HAND GESTURE RECOGNITION BASED ON ISOLATED AND CONTINUOUS SIGN LANGUAGE

AHMAD YAHYA DAWOD

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

PENGENALPASTIAN ISYARAT TANGAN BERDASARKAN BAHASA ISYARAT TERASING DAN BERTERUSAN

AHMAD YAHYA DAWOD

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

07 April 2018

AHMAD YAHYA DAWOD
P73056

ACKNOWLEDGEMENT

First and foremost praise is to Almighty Allah for all his blessings for giving me patience and good health throughout the duration of this Ph.D. research.

I take immense pleasure to express my sincere and deep sense of gratitude to my supervisor Associate Professor Dr. Md. Jan Nordin for providing support, excellent supervision, encouragement, good teaching, brilliant ideas, and motivation so that I could complete this thesis. He followed my work with keen interest and was always ready to provide valuable input.

I would like to send my thanks to Dr. Junaidi Abdullah, who has given me the opportunity to work on the topic of my choice for my Ph.D., convinced that intrinsic motivation is the best driving force for such an endeavor.

I am greatly indebted to thank Community Service Centre for the Deaf (CSCD). In the Community Service they stand with me and who have helped me in the collection of data and also in my work whenever I needed them. Especially, I appreciate Madam Cindy Leong and Madam Rose Ng who help me to test my modeling framework.

Furthermore, my sincere thanks to all postgraduate staff of FTSM for their help, friendship and creating a pleasant working environment throughout my four and half years in UKM for me.

I would like to thank all postgraduate students of UKM CAIT research group for their helpfulness, friendship and for creating a pleasant working environment throughout my studying years in UKM.

Great thanks to all my friends who made me feel happy, have fun and learn. Their friendship motivated me during this work.

Lastly, I am very grateful to thanks go to my family for their endless support my wife and two aunties. I owe both a depth of gratitude.

ABSTRACT

Hand gesture recognition from continuous sign language has attracted research interest in computer vision and human-computer interaction. The need for real-time recognition of continuous sign language has grown with the emergence of better-capturing devices such as Kinect sensors. However, the recognition of sign from the continuous input is not an easy task. One has to accurately segment the hand and fingers, track them, segment individual gestures through movement epenthesis recognition, and find the best method for classification in order to recognize the gestures. Existing approaches are mostly limited to the optical cameras that produce the low quality of input image. In addition, due to the difficult nature of the tasks involved, most works only focused on recognizing static and isolated signs rather than dynamic and continuous one. In sign language domain, very few reported work that attempts to propose a complete solution that can detect static, isolated, dynamic and continuous gestures all at the same time. This thesis presents real-time hand gesture detection and recognition from continuous sign language. The two categories in sign language recognition, namely the static gestures and the dynamic gestures were explored. For static gestures, a new technique for hand and fingertips detection using two different approaches based on *YCbCr* color space and skeletonization were presented respectively. In the first approach, adaptive skin color model in YCbCr colorspace that models the skin color was proposed. This is different from the normal approach that allocates specific range to model skin color which proves to be robust to varying background and lighting condition. In the second approach, skeletonization that is followed by gradient-based circle detection method were proposed to detect fingertips and track them. In dynamic gesture category, a new technique is proposed that employs contrast adjustment and gesture detection analysis to determine the start and end points of each individual moving gestures from a series of continuous gestures. This is also known as a process to detect the movement epenthesis. The technique is able to detect and identify the movement epenthesis with reasonable accuracy. To complete the research work, an approach has been developed that can recognize static and dynamic gestures in real time using the Kinect sensor. The contribution in the developed approach involves tracking the 3D points (X, Y, Z) of the hand centroid as features. Related works only use 2D points (X, Y). To test the effectiveness of the proposed technique, a new dataset set was developed. This new dataset is different from the existing dataset because it includes both signs performed for positive and negative sentences which have not been addressed in other dataset. For the classification and recognition steps of the gestures, four main tools were employed which include Support Vector Machine (SVM), Hidden Condition Random Field (HCRF), Hidden Markov Model (HMM), and Random Decision Forest (RDF). The proposed approach was tested for gestures that involve one and two hands and was compared with other approach and gave better accuracy. The recognition accuracy of the movement and orientation of the hands of the proposed technique is higher than 99.1%. The performance of the result is also compared with other two datasets,and the proposed method is able to obtain recognition rate that is up to 99.9%.

ABSTRAK

Pengenalpastian isyarat dari bahasa isyarat berterusan telah menarik minat penyelidikan dalam penglihatan komputer dan interaksi manusia-komputer. Keperluan pengenalpastian secara langsung dari bahasa isyarat berterusan telah berkembang dengan kemunculan peranti penangkap yang lebih baik seperti penderia Kinect. Walau bagaimanapun, pengenalpastian isyarat dari input berterusan bukanlah tugas yang mudah. Seseorang perlu tembereng tangan dan jari dengan tepat, mengesannya, mensegmen gerakan individu melalui pengenalan epentesis pergerakan, dan mencari kaedah terbaik untuk mengklasifikasi bahasa isyarat. Pendekatan sedia ada kebanyakannya terhad kepada kamera optik yang mempunyai kualiti imej yang rendah. Selain itu, disebabkan tugas sukar yang terlibat, kebanyakan kerja hanya tertumpu kepada mengenal pasti tanda statik dan terpencil dan bukannya dinamik dan berterusan. Dalam domain bahasa isyarat, sangat sedikit kerja yang telah mencadangkan penyelesaian lengkap yang dapat mengesan isyarat statik, terpencil, dinamis dan berterusan pada waktu yang sama. Tesis ini membentangkan pengesanan tangan dan pengenalpastian isyarat tangan secara langsung dari bahasa isyarat berterusan. Kedua-dua kategori dalam pengenalpastian bahasa isyarat telah diterokai, iaitu isyarat statik dan isyarat dinamik. Untuk isyarat statik, teknik baru pengesanan tangan dan hujung jari menggunakan dua pendekatan berbeza berdasarkan warna dan kerangka YCbCr telah dibentangkan. Dalam pendekatan pertama, model warna kulit yang adaptif berdasarkan warna YCbCr telah dicadangkan. Ini berbeza dengan pendekatan biasa yang memperuntukkan julat yang spesifik untuk memodelkan warna kulit. Pendekatan yang dicadangkan terbukti berkesan bagi latar belakang dan keadaan pencahayaan yang berlainan. Dalam pendekatan kedua, skeletonisasi diikuti dengan kaedah pengesanan bulatan berasaskan gradien bagi mengesan dan menjejaki hujung jari telah dicadangkan. Dalam kategori isyarat dinamik, teknik baru dicadangkan menggunakan pelarasan kontras dan analisis pengenalpastian isyarat bagi menentukan titik awal dan akhir setiap isyarat individu dari siri isyarat berterusan. Ini juga dikenali sebagai proses bagi mengesan epenthesis pergerakan. Teknik yang dicadangkan dapat mengesan dan mengenal pasti epentesis pergerakan dengan ketepatan yang munasabah. Untuk menyelesaikan kerja penyelidikan, satu pendekatan lengkap telah dibangunkan yang dapat mengenalpasti isyarat statik dan dinamik secara langsung menggunakan penderia Kinect. Sumbangan dalam pendekatan yang dibangunkan melibatkan penjejakan titik 3D (X, Y, Z) dari sentroid tangan sebagai fitur. Kerja-kerja yang sedia ada hanya menggunakan titik 2D (X, Y). Untuk menguji keberkesanan teknik yang dicadang, satu set data baru telah dibangunkan. Set data baru ini berbeza daripada set data yang sedia ada kerana ia menggabungkan dua ayat iaitu ayat positif dan negatif yang tidak terdapat dalam set data yang lain. Untuk pengkelasan dan pengenalpastian isyarat, empat metod yang merangkumi Mesin Vektor Sokongan (SVM), Ruang Rawak Keadaan Tersembunyi (HCRF), Model Markov Tersembunyi (HMM), dan Keputusan Hutan Secara Rawak (RDF) telah digunakan. Pendekatan yang dicadangkan diuji bagi isyarat yang melibatkan satu dan dua tangan dan dibandingkan dengan pendekatan lain dan memberikan ketepatan yang lebih baik. Pengenalpastian ketepatan pergerakan dan orientasi tangan bagi teknik yang dicadangkan adalah lebih tinggi daripada 99.1%. Teknik tersebut dapat memperoleh kadar pengenalpastian yang mencapai 99.9% apabila dibandingkan dengan dua set data yang lain.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 RESEARCH BACKGROUND

Standard sign languages are also known as deaf and dumb languages. Sign languages are gestural languages which contain a symbolic encoded message for communication without speech channel. They are unique in a way that they cannot be written similarly like the spoken language. They are natural languages developed and practiced in communities of deaf people around the world and varied from region to region. A sign consists of manual and non-manual components that somewhat occur in parallel but not perfectly synchronous. Manual components comprise of hand configuration, place of articulation, hand movement and hand orientation (Forster et al. 2013).

