Project Id: 2018CSEPID09

Project Report

on

HAND GESTURE RECOGNITION USING KERAS

Submitted in Partial Fulfillment of the Requirement

For the Degree of

Bachelor of Technology

In

Computer Science and Engineering

By

Moin Malik 1829010099
Taufeek Mansoori 1829010097

Under the Supervision

of

Mr Krishna Bihari Dubey
Assistant Professor

Department of Computer Science & Engineering
ABES INSTITUTE OF TECHNOLOGY, GHAZIABAD

AFFILIATED TO

Dr A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, UTTAR PRADESH, LUCKNOW

(May, 2022)

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature:
Name : Moin Malik
Roll No. :
1829010099
Date :

Signature:
Name : Taufeek Mansoori
Roll No. : 1829010097
Date :

CERTIFICATE

This is to certify that Project Report entitled "Hand Gesture Recognition using Keras" which is submitted by **Moin Malik** (1829010099), **Taufeek Mansoori** (1829010097) in partial fulfilment of the requirement for the award of degree B.Tech. in Department of Computer Science & Engineering of A.K.T.U is a record of candidate's own work carried out by him under my supervision.

Date:

Supervisor
Mr. Krishna Bihari Dubey
Department of Computer Science
& Engineering

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 been conducted on hand gestures. Different techniques have been proposed to implement hand gesture experiments. Multiple tools exist to extract features from images, as well as Artificial Intelligence, which has different classifiers to classify different data types. To extract images and classify various mini gestures and movements, 2D and 3D hand gestures require an effective algorithm. A variety of algorithms are discussed in this research. In order to extract image features, this study proposed using wavelet transforms and empirical mode decomposition to detect 2D or 3D hand gestures. Execution time, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood, negative likelihood, receiver operating characteristic, area under the ROC curve, and root mean square. Four original contributions to the field of hand gestures are discussed here. In the first contribution, we present two experiments using 2D hand gesture videos where ten different gestures are detected in small and longer distances using an iPhone 6 Plus with 4K resolution. ANN and CNN are used for classification, while WT and EMD are used for feature extraction.

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

Chapter 1
Introduction

1.1 Foreword

A Gesture is characterized as the actual development of the hands, fingers, arms and different pieces of the human body through which the human can pass on significance and data for collaboration with one another [1]. There are two distinct methodologies for human-PC connection: the glove-based method and the vision-based method. In the accompanying analyses, the vision-based approach was used to recognize and group hand signals. A hand signal is one of the most consistent ways to create an effective and highly versatile interface between gadgets and clients. In HCI frameworks, applications, for example, virtual item control, gaming, and motion recognition are conceivable. Hand following, as a hypothesis perspective, manages three basic components of PC vision: hand division, hand part discovery, and hand following. The best open strategy and the normal idea utilized in a motion acknowledgment framework is hand signals. Hand signals can be distinguished by one of these following methods: pose is a static hand shape proportion without hand developments, or a signal is dynamic hand movement regardless of hand developments. Utilizing any kind of camera will recognize any sort of hand signal; remembering that various cameras will yield different goal characteristics. Two-layered cameras can distinguish most finger movements in a steady surface called 2D [2].

As of late, gesture based communication might be accomplished by certain kinds of mechanical technology utilizing a few fitting sensors utilized on the body of a patient [3]. Another model is stroke recovery. Individuals who have encountered stroke can have paraplegia which forestalls them moving their lower appendages. Stroke restoration can assume a huge part to settle this kind of issue. Moreover, certain individuals who have stroke can't discuss sufficiently with others. Scientists introduced various examinations close by signals, for example, object recognition and item movements. Gaming takes a distinct fascination with the area of Three-Dimensional (3D) hand following. At the beginning of the 2010s, ongoing film discharges, for example, Avatar, reformed film by consolidating content creation and 3D innovation with genuine entertainers, prompting the making of the new sort [4]. After the

outcome of 3D film, the different electronic organizations zeroed in on creation of Three-Dimensional Television (3DTV) innovation. The scientists proposed the vault auto stereoscopic showcase that is utilized to see the place that is as yet restricted [4]. The two unique advances like sound system and multi-view depend on the mind to combine the two pictures to make the impact of 3D [4].

1.2 Research Aim and Objectives

The point of the examination is to foster a framework for 2D, and 3D hand signal acknowledgment utilizing any sort of camera, foundation, enlightenments or position of hand, by tracking down the most fitting calculations to execute the framework and test the approval of framework. This framework assists people with extraordinary requirements and individuals who have encountered stroke to precisely convey. Involving WT and EMD calculations for include extraction and AI for order gives various outcomes while CNN gives a precise outcome.

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 Thesis Outline and Chapters' Summary

The proposition covers eight primary parts; the presentation of the postulation, the examination writing survey for the new investigations, the presentation of signal acknowledgment, picture handling and acknowledgment hypothesis, 2D video motion acknowledgment, 3D video motion acknowledgment, stroke patients' motion acknowledgment, and end and future work.

Chapter 1 - Introduction

Chapter one is an underlying section, which presents the foundation of the examination, the point and targets, the exploration unique commitments, and the proposal diagrams and sections' rundown.

Chapter 2 - Literature Survey

Chapter two defines the literature survey 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. finally, it also discusses the hand tracking studies. These areas particularly, elaborate the state-of-the-art works on hand motion recognition.

Chapter 3 - Gesture Recognition

Chapter three presents the HCI foundation and the historical backdrop of signal acknowledgment. It makes sense of momentarily the principal kinds of motion acknowledgment. The hand signal acknowledgment and its sorts arediscussed in this part. It additionally proposes the sorts of cameras utilized for 2D and 3D pictures. The holoscopic 3D imaging framework camera procedure is likewise made sense of in section three.

Chapter 4 - Image Processing and Recognition

Chapter four incorporates a concise hypothesis of picture handling, video handling, PC vision, observational mode disintegration, wavelet changes, man-made brainpower and convolutional brain organization, and examines the usefulness of every procedure and its purposes.

Chapter 5 - 3D Video Gesture Recognition

Chapter five presents a framework produced for 3D hand signal acknowledgment utilizing CNN. Twelve 2D and 3D signal pictures, in small and significant distances, for three subjects utilized

in this trial work. The exploratory works performed to look at the execution of preparing and testing utilizing various variables.

Chapter 6 - Conclusion and Future Work

Chapter six provides an final approach of all research works as proposed in this thesis.

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

1.4 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.

Teacher Gabriel M. Lippmann proposed the utilization of microlenses cluster at the picture surface [4]. He introduced this idea to the French Academy of Sciences as La Photography Integral [4] [5]. The spatial picture with full parallax every which way recorded totally by Professor Lippmann and is known as a fly's-eye focal point cluster. Basically, the showcase framework was a screen containing various little focal points [4][5]. During the 1920s, a few researchers, like Herbert Ives, began to contemplate working on Lippmann's idea by joining a lenticular focal point sheet, which contains a signaller exhibit of circular focal points known as lenticules. It was intended to see various points of pictures to give a pixel from each miniature picture [4][5]. The focal point sheet is straightforward, and the back face makes the central plane which is level. One model is the focal points utilized in lenticular creation where the innovation is utilized to give a deception of profundity, i.e; making pictures which seem to move or change as the picture is seen from various points. This inventive innovation may likewise be utilized for creating 3D pictures on a level sheet show. Thus, on the off chance that the movement of the photos is thought of, this would bring about a 3D Holoscopic video [4][5] [6].

Recently, Film industry discharges, for example, Avatar have changed the film by joining 3D innovation and part creation with entertainers which lead to the development of another classification. The achievement of 3D film constrained some significant purchaser gadgets ventures to send off 3D TVs and telecasters to introduce 3D substance. Newsday's 3D TV innovation utilizes a sound system vision situation introducing left and right eye pictures by means of spatial and fleeting multiplexing, to watchers wearing a couple of glasses. The following stage in 3D TV frameworks will most likely be a multi-view autostereoscopic imaging framework. This framework will record and present numerous video motions on a showcase. This won't urge watchers to wear glasses [4][5] [6].

on energized stereoscopic framework. Holoscopic 3D camera connectors are normally intended for huge scope Single-focal point Reflex (SLR) cameras. Albar and Swash [12] proposed a method for prototyping a holoscopic 3D camera connector for a Visa estimated board PC, called Raspberry Pi. The two models are moderately modest to collect. This imaginative holoscopic 3D connector configuration can be utilized in an assortment of utilizations, for example, reconnaissance, clinical, diversion and furthermore in hardware where 3D profundity detecting and estimation are the principal concerns. The rule behind this development was that one can keep various fundamental photos of a 3D scene in a 2D organization sensor with every normal picture taking care of a substitute perspective of the 3D scene. This is achieved by coordinating a Microlens Array (MLA) before a sensor or use a camera group to get the 3D item according to alternate points of view.

Fatah et al. [13] designated the computerized pulling together where they proposed a strategy. This technique utilizes Michelson contrast recipe to extricate all-in-center pictures. The most elevated contrast values at various places in space can return the engaged places where the articles are at first situated, which make it conceivable to acquire all-in-center picture. . It requires another framework which successfully isolates the same arranged pixels under every Exposure Index. Oliveira [14] introduced in their proposition a 2D picture extraction procedure for holoscopic 3D pictures. They named it Disparity Assisted Patch Blending which beats existing techniques. The second commitment in their examination work is the ID of potential non-reference picture quality evaluation measurements. These measurements can gauge 2D picture extraction to extractions to contrast and human insight.

A 3D finger movement estimation framework which depends on delicate sensors are proposed by Park et al. [15]. These sensors are made from Ecoflex (delicate material), having inserted miniature channels loaded up with conductive fluid metal Eutectic gallium-indium (EGaln). These sensors have the capacity to implant in such conditions where other conventional sensors can't be implanted. Joint of thumbs are demonstrated to indicate the area of the sensors. A calculation is proposed to decouple the signs and concentrate the movements, for example, flexion, expansion, kidnapping and so on. They contrast their method and the camera-based movement catch framework.

2.2 Image Depth

Profundity assessment or extraction concerns the gathering of innovation and calculations to address the spatial design of a scene, all in all, to compute the profundity of each point in a scene precisely [16]. The premise of profundity is portrayed by two strategies, dynamic and latent. Dynamic strategies work by discharging energy into a scene before latently handling the energy reflected. These strategies were proposed before inactive techniques as miniature handling had not yet been developed [16]. The primary disservice introduced by dynamic strategies is the energy expected to work. By and by, it viewed their unwavering quality as a lot higher, and some of them are utilized to get ground-based proof. Light-based computation of profundity for distance estimation is the principal kind of energy. An illustration of this should be visible in explores different avenues regarding glowing light. In any case, many light sources can be utilized and thusly a wide range of calculations, setups and equipment are accessible too. Ultrasound based strategies use the Time of Flight (ToF) standard to quantify distances. A genuine illustration of this method is while imaging an embryo in a mother's belly [16]. The primary concern of this suggestion is the high precision and handling rates (up to 100 fps) because of CMOS and LED based light.

Aloof Methods work involving regular light in the climate where the optical information of the caught picture can be utilized to assess profundity. Such procedures catch pictures with picture sensors, which fixes the PC issue. The upside of this technique is the low measure of tasks expected to handle a solitary picture, rather than at least two. There are two previous classes in this calculation family, monocular and Multiview arrangements [17].Monocular profundity assessment is the undertaking of assessing a profundity from a RGB picture [17]. To get the profundity map, a solitary picture or a video succession might be utilized in this methodology. The benefit of utilizing this approach is that few activities is expected to handle each picture [17]. Profundity on-defocus is a methodology that utilizes monocular data to give an outright estimation of distance in light of the center properties of the picture. This approach appraises the distance of each point in a picture following the human visual centering framework by ascertaining the defocusing level of such places. This defocusing estimation is done principally with Laplacian administrators, which measure the second spatial subsidiary toward every path for each point in a N pixel neighborhood [16].

Multiview profundity assessment approaches have various calculations managing at least two pictures to ascertain the profundity map. Sound system vision is an instance of this exhibit,

utilizing two items. It can utilize sound system vision and multiview for multiple pictures for explanation purposes when two pictures are involved. Outright estimations in certain circumstances might 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." Moreover, the scene is constantly found in a specific region. The obscured picture seen at this stage is the scene's alleged projection. This projection comprises of a progression of beams that cross a little opening to the purported projection plane [16]. A portion of the profundity assessment applications incorporate smoothing obscured picture parts, further developed 3D view delivering, self-driving vehicles, mechanical getting a handle on, robot-helped a medical procedure, programmed 2D-to-3D film change and 3D PC illustrations shadow planning. There are different examinations about profundity assessments. A hand signal acknowledgment method in light of profundity information is proposed by Dominio et al [18]. To identify a more complete 3D hand motion, profundity data was utilized to quantify the stance by separating the hand profundity map into its different parts, changing the AI approach for the full body that was performed utilizing a Kinect. Right away, the hand is extricated from profundity maps procured, alongerside variety data from related sees. The following stage is division of the palm and finger area from the hand. Two element descriptors are removed, the first in view of distances of the fingertips from the hand place and the second one on the hand form arch. To perceive the motions, a multiclass Support Vector Machine (SVM) classifier is utilized. A high exactness is accomplished on profundity information.

As indicated by Liu [19], a profound CNN model is utilized to handle profundity assessment from single monocular picture issues. It likewise expects to investigate the limit of profound CNN and ceaseless Conditional Random Field (CRF). The proposed conspire learns the unary and pairwise possibilities of persistent CRF. Besides, a model in view of completely convolutional networks and a clever super-pixel pooling strategy is proposed which is multiple times quicker. This proficient model is a superior performing CNN plan. Probes indoor and outside scene information shows that the proposed technique beats the cutting edge profundity assessment approach.

