extraction second person's hand motions in long distance single (LCR)

6- Post-extraction second person's hand motions in long distance combined (LCR)

a) Sweep motion

b) Shrink motion

c) Circular motion

d) Squeeze motion

e) 2 Fingers Shrink

f) Back/Forth

g) Rub motion

h) Click motion

i) Dance motion

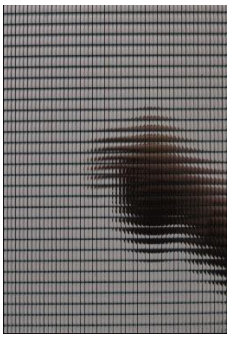
j) Pinch motion

write motion

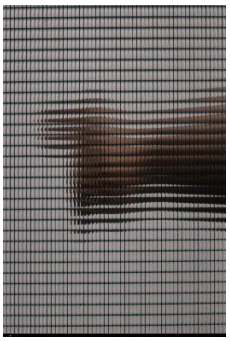
l) Click motion 2

Figure 6.16: Post-extraction second person's hand motions in long distance combined (LCR)

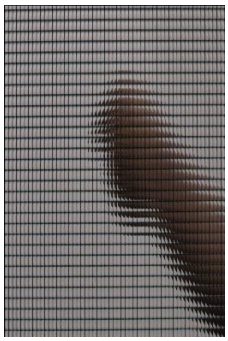
7- Pre- extraction third person's hand motions in long distance



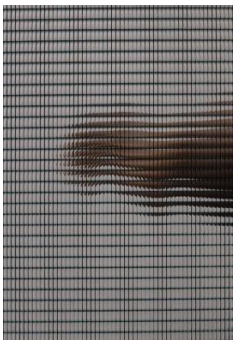
a) Sweep motion



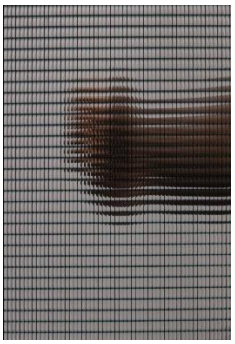
b) Shrink motion



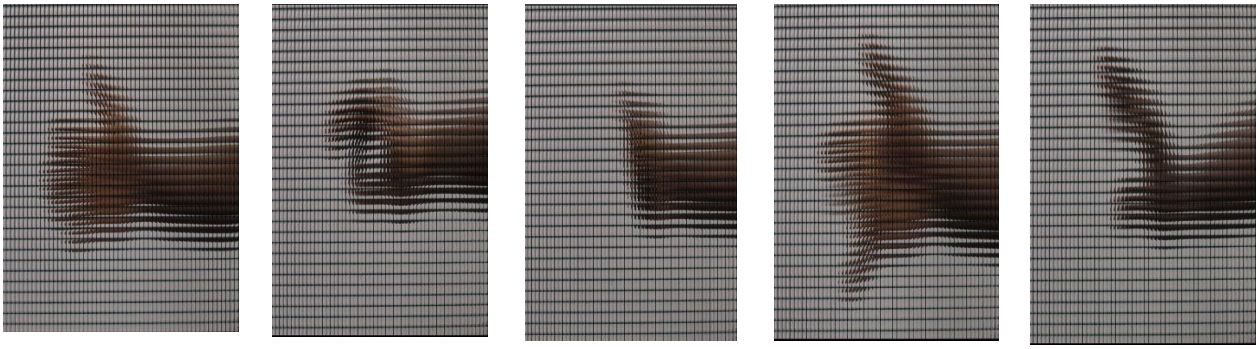
c) Circular motion



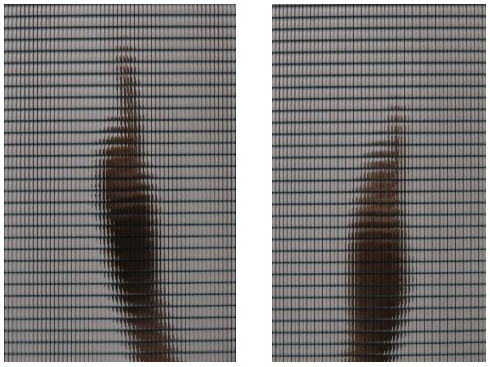
d) Squeeze motion



e) 2 Fingers Shrink



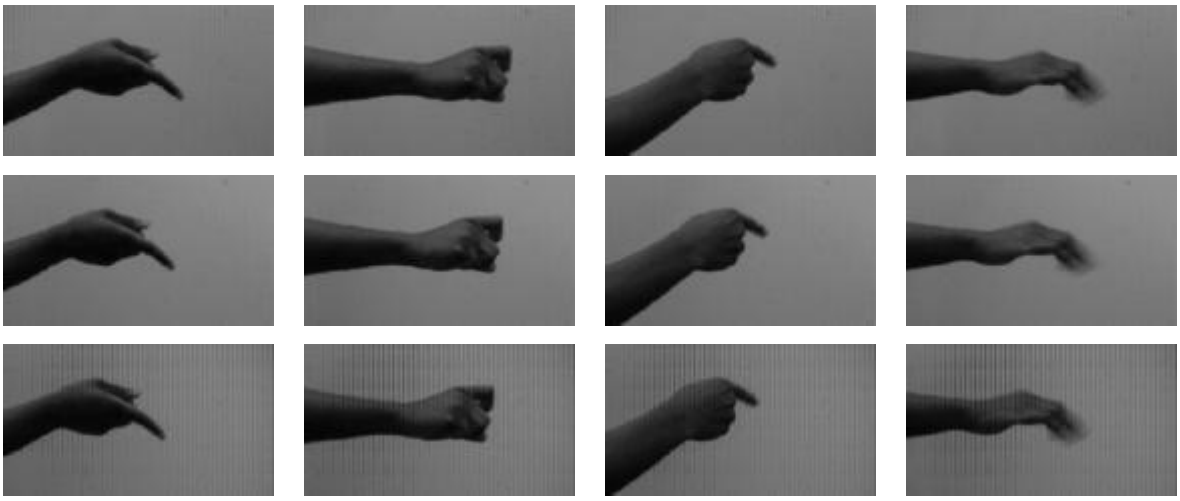
f) Back/Forth g) Rub motion h) Click motion i) Dance motion j) Pinch motion



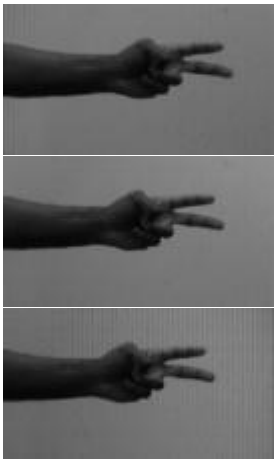
k) write motion l) Click motion 2

Figure 6.17: Pre- extraction third person' hand motions in long distance

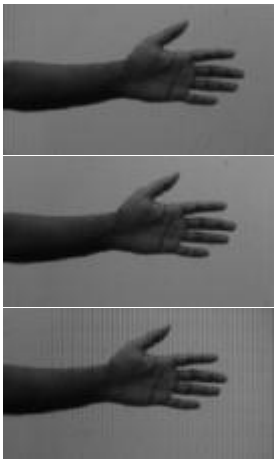
1- Post-extraction third person's hand motions in long distance single (LCR)



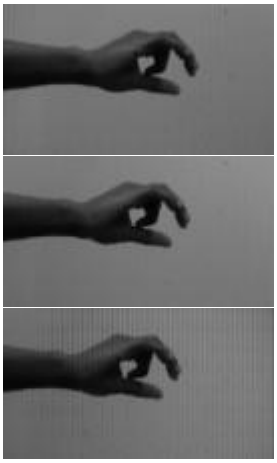
a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion



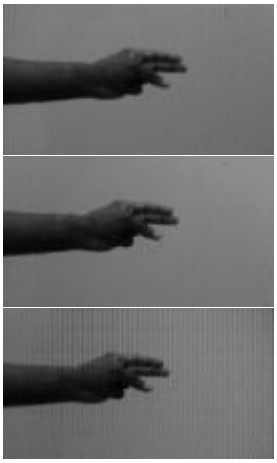
e) 2 Fingers Shrink



f) Back/Forth

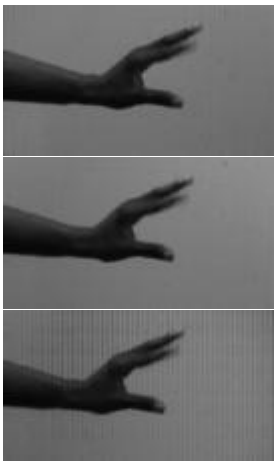


g) Rub motion



h) Click motion

i) Dance motion



j) Pinch motion

k) write motion

l) Click motion 2

Figure 6.18: Post-extraction third person's hand motions in long distance single (LCR)

8- Post-extraction third person's hand motions in long distance combined (LCR)

a) Sweep motion

b) Shrink motion

c) Circular motion

d) Squeeze motion

e) 2 Fingers Shrink

f) Back/Forth

g) Rub motion

h) Click motion

i) Dance motion

j) Pinch motion

k) write motion

l) Click motion 2

Figure 6.19: Post-extraction third person's hand motions in long distance combined (LCR)

A. Convolutional Neural Network Implementation

Convolutional Neural Network is an integral part of deep learning since it is used to train data without applying any image processing method. In this experiment, a new separate directory is created for each video of three people. The length of each video is 10 seconds for the separated and combined images. Each video will be read to generate 900 images, i.e. 300 for left, 300 for right and 300 for the centre while combined is 300 images in one directory. The images are divided for training and testing the models. The number of training frames for separate images is 390 whereas it's 210 for combined images which are %70. The CNN's topology is produced in seven layers with each layer having the following functionality and size: ImageInputLayer size [135, 75, 1] for separate images whereas combined [405, 75,1] , Convolution2DLayer Filter size [5,20], Rectified Linear Unit (ReLULayer), MaxPooling2DLayer Pool size [2,2], FullyConnectedLayer size [auto] and Output size [7], SoftmaxLayer and ClassificationOutputLayer Output size [auto]. The CNN hyperparameters are created inside the training options function. The epochs' parameter value is set to 100 epochs.