American Sign Language (ASL) is the primary means of communication for hearing impaired people. ASL is a visual language with complete grammatical structure, which uses the hand, facial and body gestures to convey meaning. ASL consists of vocabulary, grammar, and fingerspelling. Fingerspelling is a manual representation of the language that is spoken. In ASL, fingerspelling is used to represent proper names, titles, addresses, and words that have no sign representation. In general, fingerspelling is a process of spelling out English words where there is no sign equivalent in ASL (Kim et al. 2017).

Sign language, which is a kind of gestures, is one of the most natural means of exchanging information for most deaf people all over the globe. It is a visual and manual language made up of signs created with the hands, facial expressions, and body posture and movement (Tolba et al. 2013). It conveys ideas, information, and emotion with as much range, complexity, and versatility as spoken languages.

Computer science research in ASL recognition began in early 1990s. Hearing impaired people are facing some problems while communicating with others, especially with hearing people. However, they can communicate using sign languages, and hence there is a need for common and unified sign language. Most of the hearing impaired practice sign language of their own country, and their nativity and culture influence each sign language.

Signs can be static or dynamic; hence sign languages can also be categorized in terms of hand movements namely static gesture and dynamic gesture. As its name suggests, the static gesture category entails a person to keep his hand steady while performing a sign. Otherwise, such hand movement is categorized under dynamic gesture. The commonly practiced sign language is ASL. In ASL, the formation of all letters (except letter J and Z) for fingerspelling is categorized as a static gesture. Meanwhile, the formation of letter J and Z requires hand motion, thus, they are classified under dynamic gesture.

Static gestures are also called hand postures. Posture is a particular combination of hand position, orientation, and flexion observed at some time instance. Static gestures are not time-varying signals, so they can be completely analyzed using only one or a set of the hand images taken at a particular time. Meanwhile, the dynamic gesture is a sequence of postures connected by motions over a short time span. A gesture can be thought as a sequence of postures. In video signals, an individual frame defines the posture and the video sequence defines the gesture (Martins et al. 2015).

Gestures are a powerful means of communication among human. Among different modalities of the body, the hand gesture is the most simple and natural way of

communication mode. There are communication barriers between the deaf/mute and normal people, either because the signer (deaf/dumb people) is not able to hear or speak or the hearing people are not able to sign. This communication gap can cause an adverse impact on lives and human relationships. There are two traditional ways for the deaf and hearing individuals who do not know sign language to communicate with each other: through interpreters and text writing. However, both ways are ineffective. The cost of hiring an interpreter can be very expensive for daily conversations, and their involvement will result in a loss of privacy and independence of the deaf person. Text writing is not an efficient way to communicate because writing is a time-consuming task compared to both spoken and using sign language (Ben Jmaa et al. 2016). According to (Ghotkar & Kharate 2015), hearing impaired people can be included into the mainstream of society with the aid of human interpreter and computer interpreter. However, as compared to the number of deaf people, human interpreters are much less in number, and moreover, they may not be available all the time. Some professional interpreters can serve hearing-impaired people using real-time sign language interpretation, but the service is usually costly. Furthermore, such interpreters are often not available. Thus, a low-cost, more efficient way of enabling communication between hearing and deaf people is essential, and an ASLR recognition techniques are highly desirable.

In recent years, interest in developing automatic sign language recognition (ASLR) approach has been increasing, in order to enhance communication between normal hearing and deaf people. ASLR approaches mostly focus on the manual aspect of signs and recognize hand configurations including hand shape, position, orientation, and movement. The systematic change of the hand shapes produces different signs, which are well-defined in ASL dictionaries. Basically, there are three levels of sign language recognition: fingerspelling (alphabets), isolated words, and continuous sign language (sentences). Majority of the research available in the literature focused on the latter two.

In isolated sign language recognition (ISLR), a hand gesture is defined as a sequence of fixed starting and ending points as the sign boundary. On the contrary, there is no such explicit sign boundary in continuous sign language recognition (CSLR). In

CSLR, a sentence sequence consists of several true-sign sequences and non-sign sequences, also called movement epenthesis (ME), which connect the end position of the previous sign to the start position of the next sign. The main tasks of the CSLR are to split sentence sequence into true-sign sequences and ME sequences and to label each true-sign sequence with the right sign label. Many statistical methods or machine learning methods are proposed and employed to handle the crucial problems in ISLR and CSLR (Yang et al. 2016).

Although hearing impaired people communicate using sign language effortlessly among themselves, they face difficulties integrating into educational, social and work environment as a lot of hearing people do not possess any signing skills. In addition, most public services are not interpreted as sign due to expensive human translators and lack of translation tools. The purpose of any ASLR approach is to reduce, if not, eliminate these hurdles so that deaf people can lead better lives. Even though sign language recognition is a subset of gesture recognition, it is much more challenging because gestures can be defined in a way that common computer vision problems either do not occur or their impact is mitigated. Signs, however, are already defined by (Flasiński & Myśliński 2010).

- Segmentation: Hands and important facial features have to be extracted from images; these should not impose any restriction on signer or the environment.
- Spotting: Temporal sign boundaries have to be detected automatically.
- Co-articulation effect: A sign is affected by the preceding and the subsequent sign.
- Movements of the signer: Signers shift, rotate and perform other movements while signing; any good approach should be able to handle these movements.
- Signer independence: Both the signing speed and the trajectory of hands differ significantly for different signers. Even if the same person performs the same sign twice, small changes in speed and position of hands occur.
- Real-time: Processing a large amount of data, which is time-consuming for video-based methods requires efficient algorithms; algorithms for real-time processing have to be developed.

- Occlusion: While signing, some parts of the hand might be occluded; a good approach should be able to deal with occlusion issue.

In this thesis, the advent of low-cost depth cameras and high-performance sensing devices, such as the Microsoft Kinect sensor v2 for Windows has spurred renewed interest among many researchers in sign language recognition. Previously, the high cost, low reliability and complicated calibration issues with depth cameras have discouraged researchers from using them for sign language recognition. (Kumar et al. 2017a) has conducted experiments to show that the Microsoft Kinect can be viable as a hardware platform to collect data to perform ASL recognition. In sign language, there are many specificities related to sign performance such as hands shape, palms orientation, hands movement and non-manual signals such as facial expression and body expression. Microsoft Kinect motion capture allows the feature extraction of 25 joints per person. Unfortunately, this tool does not permit hand shape recognition. In this study, we intend to develop an approach that allows users to record and to add sign language specificities according to the 3D recorded motion Figure1.1

Figure1.1 Aim of the approch. Hearing impaired is making a sign, and the system outputs the corresponding word over the screen of the computer so that the normal hearing can understand him (Capilla 2012).

The proposed approach enables a real-time virtual agent animation from Kinect v2 data stream. Based on the 3D recorded scenes, an approach which is invariant to position changes is developed to recognize isolated and dynamic signs. Our methodology relies on the 3D motion analysis to extract the signature of the sign. 3D nonlinear regression is

applied to examine the 3D motion trajectory. The trajectories of both hands are fed to four statistical algorithms for classification and recognition task. Signs consist of hand shape, hand location, hand orientation, hand and arm movement and facial expression. In our study, the facial expression is not considered.

The most common approach for distance estimation is to use each pixel of the depth frame contains the Cartesian distance, in millimeters, from the camera plane to the nearest object at that particular (X,Y,Z) coordinate. The (X,Y,Z) coordinates of a depth frame do not represent physical units in the room; instead, they account for the position of a pixel in the depth frame. The range of movement test is performed for the two matched candidates with the lowest errors to increase the accuracy of pattern matching. The maximum and minimum X, Y, Z values of the captured gesture are compared to the X, Y, and Z range of the matched candidates. If the range of movement of the captured gesture is within a certain threshold of the top candidate, then a match is found. Otherwise, a comparison to the fields of the second candidate is made. Upon failing to find a match, user will be notified and requested to repeat the gesture.

1.2 PROBLEM STATEMENT

Many researchers tackled the problems in hand detection and position of fingertips, therefore, there are different types of problems appear from time to time such as skin region, extraction of the hand from the background, location, and shape of hands. Most of the hands tracking approaches today use some appropriate parts, the whole hand or parts like fingers or fingertips that match the input image (Chen et al. 2014). In addition, some signers might switch hands while performing a sign. This effectively introduces missing features in the information stream of the right and left hand. Its subsuming signs performed with left hand only and one for long gaps between signs (Forster et al. 2013). In many signs performed with both hands, the hands can be overlapped at some point. This overlap does not have to involve direct contact with the hands, because of the 2D projection of the camera. The same hand should be tracked before and after the crossing of the hands (Do & Yanai 2014).