2.3 Finger Movement Measurement

Rash et al. [20] were among the first to investigate 3D hand motions. They performed 3D video

movement examination to gauge hand movements. The objective was to represent the legitimacy of this procedure by standing out it from a two-layered development, thought about the 'greatest level'. Their examination was performed to choose if (1) markers put on the dorsal piece of the hand and fingers unequivocally measure joint edges, and (2) the three dimensional technique for assessing finger developments is precise by using a standard development examination system.
Shade et al. [21] then followed fingertip positions involving two sound system cameras to limit the blunder among planned and estimated fingertip positions, they compelled an Inverse Kinematics (IK) solver. The client wears the information glove and moves fingers straightforwardly while the wrist is fixed on the table. The vision system records a movement of finger developments and measures the particular fingertip spots of each finger. At the same time, the uncalibrated rough Image Classification

Damasio and Musse. [26] proposed a framework utilizing information glove and a fake brain network framework, for perceiving hand stances. This framework uses remarkably planned gloves with adaptable sensors to secure and communicate information to a PC consequently empowering cooperation among genuine and virtual people. Salomon and Weissmann [27] explored the planning precise estimations got from gloves (with sensors) to predefined hand signals. This is one more instance of utilizing an order approach as opposed to movement recreation. They additionally utilized brain network classifier, utilizing preparing sets involved 200 hand presents. They made a correlation between the engendering brain organization and the spiral premise practical brain organization. They presumed that just prepared back engendering brain network groups the stances better than outspread premise work brain organizations.

Plancak and Luzanin et al. [28] in their analysis, utilized 5DT information glove 5 ultra, which is modest glove. Probabilistic brain network was prepared on the information to arrange tokens of completely open and shut hands. The preparation dataset size is decreased, utilizing some famous grouping calculations, considering quicker execution time with slight misfortune in preparing quality. Assumption Maximization (EM) grouping calculation was chosen to address the center of the bunching gathering.

2.4 Image Processing

To remove some valuable data or upgrading the picture, a few tasks are performed on pictures; the strategies for playing out these procedure on pictures are called picture handling. Very much

like sign handling, the info is some picture, while the result might be some upgraded picture, or some trait connected with the information picture. Two kinds of methods are utilized for the two sorts of picture. Simple picture handling targets simple pictures like photos, while computerized picture handling centers around advanced pictures. Since the focal point of this record is advanced picture handling thus, we will limit this conversation to computerized picture handling.

Computerized picture handling can be characterized as handling a variety of genuine numbers introduced by various pieces. The field of picture handling assumes a crucial part and contributes towards the arrangement of numerous issues including security, remote detecting applications, businesses, medication, and so on [29]. There exist a few stages engaged with handling a computerized picture going from picture pre-handling, picture division, picture include extraction and picture order. The accompanying conversation remembers methods and applications for the area of 2D-3D picture and video handling. An exploration paper connected with these procedures is introduced for every strategy and application. The initial segment comprises of procedures, for example, Field-present a technique that utilizes symmetric properties from the visual information to recognize extra and stable picture highlights. A subjective balance administrator with quantitative evenness range data is utilized to frame the local highlights. This strategy effectively demonstrated that scale from-balance is effectively relevant as a measured inset for include location.

Extraction and order of nearby picture structure is talked about by Gevers et al. [50] in their examination paper. For the a large portion of picture handling and PC vision errands, for example, object acknowledgment, sound system vision and 3D remaking, extraction and grouping of nearby picture structure are vital. Utilizing the mathematical and photometric data, they proposed a strategy which characterized the actual idea of nearby picture structure. This methodology included different kinds of picture taking care of frameworks which are utilized to normalize pictures.

The Gaussian and MoG were viewed as the most suitable for highlight revelation in light of the effectiveness and precision of the outcomes.

Yalla et al. [51] fostered a framework that incorporates a 3D element discovery module and a 3D acknowledgment module. They utilized a biometric object, where the 3D component recognition

module processes the 3D surface guide of that item and decides if there is any sort of 3D element on the 3D surface guide. Assuming there exists any 3D component, they extricated it alongerside its sort. The subsequent stage is that 3D acknowledgment model coordinates the 3D element with the biometric dataset to distinguish the individual.

2.5 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.

2.5.1 Medical Image Applications

With progression in clinical pictures, for example, Magnetic Resonance Imaging (MRI), Computed Tomography (CT) output and ultrasound advances, the need to handle the picture information to extricate important data increments. This need draws in specialists from many fields like measurements, applied math, science, physical science, designing, software engineering and medication. In this part, it examines some the exploration work in the space of clinical picture handling. Mahmoudi et al. [52] introduced a detail survey of various online intuitive programming devices for 2D/3D clinical picture handling. In a previous work [53] proposed a picture handling and perception calculation that works intelligently for diagnosing moving organs like the heart during the cardiovascular cycle. The framework is tried on Magnetic Resonance (MR) Zhang [57] presents an information digging approach for movement recognition in immense reconnaissance video data sets gathered by military observation cameras. They follow totally subjective methodology, in light of signaller framework consistency examination, called QLS. This approach centers around what is important to process the arrangement subsequently diminishing the computational expense and expanding the productivity. Yaun et al. [58] proposed a method for recognition movement districts in video successions. This method orders picture pixels into movement districts by applying 2D planar homographs, famous and mathematical consistency limitations. Their primary commitment is mathematical consistency requirements got from the camera presents from three progressive casings. It is carried out inside the Plan + Parallax system.

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.

2.6 Hand Tracking

People speak with one another through voice correspondence or communicating in specific dialects. Adjacent to voice correspondence, hands are one more method for correspondence between hearing weakened individuals. Hand motion acknowledgment is a fascinating point among specialists from various fields, for example, PC vision, human PC communication and picture handling. Scientists from the software engineering spaces, for example, motion acknowledgment, virtual item control and gaming take a distinct fascination with the field of 3D hand following. Kavitha [61] introduced an overview paper on the method utilized in signal acknowledgment. As indicated by them, the fundamental point of signal acknowledgment is to foster a framework which can distinguish human activities and use them to remove significant data for gadget control. On account of human-PC communications, hand motions can assume a key part where individuals with verbal incapacities can take full benefits of PC frameworks. This will likewise help in diminishing the utilization of equipment gadgets engaged with working the PC framework and subsequently lead to less ozone harming substance producing. fingertips with utilizing sensor is drawn closer. The extrema of the hand are found utilizing diagram approach. After that the tokens of the hand are distinguished. The subsequent commitment is to distinguish the posture of the hand and district of hand; Random Decision Forest (RDF) is utilized to the component that example profundity. In the last methodology, AI and model-based approaches are mutually used to defeat the disadvantages of both when utilized in separation.

2.7 Summary

This part presents extensive writing audit of 2D/3D picture and video-related procedures and applications. Beginning with the assessment in 3D cameras research papers connected with our subject of interest were examined. In this part, the academic business related to assessment of the

holoscopic 3D camera is additionally introduced. Research papers connected with picture profundity are depicted in the following segment. The third area examines the figure development estimation procedures.

After finger development estimation procedures, the utilization of order calculation for signal location were examined, trailed by picture handling strategies and applications for 2D/3D recordings and pictures. This introduced a definite examine of picture handling strategy like FPGA, division, highlight extraction, and picture handling application like clinical and movement identification.

Writing business related to hand following calculations and procedures were introduced in the last area. Existing and cutting edge calculations were introduced in this part, remembering the flow strategies proposed by different scientists for the field of hand following. The following section characterizes the historical backdrop of signal acknowledgment including its sorts. It portrays the different hand signals and the kinds of cameras used to distinguish various motions.

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].

There are several types of **human-computer interactions (HCIs),** one of **which** is called **a gesture. A** simple definition of gesture is **the** nonverbal **communication** method **used in the HCI interface. The ultimate goal** of **gestures** is to **consciously recognize human gestures and** design specific **systems** that can use **those** gestures to convey information for device control.

Recently, HCI has **become more relevant as it has become increasingly used** in **a variety of applications,** including human motion **detection. She first needs to** define the idea of **HumanMotionAcquisition,** which **captures** the movements of **humans** or **objects** and **sends** them as 2D or 3D image data. **In order** to **realize a digital** 3D **model** or **analyze motion, it is necessary** to **thoroughly examine** the converted 3D **data. Creating** 3D digital **objects requires** special applications and **special** tools **that** are considered **specific** to **a particular enterprise** [62] [66]. **Disney, for example,** is

one of **the** most **well-known** companies **using** human motion capture as **the latest** technology in the **production of animated films.** 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 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.

3.2 Definition of Gesture Recognition

In this day and age, HCI takes great importance in our daily lives. Touch detection can be termed as a long way [69] [70] [71]. So, what is touch recognition? In the previous section Action Visibility was defined as non-verbal movement used as a means of communication in HCI interfaces [69] [70] [71]. Touch is one of the most important aspects of HCI in both human and device communication [69] [70] [71]. Another definition of touching the body movement or posture of a person's fingers, hands, arms or the whole body used to convert information. Touch, in a non-realistic system can be used to navigate, control or communicate with a computer [69] [70] [71]. The process by which touch is constructed in a certain way by a person, known in the system, is the main goal of touch recognition. Symptoms can be manifested in a number of ways through body language, for example, sign language used by deaf people. Some examples of gestures developed outside the computer screen can be seen using traffic police, construction workers, and airport controllers. Touch may be stationary, meaning that the user gets a stop, or a rotation where the movement is a signal of its own [69] [70] [71]. Attached devices such as gloves, data suits, Six Pens, 2D keyboards, mice and fixed image interaction are generally unsuitable to operate on visual systems unlike devices used to hear any part of body shape, facial expressions, sound and speech, skin response and other human behavior or regions. used to introduce communication between humans and the environment [69] [70] [71]. Touch may stop or change or both in some 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 alongerside 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].

3.3 Types of Gesture Recognition

Touch perceptions have been briefly presented in previous sections [69] [70] [71]. Symbols are created by the user and recognized by the recipient. Logical gestures include movements of the fingers, hands, arms, head, face, or body to convey meaningful information. At this stage, some important types of touch recognition are considered briefly.

Hand gesture recognition is one of the most understandable ways to produce simple, flexible communication between devices and users [69] [70] [71]. Using a series of finger and hand movements with the use of sophisticated equipment allows for the recognition of touch. This process will eliminate the need for physical interaction between the user and the device [69]. The next section introduces the most touching facial expressions.

Facial touch detection is an additional way to produce a visual connection that does not communicate effectively between users and equipment. The primary objective of machine facial recognition is to detect emotions and other signs of communication between people, despite the countless visual differences between users [69] [70] [71].

The goal of facial recognition is similar to facial recognition that identifies and identifies a person's face successfully despite their measurements, posture, posture, position and brightness [69] [70] [71]. Low face data bandwidth transfers, crime detection, lost children detection, surveillance, credit card verification, office security, telecommunications, video recording, High-Definition Television (HDTV), personal computer communications, medication and multimedia face inquiries are examples requiring an automated system of facial touch recognition [69] [70] [71].

There are other methods used for facial recognition including Wavelet Transform, practical accounting and information models or legal-based techniques such as facial code coding system. The second approach is all-encompassing and involves the comparison of a gray scale model using global recognition. To represent the entire facial image requires the use of the vector element [69] [70] [71]. Signal pathologists, ANNs, PCAs, optical flow, single volume and decay using eigenfaces are included in the complete path [69] [70] [71].

Many factors need to be understood such as general muscle tension, arm tension, student stretch and contact areas [69] [70] [71]. To define all these principles, the shape of the human body, the setting of angles, rotation, and speed-like movements need to be determined. All features can be completed with sensor devices attached to the user. Sensor devices may be magnetic trackers, data gloves or body suits. If not, the use of computer vision and camera techniques may also be called hearing aids. The sensory technician varies in dimensions, some including accuracy, size, delay, adjustment, user comfort, cost and range of motion. The user is required to wear the device and carry the wires connecting the device to the computer using glove-based connections [69] [70] [71].

They may be of the following types: hand and arm touch to detect hand position; sign language and entertainment applications, such as children who are allowed to play and participate in the real world; head and facial expressions, such as shaking or shaking the head, opening the mouth to speak or the look of happiness, anger and fear etc; and lastly, physical gesture is a complete body movement such as understanding dance movements to create similarities between music and images and tracking the movements of two people sharing outside and other [69] [70] [71].

3.4 Overview of Hand Gesture Recognition

The hand is often best known as the natural and natural phenomenon of human interaction. In the world of HCI, proper hand tracking is the first step in developing natural HCI systems that can be used in programs such as, material manipulation, games and touch detection. In addition, hand tracking is an interesting systematic point that deals with the three main components of computer vision namely hand splitting, hand detection, and hand tracking. Touch gestures are often the most obvious and widely used method of touch sensitivity, which involves standing with a standing finger without a hand movement and a touch with a flexible hand movement or without finger movement [69] [70] [71].

Touch of the hand requires a 27-degree follow-up of the hand which includes two main stages, the position of the hand standing upright without any movement; While hand movements are hand movements, either full hand or fingers. Hand gestures consist of three main types of data-based gloves, based on the visual acuity and sensation of the electric field. Measuring a person's body or limbs requires an electric field sensor, and this device is officially used to measure the distance a person's hand or other body part from a device. Currently, most of the key types almost all researchers are interested in studying, data-based technologies and data-based theory. A data-based glove is a glove with a variety of sensors used to detect hand and finger movements. There are many styles of data gloves and each has its own uses, such as MIT Data Glove, CyberGlove III, CyberGlove II, Fifth Dimension Sensor Glove Ultra, X-IST Data Glove and P5 Glove. Every of these types will be presented in more detail in Section 3.5 [70].

As always, there is no perfection in the physical world. It means that there are serious problems that researchers face such as, self-closing, hand rotation, abnormal movement and visual similarity that makes 3D hand tracking a challenging task [62]. The proposed 3D hand tracking method in this thesis can be used to extract precise hand-to-hand features and to enable complex human machine interactions such as to play and manipulate the visual object [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.