6.3.2 Results

Table 6.2 shows the comparison between three people to obtain the best results. Single, combined and all three combined results presented in terms of execution time, training, testing, sensitivity, specificity, PPV, NPV, LR+ and LR−.

In the single experiment, the execution time of the third person is quite lower than the second and third persons. The second person has the lowest value compared to the first and third persons in training. The result of testing for the first person is the best while other results are not as good. Sensitivity results for all persons are equalled whereas the result of the third person is less than other results in specificity. The PPV results in this experimental work for the third

person is slightly less than other results whereas the result of NPV is equalled for all persons. The LR+ result for the third person is the highest value compared to the first- and second persons value. The result of LR- is equalled for all persons.

In the combined images, the result of execution time for the first-person experiment is the highest. The training result for the second person is lower than the first and third persons. The third person has the best testing result at %99. The results of sensitivity and specificity are equalled for all persons. The PPV and NPV results are also equalled for all persons. LR+ and LR- has the best values for all persons.

The ALL combined experiment presented the performance of all three persons. The execution time of the ALL experiments is the highest. The result of training for ALL three persons is slightly higher than the second person in the single experiment. ALL has the lowest result in testing compared to other results. The sensitivity result for ALL is the lowest while the specificity result is 100%, which is higher than the third person result. The result shown in PPV is 100%, whereas NPV for ALL is the lowest. The LR+ value in ALL is lower than the third person result while LR- for ALL has the highest result.

The summary of the comparison shows that the first person has the best results in all categories in the single experiment compared to the other persons' results, except for the execution time which is the highest. The results of the second person in combined are better than the first and second's results. However, the result of testing for the second person in combined is slightly lower than the others' results. The values of the ALL experiment in categories are slightly lower than the other experiments. Except the value of specificity and PPV which are 100%. Overall, the single experiment of the first person has the best values in most parameters. Using the CNN algorithm in this study offers many advantages. The first advantage is being able to detect features of images without human observation. It also has the capability to learn from image or video data faster than ANN. Finally, CNN surpasses ANN in conventional image recognition. However, the implementation time of CNN is longer than that of ANN.

Table 6.2: Comparison Between first person, second person and third person in CNN

	First person		**Second person**		**Third person**		**ALL**
	Single (LCR)	**Combined**	**Single (LCR)**	**Combined**	**Single (LCR)**	**Combined**	**Combined**
Execution Time (H:M:S)	02:35:19	02:37:07	00:47:09	00:25:00	00:45:28	00:46:37	02:48:06
Training	1	1	0.99	0.99	1	1	0.99
Testing	1	0.97	0.99	0.99	0.98	0.99	0.91
Sensitivity	1	1	1	1	1	1	0.90
Specificity	1	1	1	1	0.99	1	1
Positive Predictive Value (PPV)	1	1	1	1	0.96	1	1
Negative Predictive Value (NPV)	1	1	1	1	1	1	0.99
Positive Likelihood (LR+)	0	0	0	0	276.63	0	0
Negative Likelihood (LR−)	0	0	0	0	0	0	0.09

6.3.3 Summary

Hand gesture detection is elementary to provide a natural HCI skill. The most essential aspects in gesture recognition are segmentation, detection and tracking. This experiment is performed for hand gestures recognition using features extraction and classification using CNN technique. In this experimental work, twelve 3D motions are recorded within long distance for three different people. Experiments were conducted to compare performance of CNN method in terms of multi factors like execution time, training, testing, sensitivity, and specificity, PPV, NPV, LR+ and LR−. The results showed that single experiment for the first person provided better results in all categories.

6.4 Disparity

The apparent motion in pixels for every point can be measured in a pair of images derived from stereo cameras. Such an apparent pixel difference or motion between a pair of stereo images is called Disparity. This phenomenon can be experienced by trying to close one of your eyes and then rapidly close it while opening the other. The objects closer to us will be moved to a significant distance from the real position and objects further away move little. This type of motion is disparity. A case where disparity is most useful is for calculation of depth / distance. Distance and disparity from the cameras are inversely related [107]. As distance from the cameras increases, the disparity decreases. This can help for depth perception in stereo images [107]

6.4.1 Disparity Systems

A new technique for 3D rigid motion estimation from stereo cameras is proposed by Demirdjian and Darrell [108]. The technique utilises the disparity images obtained from stereo matching. Some assumptions like the stereo rig has parallel cameras and, in that case, the topological and geometric properties of the disparity images. A rigid transformation (called d-motion) is introduced whose function is mapping two disparity images of a rigidly moving object. The relation between motion estimation algorithm and Euclidean rigid motion is derived. The experiment shows that the proposed technique is simpler and more accurate than standard methods.

According to Pyo et al [109], the CNN method used to analyse and evaluate hand gestures recognition. CNN can deal with multi-view changes of hand gestures. The paper also shows how to use depth-based hand data with CNN and to obtain results from it. The evaluation is

made against a famous hand database. The results show that CNN recognises gestures with high accuracy and the technique is suitable for a hand gesture dataset. The CNN structure of three convolutional layers and two fully connected layers has the best accuracy.

A feature match selection (FMS) algorithm is presented by [110], with an aim to extract and estimate an accurate full parallax 3D model form from a 3D omni-directional holoscopic imaging (3DOHI) system. The novelty of the paper is based on two contributions: feature blocks selection and its corresponding automatic optimisation process. The solutions for three primary problems related to depth map estimation from 3DHI: dissimilar displacements within the matching block around object borders, uncertainty and region homogeneity at image location, and computational complexity.

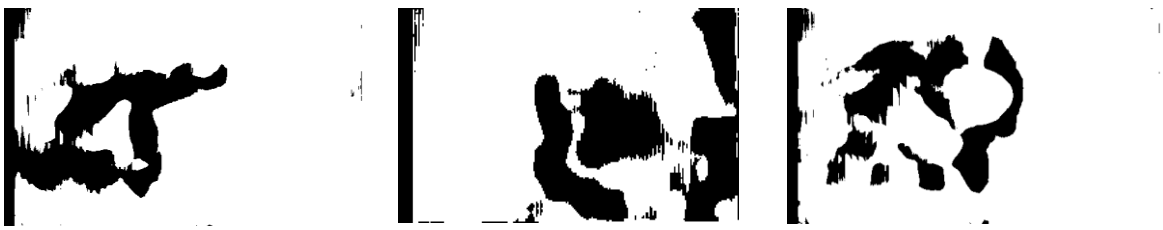
6.4.2 Implementation

A. ***Hand Gestures Input***

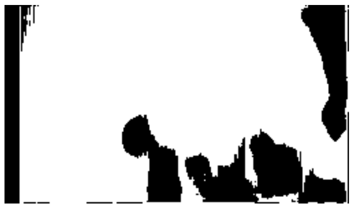
Figure 6.19 shows the disparity of left and right images taken from the previous experimental work. The images are pre-processed using the same method as previous experimental work except for a few steps that are applied to find the disparity. The images generated are converted from RGB to grey. The default size of the image needs to be 550 × 310. Create twelve directories for three different people. The first disparity image size is 59 and the window size 20 while the disparity of second and third images is 49 and the window size is 31. The stereo match function in Matlab software is used to find the disparity of left and right images.