Gesture segmentation is formulated as a multi-objective optimization problem, we can obtain the hand region based on its own color. As a result, formulating gesture segmentation as a multi-objective optimization problem can obtain the hand region, which cannot be obtained by the existing methods (Liu et al. 2017). Gesture segment of the same class, then the gesture is said to be recognized. However, if this also holds for a gesture segment of another class, the prediction is said to be incorrect. Otherwise the gesture is rated as undetected (Wu et al. 2016). Several approaches have been explored to deal with the problem of gesture segmentation, some of them simply based on the definition of a threshold value, while others are more complex. For continuous sign language recognition, one of the challenges is the absence of sign boundaries. The start and end of each sign have to be estimated in order to classify them into single sign. (Nayak et al 2012; Hernández et al. 2014).

Gestures are performed with one hand. Nowadays, the model includes an operator for hand preference to allow analyzing tasks assuming users will use either the left or the right hand. However, this operator is not enough to model two-handed interactions. For example, one hand could be used to perform gestures while the other hand is used as reference, or both hands could be used to perform gestures symmetrically or asymmetrically (Erazo & Pino 2017).

Several problems with sign language alphabets and numbers exist, where the background of a signer does not play a significant role in the grammatical distribution of fingerspelled words. Sign language is the primary means of communication in the deaf and dumb community. The problem arises when deaf or dumb people try to communicate with hearing people using sign language but hearing people are usually unaware of sign language. Consequently, it has been seen that communication of hearing-impaired people is only limited to his or her family member or deaf community (Pramada et al. 2013). A few types of research have been done on this issue, and some of them are still ongoing. However, it is still an open problem, and a full-fledged solution to it is essential.

Today, CSLR approaches are highly demanded; yet continue to be challenging at the same time. Recognizing a sign language gestures from continuous gestures is a very challenging research issue (Tripathi & Nandi 2015). Several studies have focused on the challenges of the real-time environment and camera location, for example, whether a hand can temporarily disappear or not. This is very significant in the case of starting and ending of a sentence when the hand is at an ideal location for continuous sign language. Related issues such as verification of ASL sentences and classifier-based grammatical structures remain relatively underexplored ASL recognition field (Gameiro et al. 2013). Therefore, research and proposal of ASLR approach development are very significant as it can serve a lot of benefits to mankind. ASLR approaches are not only for the hearing-impaired people but also for people without hearing impairments. There are some circumstances where people cannot talk to each other because of their working environment, for example, individuals in the laboratory are wearing face shields that hinder them to communicate verbally (Kim et al. 2017).

In the automatic sign language recognition ASLR known to be successful for recognizing sign language but require an expensive cost to be commercialized. Nowadays, researchers have gotten more attention for developing Sign Language Recognition that can be used commercially. Researchers do their researches in various ways. It starts from the data acquisition methods. The data acquisition method varies because of the cost needed for a good device, but cheap method is needed for the Sign Language Recognition approach to be commercialized. The methods used in developing Sign Language Recognition are also varied between researchers. Each method has its own strength compare to other methods and researchers are still using different methods in developing their own Sign Language Recognition. Each method also has its own limitations compared to other methods (Anderson et al 2017). There are a limited number of ordinary people who are completely able to communicate in sign language. Some people hire sign language interpreters to aid the communication between the deaf and normal people, but this is often difficult due to limited availability and the high cost of interpreters. These difficulties in communication can often lead to problems in the integration of deaf people into society and conflicts with an independent and self-determined lifestyle. Therefore, depending on

the context of an utterance, a specific component may or may not be required to interpret the sign (Koller et al. 2015).

1.3 RESEARCH QUESTION AND MOTIVATION

A number of research on sign language translation has been introduced before, and gesture recognition has always been an active research area for decades. A wide number of authors have proposed new approaches to solve this problem and almost all the time, they end up inventing complex implementations based on statistical descriptors, which only increase the computation complexity.

Hearing impaired people cannot hear, and most of them are unable to use written languages, having serious problems when expressing themselves in these languages or understanding written texts. They have problems with verb tenses and concordances of gender and number. In addition, they also have difficulties creating a mental image of abstract concepts. These facts cause them to face some difficulties when accessing information, education, job, social relationship, culture, and so on. They use sign language (their mother tongue) for communication and conveniently practice it using hand gesture. However, there are not enough sign language interpreters and ASLR techniques, thus, communicating with deaf people is challenging for non-sign language speakers (Edwards 2014; Mann et al. 2014)

This thesis aims to address the problem of sign language as the primary alternative communication method between hearing impaired people and hearing people. Several dictionaries of words or sentences have been defined to make this communication possible. It is hard for most people who are not familiar with sign language to communicate without an interpreter. Thus, a method that transcribes symbols in sign languages into plain text can help with real-time communication. This study intends to develop an automatic sign language translator, where a computer will display the corresponding word to a sign executed by a hearing impaired in front of a camera. A deaf person at one end could communicate with another person, who does not understand sign language. It is hard for

most people who are not familiar with sign language to communicate without an interpreter. Thus, a software that able to transcribe gesture in sign language into plain text can help with real-time communication, and it may also provide interactive training for people to learn sign language. An ideal gesture recognition technique should be able to translate sign into text effectively and accurately.

Gesture recognition has become an important research field with the current focus on interactive emotion recognition and hand gesture recognition (HGR). Conventionally, gesture recognition involves high-quality stereoscopic cameras, and complicated computer vision approaches to recognize hand signals; the approach softens turn out to be expensive and requires extensive setup. Microsoft Kinect provides an inexpensive and easy way for real-time user interaction. It unitizes color and depth sensor to capture color RGB images and the associated depth (distance) data. It also allows the development of algorithms that classify and perform recognition of the image data. The software driver released via Microsoft called Kinect Software Development Kit (SDK) with Application Programming Interfaces (API) offers access to raw sensor data streams as well as skeletal tracking. Although it consists of information of the joints between hand and arm, there is no specific for hand data available for gesture recognition (Network 2015).

Previous research on computer vision and hand detection have established a solid groundwork for gesture recognition. However, only a few Kinect based techniques were developed for HGR. 3D Hand gesture recognition, which is directly dependent on the accuracy of hand tracking, remains a challenging problem due to the hand's deformation, appearance similarity, high inter-finger occlusion and complex articulated motion. Meanwhile, 3D hand tracking is an interesting process from a theoretical point of view, as it deals with the following principal areas in computer vision: segmentation (of hand), detection (of hand parts), tracking (of hand), and occlusion handling (of hand Regions). The research questions in this study are as follows:

a. How to implement hand detection and fingertips detection in a robust manner?
b. How to improve a fast hand tracking in real-time?

c. How to solve the issue of the start and end position of a sentence in a continuous sign language problem?
d. How to develop a highly accurate HGR for a real-time application using machine learning or statistical method for classification?
e. How to create an ASLR approach that includes both the isolated and continuous sign language recognition and involves both hands?

1.4 RESEARCH OBJECTIVES

The three major objectives of this study are as follows:

1. To propose a new method for static sign language recognition which consists of hand detection, fingertips detection and gesture segmentation.

2. To propose an efficient method for dynamic sign language recognition which includes Efficient method for 3D movement trajectory of the hands and an ASLR approach for ASL recognition and interpretation.

3. To evaluate the performance of the sign classification of the raw dataset using four machine learning algorithms which are SVM, RDF, HMM, and HCRF.

1.5 CONTRIBUTION OF RESEARCH

There are several contributions of this thesis. Major contribution of this study is the development of Automatic Sign Language Recognition algorithms for both static and dynamic cases, where Microsoft Kinect sensor v2 and webcam are used to provide a dual mode of communication between hearing impaired and hearing people. Some techniques are proposed in order to detect, classify and recognize signs, which are performed with different velocities in real time. In this case, the problem of differentiating a set of gestures with changing velocities performed by one hand, as well as two hands in real time setting is able to be addressed.

Through the development of the ASL recognition technique, the following contributions in terms of new or improved techniques in the extraction, tracking, detection, and classification stage have been achieved:

1. A method to overcome the problem of fast hand and fingertips detection in static sign language.
2. A new algorithm for gesture segmentation to mark the start and end of the word from a continuous sign language performed by the deaf.
3. A method to overcome the problem of fast hand trajectory detection for both hands in continuous sign language recognition.
4. A method for hand gesture recognition that uses position based on 3D coordinates (X, Y, Z) for each hand that work robustly for both positive and negative sentences in continuous sign language recognition.
5. An improved method for sign language classification and also enhance performance recognition of isolated sign language and continuous sign language.

1.6 RESEARCH SCOPE

This section describes the scope of the proposed ASLR approach. This study used Microsoft Kinect sensor v2 for Windows for sign language recognition. Ten hearing impaired (Six males and Four females) from Community Service Centre for the Deaf (CSCD) are selected to perform signs of 230 dynamic alphabet, 20 static number, and 595 sentences. In this study, there are two different sets for training and testing: i) native who are deaf or mute and familiar with alphabet sign language and; ii) non-native or non-sign language users. The scope of the experiment limited to fingerspelling of letters (A-Z) and numbers ranging from 1-20.