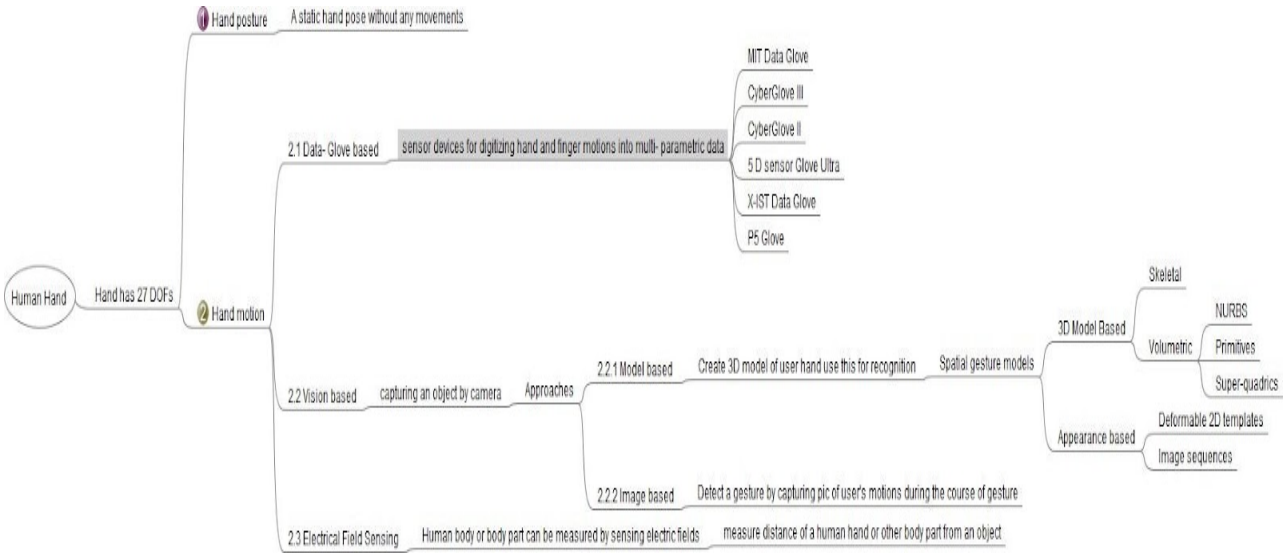


Figure 3.1: Hand Gesture Recognition Map

3.5.1 Data Glove

The history of hand-touch recognition began with the introduction of data gloves. Some researchers understand that sign language promotes touch and that it can be used to propose simple computer commands [70] [71]. Data glove is a special corded glove with touchable switches or sensors attached to the fingers or joints of the glove and worn by a person. Visual goniometers and touch switches or resistance sensors measure the distortion of different joints when present with basic measurements and that determines whether the hand is open or closed or other joints are straight or twisted. The computer is provided with the results drawn from different sources and translated. The advantage of a simple device was no need for any kind of pre-processing. With limited computer processing power back in the 1990s, these types of systems showed promise by ignoring the control limit due to the cables used to connect a data glove to a computer [70].

Gloves are a variety of types, named from 1977 until later, each glove has specific capabilities and functions. One of the first improved gloves was Sayre Glove, built in 1977 [70]. It used light-emitting tubes at the end and a photocell on the other side mounted near each finger of the glove. In 1983, another glove that uses multiple sensors was called the Digital Data Entry glove and developed by Gary Grimes. This glove uses a variety of sensors that carry something. The first commercially available data glove, introduced in 1987, was

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].

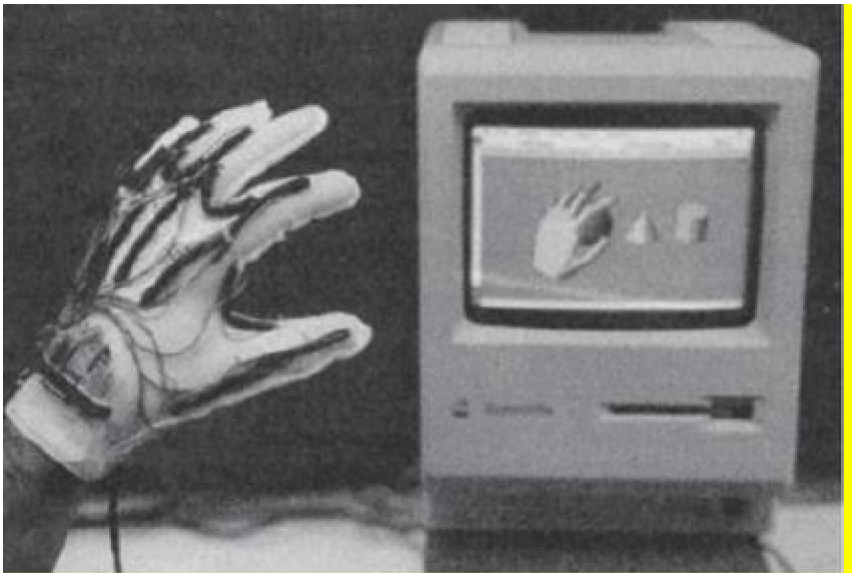


Figure 3.2: The ZTM Glove [73][74].

3.5.1.1 MIT Data Glove

MIT Data Glove has been an amazing development that introduces a variety of capabilities compared to different models. The glove was made by MIT spinoff company AnthroTronix. AcceleGlove as shown below in Figure 3.3 [73] [74], is a flexible glove that records hand and finger movements in 3D. This AcceleGlove is used in video games for sports training, or bodybuilding [70]. The accelerometer sits under each finger and behind the hand as shown in Figure 3.3 [73]. Accelerometers may receive 3D finger position and a respectful palm for the importance that any movement can be made by hand or fingers. The accuracy of these measurements is within a few degrees to allow the systems to detect small changes in the manual area. The glove will 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

CyberGlove III is also known as the MoCap Glove developed by CyberGlove Systems. The purpose of this device is to record the precise movement of the motion pictures used in the film and the animation industry, as shown in Figure 3.4 [73] [74]. Also, the glove contains Wi-Fi used for data communication with a transmission distance of 30 m. One unit has 22 sensors and can operate for two to three hours with a rechargeable battery. Similarly, a Secure-Digital memory (SD) memory card may provide action recording options for recording motion. However, the glove is not intended for computer or other peripheral controls [71].

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

3.5.1.3 Cyber Glove II

CyberGlove was developed to handle data inputs due to the different flexibility of the hand members. As shown in Figure 3.5 [73] [74], the glove has eighteen sensors that include two twin sensors on each finger, four imaging sensors, and nerves used to measure the sixth extremity, palm arch, hand extension, and wrist scan. Another version of this device includes 22 sensors and has three rowing sensors per finger, four scanning sensors, a palm arch sensor, and sensors used to measure wrist flexion and grip. One version of the glove raises open fingers that allows the user to write and type and hold simple objects. The CyberGlove motion capture system has been continuously used in many applications such as virtual reality, biomechanics, animation and

digital model testing [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 touch recognition device based on gloves with highly flexible adjustable gloves. The glove contains sensor lists that provide 10-bit flexor adjustment that is expected to provide capturing natural motion of the film industry. The glove is best known for making high-quality data with low cross-linking between different sensors for real-time animation using Bluetooth. Figure 3.6 [73] [74] shows the original and current version of the 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. Every 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 upgraded the P5 Glove to offer the cheapest option on the market and can be used for gaming. As shown in Figure 3.8 [73] [74], P5 Glove incorporates a twist sensor and remote tracking technology that provides users with natural interaction with virtual reality and 3D applications such as educational software, games and websites. It is one of the rarest technologies currently available to the user as a peripheral controller rather than a mouse, keyboard or toy. Some gloves are used for computer communication in games and communications. Some gloves are used for 3D movie animation and others are used for health care applications such as active symptom control, rehabilitated hands or treatment [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].

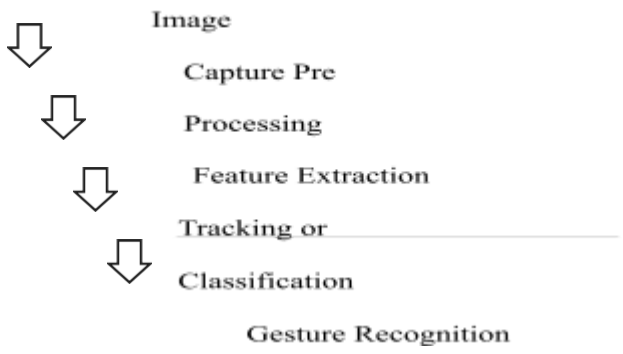


Figure 3.9: Typical computer vision-based gesture recognition approach

Using vision-based strategies requires addressing other issues related to users' genital mutilation. Although tracking devices had the ability to detect hand movements quickly while the human body was moving. Vision-based devices are able to capture features such as color and texture for touch analysis, while conventional tracking devices may not be able to handle this [4].

Vision-based techniques may also vary between the number of cameras used, the speed and the delay, the geographical location, such as speed and brightness, user requirements - each user must wear something different - the minimum features used, such as region, edges, histogram, silhouette and moment; and even if 2D or 3D presentation is not used and at any time is represented [4].

3.6 Types of Cameras

Currently, touch is available on a variety of devices while cameras become the first device to receive multiple touches. This section will introduce most of the current cameras used in the world of touch detection. The type of cameras used for touch detection, with a brief description of each type, is shown in Figure 3.4..

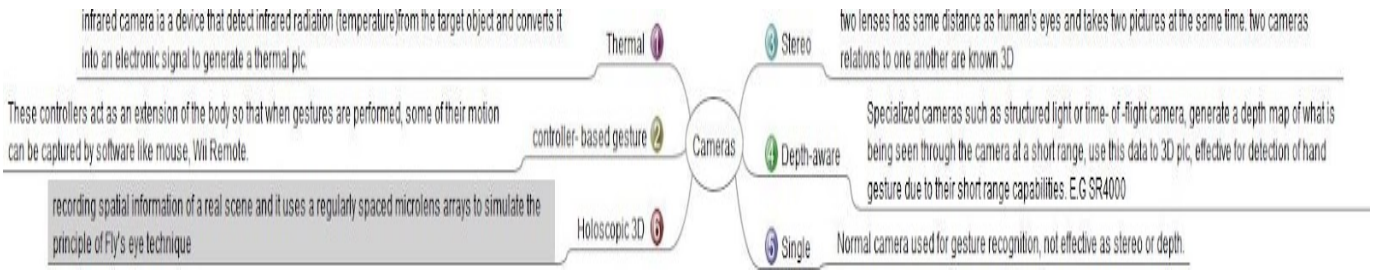


Figure 3.10: Typesof Cameras used in gesture recognition

As shown in Figure 3.6 [76], a stereo camera is a dual-lens camera with the same distance that separates them, like human eyes. It takes two pictures at once. This copies the way we humans use to see and therefore produces a 3D effect when viewed. By using two cameras that monitor their interaction with each other, 3D representation can be approached by camera output [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 small range. This data used to estimate a 3D representation of what is being viewed. These cameras may detect hand gestures effectively because of their small-range skills [75].

Figure 3.12: Depth- aware camera [77].

A hot camera is an infrared camera that receives infrared radiation similar to the temperature of an object as shown in Figure 3.13 [78]. It converts the temperature of an object into an electronic signal to create a hot image on the screen; Or, to do temperature calculations on it. Infrared cameras can take temperature and can measure or measure accurately. However, it does not work as well in detecting hand touches as other cameras and is adversely affected by the weather. Therefore, thermal behavior can be observed but also the degree of temperature-related issues can be identified and categorized 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 a single camera defined as a standard camera that can be used for touch attention where the environment or resources may not be suitable for other types of image-based recognition. A single camera may not work as deep surveillance cameras or stereo cameras despite being challenged in this Flutter vision. This is a downloaded app that 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

which aims to double the horizontal adjustment of the 3D Holoscopic 3D camera with horizontal and vertical resolutions [82]. As shown in Figure 3.17 [83], the holoscopic camera can be in the position of a distributed power supply, using the MLA [83] [82]. Despite using the same features of the holographic process, it records 3D imagery in 2D form and views it in full 3D with optical component, without the required light source and prevents dark matter. In addition, it enhances post-production processing as a 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 Holoscopic 3D camera design L0 = Nikon wide angle lens 35 mm F2, NF = Nikon F-mount, AP = adapter plate, ER = 6 mm wide expansion rods, RM = <5arcminute precision rotation mount, MLA = MLA plane, inclined process, T0-T2 = expansion tubes, L1 = Rodagon 50 mm F2.8 relay lens × 1.89, C5D M2 = Canon 5D Mark2 DSLR. The arrow indicates the center of gravity, SA = square area with a mouth to L0.

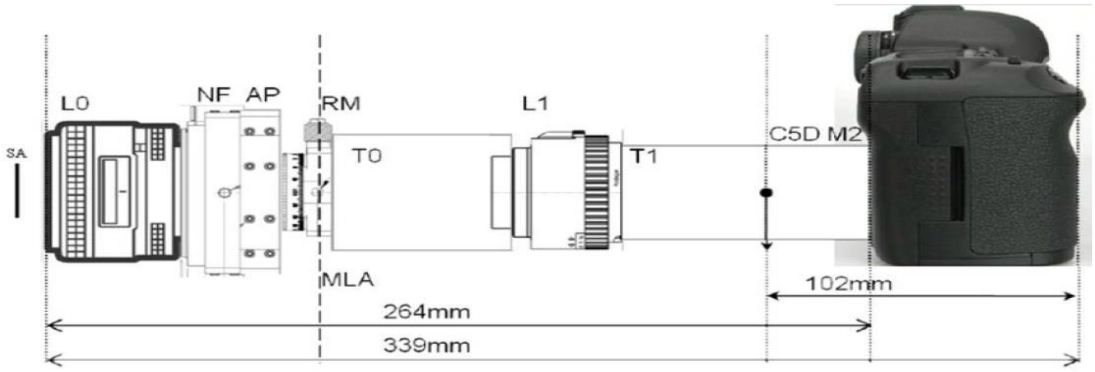


Figure 3.18: Square Aperture Type 2 camera integration with canon 5.6k sensor [82].