1- The disparity of Left and Right images for 3 people

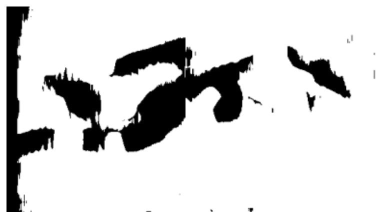
a) Sweep motion a) Sweep motion a) Sweep motion



b) Shrink motion b) Shrink motion b) Shrink motion

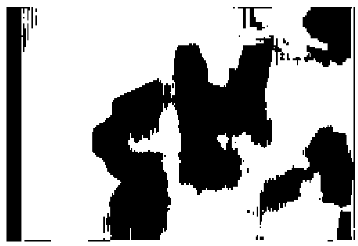
c) Circular motion c) Circular motion c) Circular motion

d) Squeeze motion d) Squeeze motion d) Squeeze motion

e) 2 Fingers Shrink e) 2 Fingers Shrink e) 2 Fingers Shrink

f) Back/Forth f) Back/Forth f) Back/Forth

g) Rub motion g) Rub motion g) Rub motion

h) Click motion

h) Click motion

h) Click motion

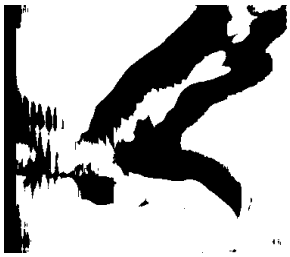
i) Dance motion

i) Dance motion

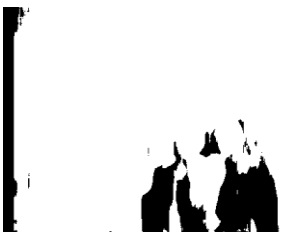
i) Dance motion

j) Pinch motion

j) Pinch motion



j) Pinch motion

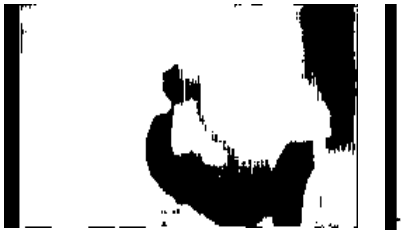


k) write motion

k) write motion

K) write motion

l) Click motion 2



l) Click motion 2

l) Click motion 2

Figure 6.19: The disparity of Persons 1, 2 and 3

A. Convolutional Neural Network Implementation

Convolutional neural network is an integral part of deep learning since it is used to train data without applying any image processing method. In this experiment, a new separate directory is created for each video of three people. Each video will be read to generate 300 disparity images in a directory. The images are divided for training and testing. The number of training frames for separate images is 390 whereas it is 210 for combined images which are %70. The CNN's topology is produced in seven layers with each layer having the following functionality and size: ImageInputLayer size [310, 550, 1], Convolution2DLayer Filter size [5, 20], Rectified Linear Unit (ReLULayer), MaxPooling2DLayer Pool size [2,2], FullyConnectedLayer size [auto] and Output size[7],SoftmaxLayer and ClassificationOutputLayer Output size [auto]. The CNN hyperparameters are created inside the training options function. The epochs' parameter value is set to 50 epochs.

6.4.3 Results

Table 6.3 presents the comparison between three people to achieve the best results. Single, ALL combined three combined results displayed in terms of execution time, training, testing, sensitivity, specificity, PPV, NPV, LR+ and LR−.

In the single images experiment, the execution time of the second person is higher than the first and third persons. The results of training for all persons are equalled at 100%. The first person has the best testing result which is 100%, while the second and third persons have lower results. The sensitivity result for the second person is slightly lower than the second and third results, while all results for three persons are equalled in specificity. The PPV result for second person is lower than other results, while the result of NPV for all three persons is equalled. LR+ for the second person is the highest which at 933 while LR- results for all three persons is equalled.

In combined, the execution time is the lowest compared to other results. The training result is 100% whereas the testing result is 0.9803. The result is decreased in sensitivity more than the three people results whereas the result of All combined is equalled in specificity. The PPV result of 100% is better than the second person result. The result recorded for NPV is the lowest compared to other results. LR+ is zero compared to the result of the second person. The highest value for LR- is 0.1333 while other results are zeros.

The summary of the comparison is the first person has the best values in all categories in the single experiment compared to other persons' results, except the execution time for the second

person is the highest. The values of the ALL experiment in categories is slightly better than the second person result. Overall, the single experiment of the first person has the best values in most parameters.

Table 6.3: Comparison the disparity Between first person, second person and third person in CNN

	First person	**Second person**	**Third person**	**All**
	Single (LCR)	**Single (LCR)**	**Single (LCR)**	**ALL Combined**
Execution Time (H:M:S)	00:29:31	00:32:09	00:31:18	00:27:28
Training	1	1	1	1
Testing	1	0.99	0.99	0.98
Sensitivity	1	0.99	1	0.96
Specificity	1	1	1	1
Positive Predictive Value (PPV)	1	0.98	1	1
Negative Predictive Value (NPV)	1	1	1	0.99
Positive Likelihood (LR+)	0	933	0	0
Negative Likelihood (LR−)	0	0	0	0.03

6.4.4 Summary

The obvious motion in pixels for each point can be calculated in a pair of images obtained from stereo cameras. Disparity is defined as an apparent pixel difference or motion between a pair of stereo images. This experimental work is performed for the disparity of hand gestures using features extraction and classification using CNN technique. This experimental work includes twelve 3D motions recorded within short distance for three different people. Experiments were implemented to compare performance of CNN technique in terms of different factors like execution time, training, testing, sensitivity, specificity, PPV, NPV, LR+ and LR−. The results

showed that the single experiment for the first person provided better results in all categories. The next chapter presents the experiment relating to gesture recognition for stroke patients using the CNN algorithm.

Chapter 7
Stroke Patients Gesture Recognition

7.1 Introduction

Gesture recognition can be defined as the ability of a computer to understand the gestures and perform certain commands based on those gestures. The essential goal of gesture recognition is to develop a system that can identify and understand specific gestures and communicates information from them [111]. Gesture based communication is a powerful approach in enabling stroke patients to convey information including for device control purposes, post recovery. The use of gestures for human computer interactions is currently in its infancy. Stroke patients in particular may suffer from limited control of their bodies with different patients being affected in different ways. Gesture recognition methods may be used to develop a patient specific means of communicating via cameras and computer vision algorithms depending on the effects suffered by the patients.

Hand gestures are a form of gestures used in healthcare, entertainment and the education industries. It's used to assist users with special needs and the elderly. Hand tracking is vital to perform hand gesture recognition, involves undertaking various computer vision operations including hand segmentation, detection, and tracking. Sign language uses hand gestures to convey feelings or information within the hearing impairment communication. The main issue is that an ordinary person would easily misunderstand the meaning conveyed. The advancement in AI and computer vision can be adapted to recognize and learn the sign language [112]. The modern systems can help an ordinary person to recognize and understand the sign language. This chapter presents a technique which is associated to the recognition of hand gestures using a method of deep learning.

Stroke is a disease which affects arteries leading to and within the brain. Stroke is the fifth leading death cause as well as a cause of disability [113]. A stroke happens when a blood vessel carries oxygen and nutrients to the brain is either blocked by a clot or bursts. This chapter shows that hand gesture is a very beneficial way to convey information and a very rich set of feelings and facts can be interpreted from gestures.

The objective of this study is to present the effectiveness of CNN technique to extract features and classify various images. In this study, CNN method is evaluated and compared between

training and testing. A hand gesture recognition system was developed based on the deep learning method. The performance of hand gesture recognition method was evaluated and compared using several factors such as execution time, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood, negative likelihood and root mean square

The remaining of the chapter is structured as follows: The Stroke Recognition Systems techniques and methods used in this chapter found in Section 7.2 The details of the proposed system's implementation used are shown in Section 7.3. Section 7.4 concentrates on the discussion and presentation of the results obtained. The conclusion is presented in section 7.5.

7.2 Stroke Recognition Systems

Convolution Neural Networks (CNNs) are used to evaluate hand gesture recognition, where depth-based hand data was employed with CNN to obtain successful training and testing results [114]. In this study, Soodtoetong and Gedkhaw present their study in process and methods related to sign language recognition using Deep Learning. 3D CNN was applied for recognising images received through Kinect sensor. The method using 3D CNN was found to be very effective and the highest accuracy was found to be 91.23% [112].

Lin et al [114] proposed another CNN method that uses a skin model, hand position calibration and orientation to train and test the CNN. A hand gesture recognition sensor is introduced in this study uses ultra-wide band impulse signals [114]. Each gesture has its own reflected waveform and CNN is used for gesture classification. Six gestures from American Sign Language (ASL) have been used for the experiment. The results show a 90% accuracy using CNN and proved its effectiveness recognising the gesture [115].

According to Kim et al [116], A Pattern Recognition model is proposed for dynamic hand gesture recognition which combines CNN with weighted fuzzy min-max neural network. The model also presents feature extraction, feature analysis and spatiotemporal template data representation based on the motion information of target is designed. The efficiency of classifier is increased by performing feature analysis technique using weighted fuzzy min-max neural network. The results display the influence caused by feature point’s spatial and temporal variation which can be reduced using the proposed implementation [116].