Sign language can be different from country to country. In this research, for the interpreter, we focus only on ASL, since the data can be easily acquired from previous studies or the Internet. To maintain the simplicity of the interpreter and recognition technique, only one feature for both hands namely the 3D location of the hand centroid is

focused. This feature can be further processed to extract other information such as velocity, though it is not considered in this study.

In terms of dataset, the sentences are categorized into positive and negative sentence. The normal sentence can be considered as positive sentence, while negative sentence can be considered as a sentence with the inclusion of negative word such 'not,' 'dislike,' 'disagree' etc. Negative sentences have never been considered in previous research.

1.7 CONCEPTUAL FRAMEWORK

The framework of our proposed ASLR approach consists of three phases such as Static Hand sign language, Dynamic Hand gesture and Continuous sign language, where each phase solves the issues related to it as illustrated in Figure 1.2. The explanation on details for this framework in chapter III.

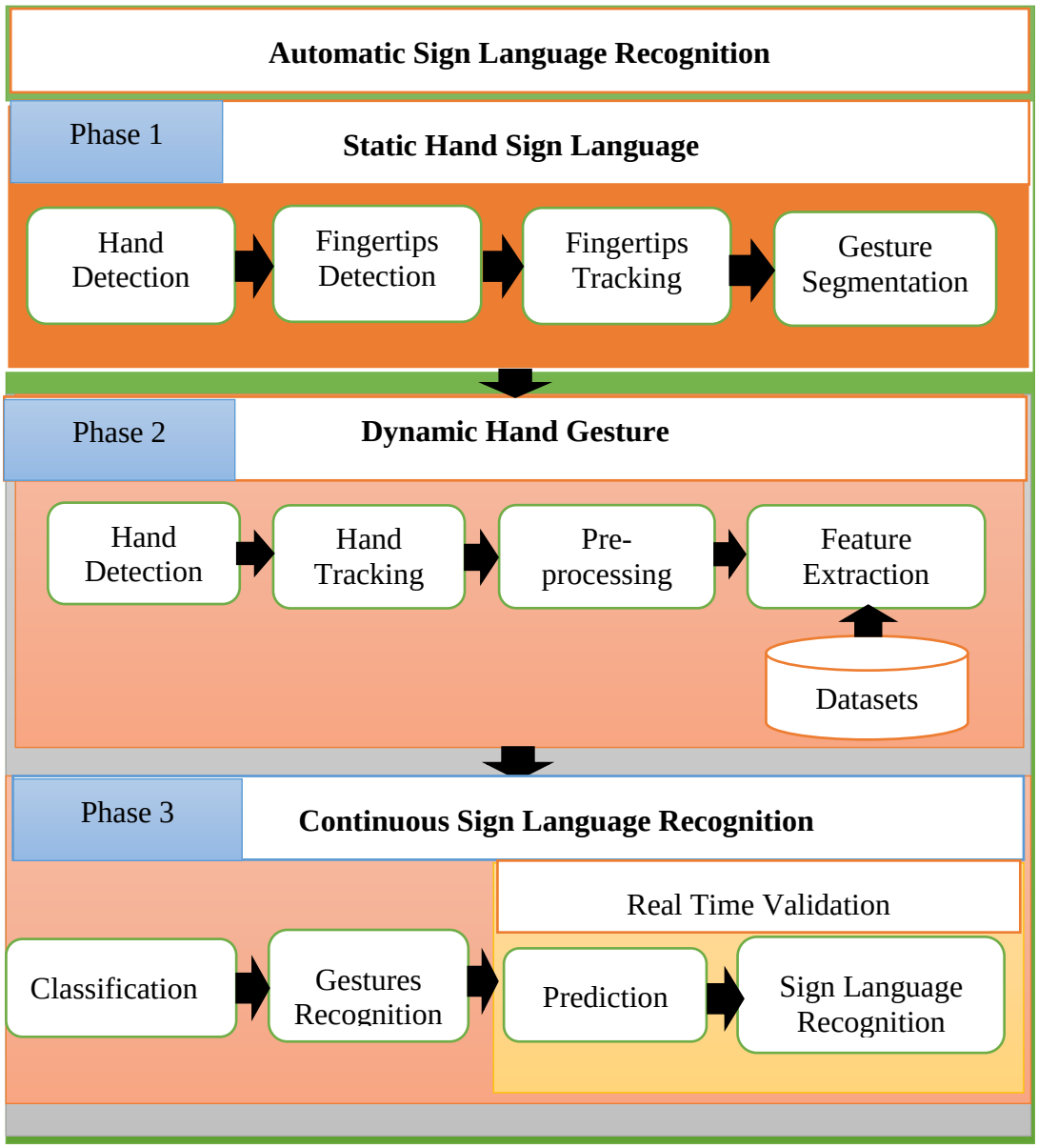


Figure 1.2 Framework of automatic sign language recognition approach is proposed

1.8 THESIS OUTLINE

This thesis consists of several chapters, and each chapter describes the conduct of the study methodically.

Chapter I presents a brief introduction to the ASLR modelling framework and the overall design of the approach, problem statement, objectives, motivation, research scope, and contribution.

Chapter II presents a comprehensive review of related studies. It consists of a literature review of previous and latest work on static, dynamic gesture recognition and continuous sign language recognition.

Chapter III explains the methodology of the study in detail. It also provides an introduction to basic techniques used in static gestures which are called hand postures, i.e., gestures, and dynamic gestures in SLR.

Chapter IV describes a likelihood modeling scheme for 2D static gesture methods. The focus of this chapter is the experimental result of hand detection and fingertips detection, as well as for fingertips tracking and gesture fingertips. Also, a selection of gesture segmentation algorithm and static gesture applied are presented.

Chapter V presents the experimental results for 3D dynamic gesture methods such as hand tracking method, hand gesture method and continuous sign language recognition.

Chapter VI explains the details of the evaluation measures and results analysis. The evaluation of experimental results using confusion matrices with four machine learning methods (SVM, RDF, HMM, and HCRF) is also presented in this chapter.

Chapter VII concludes the thesis. Discussion is presented, and the conclusion is drawn in this chapter. Last but not the least, it provides some suggestions for possible future work directions.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

Sign language is a structured gesture that is used by the hearing impaired community to communicate with each other or with hearing people. While sign language is a form of visual communication, hearing people use verbal communication and very few hearing people can communicate in sign language. Therefore, the need to improve a recognition system to bridge the communication gap between the hearing and hearing impaired communities is crucial, and for that reason, sign language recognition has been studied since 1990s. In general, sign language recognition can be categorized into isolated sign recognition and continuous sign recognition (Lim et al. 2016a).

Basically, sign language approach is developed based on five parameters which include hand recognition, hand orientation, hand movement, the shape of the hand and location of the hand (depending on the background). Among these parameters, hand orientation and hand movement in a particular direction are the most important. The method helps in recognizing the sign languages with better accuracy. The classified feature vectors are used for the training in machine learning. By advancing some of the algorithms in sign language recognition, better research outcomes can be achieved for human-computer interface. The primary tasks in sign language recognition are signer identification, hand shape extraction, hand position extraction and body posture of the signer to identification. These attributes can be used as an input in the sign language

recognition approach (Rao & Kishore 2017).

Hand gestures can also be classified into static and dynamic gestures. Static gestures, often called hand postures, are gestures where the hand is not moving during the gesturing period. Dynamic gestures, on the contrary, are gestures where the hand is in continuous motion during the gesturing period. Thus, static gestures rely mainly on shape and finger articulation while dynamic gestures rely on hand trajectories, scales, and orientations. Dynamic gestures may be seen as temporal sequences of static gestures. In the earlier days, static gestures were the primary focus of sign recognition research. However, dynamic gestures have also been actively studied since the last decades. Hand gestures can be considered as both an independent way of communication and also a complementary modality to speech. Gestures are both conscious and unconscious, and they are used in every aspect of human communication. Hearing impaired communication forms the basis of sign languages and uses natural media (Prabu et al. 2014).

Introduction to the low-cost depth sensors Kinect has made capturing of 3D data more convenient. The Kinect sensor comes with its associated Software Development Kit (SDK) library that enables the acquisition of 3D skeleton of the whole body and can be used in various applications including gaming, robotics, and gesture recognition. However, the device is not able to localize small body parts such as fingers or palm and the system is also limited to recognizing motion gestures like waving and movement such as up or down, left or right, and forward or backward. SDK developers have already provided interfaces that can be combined with machine learning algorithms during feature extraction to extract the hand and finger movements (Kumar et al. 2017b).