3.7 Summary

This Chapter has introduced the foundation of HCI as a key signal recognition field. HCI is how people can use gadgets, and how PCs respond to people's requests. There are different types of HCI. One of them, called signals, talks about the non-verbal interaction used in the UIs. Symptoms are also seen as the actual development of any part of the human body such as a finger, a hand, or a full body. For example, sign communication is a movement model used by people who are sensitive to hearing. The framework for approving hand gestures is assessed at this stage along with your types. Hand gestures are seen as a form of signal recognition, which is a way of moving a hand randomly and identifying it with certain gadgets. It has two types, information gloves and a PC view. Information glove A glove glove with connected nerves connected to the fingers or glove parts worn by a person. Many scientists design various gloves. Every knowledge glove has another reason. For models, the MIT information glove is used for computer games, bodybuilding or game preparation. However, PC vision is a common way of interacting with people and gadgets in the visual environment. In the PC vision section, it also introduces a variety of cameras such as the types of gadgets used for detection, such as sound system cameras, deep cameras, warm cameras, standalone cameras and a Holoscopic imaging framework camera. The next section introduces a hypothesis for image and video handling tools such as WT and EMD. It also provides an AI foundation that includes ANN, in-depth learning and CNN.

Chapter 4
Image Processing and Recognition

4.1 Image and Signal Processing

4.1.1 Introduction to Image Processing

Picture handling is a method for changing a photograph into a computerized structure and perform different procedure on it to create a superior picture or gain valuable data from it. The information is a picture, for example, a video outline and the result can likewise be an item or picture. This is a kind of sign engendering [85]. The arrangement of picture handling contains various devices, for example, picture securing, picture improvement, picture rebuilding, variety picture handling, wavelet and multiresolution handling, division and item acknowledgment.

Every instrument will be characterized as follows: picture securing is catching a picture and digitizing it and afterward investigating the issue space then follow the means as indicated by the issue. Picture upgrade is executed through time and recurrence to improve the picture as per the prerequisites. Picture reclamation is putting away unambiguous pieces of a picture utilizing a Point Spread Function. The variety picture handling device is utilized when the picture is high contrast. Wavelets and multiresolution handling are utilized assuming the pictures are to be delivered in various degrees or wavelets of goal. Pressure decreases the size of pictures utilizing explicit capacity. Morphological handling is an outer design of the picture utilizing widening and disintegration. Division is executed by parting a picture into various parts. Object acknowledgment used to perceive and save the picture depiction [85][86].

Amazon Rekognition is a cloud programming which is utilized to consolidate wavelets are the Daubechies' which will advance in wearable innovation. The Daubechies symmetrical is known as dbN wavelets where N is the quantity of blurring minutes. The application is normally utilized for sound, discourse, picture, video, and bio-clinical imaging. The Daubechies wavelets are characterized as follows:

$$\int x^n \psi(x) dx = 0, n = 0,1, \dots, K \quad (4.2)$$

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

PC vision is a field that expects to empower PCs to decipher, perceive and handle objects in a similar way as human vision. It is like giving knowledge and senses to a human PC. As a matter of fact, it is a troublesome undertaking to perceive PC pictures of various items. PC vision is firmly connected with man-made reasoning since machines need to comprehend what they see and afterward decipher or act suitably [95].

PC vision engineering includes handling advanced pictures by means of various stages progressively. The primary stage is picture procurement which catches a picture and digitalizes it and afterward investigations it as per the issue area. Picture Processing is the second stage which is a strategy to change an article into a computerized structure and play out specific procedure on it to deliver an improved photograph or get helpful data from it. Picture handling is likewise a type of sign administering where the info is a picture, for example, a video casing or picture, and the result can be an item or picture related highlights. The third stage is picture investigation, which extricates a snippet of data, and information handling. This strategy is commonly important to guarantee that specific presumptions proposed by the framework are fulfilled before a PC vision approach can be applied to picture information. Highlight Extraction is the fourth stage in PC vision frameworks which concentrate elements of the article and are gotten from the picture information at various degrees of intricacy. Significant level handling is the 6th stage where the information is typically a little arrangement of information at this level, for example, a bunch of focuses or a picture region that ought to contain a particular item. Ultimately, dynamic comprise of delivering an official choice required for the application [95]. A meager 3D point model of an enormous complex scene can be recreated from many somewhat covering photos [95]. Sound system matching calculations can make a definite 3D model of a structure veneer comprising of many photos taken from the web.

4.3 Artificial Intelligence

Simulated intelligence is the capacity of a machine to shrewdly perform mental undertakings and act. The field of AI attempts to comprehend astute substances [1]. Computer based intelligence is another discipline that started in 1956. With an assistance of AI, it is feasible for machines to gain from their own insight, adjust to new sources of info and perform human-like undertakings. Simulated intelligence is broadly utilized in finance, training, medical care, transportation fields and in different ventures, for example, PC vision, clinical determination, mechanical technology and remote detecting [96].

The dad of software engineering and AI is Alan Turing who proposed a 'Turing test' in 1950 which was intended to give a functional meaning of knowledge. In the event that a machine breezes through this Turing assessment, it is supposed to be smart. Be that as it may, no machines have totally finished this assessment at this point. There are different marks of knowledge like Intelligence Quotient (IQ) tests and mind size, however not even one of them convey insight in machines. As per Daniel Gilbert, there is one basic component in which our brains contrast from the personalities of creatures and PCs; it can encounter something that has not yet happened [96].

PCs and robots can surpass the human capacity at certain undertakings that are viewed as 'smart' utilizing procedures, for example, information mining and example acknowledgment and so on. Lower mental errands that are normal for people can be very intricate for machines. For instance, a dream framework and item acknowledgment, to some degree disguised objects, same item, unique shape, variety, surface and size consistency [96].

Powerless AI are machines that can go about as though they are smart, however their reasoning is mimicked thinking and not genuine. Solid AI are machines that go about as though they are smart and they are thinking. Tragically, we still just have Weak AI. On the off chance that a machine breezes through the Turing assessment, 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 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

$R^{H\times W\times C}$ 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 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 described in this chapter. Image processing is a method of applying certain techniques to digital photography. The applications used for image processing are remote sensing, entertainment and geography processes. The second theory discussed in this chapter is video analysis which is an analysis method used in video and time-sharing data to achieve important processes. Image processing algorithms such as empirical decay and Wavelet Transforms are discussed in this chapter. EMD is a method of analyzing random and indirect data while WT uses signal analysis where signal frequency changes over time.

Chapter 5

3D Video Gesture Recognition

5.1 Introduction

The main method of interaction between user and machine is direct communication. Devices such as mouse, keyboard, touch screen, remote control, and other direct communication systems act as a communication channel. Person-to-person communication is achieved through precise and natural forms of communication, e.g. body language and noise. The effectiveness and versatility of these communication methods has led many researchers to consider using them to support communication between humans and computers. Touch forms a large part of human language and is an important means of communication. Historically, in order to capture the shapes and angles of all members in a user action, wearable data gloves were commonly used. The cost and complexity of the wearable sensor limit the widespread use of this method. The computer's ability to sense touch and to execute certain commands based on that touch is called touch recognition. The primary goal of such touch recognition is to develop a system that is able to recognize and understand specific touches and transmit information.

At present, touch-based observation methods based on untouched visual inspections are popular. The reason for such popularity is their low cost and ease of use. Touch is a clear form of communication used extensively in the entertainment, health, and education sectors.

Herbert Ives, in the 1920's, began to think of a way to simplify Lippmann's concept by assembling a lens lens that contained a series of circular lenses called lenticules. A range of magnifying lens signals is designed to detect various angles and images are constantly exaggerated to provide a pixel for each small image. The lens sheet shows across and the old rear face of the plane focused on it is flat. For example lenses are used in lenticular production where technology is used to show the illusion of depth that creates moving or changing images as the image appears at different angles. This new technology can be used to produce 3D images on a flat sheet display. Therefore, when photographic motion is considered, this results in 3D holoscopic video [5] [6].

5.2 3Dsmall Distance Gesture Recognition Systems

Ge et al [105] proposed a 3D CNN technique to gauge constant hand presents from single profundity pictures. The elements extricated from pictures utilizing 2D CNN are not appropriate for assessment of 3D hand act like they need spatial data. The proposed technique accepts input as a 3D volumetric portrayal of the hand profundity picture and catches 3D spatial construction and precisely relapse full 3D hand present in a solitary pass. 3D information increase is performed to make the CNN technique strong to worldwide directions and hand size varieties. The aftereffects of the examination show that the proposed 3D CNN outflanks the best in class strategies on two testing hand present datasets. The execution runs at north of 215 fps on a standard PC with a solitary GPU which is demonstrated to be extremely viable.

A strategy involving a profundity camera in a savvy gadget for hand signal acknowledgment is proposed by Keun and Choong [107]. The acknowledgment is made through the acknowledgment of a hand or location of fingers. For identifying the fingers, the hand skeleton is recognized by means of Distance Transform and fingers are distinguished by utilizing Convex Hull calculation. To perceive a hand, a recently produced motion is contrasted and as of now educated information utilizing the Support Vector Machine calculation. The hand's middle, finger length, hub of fingers, hand hub and arm focus are surveyed for this. A real shrewd gadget was carried out for the assessment of this trial.

5.2.1 System Implementations

A. Hand Gestures Input

In this experiment, hand gestures are fed as input into CNN. Figure 5.1, Figure 5.4 and Figure 5.7 show twelverandom hand gestures recorded in small 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 5.1, the resolution of the camera used is full High Definition (HD) while for Figure 5.4 and Figure 5.7 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 5.1is 47 for x-axis whereas for y-axis it is 84. For Figure 5.4 and Figure 5.7, the number of lenses is decreased to 31 on thex- 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 5.2, the image is rotated 0.30 degrees to adjust the image position while for figure 5 and figure 8 are rotated 180.20degrees.

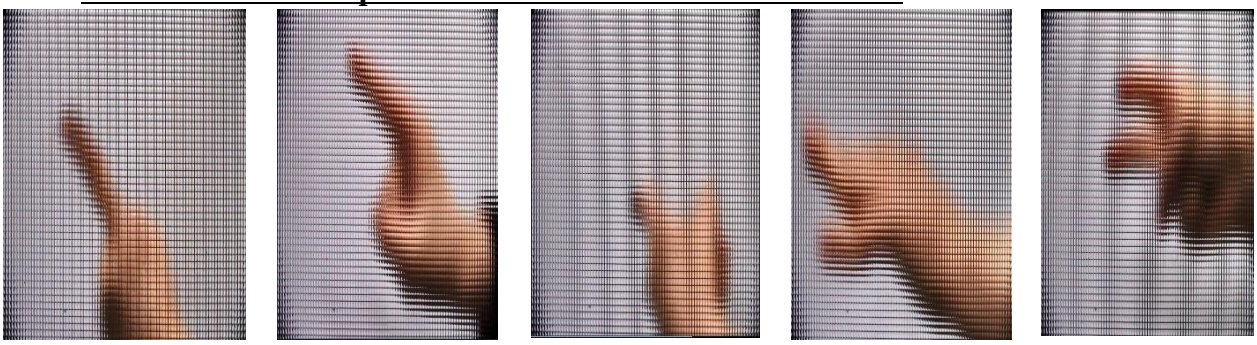
5- Divide lens into seven segments i.e. 7 ×7, the X segment is a constant of 4 while Y is changeable to 2, 4 and 6.

6- Create twelveseparate 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.

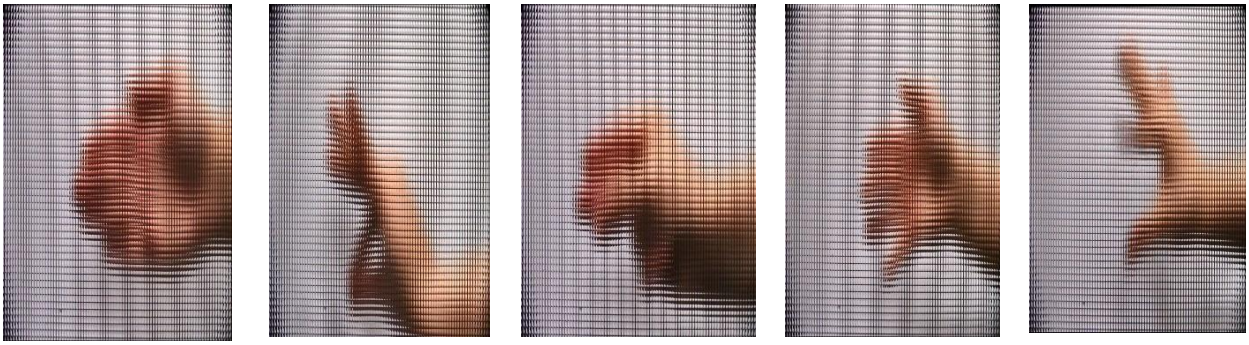
7- Combine each left, centre and right images for three people in one directory.

8- Combine the three images i.e. left, centre and right to get one image with a size 405 × 75 in Figure 5.3, Figure 5.6 and Figure 5.9.