Rao et al suggest a CNN algorithm to recognise the Indian sign language gestures. The capture method used was Selfie mode continuous sign language video, where a hearing-impaired person would use the sign language recogniser mobile individually. The datasets were not

available for mobile use, hence, the authors created datasets with five subjects which performed two hundred signs in five various viewing angles under several background environments. Various CNN architectures were designed and tested and 92.88% was the best recognition rate obtained on the dataset [117].

Multi-class SVM and k-NN classifier are used to observe seven gestures for residential rehabilitation of the patients who have had stroke [118]. These seven gestures were implemented on seventeen young people. The results were evaluated using k-fold cross validation method. The results show that multi-class k-NN and SVM classifier achieved an accuracy of 97.29% and 97.71%, respectively [118].

Chung et al use webcam to track region of interest (ROI) which is hand region gestures [119]. The Kernelized Correlation Filters (KCF) method is used to track the detected ROI. The image is resized and input to a deep CNN to recognize different hand gestures. CNN and back propagation methodologies were used by Varun et al to recognize gestures to help disables [120]. The machine can understand the images and identify what the images are which is valuable [120].

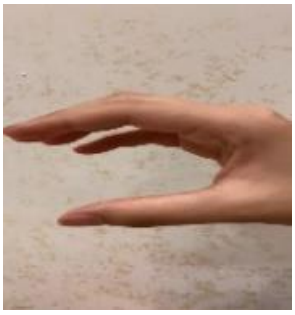
In this research, Alani et al [121] propose Adapted Deep Convolutional Neural Network (ADCNN) to recognize the hand gestures. Data augmentation is applied to increase robustness of deep learning and rise the size of dataset. The images are input into ADCNN in presence of RELU and Softmax, L2 regularization is used to remove overfitting. This method has been proved to be efficient to recognize hand gestures. The model is first trained using 3750 images with several variations in features like rotation, translation, scale, illumination and noise. Compared to baseline CNN, ADCNN had an accuracy of 99.73%, and a 4% improvement over the baseline CNN model (95.73%) [121].

7.3 System Implementation using CNN

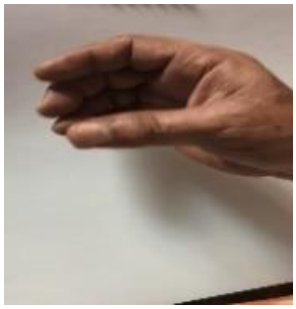
A hundred and forty hand gestures were used comprising seven different gestures created by each person with twenty people in total. These gestures were inputted into the gesture recognition method before being evaluated and compared in this study. Figure 7.1 illustrates three examples of twenty people showing seven 2D and 3D universal common hand gestures with three different mobile cameras, backgrounds, illumination, and the position of the hand and the shape of the hand. The mobile camera used to record the first gesture are iPhone 8 and Samsung Galaxy S10 is used to record second gesture and the last gesture is recorded using iPhone 8. The first background is light blue, the second background is lightly floral and the last

one is plain. The illumination of the first example is less than the second and third example. The position of the hand is also slightly different as well as the shape of the hands. A first-hand gesture is for a young woman in the late of twenty, another hand gesture is for a young woman in the mid of thirty, the last one is for an old man in the mid of seventy. They are recorded within short distances and used in the study's experimental work. The hand gestures signs are referenced from Simple hand sign communication cards used by Single hand communications and are shown in Figure 7.2 [122].

a) Drink



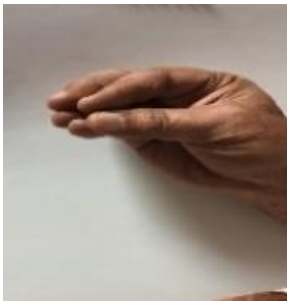
a) Drink



a) Drink

b) Eat

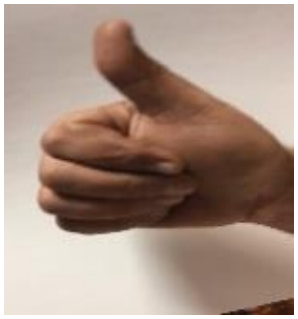
b) Eat



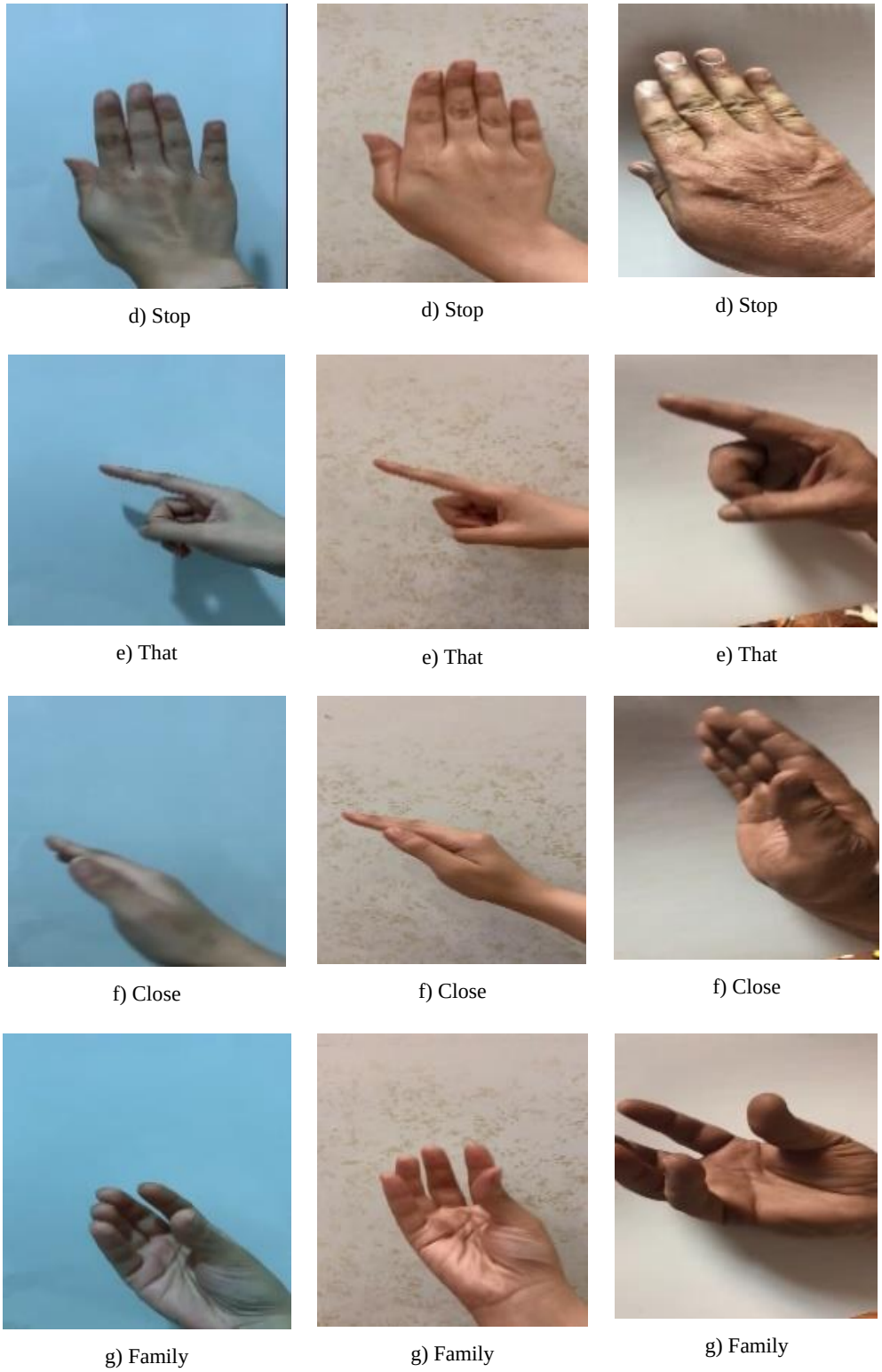
b) Eat

c) Good\Bravo

c) Good\Bravo



c) Good\Bravo



d) Stop d) Stop d) Stop

e) That e) That e) That

f) Close f) Close f) Close

g) Family g) Family g) Family

Figure 7.1: Three examples of seven universal hand gestures for three different hands

:

Figure 7.2: Simple hand signs cards [122].

The framework model in Figure 7.3 illustrates the system implementation steps. Using different high standard mobile cameras with two resolutions which are HD and 4k, the hand motions shown in Figure 7.4 are recorded. Each recording lasts from 1 to 10 seconds, and the recorded video resolution is varied.

Figure 7.3: Framework Model of System Implementation

7.3.1 Computing Specification

The experimental work was implemented using a Dell desktop C2544404 with processor Generation Intel ® Core™ i7-6700 CPU @ 3.40GHz, DDR4 16 GB memory type, storage hard drive 512 GB, 24 inch (60.9 cm) Full HD (1920 X 1200) Widescreen Ultrasharp IPS Panel display w/Adobe RGB colour space. Windows 10 (64 bits) operating system was used and the system is implemented using MATLAB R20187bV language.