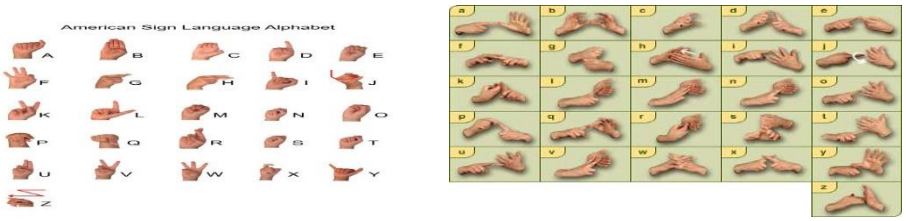
The research architecture involves the design and implementation of proposed framework and algorithms. In this thesis, algorithms for static gesture for sign language, dynamic gesture for sign language, continuous sign language and dataset have been designed and implemented. There is a limited number of hearing people who are completely able to communicate in sign language. Some people hire sign language interpreters to aid the communication between deaf and normal people, but this is often

difficult due to limited availability and the high cost of interpreters. The system is proposed not only for normal people to understand the hearing impaired but also for some particular scenario or working environment such as in the laboratories where people cannot talk to each other as they are required to put on face shields/mask which hinders them to communicate verbally. This research addresses the first step needed to make what could ultimately be a product for real-time ASL sign recognition and allows future researchers to continue the work towards a goal that would assist in resolving this communication barrier present in the world today. Various approaches presented in previous research is also presented in this chapter.

2.1.1 American Sign Language (ASL)

American Sign Language (ASL) is the predominant sign language for hearing impaired and mute communities in the United States. For decades, Christian missionaries have been spreading dialects of ASL and ASL-based creoles to several countries around the world, which include large regions in West Africa and some part of Southeast Asia. ASL is widely learned as a second language, serving as a unifying language to bring many existing sign languages in the world together and it is also closely related to French Sign Language (FSL). ASL has also been proposed on The Creole language, although ASL shows feature a typical of Creole languages, such as agglutinative morphology (Snow 2017). Similar to the origin of Auslan, ASL originated in the early nineteenth centuries at the American School of Deaf (ASD) in Hartford, Connecticut. It is estimated around 250,000–500,000 people practicing it, which is mainly by non-disabled children of hearing impaired and mute person. ASL is a distinct language and completely different from English. It contains all the essential features of language such as its own rules for pronunciation, complex grammar, and word order. As every language has its way of signaling different purposes, such as asking a question rather than creating a statement, languages differ in how they are to be expressed. For example, while asking a question, English speakers would raise the pitch of their voice, whereas, ASL users would raise their eyebrows, widening their eyes, and tilting their bodies forward. ASL today is considered as a very structured language. However, it is very interesting to note that even before its present form, sign language had

been practiced by various communities in the United States. In the United States, hearing families with hearing impaired or mute children have historically practiced hand signs as a way of communication and delivering information, which is different from the gestures used by nondisabled people in spoken conversations (Premarat 2014). Figure 2.1 depicts the fingerspelling alphabet.



(a) (b)

Figure 2.1 Fingerspelling Alphabet (a) One hand Alphabet, (b) Two hands Alphabet (Premarat 2014)

2.1.2 Sign Language Recognition System in Computer Vision

Research on hand gesture recognition can be categorized into two major categories: i) static gesture (hand posture) recognition and ii) gesture recognition (dynamic hand movement). Static gesture or hand posture recognition refers to an isolated single hand pose. On the other hand, gesture recognition or dynamic gesture recognition uses a variety of static gestures and hand movements, sometimes accompanied by body or facial movements. In ASL, letter J and Z involve dynamic (continuous) gestures, where the letter is drawn using a tiny finger (J) and index finger (Z), as shown in Figure 2.2. Gesture spotting is one of the issues in hand gesture recognition in controlled environments, where user has to differentiate the meaningful gestures from the independent ones. Meanwhile, in sign language recognition, the issue in continuous recognition problem includes the co-articulation, such that the previous sign affects the succeeding one which consequently complicates the recognition task as the transition between the signs should be explicitly modeled and incorporated in the recognition system. Moreover, language models are needed for the system to be able to perform on large-vocabulary databases (Liu & Wang 2017a).

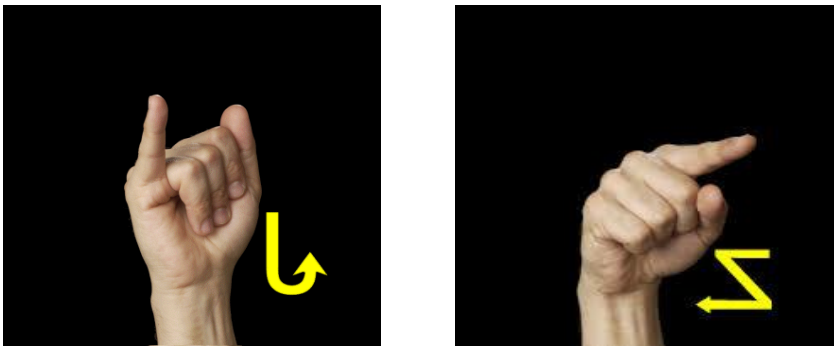


Figure 2.2 Gesture sequence for letter J and Z (Premarat 2014)

2.2 REVIEW OF TWO DIMENSIONS (2D) STATIC GESTURE FOR SIGN LANGUAGE ALGORITHM

Gestural interface based on vision skills is one of the developments in advanced man-machine interface. However, researchers in this field of study have been dealing with three major and significant issues: i) hand segmentation, ii) tracking, and iii) hand gesture recognition (feature extraction and classification). Hand gestures can be achieved by one or two hands in several of classes of gestures due to the ability of the hand to obtain a huge number of clearly discernible shapes, the fact of significance for the sign languages (Kumar et al. 2017a).

2.2.1 Hand Segmentation and Detection

Up to now, skin detection is one of the most challenging problems in image processing field and has been widely studied over the years. However, there is no satisfactory solution has been proposed so far. The existing techniques are based on the idea that the skin color can be successfully modeled in various color spaces, which in turn allows the segmentation of the skin regions. Although the color features constitute the primary source of information for skin detection, the value of color-based classification is unsatisfactory due to a significant overlap between the skin and non-skin pixels that appear in all of the commonly

used color spaces. Human skin color varies based on some factors such as variation of ethnic groups, changing of lighting conditions and complex background. Furthermore, the background objects often possess skin-like color, which results in the false positive errors in the segmentation outcome (Shaik et al. 2015).

Hand detection in computer vision field has been a topic of study for many years due to its apparent application in several important domains such as sign language recognition, gesture recognition, and human-computer interfaces. Accurate hand detection in images or video is still a challenging task due to the variation of hand appearance, and as hands do not have a fixed shape, it is hard to describe their form computationally. There are different algorithms have been proposed for hand segmentation and image classifier of two homogeneous parts, which are the foreground (image of the hand) and the background (the rest of the image). There are two types of background which are uniform and non-uniform background. It is essential to locate the hand successfully in the image with noisy and cluttered background as uniform background is possible only in lab-scale trials. The following algorithms are some of the most popular algorithms for hand segmentation. Hand detection can be referred as a procedure of locating the hand position of still image or sequence of images .i.e. moving images. In the case of moving sequences, it can be followed by tracking the hand in the scene, but this is more relevant in the applications such as sign language. The underlying concept of hand detection is that human eyes can detect objects which machines cannot, with much accuracy as that of a human (Palacios et al. 2013). There are several basic techniques related to hand segmentation and detection such as follow:

a. Thresholding

In this method, the image is divided into two regions of interest, i.e., background and foreground, based on a particular value known as thresholding value. Thresholding relies on different real-time parameters such as range/depth thresholding (Ren et al. 2011) and color $(RGB, HSV, YCbCr)$ thresholding (Rajesh et al. 2012). Otsu thresholding (Guo & Li 2011) proposed an Otsu thresholding which performs histogram shape-based image

thresholding or reduction of the gray level image into a binary image, assuming that the image contains two types of the pixel.

b. Skin Information

Human skin color can be used to separate the hand, head or body from the background using RGB, Grey, HSV and *YCbCr* color space representation (Shaik et al. 2015). It can be used to separate the foreground from the background concerning the predefined range of the color.

c. Subtraction

Test image can be separated from its background concerning a previously static image to find any changes in the test image (Lee & Park 2012). The static camera is useful in finding moving objects in videos. A robust background subtraction method should be capable of handling lighting or changing light condition and repetitive motion from clutter and long-term scene changes.

d. The distribution

The distribution of image color values affects the illumination. These values vary depending on cameras and lighting conditions. Color normalization allows object recognition methods based on RGB color, to compensate these variations (Sarkar et al. 2013).

e. Edge detection

Edge detection is an algorithm which attempts to capture the significant properties of the image of an object (Zhu et al. 2015). Photometrical Discontinuities is the geometrical and physical properties of the objects that contribute for such detection. These can be obtained by finding out the variations regarding discontinuities edges step, local extrema of line

edges and 2D features shape where at slightest two meets (junctions) in the gray level images. Edge detection helps to localize these differences and examines the reason behind. Some edge detection techniques can be classified as gradient and Laplacian-based edge detection (Chandrakar & Bhonsle 2012). The gradient technique computes the first derivative of the image for the edges by finding the minimum and maximum value, whereas the Laplacian method decides the zero intersections in the second subordinate of the image to detect edges. Meanwhile, some researchers have classified the edge detectors as follow (Shrivakshan & Chandrasekar 2012).

f. Morphological operations

Morphological filtering is a broad set of nonlinear image processing operations that can be used to process images based on shapes. These operations apply structuring element of different shapes on an input image. The output image usually retains its original size. The structural element denotes the size of the window that would operate in a neighborhood of a pixel to create the output (Hasan & Ko 2016). The size and shape of the neighborhood can be chosen to construct a morphological operation that is sensitive to a particular shape (s) in the input image. Extraction of edge or contour information has been successfully achieved, in which erosion and image subtraction have been applied. However, the drawback of this algorithm is that it is a time-consuming technique.