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



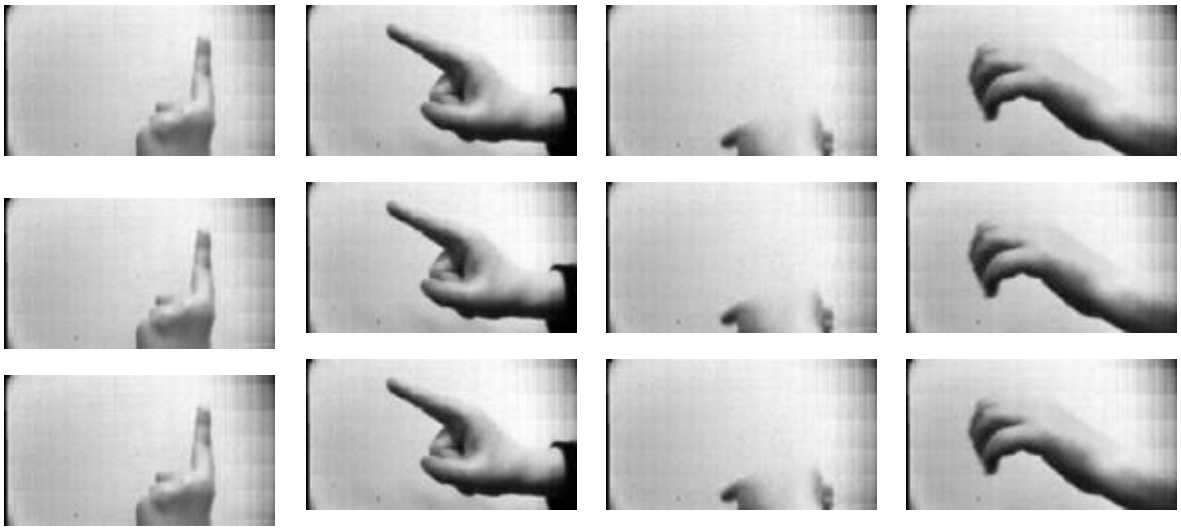
a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion e) 2 Fingers Shrink



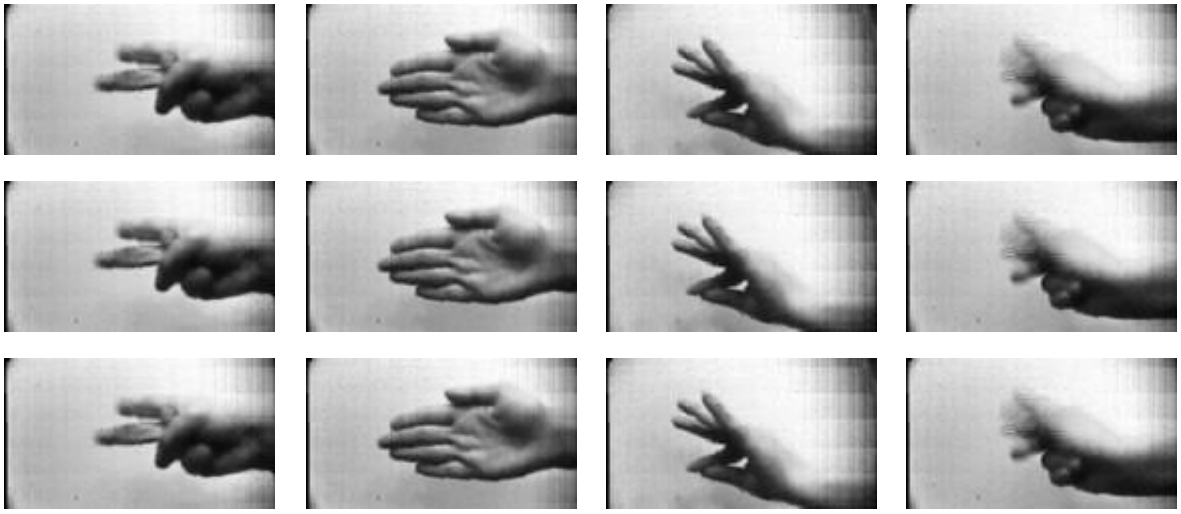
f) Back/Forth g) Rub motion h) Click motion i) Dance motion j) Pinch motion

Figure 5.1: Pre-extraction first person's hand motions in small distance

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



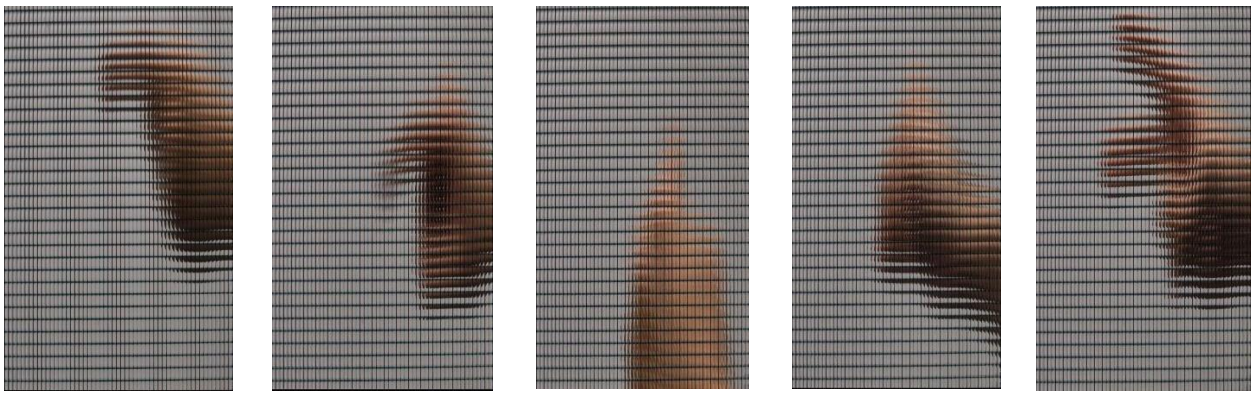
a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion



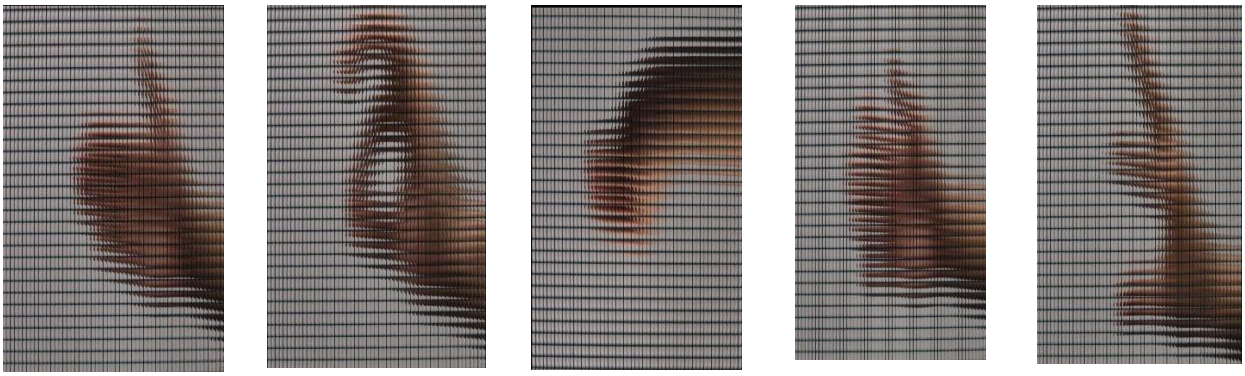
e) 2 Fingers Shrink f) Back/Forth g) Rub motion h) Click motion

Figure 5.2: Post-extraction first person's hand motions in small distance

3- Pre- extraction second person's hand motion in small distance



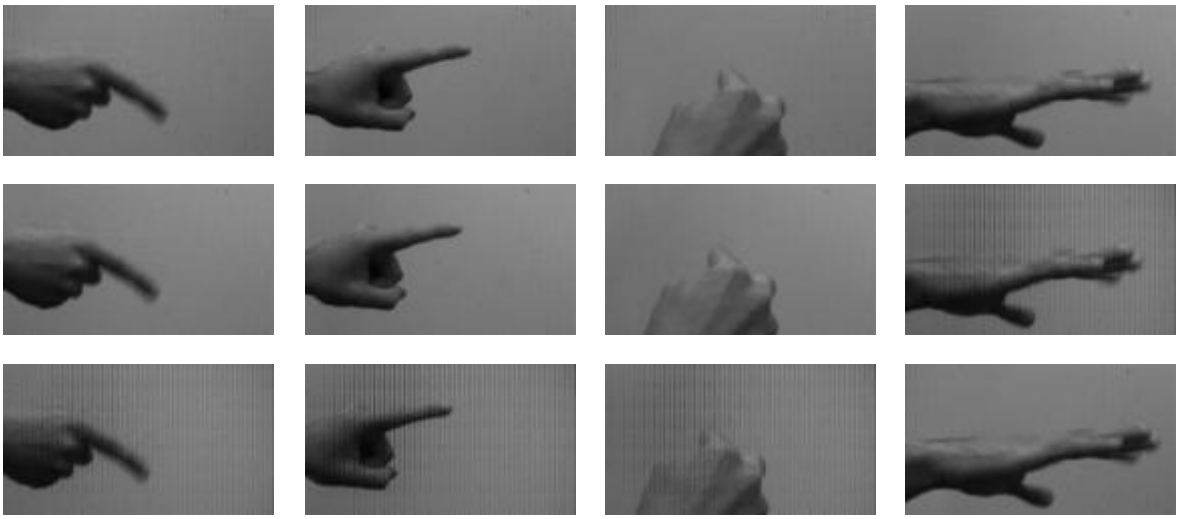
a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion e) 2 Fingers Shrink



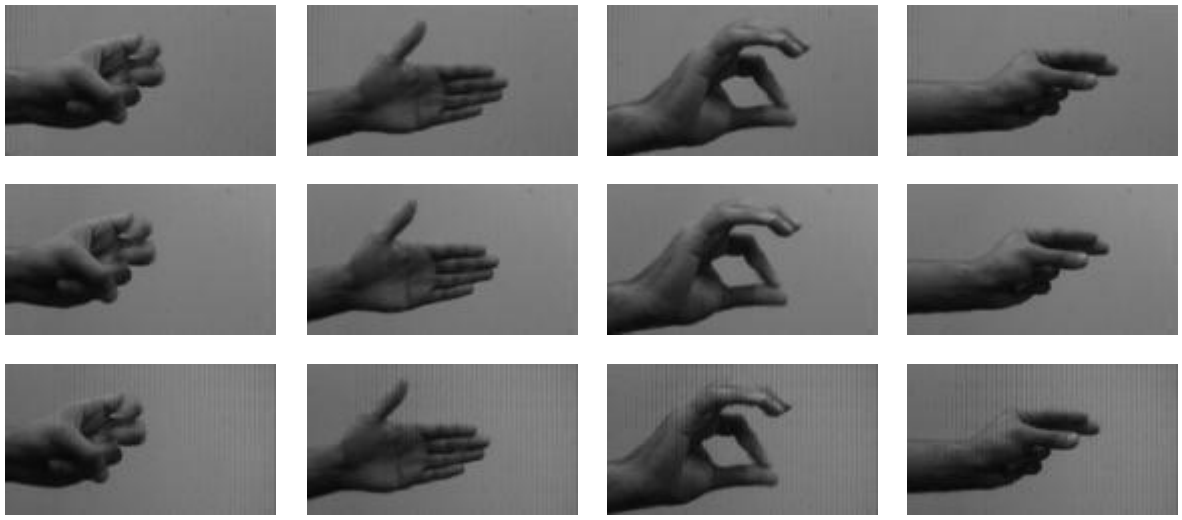
f) Back/Forth g) Rub motion h) Click motion i) Dance motion j) Pinch motion

Figure 5.3: Pre-extraction second person's hand motion in small distance

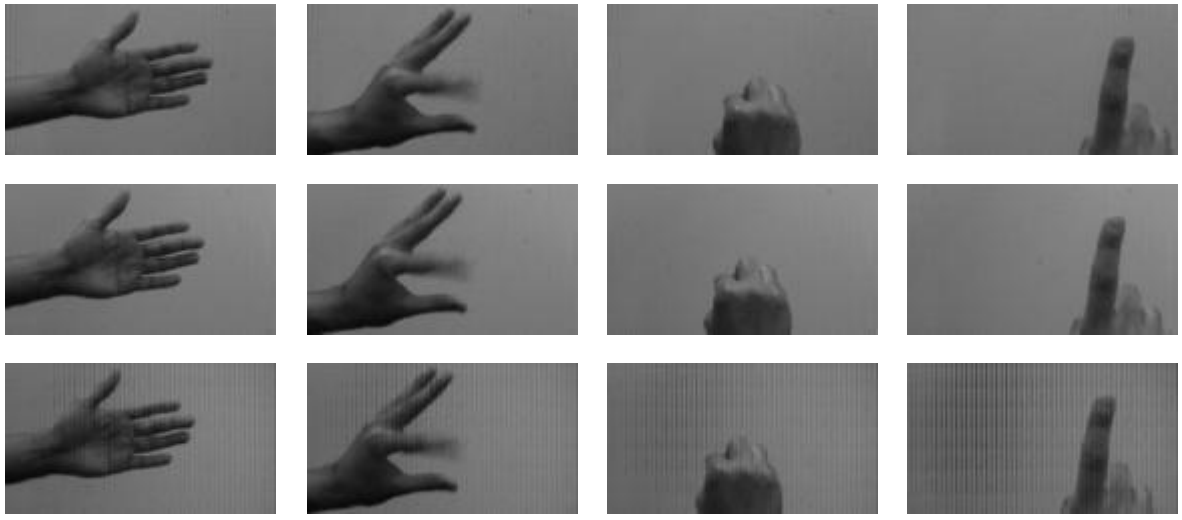
1- Post- extraction second person's hand motion in small distance single (LCR)