7.3.2 Convolutional Neural Network Implementation

CNN forms an integral part of deep learning as is applied to train data without using any image processing technique. In this experimental work, a new directory is created for each video. Hundred and forty videos are read to generate 24,698 image frames. The method to convert the image frame from RGB colour to grey and resize it to 227×227 from the original image size is

shown in Figure 7.4. Each recorded video has a various number of frames between 3394 to 3670 frames. The data of images is divided into training and testing datasets. The number of training frames is 2485 which is 70%. The topology of CNN is generated in seven layers with each layer having the following functionality and size: ImageInputLayer size [227,227,1], Convolution2DLayer Filter size [5,20], Rectified Linear Unit (ReLULayer), MaxPooling2Dlayer Pool size [2,2], FullyConnectedLayer size [auto] and Output size[7], SoftmaxLayer and ClassificationOutputLayer Output size [auto]. The CNN hyperparameters are produced inside the training options function. The value of epochs parameter is set to 50 epochs.

a) Drink

a) Drink

a) Drink

b) Eat

b) Eat

b) Eat

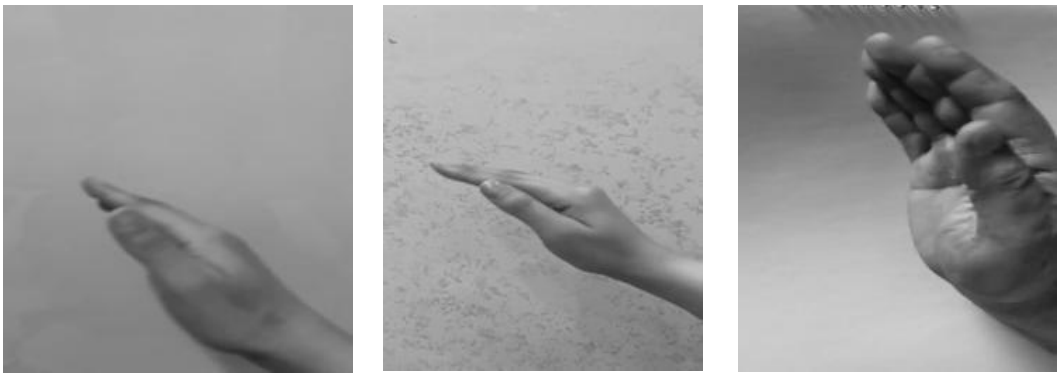
c) Good\Bravo

c) Good\Bravo

c) Good\Bravo

d) Stop d) Stop d) Stop

e) That e) That e) That



f) Close f) Close f) Close

g) Family g) Family g) Family

Figure 7.4: Three examples for seven universal common hand gestures for three different hands post extraction

7.4 Results and Discussion

The performance of CNN algorithm is compared between training and testing using several parameters including execution time, which is the duration taken by the software to implement the task. Sensitivity measures the percentage of positives that are appropriately identified. Specificity is a measure of the false positive rate. The PPV and NPV are the percentages of positive and negative results in diagnostic and statistics tests which also describe the true positive and true negative results. The LR+ and LR− are identified measures in diagnostic accuracy.

The experiments were executed ten times to acquire the mean of seven-hand gestures. Two different training and testing modes were presented and compared to find the best result. Training accuracy is accomplished by applying a prototype on the training data and determining the accuracy of the algorithm.

A summary of the values obtained for various parameters in training and testing approach is listed in Table 7.1. It can be noticed that the execution times for training and testing are the same. The accuracy result of training is 100% compared to that of testing. The value of sensitivity in training is a bit higher than testing. Specificity in training is 100% whereas in testing is 0.9989. The PPV and NPV of testing is lower than training. The best value for LR+ and LR− are recorded for training. For RMS, the value of training and testing are matched.

The training parameter values in CNN are fixed for all categories. The execution time is approximate 15,598 seconds, which is duration to train and test the system using seven hand gestures which are used in the experiment. Overall, training has the best values in most parameters. The implementation of CNN algorithm in this study has many advantages. Firstly, CNN able to capture the features of image without any human intervention. It is capable to learn the image or video faster than ANN. CNN surpass ANN on conventional image recognition. On the other hand, the execution time of CNN is longer than ANN.

Table 7.1: CNN Training and Testing Approach

Factors	**CNN**	
	Training	**Testing**
Exe Time ± SD (sec)	15,598 ± 244.9784	15,598 ± 244.9784
Accuracy ± SD	1 ± 0	0.9912 ± 0.0086
Sensitivity ± SD	1 ± 0	0.9934 ±0.0042

Specificity ± SD	1 ± 0	0.9989 ± 0.0023
Positive Predictive Value (PPV) ± SD	1 ± 0	0.9934 ± 0.0040
Negative Predictive Value (NPV) ± SD	1 ± 0	0.9989 ± 0.0021
Positive Likelihood (LR+) ± SD	1 ± 0	884.4175 ± 37.5328
Negative Likelihood (LR−) ± SD	1 ± 0	0.0066 ± 1.7920
RMS ± SD	1 ± 0	1 ± 0

7.5 Summary

Hand gesture detection is fundamental to provide a natural HCI skill. It is now known that in gesture recognition, the most essential aspects are detection, segmentation and tracking. This experiment is a system which has been created for hand gestures recognition using features extraction and classification in CNN technique. Seven 2D and 3D motions with different mobile cameras, backgrounds, illumination, position of hand and the shape of hand are recorded within short distances. The experiments were performed to compare the performance of training and testing in CNN method. The results showed that training provides better accuracy compared to testing. In future work, the number of gestures will be extended to ten common gestures using 3D holoscopic imaging technique camera.

Chapter 8
Conclusion and Future work

8.1 Conclusion

This thesis shows six essential experiments in the field of hand gesture. The first two experiments are 2D video detection which consist of detecting ten different gestures in short and long distances using an iPhone 6 Plus camera. The aim of these two experimental works is to compare different algorithms in training and testing approaches to discover the best algorithm to extract and classify hand gesture recognition. These algorithms were evaluated in terms of different parameters such as execution time, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood, negative likelihood, receiver operating characteristic, area under ROC curve and root mean square. After the pre-processing phase, both studies were implemented using two image processing tools which are WT and EMD. WT is one image processing technique which performs signal analysis with one signal frequency differing at the end of time. An innovative technology used in both non-stationary and non-linear data namely EMD. The primary function of this method is decomposing a signal into intrinsic mode functions consistently through the domain. For classification, ANN is used for both experiments which is defined as a system that processes information and has structure much like that of the biological nervous system. CNN is a multi layers neural network which is one of the deep learning techniques used efficiently in the field of gesture recognition. In system implementation, WT and EMD algorithms are used to extract image features which are later fed into ANN for gesture classification. Applying CNN in both experiments reduces two phases which are image extraction and classification to one phase only. Comparing the results showed that CNN is clearly the most appropriate method to be used in hand gesture system.

In the third and the fourth experimental works, the number of hand gestures is extended to twelve for three different people and all of them have been recorded using a 3D holoscopic imaging system camera. In the pre-processing phase, the twelve 3D videos were extracted in single left, centre and right images, and combined (LCR) images. The 3D holoscopic concept is based on the imagining system which represents a true volume spatial optical model of the object scene. The significant aim of the 3D vivant project is to analyse and investigate the possibility of displaying the 3D holoscopic content on the auto stereoscopic display. The aim of the third and fourth experiments is to use the CNN method to discover the best results in

single and combined images between three people in terms of multi parameters. The parameters are execution time, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood and negative likelihood. Comparing the results shows that the hand gestures of the first person in single had the best results compared to the other two people except the execution time which takes longer than for the other two people.

The fifth experimental work is finding the disparity for hand gestures in short distance for three people. Disparity is defined as the obvious motion in pixels for each point and can be measured in a pair of images obtained from stereo cameras. The pre-processing used in this experimental work is similar to the third and fourth experiments. The left and right images for three people are taken from the previous experimental work. The aim of this experiment is applying a CNN algorithm to find the best results in single images between three people in terms of multi factors. Therefore, the results presented for hand gestures of the first person in single were the best results whereas the results of the other two people were lower, except for the execution time.. The combined images had less execution time compared to the single image experiments.

The last experimental work is detecting hand gestures to assist people who have experienced stroke. This experiment is implemented by detecting a hundred and forty gestures composed of seven different gestures for twenty people. These gestures were recorded using different mobile cameras, backgrounds, illumination, position of the hand and shape of the hand. The aim of this last experiment is to use a CNN method to display the best results between training and testing modes for a hundred and forty gestures in terms of multi parameters. Overall, the CNN demonstrated the ability to classify 2D and 3D images.