Meanwhile, Jalilian et al. (2013) propose a method for hand and face segmentation that helps to create a better vision based sign language recognition system. The method proposed is based on $YCbCr$ color space and single Gaussian model. It detects regions of face and hands in an image with complex background and non-uniform illumination. Hand and face segmentation and detection reduce processing time and increase the precision of recognizing postures in sign language recognition systems. One of the biggest challenges in vision-based approaches is the segmentation of face and hand with a complex background and non-uniform illumination. The background color resembles the skin color or black pixels on the hand region. These noises may be caused by the lighting conditions or existing pixels similar to skin pixels regions.

Nayak et al. (2013) proposed a new method for hand segmentation using different color space models with the application of required morphological processing. Challenges such as skin color location, cluttered background evacuation, and lighting conditions are found to be viable dealt with this system. The segmented image noise due to dynamic background can be removed with the help of this adaptive technique. The segmented color result is to make sure that the background will not have the same color as the object or ROI so that the final output will contain less noise. However, the system cannot get a proper gesture input.

In addition, Rasim et al. (2013) proposed a simple texture analysis method in a supplement to skin detection that allows the increment of the confidence of finding a piece of hand skin in an image. It is developed using adaptive skin tone segmentation and identification of image of hand key points specifically. The details of approximately skin detector implementations are provided in the study. Application of different feature detectors and hand detection descriptors is also considered. This approach is less dependent on the shape of skin region locus in color space for illumination component as it is not utilized for the development of the skin tone model. The overlapping of skin and non-skin colors are also taken into account.

Stergiopoulou et al. (2014) presented a vision-based gesture recognition system which applied hand segmentation and detection. Segmentation can be a particularly challenging task when it comes to varying illumination and complex background. In such environments, most hand detection methods fail to obtain the exact region of the hand shape, especially in cases of dynamic gestures. Meeting these requirements becomes even more challenging due to real-time operation demand. Ideally, the hand should not merge with the background, and the objects consist of at least one region. In this method, there is some noise when it is applied to cluttered environment.

A new method based on $YCbCr$ color space and K-means clustering algorithm for hand gesture segmentation has been proposed by Qiu-yu et al. (2015) Firstly, the image RGB color space is converted into $YCbCr$ color space. The image $YCbCr$ color space is

then distributed into luminance Y and chrominance Cb and Cr. The binary image is achieved by clustering the values of chrominance using a k-means clustering algorithm and finally, hand gesture segmentation is completed by conducting the morphological process on the obtained binary image. $YCbCr$ Color space transformed from RGB Color space is not suitable for clustering algorithm because it is hard to digitize the details as RGB Color space combines hue, luminance, and saturation and it is not appropriate to be clustered as well.

Meanwhile, Zou et al. (2016) describe hand detection as a challenging task in hand gesture recognition system, and the detection results can be easily affected by changes in hand shapes, viewpoints, lightings or complex backgrounds. Localizing human hands in static images against complex backgrounds, especially one hand gesture, is one of the most challenging problems of vision-based static hand gesture recognition.

In most recent study, Betancourt et al. (2017) proposed an algorithm for segmenting hand as a foreground and background problem. The study disregards two significant facts: i) hands are not a single "skin-like" moving element but a pair of interacting cooperative objects and ii) close hand interactions may lead to hand-to-hand occlusions and as a consequence, creates a single hand-like segment. These facts complicate hand movements and need proper understanding. However, it is intuitive to assume that in the presence of two important hand-like segments, they cannot be at the same time left or right hand. Due to the fast camera and hand movements, not all the pixels inside the occluded hand-segment is able to be solved this way. Table 2.1 shows a summary of hand segmentation methods. The limitation of researchers works show some of the drawbacks for hand segmentation, desired segmentation result could not be achieved well.

Table 2.1 Summary of hand segmentation methods in the literature

Author and years	Technique	Background	Limitation
Jalilian et al. 2013	*YCbCr*	Complex	Bad lighting condition may cause noise
Nayak et al. 2013	*YCbCr* , HSV	Complex	The system performs segmentation poorly
Rasim et al. 2013	Thresholding	Uniform	Illumination component is not used for the development of skin tone model
Stergiopoulou et al. 2014	HSV	Complex	The hand merges with the background.
Qiu-yu et al. 2015	*YCbCr*	Complex	RGB color space is not suitable for clustering algorithm
Zou et al. 2016	Thresholding	Complex	Used single hand only
Betancourt et al. 2017	-------	Complex	The occluded hand-segment can be solved but not all the pixels included

2.2.2 Fingertips Detection

The main objective of this research is to improve the recognition rate and the accuracy for fingertips detection. The detection of fingertips and finger types is significant in some applications such as sign language, vision-based finger guessing games, applications related to real-time systems and virtual reality, and pointing gestures recognition in the context of human-robot interaction (Bhutan et al. 2012). Although there a lot of research that has been conducted on the detection of finger and finger types, there are still some gap and challenging issues that need to be addressed. Thus, research on fingertips detection is still significant, especially, in a situation where silent communication is preferred. For example, hand gesture recognition is a great help to the communication between deaf people and normal people.

Park et al. (2013) Presented a new method for finger-gesture-based mobile client interface (UIs). The proposed method utilized edge detection for fingertip detection, which is flexible to changes of lighting condition of mobile UIs based on finger gestures. Canny edge detection algorithm is applied to overcome the shortcoming of some common

algorithms in dealing with lighting condition.

Meanwhile, Tsagaris & Manitsaris (2013) compared the RGB, RG normalized, HSV and $YCbCr$ Color space in terms of detection of the fingertip position and gesture recognition accuracy. The research aims to find the most appropriate color space for finger gesture recognition based on skin detection. The hands of the users are in a semi-closed pose with a distance of 15 cm from the optical camera and with a maximum rotation of 10°. The skin model is not perfect; the system considers the small region of the background as foreground and vice versa. After the detection, some pixels do not correspond to dermal pixels.

According to Alam et al. (2014) fingertips positions in hand images are crucial for creating a human-computer interaction framework. Fingertips detection algorithm of a hand image analyzed the concept of the primary geometrical data of fingers. The proposed method includes three parts. Firstly, the hand image is segmented in order to locate the hand region. Secondly, invariant highlights (curvature zero-crossing points) are extracted from the boundary of the hand and finally, fingertips are recognized. The drawback of the method is that it cannot detect fingertips in the case of numerous hands in one image or if the fingertip is covered with the skin colored object.

Brancati et al. (2015) presented a new method for fingertip and hand detection on various backgrounds. The method automatically selects between color and depth features according to the illumination condition, allowing it to detect the hand and then the user's fingertip robustly, for both indoor and outdoor scenarios. The special requirements for considering mobile applications (real-time processing, low-complexity, robustness, etc.), such gesture recognition technologies are not easy to apply to wearable AR systems. Most skin color based approaches do not classify background objects in color similar to the user's hand skin. Therefore, choosing the most appropriate features according to the environmental conditions is significant.

Chen et al. (2016) presented a new approach to modify hand palm for fingertips detection. In this method, two-dimensional Gaussian probability distribution in *YCbCr* color space is introduced to segment the hand region from the background. Harris corner detection is applied to detect the feature points in the binary image. With prior information of feature points, it utilized global optimization approach of the centroid and convex hull analysis to locate fingertips. The method, however, should modify the value of threshold to increase the accuracy of the detection.

In another research, Tan et al. (2016) proposed a hybrid convex hull algorithm to reduce the computational cost of fingertips detection. The hybrid convex hull algorithm requires only four global extreme points. Thus the preprocessing step is simpler. An additional process is added to create a polygon monotonically in both horizontal and vertical directions before the concavity tree technique is applied. However, this modified algorithm sometimes possess some issues with a convex polygon. Different research has been used various methods to provide for fingertips detection by employed different techniques also reach to different accuracy. From the comparison between techniques of researchers, it was observed that fingertips detection still have some limitation how is show in Table 2.2 depicts a summary of the fingertips detections.

Table 2.2 Summary of literature for fingertips detection

Author and years	**Method**	**Feature and accuracy**	**Disadvantage**
Park et al. 2013	Adopts edge detection	Haar-like, 96.4 %	In this method not used all fingers in hand but them using only one finger in images.
Tsagaris & Manitsaris 2013	Tested with the PianOrasis	Low-cost optical camera. 89,86%	After the skin detection, some pixels that do not correspond to dermal pixels
Alam et al.2013	Geometrical structural information	curvature zero-crossing points	whether a consecutive pair of zero-crossing points holding a fingertip or not
Brancati et al. 2015	Colour and depth features	Touchless point-and-click interface,	The RGB information being considered not reliable. Choosing the most appropriate according to the environmental conditions.