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



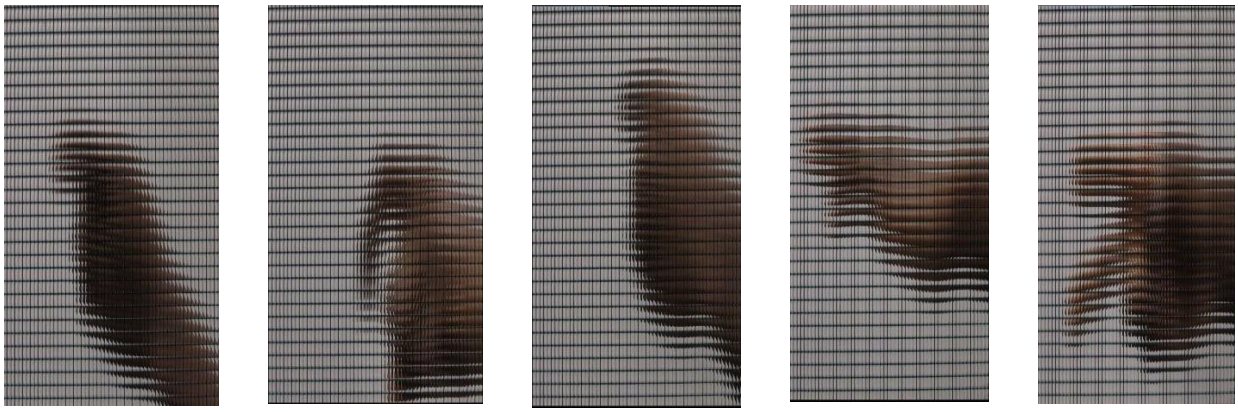
e) 2 Fingers Shrink f) Back/Forth g) Rub motion h) Click motion



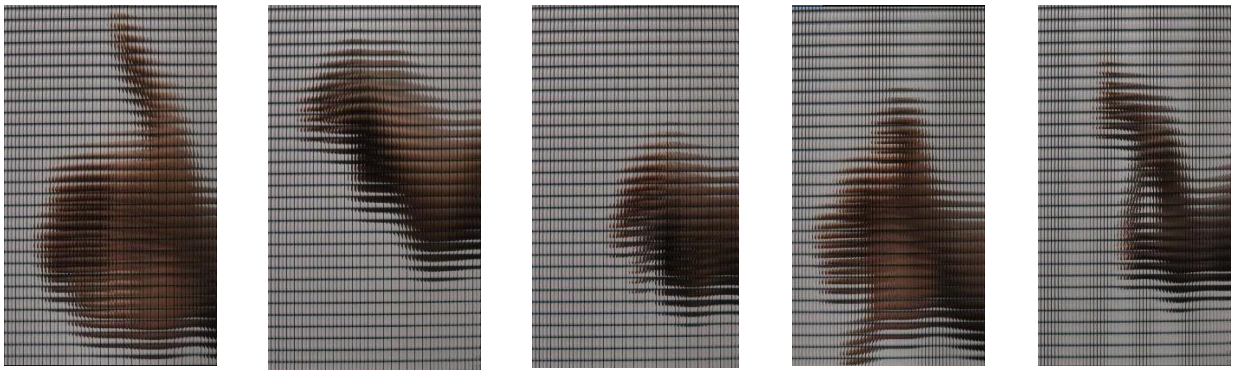
i) Dance motion j) Pinch motion k) write motion l) Click motion 2

Figure 5.5: Post-extraction second person's hand motion in small distance single (LCR)

1- <u>Pre- extraction third person's hand motion in small distance</u>



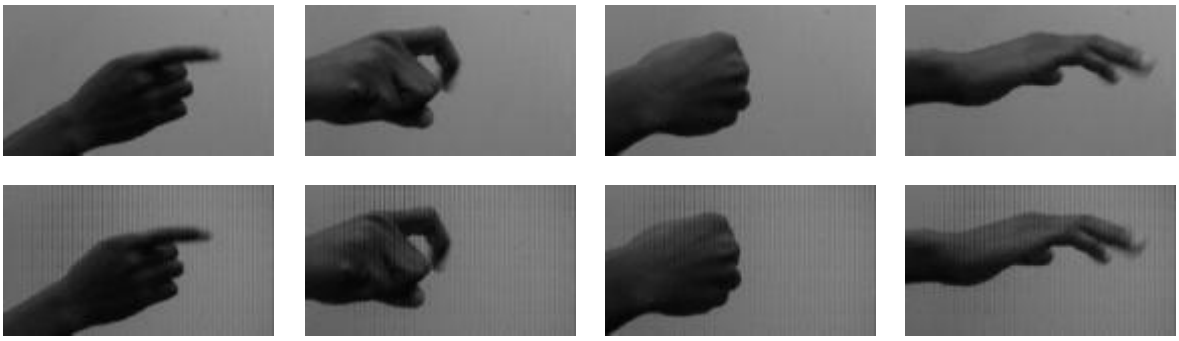
a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion e) 2 Fingers Shrink



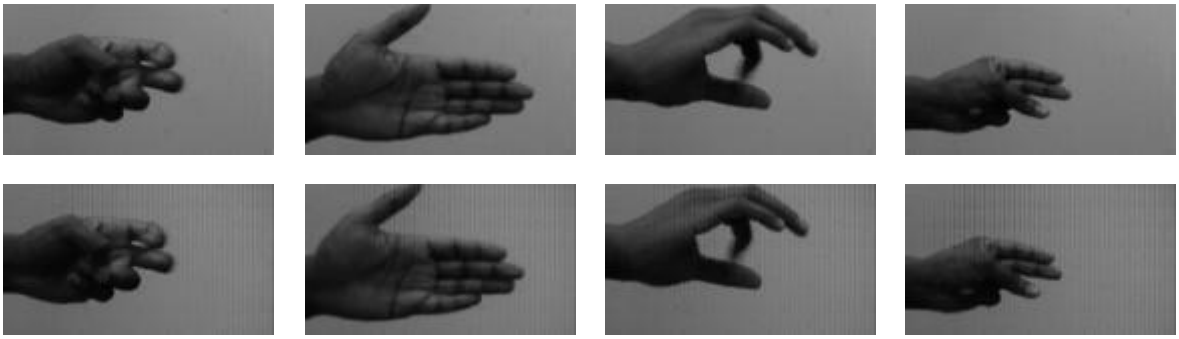
f) Back/Forth g) Rub motion h) Click motion i) Dance motion j) Pinch motion

Figure 5.6: Pre-extraction third person's hand motion in small distance

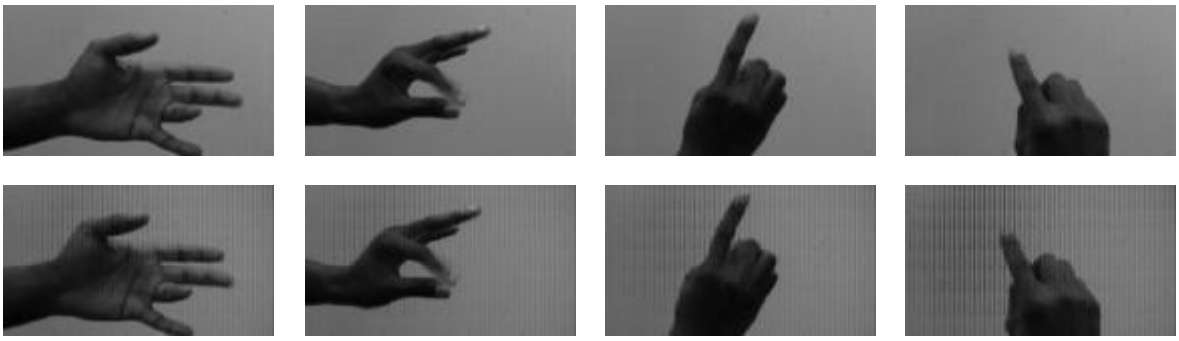
1- Post- extraction third person's hand motion in small distance single (LCR)



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 5.7: Post-extraction third person's hand motion in small distance small distance single (LCR)

5.2.2 Result

A. Convolutional Neural Network Implementation

The Convolutional neural network is an important part of in-depth learning as it is used to train data without the use of any imaging processing techniques. In this experiment, a whole new list of texts was created for each three-person video. CNN topology is shown in Figure 6.10. The length of each video is 10 seconds in split and merged images. Every video will be studied to produce 900 images, namely 300 on the left, 300 on the right and 300 at the center, while 300

images are combined in a single directory. Images are divided into training and testing models. A total of 390 different photo frames and 210 of the 70% integrated images. CNN topology is produced in seven layers each layer has functionality with the following size: ImageInputLayer size [135, 75, 1] Linear Unit (ReLULayer), MaxPooling2DLayer Pool size [2,2], FullyConnectedLayer size [default] and Output size [7], SoftmaxLayer and ClassificationOutputLayer size [default]. CNN hyperparameters are created within the training options function. The parameter value of epochs is set to 100 epoch.

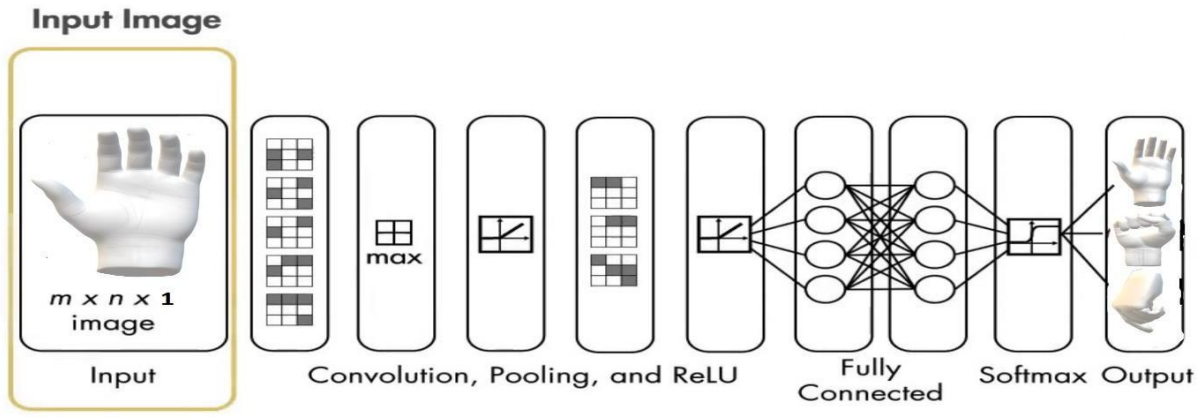


Figure 5.8: CNN topology

B. Parameters Comparison

The performance of the CNN algorithm can be compared using a number of performance integration parameters (H: M: S), i.e. the length of time taken by the software to execute the work, training accuracy is calculated using the training data model and obtaining algorithm accuracy. Testing accuracy accuracy test data. Sensitivity estimates the appropriate percentage of the positive rating, the exact estimates of the false positive value, PPV and the amount of training is slightly better than the result of one second person. Overall, the first one has the best value in most parameters.

Table 5.1: 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
+Ve Predictive Value (PPV)	1	1	1	1	1	0.97	0.94
-Ve Predictive Value (NPV)	1	0.98	1	1	0.99	1	0.98
+Ve Likelihood (LR+)	0	0	0	0	0	506	212.58
-Ve Likelihood (LR−)	0	0.13	0	0	0.04	0	0.20

5.2.3 Summary

5.3 3DLong Distance Gesture Recognition Systems

The main method of interaction between user and machine is direct contact. Devices such as mouse, keyboard, touch screen, remote control, and other direct communication systems act as a communication channel. Human-to-human communication is achieved through precise and natural forms of communication, e.g., body language and sound. The effectiveness and versatility

of these methods of communication have led a number of researchers to consider using them to support human communication. Touch forms a large part of human language and is an important means of communication. Historically, in order to capture the shapes and angles of all members in a user action, wearable data gloves were commonly used. The cost and complexity of the wearable sensor limit the widespread use of this method. The computer's ability to sense touch and to execute certain commands based on that touch is called touch recognition. The primary goal of such touch recognition is to develop a system that is able to recognize and understand specific touches and transmit information.

5.3.1 System Implementations

In this test function, hand gestures are given as input into a few touch algorithms. Figures 5.11, 5.14 and 5.17 show twelve hand gestures recorded at the empty back distance using the holoscopic imaging camera system. It is evident that one movement is 2D and the other is 3D. Photos are pre-processed before releasing videos according to certain steps. The steps used in all the other calculations are similar to those of small-scale hand gestures.