8.2. Suggestions for Future Work

After performing all the experiments and reviewing the results, the following are suggestions for future work:

1- Extend the number of hand gestures to cover all universal common gestures like victory/peace, hungry, cold, luck and more to build a strong model for people who have experienced a stroke and people with hearing impairments. These gestures could be learnt easily to communicate better with other people.

2- Record gestures with different objects and backgrounds. For example, recording hand gestures while holding a pen or a stress ball in an office.

3- Extend gestures to include different parts of the body such as hands and lips for gesture recognition to cover universally common gestures. For instance, developing a gesture to represent drinking using both a hand and the lips.

4- Implement real time object detection using OpenCV. This method could be implemented using a webcam. The advantage of using this method is detecting gestures or objects and showing them on a screen in real time.

5- Design a mobile application by building a system which has some functionalities to translate common gestures to meaningful words. The output could be words shown on a screen or be dictated by sound. Include educational games to the application to aid learning for children with hearing impairments and people who have experienced strokes.

6- Record gestures with a high-resolution size such as 6K to examine the effectiveness of a CNN algorithm.

7- Record gestures using different cameras and lenses such as Kinect, stereo cameras, depth cameras, thermal cameras and single cameras and implement the recorded gestures using a CNN algorithm.

8- Implement different algorithms in deep learning like RNN because data used in the current experiments is video which is time based. Compare the efficiency of CNN and RNN in terms of training and testing accuracy.

9- Applying the proposed prototype in different fields like education to help children with hearing impairments. This prototype could be applied in schools and universities as a part of learning.

References

[1] Y. Li, J. Huang, F. Tian, H.-A. Wang, and G.-Z. Dai, "Gesture interaction in virtual reality," *Virtual Reality & Intelligent Hardware*, vol. 1, no. 1, pp. 84–112, Jan. 2019.

[2] V. Pavlovic, R. Sharma, and T. Huang, "Visual interpretation of hand gestures for human-computer interaction: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 677–695, 1997.

[3] H. Hasan and S. Abdul-Kareem, "Human–computer interaction using vision-based hand gesture recognition systems: a survey," *Neural Computing and Applications*, vol. 25, no. 2, pp. 251–261, 2013.

[4] 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, vol. 20, no. 1, pp. 28–37, 2013.

[5] M. G. Lippmann, "La photographie integrale," Comptes-Rendus Acad. *Sci.*, vol. 146, pp. 446–551, 1908.

[6] A. Agooun, O. A. Fatah, J. C. Fernandez, C. Conti, P. Nunes, and L. D. Soares, "Acquisition, processing and coding of 3D holoscopic content for immersive video systems," *2013 3DTV Vision Beyond Depth (3DTV-CON)*, 2013.

[7] S. Adedoyin, W. Fernando, A. Aggoun, and K. Kondoz, "Motion and Disparity Estimation with Self Adapted Evolutionary Strategy in 3D Video Coding," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 4, pp. 1768–1775, 2007.

[8] V. M. Bove, "Display holography's digital second act," *Proc. IEEE*, vol. 100, no. 4, pp. 918–928, 2012

[9] X. Xiao, K. Wakunami, X. Chen, X. Shen, B. Javidi, J. Kim, and J. Nam, "Three-Dimensional Holographic Display Using Dense Ray Sampling and Integral Imaging Capture," *Journal of Display Technology*, vol. 10, no. 8, pp. 688–694, 2014.

[10] J. Wang, X. Xiao, H. Hua, and B. Javidi, "Augmented Reality 3D Displays With Micro Integral Imaging," *Journal of Display Technology*, vol. 11, no. 11, pp. 889–893, 2015.

[11] O. A. Fatah, "Generating stereoscopic 3D from Holoscopic 3D," in *2013 3DTV Vision Beyond Depth (3DTV-CON)*, 2013, pp. 1–3.

[12] A. Albar and M. Swash, "Portable Holoscopic 3D camera adaptor for Raspberry Pi - IEEE Xplore Document", *Ieeexplore.ieee.org*, 2016. Available: http://ieeexplore.ieee.org/document/7574929/. [Accessed: 19- Nov- 2019].

[13] O. A. Fatah, P. Lanigan, A. Aggoun, and M. R. Swash, "Digital Refocusing: All-in-Focus Image Rendering Based on Holoscopic 3D Camera," *Journal of Computer and Communications*, vol. 04, no. 06, pp. 24–35, May 2016.

[14] J. F. Oliveira, "2D Image Rendering for 3D Holoscopic Content using Disparity-*Assisted Patch Blending*," 2013.

[15] W. Park, K. Ro, S. Kim, and J. Bae, "A Soft Sensor-Based Three-Dimensional (3D) Finger Motion Measurement System," *Sensors*, vol. 17, no. 3, p. 420, 2017.

[16] P. R. Sanz, Mezcua Belén Ruiz, and Pena José M. Sánchez, Depth Estimation - An Introduction. *INTECH Open Access Publisher*, 2012.

[17] A. Bhoi, "Monocular Depth Estimation: A Survey," *ArXiv 2019*, 2019.

[18] F. Dominio, M. Donadeo, G. Marin, P. Zanuttigh, G. M. Cortelazzo "Hand Gesture Recognition with Depth Data," ARTEMIS *'13 Proceedings of the 4th ACM/IEEE international workshop on Analysis and retrieval of tracked events and motion in imagery stream*, pp. 6–19, Oct. 2013.

[19] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields*," IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2024–2039, Jan. 2016.

[20] G. S. Rash, P. Belliappa, M. P. Wachowiak, N. N. Somia, and A. Gupta, "A demonstration of the validity of a 3-D video motion analysis method for measuring finger flexion and extension," *Journal of Biomechanics*, vol. 32, no. 12, pp. 1337–1341, 1999.

[21] H. Hu, X. Gao, J. Li, J. Wang, and H. Liu, "Calibrating human hand for teleoperating the HIT/DLR hand," *IEEE International Conference on Robotics and Automation, 2004. Proceedings*. ICRA 04. 2004, Apr. 2004.

[22] S. Sridhar, A. Oulasvirta, and C. Theobalt, "Interactive Markerless Articulated Hand Motion Tracking Using RGB and Depth Data," *2013 IEEE International Conference on Computer Vision*, 2013.

[23] A. Majkowska, V. Zordan, and P. Faloutsos, "Automatic splicing for hand and body animations," *ACM SIGGRAPH 2006 Sketches on* - SIGGRAPH 06, 2006.

[24] J. C. V. D. Noort, H. G. Kortier, N. V. Beek, D. H. E. J. Veeger, and P. H. Veltink, "Measuring 3D Hand and Finger Kinematics—A Comparison between Inertial Sensing and an Opto-Electronic Marker System," *Plos One*, vol. 11, no. 11, Mar. 2016.

[25] R. Krupicka, Z. Szabo, S. Viteckova, and E. Ruzicka, "Motion capture system for finger movement measurement in Parkinson disease," *Radioengineering*, vol. 23, no. 2, pp. 659–664, 2014.

[26] F. Damasio and S. Musse, "Animating virtual humans using hand postures," *Proceedings. XV Brazilian Symposium on Computer Graphics and Image Processing*, p. 437, 2002.

[27] J. Weissmann and R. Salomon, "Gesture recognition for virtual reality applications using data gloves and neural networks," *Neural Networks, 1999. IJCNN'99. ...*, vol. 3, pp. 2043–2046, 1999.

[28] O. Luzanin and M. Plancak, "Hand gesture recognition using low-budget data glove and cluster-trained probabilistic neural network," *Assembly Automation*, vol. 34, no. 1, pp. 94–105, 2014.

[29] M. V. G. Rao, P. R. Kumar, and A. M. Prasad, "Implementation of real time image processing system with FPGA and DSP," *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, pp. 1–4, 2016.

[30] J. Batlle, Joan and Mart, J and Ridao, Pere and Amat, "A new FPGA/DSP-based parallel architecture for real-time image processing," *Real-Time Imaging, Elsevier*, vol. 8, no. 5, pp. 345--356, 2002.

[31] C. Desmouliers, F. Vallina, S. Aslan, J. Saniie, and E. Oruklu, "Image and video processing platform for field programmable gate arrays using a high-level synthesis*," IET Computers & Digital Techniques*, vol. 6, no. 6, pp. 414–425, Jan. 2012.

[32] L.-F. Liu, H.-M. Ma, and M.-Q. Lu, "A FPGA and Zernike Moments Based Near-Field Laser Imaging Detector Multi-scale Real-Time Target Recognition Algorithm," *2010 Third International Symposium on Information Science and Engineering*, 2010.

[33] H. M. Rummele W, Braun L, "A FPGA based fast runtime reconfigurable real-time multi-object-tracker," in *Circuits and Systems (ISCAS), 2011 IEEE International Symposium*, 2011, pp. 853--856.

[34] M. I. Alali, K. M. Mhaidat, and I. A. Aljarrah, "Implementing image processing algorithms in FPGA hardware," *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, 2013..