To be Continued…

…Continuation

Chen et al. 2016	Two-dimensional Gaussian	Harris corner detection	Should modify the value of threshold to be confident.
Tan et al. 2016	Bresenham algorithm embedded,	Convex hull algorithm	This modified algorithm does not always work with the convex polygon.

2.2.3 Fingertips Tracking

Human-Computer Interaction (HCI) is a field in which developers develop a user-friendly system. The user can interact with a computer system without using any conventional peripheral devices. For fingertips tracking, the proposed method modifies the randomized circle detection method presented in (Jia et al. 2011). Fingertips tracking is an active area of research in computer vision field as one of the important hand features is the position of the fingertips.

Fogelton (2011) Proposed static initialization algorithm to process a frame using histogram back projection of the skin color before applying 2D hand tracking with flocks of features (FoF). This modification provides better results with a lower false positive error. The object being tracked for individual features can latch onto arbitrary artifacts, such as fingers. Their movement is free along with the artifact, without exasperating other features. In this work, some strict conditions are applied such as the captured image must possess the same skin color characteristics, lighting condition has to be consistent, and the individual has to be of a particular skin color. Hence, dark people do not have the same run of edges as fair individuals. These set conditions are considered as limitations and thus, the proposed method is not robust. In addition, the background is not invariant, and it fails to track human hand during various movements when the gesture is not predefined.

Do et al. (2011) presented a tracking technique of fingertip by combining mean-shift and particle filter. Morphological operators are applied to the segmented image in order to remove noise. Image gradients are calculated based on different scales. One of the method problems is the error on the fingertip localization increases across time. Unfortunately, curves around the fingertips are not always perfect circular arcs features.

The tracking procedure fails if the finger lost, hence, the hand movements should be slow for the finger to be tracked.

In another research, Kasprzak et al. (2012) suggested a method which includes capturing color image sequences, recognizing static hand posture letters and interpreting sequences in terms of gestures (i.e., "words"). These conveyances are subtracted one from the other and length-normalized, and all negative values are reset to zero. The tracking of the hand posture in a brief sequence permits detecting "modified poses" like diacritic letters in national alphabets. The static hand pose set corresponds to hand sign of a thumb alphabet. The fingers are detected based on the position of fingertips with regard to the projection point of the center of mass onto the palm rectangle side, opposite to the tips. However, the fingers are not separated correctly from each other. The inner contour is not sufficiently and "deeply" move between the fingers, leading to insignificant local maxima of the two-contour-difference function.

From the comparison between three techniques it was observed the fingertips tracking shown in Table 2.3 still have some of the gap for static images to show the good result for fingertips tracking. Therefore some of the method using one finger for tracking. However, this research study still has some limitation for research. Table 2.3 shows a summary of fingertips tracking methods.

Table 2.3 Summary of literature for fingertips tracking

Author and year	Method	Feature extraction	Disadvantage
Fogelton 2011	histogram back projection	flocks of features	does not background invariant also not able to track human hand
Martin et al. 2011	mean-shift	skin color segmentation is performed	observe that the error on the fingertip localization increases
Kasprzak et al. 2012	a double-active contour-based	negative gradient	The fingers are not correct separated from each other.

2.2.4 Gesture Segmentation

Movement epenthesis refers to the transition segment which connects two adjacent signs. This is formed when the hands move from the ending location of one sign to the starting location of another sign and do not carry any information on the signs. Linguistic studies of movement epenthesis in the literature are limited, and there is no well-defined lexicon for movement epenthesis. On another hand, it is shown that movement epenthesis had no phonological representation. Though movement epenthesis may not carry meaning, it can have a significant effect on computer recognition of continuously signed sentences as the transition period of the segment can even be as long as a sign segment. Therefore, problems related to movement epenthesis need to be addressed explicitly for robust sign language recognition (Elakkiya & Selvamani 2015).

Hand gesture segmentation is the major premise to hand gesture analysis and identification. The research of segmenting gesture accurately from a complex background in a sequence of images is significant for the understanding of human gestures. The effective use of various information such as color, motion, geometric information is the main key to the study. The task of gesture segmentation is to determine, from continuous data, the starting and ending of a communicative gesture. Gesture segmentation can be referred as a separation of hand gestures from a continuous image sequence containing gesture. Gesture segmentation can be a complicated task due to gesture variation in the duration of both static and dynamic gesture. There are several methods that have been proposed to overcome the issues in gesture segmentation (Athavale & Deshmukh 2014). Some of the proposed methods are simply based on the description of the threshold value for gesture data, while other methods are more complex (Rajaganapathy et al. 2015). In addition, some studies suggested to separate the pixels in the hand gesture region from the background on a particular image by setting the threshold in RGB color space.

Tao et al. (2013) proposed a framework for recognition and joint segmentation of surgical gestures from kinematic and frame sequences. In comparison to other previous works that relied on either frame-level kinematic cues or segment-level kinematic, the

method exploits both signals by using a combined Markov or semi-Markov conditional random field (MsM-CRF) model. However, this is possible only because the temporal segmentation of the video into gestures is assumed to be known. The gestures sequence is obtained by minimizing the energy of the CRF.

Madeo et al. (2013) presented an initiative for segmenting gesture phases which focusing on the gesture segmentation units, that is, classifying frames within a video in rest position or gesture, which represents the first step towards the segmentation. It uses Support Vector Machines (SVM) and applies several pre-processing methods to extract time-domain and frequency-domain to accomplish this task. Time-frequency domain features aims at benefiting from temporal aspects of the problem. Some issues related to gestures are not well-defined in linguistics researchers, such as if it should analyze the two hands movements as a unique information unit or if each hand produces distinct information. It decides to consider only previous frames since this approach requires a smaller window and results.

Meanwhile, Ju et al. (2014) suggested refining the edge of the tracked hand gesture, which is used to segment the RGB image of the hand gestures by applying a modified Expectation-Maximisation (EM) algorithm based on Bayesian networks. Precise calibration is adequate for human body movement analysis or casual use, but it lacks accuracy in hand gesture segmentation and recognition. Hand gesture segmentation is evaluated based on the proposed alignment method. Improvements have been achieved, and segmentation results of hand gesture "five" are shown with the comparison between the official calibrations. However, the finger color cannot be recognized, and the mismatch between the depth and color images is huge.

Favorskaya et al. (2015) suggested a method based on two-level different classifiers consist of the trajectory classifiers in selected time instants for the posture classifiers of sub-gestures. The direction classifiers contain skin detector, normalized skeleton outline of one or two hands, and development history speaking to by movement vectors normalized through foreordained bearings. As a result, a background

model can be constructed from the hand. An accurate background model is not required for hand capturing meaning that there is a possibility of using online background model. The hand motions are extracted using a background subtraction model. Then, the hand boundaries are computed based on the detection and merger of contour components in order to obtain the closed boundaries of hand motions. However, not all gestures may have a full description because some fingers may be invisible.

Elakkiya & Selvamani (2015) proposed sign recognition technique that does not require explicit modeling of movement epenthesis. Gesture base of signs is to be recognized by the proposed system and not the type of movement epenthesis. The classic Level Building algorithm (CLB) is enhanced to meet the above criteria as well as to match the continuous sign sentences without accurate movement epenthesis models. This system utilizes both motion and skin cues to recognize gestures in the cluttered background, and it also uses a set of keyframes to shape the environment. The detection algorithm exploits the fact that the hand is changing its orientation faster than other parts during the signing. To create the model base during training, it needs the sign keyframes in continuous sentences without considering its associated movement epenthesis. The system is flexible since it does not use any constraint to model the hand as well as background environment or motion information.

Wu et al. (2016) presented a method called Deep Dynamic Neural Networks (DDNN) for multimodal gesture recognition. A hierarchical dynamic structure based on a Hidden Markov Model (HMM) is suggested for gesture segmentation and recognition, where skeleton joint information, RGB images, and depth are the multimodal input observations. The modeling and learning of the HMM required to conclude the gesture sequence is achieved. The gestures must be detected in a period of continuous data stream even the test sequences are not segmented in advanced. Some of the gestures differ primarily in hand pose but not in the arm movement.

Recently, Liu et al. (2017b) presented a multi-objective optimization model to tackle some issues in gesture segmentation. An algorithm for solving a two-phase

distribution based on estimation system is presented in the study. During the development of the model, the position of a series of pixels are considered as the decision variable, and the variances between the pixels colors and those of a hand are considered as objective functions. The gesture segmentation technique based on a two-phase estimation of distribution algorithm is presented according to the correlation among the positions of pixels. The method divides the solution of the problem into two phases based on evolutionary optimization and uses a different estimation of distribution algorithms in various stages. The dataset used in the study does not take advantage of the correlation among the positions of hand pixels to generate a temporary population.