1- Pre- extraction first person's hand motions in longer distance

a) Sweep motion

b) Shrink motion

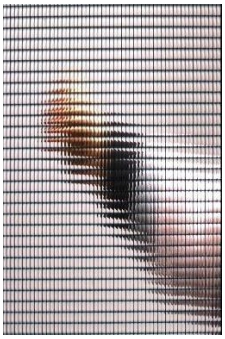
c) Circular motion

d) Squeeze motion

e) 2 Fingers Shrink

f) Back/Forth

g) Rub motion



h) Click motion

i) Dance motion

j) Pinch motion

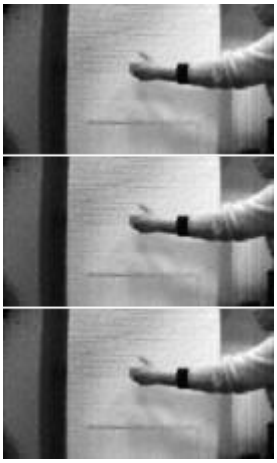
Figure 5.9: Pre-extraction first person's hand motions in longer distance

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

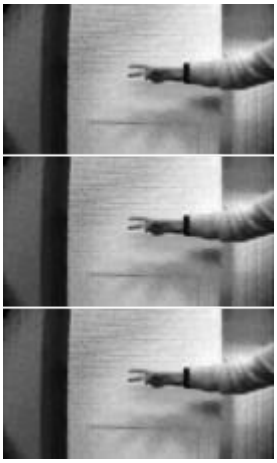
a) Sweep motion

b) Shrink motion

c) Circular motion



d) Squeeze motion

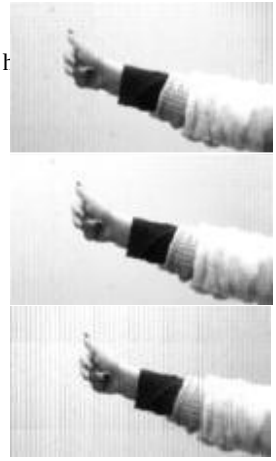


e) 2 Fingers Shrink

f) Back/Forth

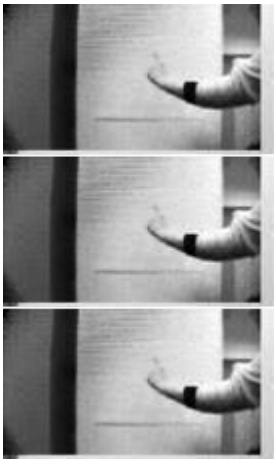
g) Rub motion

h

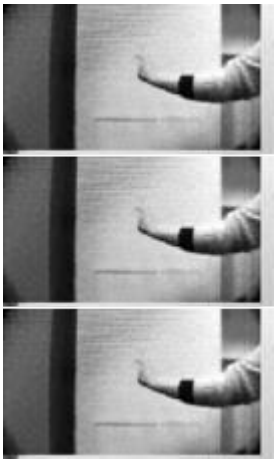


i) Dance motion

j) Pinch motion



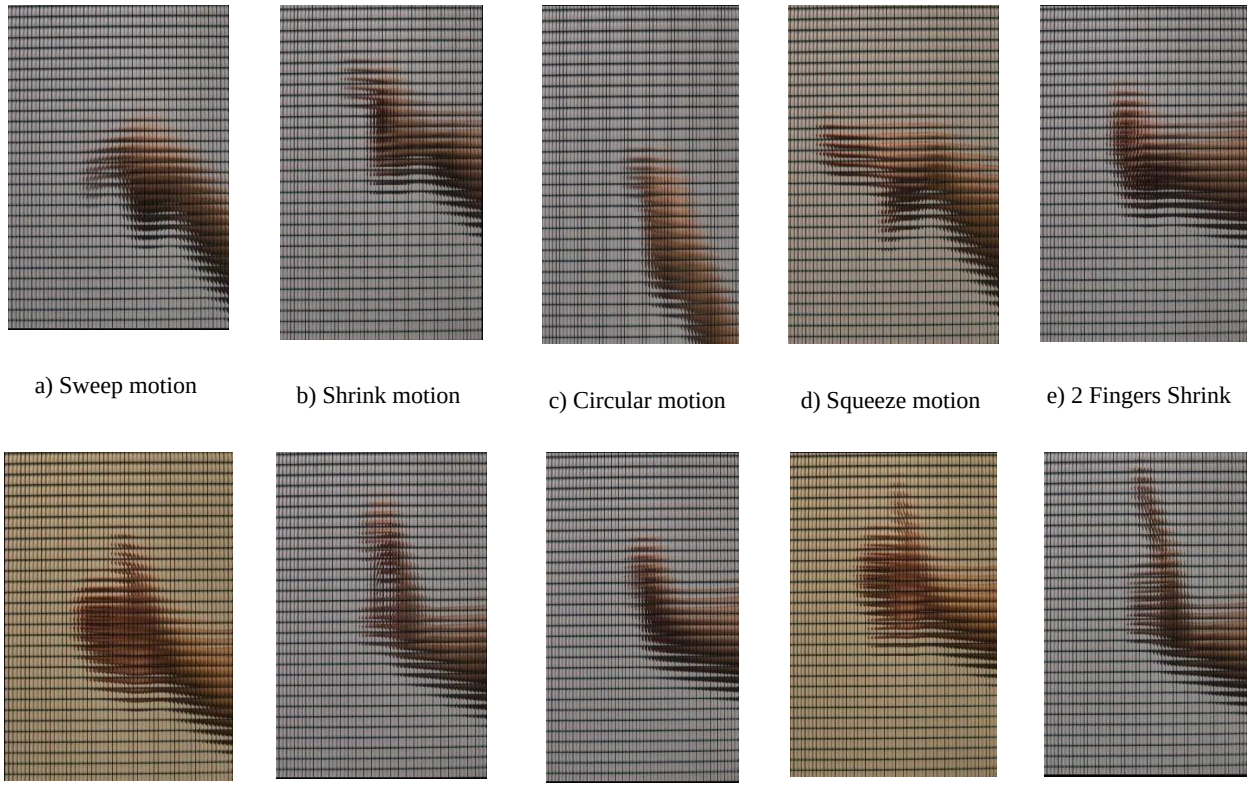
k) write motion



l) Click motion 2

Figure 5.10: Post-extraction first person's hand motions in longer distance single (LCR)

3- Pre- extraction second person's hand motions in longer distance

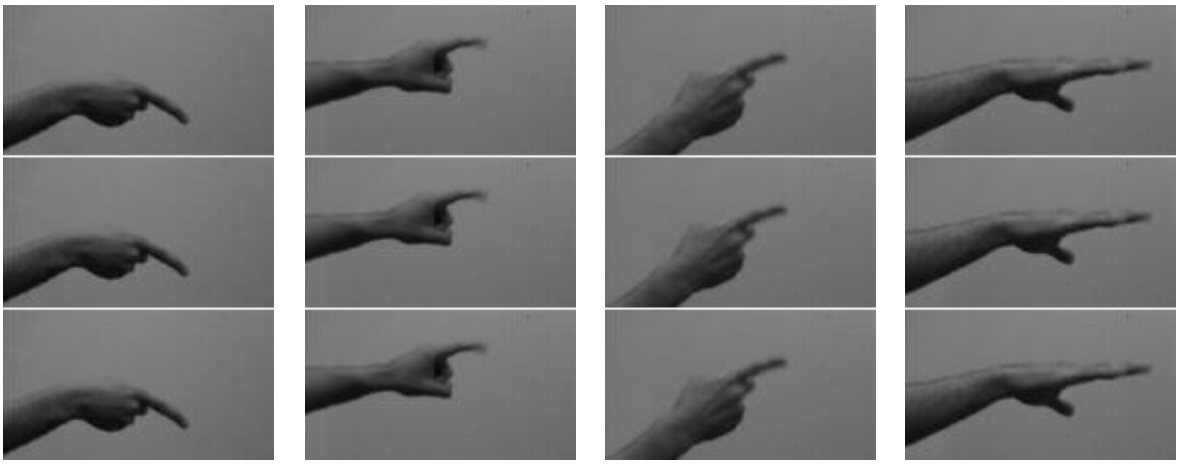


a) Sweep motion b) Shrink motion c) Circular motion d) Squeeze motion e) 2 Fingers Shrink

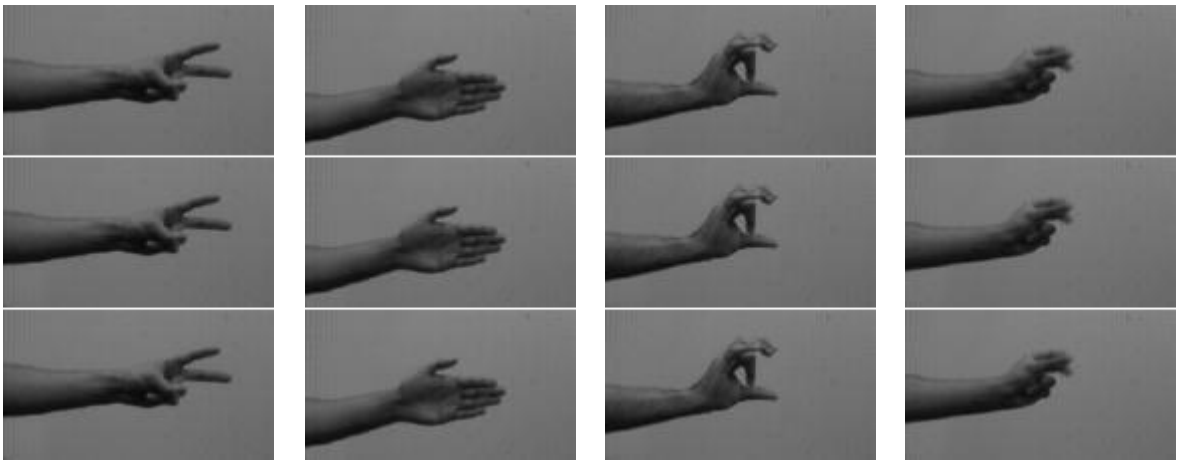
f) Back/Forth g) Rub motion h) Click motion i) Dance motion j) Pinch motion

Figure 5.11: Pre-extraction second person's hand motions in longer distance

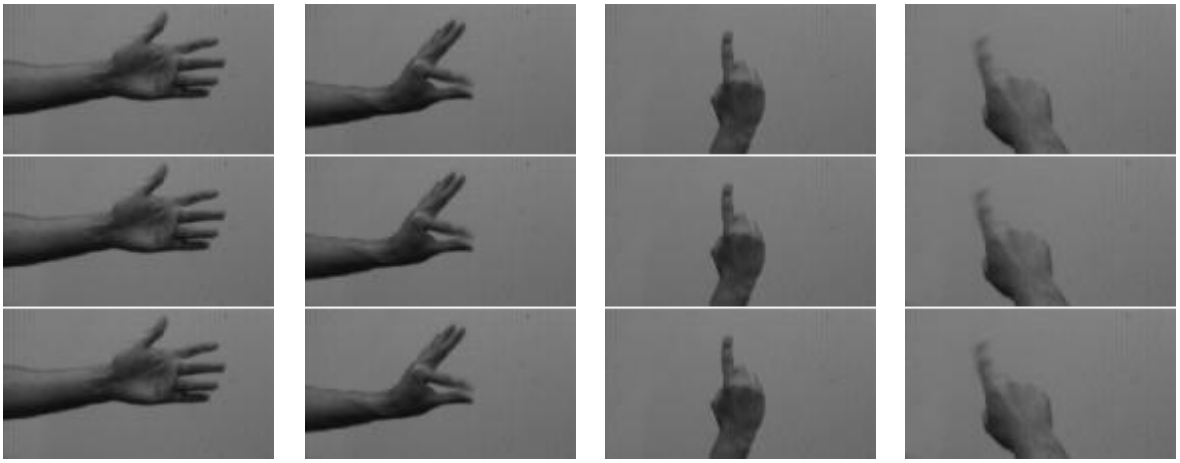
4- Post- extraction second person's hand motions in longer distance single (LCR)



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



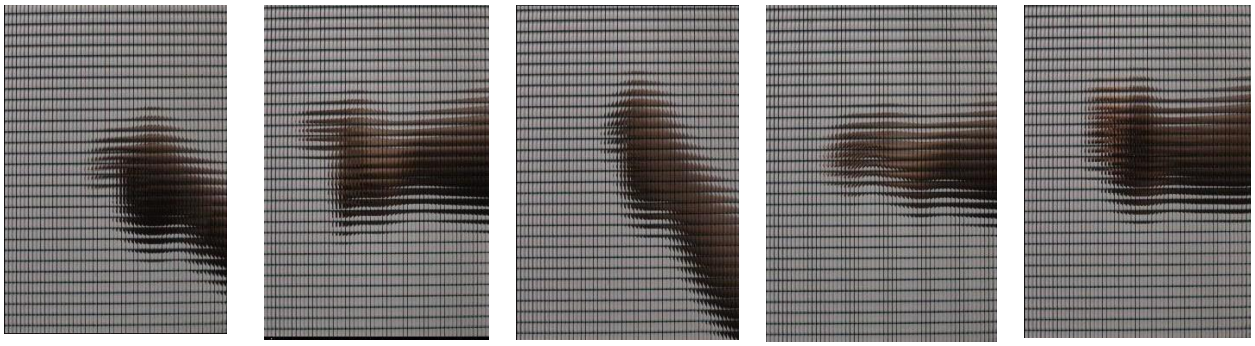
e) 2 Fingers Shrink f) Back/Forth g) Rub motion h) Click motion



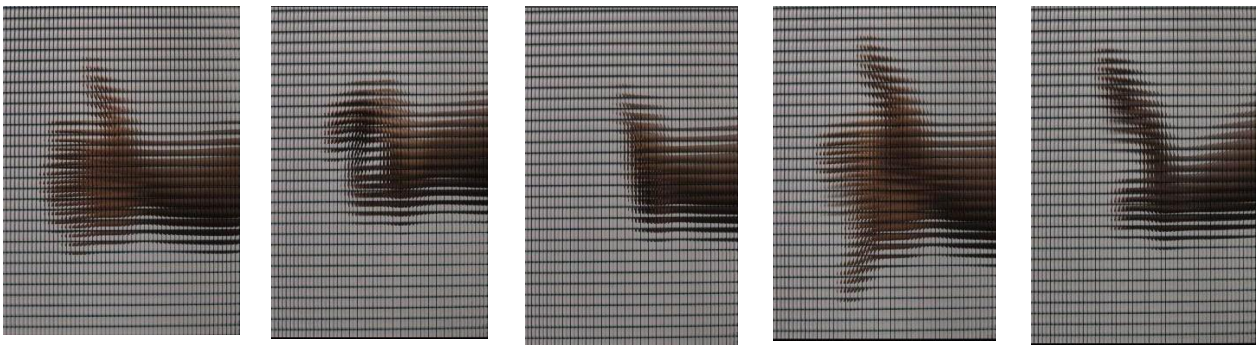
i) Dance motion j) Pinch motion k) write motion l) Click motion 2

Figure 5.12: Post-extraction second person's hand motions in longer distance single (LCR)