[35] N. M. Zaitoun and M. J. Aqel, "Survey on Image Segmentation Techniques," *Procedia Comput. Sci.*, vol. 65, no. Iccmit, pp. 797–806, 2015.

[36] N. Dhanachandra and Y.J Chanu, "A Survey on Image Segmentation Using Clustering Techniques," *EJERS, Eur. J. Eng. Res. Sci.*, vol. 2, no. 1, pp. 51–55, 2017.

[37] Y. Xu and E. C. Uberbacher, "2D image segmentation using minimum spanning trees," *Image and Vision Computing*, vol. 15, no. 1, pp. 47–57, 1997.

[38] R. Sandhu, S. Dambreville, A. Yezzi, and A. Tannenbaum, "Non-rigid 2D-3D pose estimation and 2D image segmentation," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[39] A. J. Cuadros-Vargas, L. C. Gerhardinger, M. D. Castro, J. B. Neto, and L. G. Nonato, "Improving 2D mesh image segmentation with Markovian Random Fields," *2006 19th Brazilian Symposium on Computer Graphics and Image Processing*, 2006.

[40] R. Hemalatha, N. Santhiyakumari, M. Madheswaran, and S. Suresh, "Segmentation of 2D and 3D Images of Carotid Artery on Unified Technology Learning Platform," *Procedia Technology*, vol. 25, pp. 12–19, 2016.

[41] I. Nyström, F. Malmberg, E. Vidholm, E. Bengtsson "Segmentation and Visualization of 3D Medical Images through Haptic Rendering," *Segmentation Vis. 3D Med. Images through Haptic Render.*, pp. 43–48, 2009.

[42] J. Edwards, P. Egbert, and B. Morse, "Live Mesh: An Interactive 3D Image Segmentation Tool," vol. d, no. December, p. 49, 2004.

[43] Tian Shen, Hongsheng Li, Zhen Qian, and Xiaolei Huang, "Active volume models for 3D medical image segmentation," *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. M, pp. 707–714, 2009.

[44] A. Top, G. Hamarneh, and R. Abugharbieh, "Active Learning for Interactive 3D Image Segmentation," *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*, pp. 603–610, 2011.

[45] S. J. F. Guimarães, M. Couprie, A. de Albuquerque Araújo, and N. J. Leite, "Video segmentation based on 2D image analysis," *Pattern Recognit. Lett.*, vol. 24, no. 7, pp. 947–957, 2003.

[46] F. Galasso, R. Cipolla, and B. Schiele, "Video segmentation with superpixels," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7724 LNCS, no. PART 1, pp. 760–774, 2013.

[47] H. Jiang, G. Zhang, H. Wang and H. Bao "Spatio-Temporal Video Segmentation of Static Scenes and Its Applications," *IEEE Trans. Multimed.*, vol. 17, no. 1, pp. 3–15, 2015.

[48] R. Zeliski, *Computer vision: algorithms and applications*. Springer, 2010.

[49] K. Huebner and J. Zhang, "Stable Symmetric Feature Detection and Classification in Panoramic Robot Vision Systems," *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 7–7, 2006.

[50] T. Gevers, S. Voortman, and F. Aldershoff, "Color feature detection and classification by learning," *IEEE International Conference on Image Processing 2005*, 2005.

[51] R. K. Yalla, C. Boles, R. C. Daley, R. K. Fleming "System and method for three-dimensional biometric data feature detection and recognition," US Patent 9,141,844, 2015.

[52] S. E. Mahmoudi, A. Akhondi-Asl, R. Rahmani, S. Faghih-Roohi, V. Taimouri, A. Sabouri, and H. Soltanian-Zadeh, "Web-based interactive 2D/3D medical image processing and visualization software," *Computer Methods and Programs in Biomedicine*, vol. 98, no. 2, pp. 172–182, 2010.

[53] M. Santarelli, V. Positano, and L. Landini, "Real-time multimodal medical image processing: a dynamic volume-rendering application," *IEEE Transactions on Information Technology in Biomedicine*, vol. 1, no. 3, pp. 171–178, 1997.

[54] M. Sameti, R. K. Ward, J. Morgan-Parkes, and B. Palcic, "Image Feature Extraction in the Last Screening Mammograms Prior to Detection of Breast Cancer," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 1, pp. 46–52, 2009.

[55] A. Patel and K. Mehta, "3D Modeling and Rendering of 2D Medical Image," *2012 International Conference on Communication Systems and Network Technologies*, pp. 149–152, 2012.

[56] M. Irani and P. Anandan, "A Unifed Approach to Moving Object Detection in 2D and 3D Scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 6, pp. 577--589, 1998.

[57] Z. Zhang, "Mining surveillance video for independent motion detection," *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pp. 741–744, 2002.

[58] C. Yuan, G. Medioni, J. Kang, and I. Cohen, "Detecting Motion Regions in the Presence of a Strong Parallax from a Moving Camera by Multiview Geometric Constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1627–1641, 2007.

[59] N. Verbeke and N. Vincent, "A PCA-based technique to detect moving objects," *Image Anal.*, pp. 641–650, 2007.

[60] R. Fablet and M. J. Black, "Automatic detection and tracking of human motion with a view-based representation," *Comput. Vison - Eccv 2002, Pt 1*, vol. 2350, pp. 476–491, 2002.

[61] R. Pradipa and M. S. Kavitha, "Hand Gesture Recognition – Analysis of Various Techniques, Methods and Their Algorithms," *2014 Int. Conf. Innov. Eng. Technol.*, vol. 3, no. 3, pp. 2003–2010, 2014.

[62] R. P. K. Poudel. "3d Hand Tracking,", ProQuest Dissertations Publishing, 2014.

[63] S. Park, S. Yu, J. Kim, S. Kim, and S. Lee, "3D hand tracking using Kalman filter in depth space," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, 2012.

[64] C. Manresa, J. Varona, R. Mas, and F. J. Perales, "Hand Tracking and Gesture Recognition for Human-Computer Interaction," *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, vol. 5, no. 3, p. 96, Jan. 2005.

[65] P. Krejov, "Real Time Hand Pose Estimation for Human Computer Interaction," no. September 2015.

[66] P. Ramachandra and N. Shrikhande. Hand gesture recognition by analysis of codons. Sep 9, 2007, Available: http://dx.doi.org/10.1117/12.733193. DOI: 10.1117/12.733193.

[67] M. Kitagawa and B. Windsor, *MoCap for artists: Workflow and techniques for motion capture*. Elsevier/Focal Press: Taylor & Francis, 2008. Available: https://books.google.co.uk/books/about/MoCap_for_Artists.html?id=pJFowfd5EtkC. Accessed: Oct. 30, 2016.

[68] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A Survey on Human Motion Analysis from Depth Data," *Lecture Notes in Computer Science Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pp. 149–187, 2013.

[69] N. A. Ibraheem and R. Z. Khan. "Survey on various gesture recognition technologies and techniques". *International Journal of Computer Applications 50(7)*, 2012. Available: http://search.proquest.com/docview/1032039156. DOI: 10.5120/7786-0883.

[70] P. Premaratne, *Human Computer Interaction Using Hand Gestures*, 2014th ed. Singapore: Springer, 2014.

[71] S. Mitra and T. Acharya. Gesture recognition: A survey. *Tsmcc 37(3)*, pp. 311-324. 2007. Available: http://ieeexplore.ieee.org/document/4154947. DOI: 10.1109/TSMCC.2007.893280.

[72] M. Al-Rajab, "Hand Gesture Recognition for Multimedia Applications.", ProQuest Dissertations Publishing, School of Computing, University of Leeds, 2008.

[73] P. Premaratne, *Chapter2: Human Computer Interaction Using Hand Gestures*, 2014th ed. Singapore: Springer, 2014.

[74] P. K. Sharma and S. Sharma, "Evolution of Hand Gesture Recognition: A Review," *International Journal of Engineering and Computer Science*, vol. 4, no. 1, p. 9963, Jan. 2015.

[75] P. Wadekar, "Gesture recognition," *LinkedIn SlideShare*, 15-May-2013. [On signal]. Available: https://www.slideshare.net/PrachiWadekar/gesture-recognition-21207480. [Accessed: 12-Apr-2017].

[76] F. P. Supplies, "Holga 120-3D Stereo Camera," *Freestyle Photographic Supplies*. [Online]. Available: https://www.freestylephoto.biz/194120-Holga-120-3D-Stereo-Camera. [Accessed: 28-Nov-2019].

[77] T. Deyle, "Low-Cost Depth Cameras (aka Ranging Cameras or RGB-D Cameras) to Emerge in 2010?," *Hizook*, 29-Mar-2010. [Online]. Available: http://www.hizook.com/blog/2010/03/28/low-cost-depth-cameras-aka-ranging-cameras-or-rgb-d-cameras-emerge-2010. [Accessed: 28-Nov-2019].

[78] "FLIR E60 Infrared Compact Thermal Camera," *tequipment.net*. [Online]. Available: https://www.tequipment.net/FLIRE60.html. [Accessed: 28-Nov-2019].