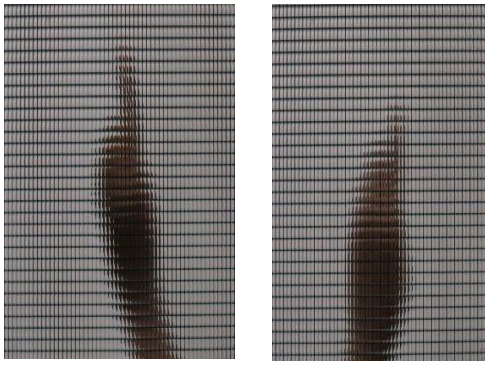
4- Pre- extraction third person's hand motions in longer distance



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



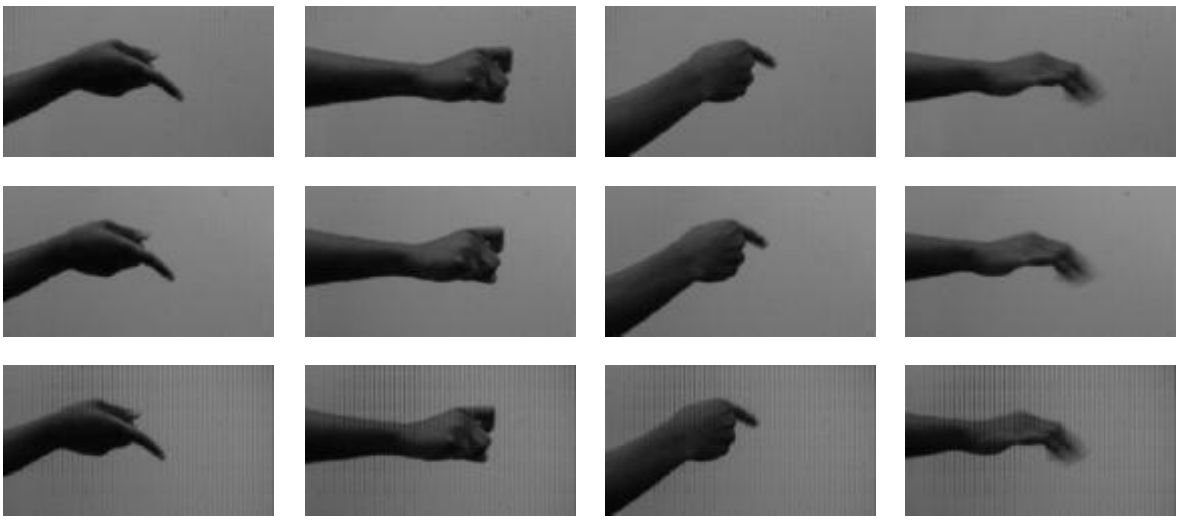
f) Back/Forth g) Rub motion h) Click motion i) Dance motion j) Pinch motion



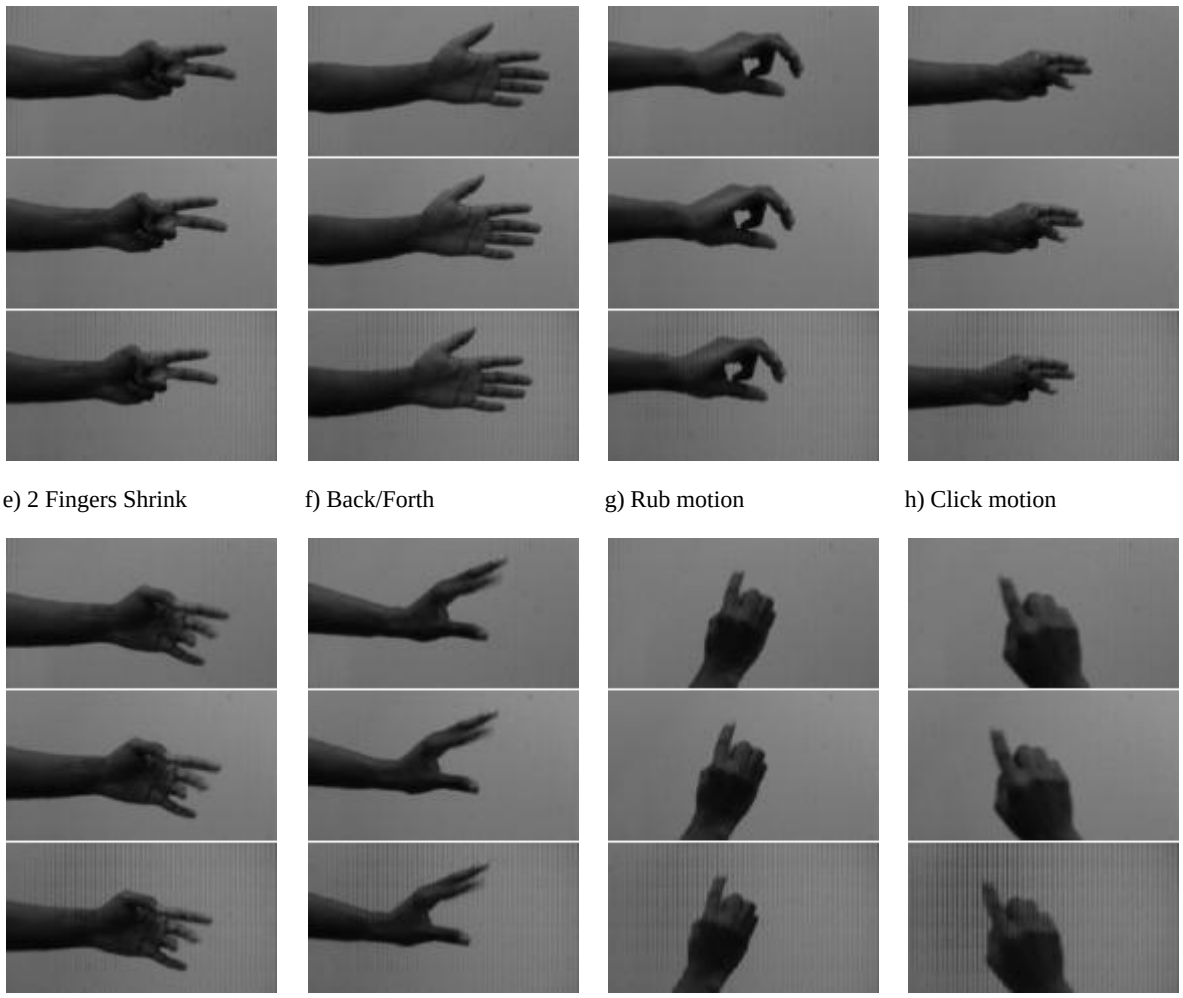
k) write motion l) Click motion 2

Figure 5.13: Pre-extraction third person' hand motions in longer distance

1- Post-extraction third person’s hand motions in longer 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 5.14: Post-extraction third person's hand motions in longer distance single (LCR)

A. Convolutional Neural Network Implementation

The Convolutional Neural Network is an important part of in-depth learning as it is used to train data without using any image processing method. In this experiment, a whole new list of texts was created for each three-person video. The length of each video is 10 seconds in split and merged images. Every video will be studied to produce 900 images, namely 300 on the left, 300 on the right and 300 at the center while combining 300 images into a single directory. Images are categorized to train and test models. The total number of separate photo training frames is 390 and 210 of the combined 70% photographs. CNN topology is produced in seven layers per layer that can work with the following size: ImageInputLayer size [135, 75, 1] Line Unit (ReLULayer), MaxPooling2DLayer Pool size [2,2], FullyConnectedLayer [default] size and Output size [7], SoftmaxLayer and ClassificationOutputLayer [default] size. CNN hyperparameters are created within the training options function. The parameter value of epochs

is set to 100 epoch.

5.3.2 Results

Table 5.2 shows the comparison between the three individuals for the best results. One, combined and three-dimensional outcomes presented in terms of performance, training, testing, sensitivity, specificity, PPV, NPV, LR + and LR-.

In one study, the timing of third-person homicide was much lower than that of second- and third-party people. The second person has a lower value compared to the first and third person in training. The first person test result is the best while the other results are not so good. The effects of empathy on all people are equal while the effect on the third person is less than the other effects on its specificity. PPV results in this third test activity

Table 5.2:ComparisonBetween 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
+Ve Predictive Value (PPV)	1	1	1	1	0.96	1	1
-Ve Predictive Value (NPV)	1	1	1	1	1	1	0.99
+Ve Likelihood (LR+)	0	0	0	0	276.63	0	0
-Ve Likelihood (LR−)	0	0	0	0	0	0	0.09

5.3.3 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 longer 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.

5.4 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]

5.4.1 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.

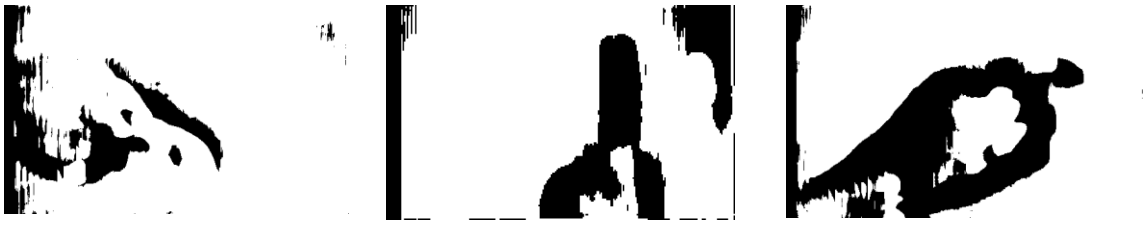
A feature match selection (FMS) algorithm is presented by [110], with an aim to extract and estimate an accurate full parallax 3D model form from a 3D omni-directional holoscopic imaging (3DOHI) system. The novelty of the paper is based on two contributions: feature blocks selection and its corresponding 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.

5.4.2 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

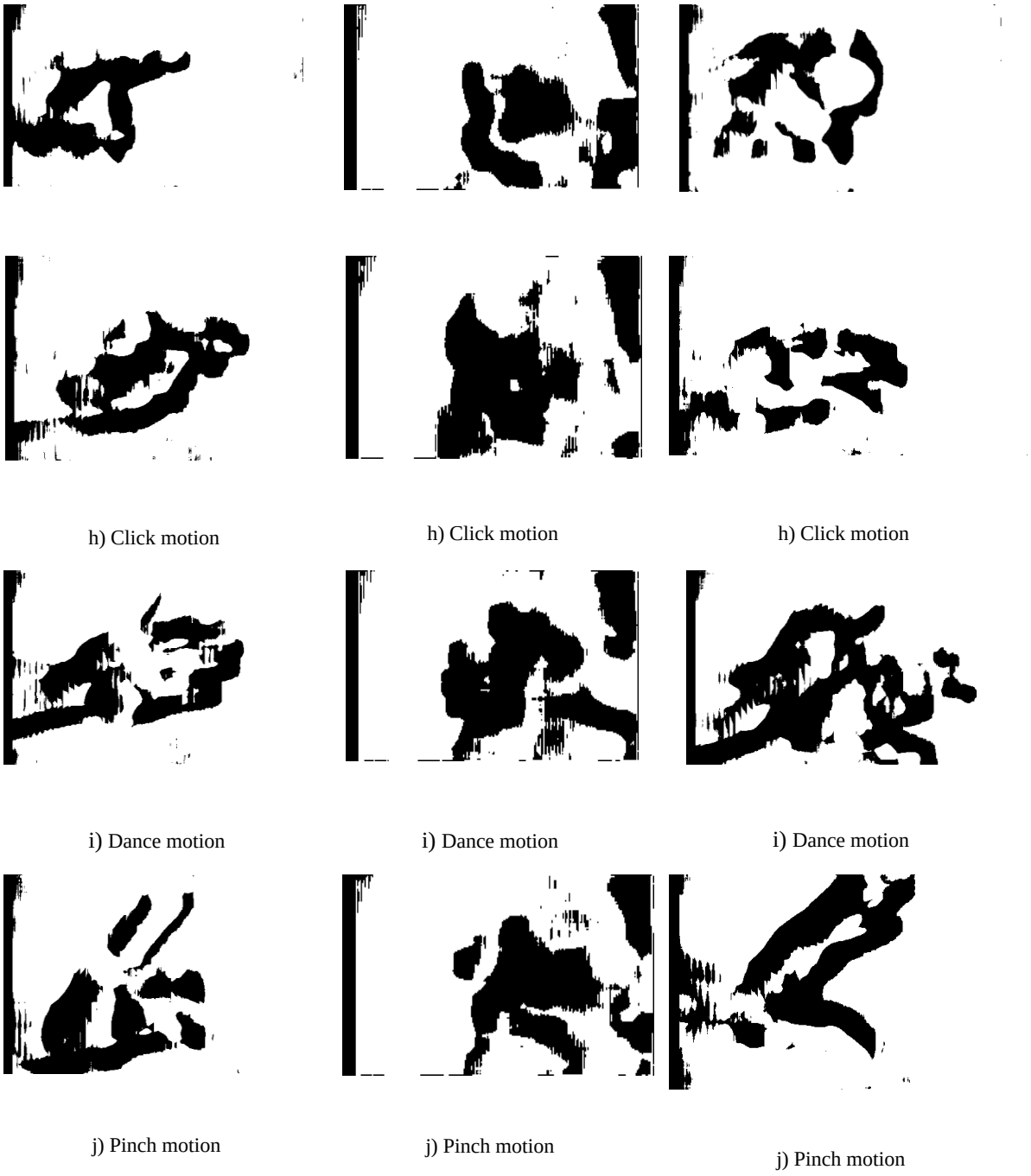


Figure 5.15: The disparity of Persons 1, 2 and 3

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 5.3: 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
+Ve Predictive Value (PPV)	1	0.98	1	1
-Ve Predictive Value (NPV)	1	1	1	0.99
+Ve Likelihood (LR+)	0	933	0	0
-Ve Likelihood (LR−)	0	0	0	0.03

5.4.4Summary

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 small 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

Chapter 6
Conclusion and Future work

6.1 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 small and longer 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 longerer than for the other two people.

The fifth experimental work is finding the disparity for hand gestures in small 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.

6.2. 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.

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**