[79] V. Turk, "The VR Controller of the Future Could Be Your Own Hands," *Vice*, 27-Sep-2016. [Online]. Available: https://www.vice.com/en_us/article/bmv5za/the-vr-controller-of-the-future-could-be-your-own-hands. [Accessed: 28-Nov-2019].

[80] R. M. Ltd, "SINGLE CAMERA SHOOTING," *Single Camera Shoot | Top Telly*. [Online]. Available: http://www.toptelly.co.uk/single-camera-unit. [Accessed: 28-Nov-2019].

[81] "Holoscopic 3D Vision," *Holoscopic 3D Vision | Brunel University London*. [Online]. Available: https://www.brunel.ac.uk/research/Projects/Horoscopic-3D-Vision. [Accessed: 28-Nov-2019].

[82] R. M. Swash, "Holoscopic 3D Imaging and Display Technology : Camera / Processing / Display By Mohammad Rafiq Swash A thesis submitted for the degree of Doctor of Philosophy in Electronic & Computer Engineering School of Engineering and Design Brunel University," no. November, 2013.

[83] A. Agooun, O. A. Fatah, J. C. Fernandez, C. Conti, P. Nunes, and L. D. Soares, "Acquisition, processing and coding of 3D holoscopic content for immersive video systems," *2013 3DTV Vision Beyond Depth (3DTV-CON)*, 2013.

[84] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz and P. Hanrahan"Light field photography with a hand-held plenoptic camera*," Stanford University Computer Science Tech Report CSTR 2005-02*, pp. 1–11, 2005.

[85] J. C. Russ, Image Processing Handbook – *5th Edition*. CRC Press, 2006.

[86] *Segmentation - an overview | ScienceDirect Topics*. [Online]. Available: https://www.sciencedirect.com/topics/computer-science/segmentation. [Accessed: 28-Nov-2019].

[87] PatrickFarley, "Computer Vision documentation - Quickstarts, Tutorials, API Reference - Azure Cognitive Services," *Quickstarts, Tutorials, API Reference - Azure Cognitive Services | Microsoft Docs*. [Online]. Available: https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/. [Accessed: 28-Nov-2019].

[88] A. Mishra, "Machine learning in the AWS Cloud add intellegence to appllcatgions with Amazon SageMaker and Amazon Rekognition," *Amazon*, 2019. [Online]. Available: https://aws.amazon.com/rekognition/. [Accessed: 28-Nov-2019].

[89] SimpleCV [Online]. Available: http://simplecv.org/. [Accessed: 28-Nov-2019].

[90] A. C. Bovik, Handbook of image and video processing. Boston, MA: Academic Press, 2005.

[91] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, H. H. Liu "The Empirical Mode Decomposition and Hilbert Spectrum for Nonsignalar and Nonstationary Time Series Analysis. Proceedings of the Royal Society London A., 454:903–995, 1998.

[92] M. Lambert, A. Engroff , M. Dyer, B. Byer "Empirical Mode Decomposition," *Rice University*. [Online]. Available: https://www.clear.rice.edu/elec301/Projects02/empiricalMode. [Accessed: 13-Jun-2017].

[93] E. U. Nathaniel, N. J. George, and S. E. Etuk, "Determination of Instantaneous Frequencies of Low Plasma Waves in the Magnetosheath Using Empirical Mode Decomposition (EMD) and Hilbert Transform (HT)," *Atmospheric and Climate Sciences*, vol. 03, no. 04, pp. 576–580, 2013.

[94] Daubechies, I., "Ten lectures on wavelets. SIAM, Philadelphia," PA, 1992.

[95] R. Szeliski, *Computer vision: algorithms and applications*. London: Springer, 2011.

[96] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Hoboken: Pearson, 1995.

[97] I. Aleksander and J. Taylor, *Artificial neural networks, 2*, Amsterdam, North-Holland, 1992.

[98] B. Fritzke, "Growing Cell Structures – a Self-organizing Network in k-Dimensions," *Artificial Neural Networks*, pp. 1051–1056, 1992.

[99] "What Is Deep Learning?: How It Works, Techniques & Applications," How It Works, Techniques & Applications - MATLAB & Simulink. [Online]. Available: https://www.mathworks.com/discovery/deep-learning.html. [Accessed: 25-Feb-2020].

[100] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. Cambridge, MA: The MIT Press, 2017.

[101] S. Pang, J. J. D. Coz, Z. Yu, O. Luaces, and J. Díez, "Deep learning to frame objects for visual target tracking," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 406–420, 2017.

[102] *Convolutional Neural Network - MATLAB & Simulink*. [Online]. Available: https://uk.mathworks.com/solutions/deep-learning/convolutional-neural-network.html. [Accessed: 28-Nov-2019].

[103] "trainNetwork," *Learn About Convolutional Neural Networks - MATLAB & Simulink - MathWorks United Kingdom*. [Online]. Available: https://uk.mathworks.com/help/deeplearning/ug/introduction-to-convolutional-neural-networks.html. [Accessed: 28-Nov-2019].

[104] C. Junsheng,, "Research on the Intrinsic Mode Function (IMF) Criterion in EMD Method," *Mechanical Systems and Signal Processing*, vol. 20, pp. 817-824, 2016.

[105] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[106] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3D Hand Pose Estimation in Single Depth Images: From Single-View CNN to Multi-View CNNs," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4422–4436, May 2018.

[107] A. Y. Keun and P. Y. Choong, "The Hand Gesture Recognition System Using Depth Camera," *The Tenth International Conference on Advances in Computer-Human Interactions*, 2017.

[108] D. Demirdjian and T. Darrell, "Motion estimation from disparity images," *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*.

[109] J. Pyo, S. Ji, S. You, and T. Kuc, "Depth-based hand gesture recognition using convolutional neural networks," *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, 2016.

[110] E. Alazawi, A. Aggoun, M. Abbod, M. R. Swash, O. A. Fatah, and J. Fernandez, "Scene depth extraction from Holoscopic Imaging technology," *2013 3DTV Vision Beyond Depth (3DTV-CON)*, 2013.

[111] V. Vardhan1 and P. Prasad, "Hand Gesture Recognition Application for Physically Disabled People," International Journal of Science and Research, vol. 3, no. 8, pp. 765–796, Aug. 2014.

[112] N. Soodtoetong and E. Gedkhaw, "The Efficiency of Sign Language Recognition using 3D Convolutional Neural Networks," 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Rai, Thailand, 2018, pp. 70-73.

[113] "About Stroke," *www.stroke.org*. [Online]. Available: https://www.stroke.org/en/about-stroke. [Accessed: 28-Nov-2019].

[114] H.-I. Lin, M.-H. Hsu, and W.-K. Chen, "Human hand gesture recognition using a convolution neural network," *2014 IEEE International Conference on Automation Science and Engineering (CASE)*, 2014.

[115] S. Y. Kim, H. G. Han, J. W. Kim, S. Lee, and T. W. Kim, "A Hand Gesture Recognition Sensor Using Reflected Impulses," *IEEE Sensors Journal*, vol. 17, no. 10, pp. 2975–2976, 2017.

[116] H.-J. Kim, J. S. Lee, and J.-H. Park, "Dynamic hand gesture recognition using a CNN model with 3D receptive fields," *2008 International Conference on Neural Networks and Signal Processing*, 2008.

[117] G. A. Rao, K. Syamala, P. V. V. Kishore, and A. S. C. S. Sastry, "Deep convolutional neural networks for sign language recognition," *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*, 2018.

[118] W.-J. Li, C.-Y. Hsieh, L.-F. Lin, and W.-C. Chu, "Hand gesture recognition for post-stroke rehabilitation using leap motion," *2017 International Conference on Applied System Innovation (ICASI)*, 2017.

[119] H.-Y. Chung, Y.-L. Chung, and W.-F. Tsai, "An Efficient Hand Gesture Recognition System Based on Deep CNN," *2019 IEEE International Conference on Industrial Technology (ICIT)*, 2019.

[120] K. S. Varun, I. Puneeth, and T. P. Jacob, "Hand Gesture Recognition and Implementation for Disables using CNN'S," *2019 International Conference on Communication and Signal Processing (ICCSP)*, 2019.

[121] A. A. Alani, G. Cosma, A. Taherkhani and T. M. McGinnity, "Hand gesture recognition using an adapted convolutional neural network with data augmentation," 2018 4th International Conference on Information Management (ICIM), Oxford, pp. 5-12, 2018. doi: 10.1109/INFOMAN.2018.8392660

[122] "SHC bundle," *pressport*. [Online]. Available: https://www.pressport.com/uk/news/files/shc-bundle-22665. [Accessed: 28-Nov-2019].

Appendix A

Seven hand motions for the remaining seventeen subjects

1	2	3	4	5	6	7
Drink	**Eat**	**Good\ Bravo**	**Stop**	**That**	**Close**	**Family**