There are different researchers for improve the gesture segmentation but still have some limitation for them works show in Table 2.4. Several methods for modifying the result and accuracy, furthermore many of researchers reach to good accuracy how mention above but still have gap such as how they know the start and end of each word in a sentence, however, this research needs to be improved. Table 2.4 shows a summary of gesture segmentation methods.

Table 2.4 Summary of review of gesture segmentation

Author and years	**Algorithm**	**Accuracy**	**Limitation**
Tao et al. 2013	Joint segmentation + HMM	-------	Limited result after modification with kinematic data
Madeo et al. 2013	Pre-defined set + SVM	68.8%	Consider only previous frames since this approach requires a smaller window.
Zhaojie et al. 2014	Expectation-Maximisation (EM)	-----	The cameras on the NAO are not suitable to recognize human hand gestures
Favorskaya et al. 2015	Trajectory classifiers	91%	Some fingers may be invisible
Elakkiya et al. 2015	Boosting-based framework	99.6%	Does not explicitly model movement epenthesis and the demand on annotated training video data is little.

To be Continued …

…Continuation

Wu et al. 2016	Deep Dynamic Neural Networks (DDNN)	81%	RGB-D features allow a reliable recognition but not useful for segmentation.
Liu et al. 2017b	Two-phase estimation of distribution	92.24%.	The dataset used does not take advantage of the correlation among the positions of hand pixels to generate a temporary population.

2.3 THREE DIMENSIONS (3D) DYNAMIC GESTURE FOR SIGN LANGUAGE RECOGNITION

The majority of the research on development of sign language recognition (SLR) systems have been carried out using single depth sensor, i.e., Kinect sensor. The Kinect sensor comes with its associated Software Development Kit (SDK) library that enables the acquisition of 3D skeleton of the whole body and has been used in various applications including gaming, robotics, and gesture recognition. This section discusses 3D dynamic gesture for sign language interpretation and recognition system in details, which includes hand detection and tracking and hand gesture recognition in details. An explicit description on Microsoft Kinect sensor v2 for Windows is also included in this section.

2.3.1 Sources of Kinect Data

There are three sources of Kinect data: color sensor, depth sensor, and an infrared sensor. The data is collected by subscribing to occasions when new data is accessible.

a. Color Sensor

The color sensor provides image data from the Kinect's RGB camera. The RGB camera captures color information with a resolution of 1920 x 1080 pixels. Image acquisition can be carried out with an outline rate up to 30 Hz. The final feature to be specified in the field of view for profundity detecting of 70 degrees horizontally and 60 degrees vertically. To empower the utilization of the sensor for designers and analysts, an official Microsoft SDK

2.0 is available to be downloaded. It provides set of capacities or code tests that can be utilized for related purposes. It also offers assistance in scenarios with little light or brightly lit scene and allows the adjustment of hue, brightness or differentiation to improve visual clarity. Furthermore, the color stream can be utilized as an infrared stream by setting the color image to infrared format.

b. Depth Sensor

The depth sensor provides depth data of almost each point seen by the Kinect's infrared recipient. The depth stream contains data from the Kinect's depth camera (Khoshelham & Elberink 2012). It is represented as an outline made up of pixels that incorporate the separation of the camera plane to the closest object as described in Figure 2.3.

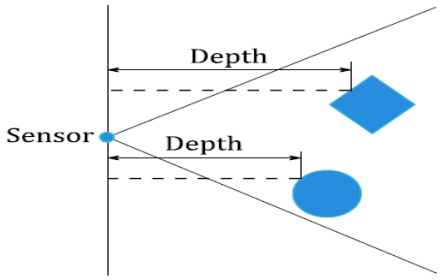


Figure 2.3 An illustration of the depth stream values (Altman 2013)

The depth frame is presented in different resolutions. The maximum resolution is 640x 480 pixels. Resolution of 320 x 240 and 80 x 60 pixels are also available. Depth frames are captured at 30 frames per seconds for all resolutions, and the depth camera can acquire in two extended modes, default and close mode. If the run mode is set to default, the sensor captures depth values from 0.8 meters to 4.0 meters. If the range mode is configured to approach value, it captures depth values from 0.5 meter to 4.5 meters. According to the description of depth, space run portrayed in (Kirac et al. 2014), the maximum captured depth value can be up to 8.0 meters in both range modes. However, the quality of depth value surpassing a limit value of 4.0 meters in default mode and 3.0 meters in close mode may be degraded with distance.

c. Infrared Sensor

This source provides image data from the Kinect's infrared camera sensor. The corners of the board can be easily detected in the captured infrared images within sub-pixel precision (Caruso et al. 2017). The Kinect sensor v2 for Windows contains a TOF camera and determines the depth by measuring the time emitted light taken from the camera to the object and back. Therefore, it continuously emits infrared light with modulated waves and detects the shifted phase of the returning light (Caruso et al. 2017). In comparison to triangulation methods in passive or active stereo vision, the ToF camera is very compact, where the illumination (pattern projector) is placed just next to the camera lens, whereas other systems need a certain minimum baseline. However, the accuracy of TOF measurements is limited by several factors, such as the properties of the target and the intensity of background illumination (Luna et al. 2017). An individual body possesses depth information, where it is discoverable for the Kinect's field of view. Once a person is found, the 3D position of joints is assembled up and become the source of data. This is an essential source for this venture, as it offers a significant level of reflection. The body source also provides additional information, where the center of this source is not body joints, but to distinguish the part of people's bodies from the background.

2.3.2 Hand Detection and Tracking

Hand tracking for SLR is a difficult task since sign gestures are not designed for automatic recognition purposes, and they comprise of quick hand movements and occluded body parts. The latest Microsoft Kinect sensor is capable of identifying and tracking both hands. Many previous methods follow the same approach as Microsoft's but apply it to detailed hand recognition instead of the body.

Palacios et al. (2013) presented a new method for hand gesture recognition based on an RGB-D sensor. The proposed method takes advantage of depth data in order to cope with the most common issues of conventional video-based hand division strategies. The method put away the 3D position of the palm center of the hand. For each outline of a

video sequence, user must not hold her/his hand closer than a particular distance, or otherwise, the system could not return any estimation. The hand can be the front-most portion of the cluster but, generally, it cannot be isolated from the arm as it is with depth data.

Meanwhile, Czarnuch & Mihailid (2014) presented the improvement and assessment of a hand tracking method based on single profundity images captured from an overhead viewpoint unitized in COACH framework. Random Decision Forest (RDF) is applied where a body part is classified using approximately 5,000 manually labeled, unbalanced and partially labeled training images. Existing depth-based body part, pose, and posture tracking methodologies are not suitable for this application as frame-by-frame tracking from a fixed overhead perspective using only depth images have not been attempted.

In the same year, Joo et al. (2014) presented a hand detection and tracking method based on real-time depth information. The method used a classifier that combines a boosting and a cascade structure for hand detection. The classifier uses depth-difference features for learning. Each candidate segment features are to be computed by subtracting the depth values midpoints of subblocks from the central depth value of the segment. The focalized point may drop in the hand region Ψ, though it seldom happens. It happens because border pixels are included in the preparation of tracking point decision.

Yang (2014) employed CRF to detect candidate segments of signs based on hand movement. BoostMap algorithm is applied to confirm the hand shapes of the segmented signs. The face and hand positions are vigorously identified using the tracking function in Microsoft Kinect SDK. The skeletal comprises of 10 feature points, where six and one features are extracted in 3D and 2D space, respectively, from the detected hand and face. The primary objective of the hand shape-based sign confirmation is to decide whether or not a sign spotted by the H-CRF should be accepted as a sign. However, this method has trouble recognizing different hand shapes.

In another research, Kim et al. (2015) proposed a hand gesture control system employing immersive tactile feedback to user's hand. In this study, the modified local binary pattern (MLBP) is suggested to analyze 3D shapes in-depth images efficiently. LBP is the pattern of features, also called as texture descriptor, intensively used for classification of grayscale images. Since depth image does not contain texture and color information, it is challenging to detect and search an object without such information.

Karbasi et al. (2015) presented a distancing technique for hand detection based on depth image information obtained from Kinect sensor. The monent, background subtraction and iterative method for shadow removal are combined to remove noise from the profundity images. Microsoft SDK is applied for the extraction process. In this method, the two hands could be segmented on a diverse color in particular distance.

Meanwhile, Liu et al. (2015a) introduced a method for tracking and gesture trajectory recognition based on synchronized color and depth video. In the aspect of hand tracking, a joint observation model with the hand cues of skin saliency, motion and depth is integrated into particle filter in order to move particles to a local peak in the likelihood. However, the complete rotation invariance impedes expressing the original meaning of the gesture. Hence, it should be emphasized that the complete rotation invariance for the gesture trajectory shape is not suitable for gesture recognition.

In addition, Wu et al. (2015) presented an object tracking on depth image sequences which crop 3D motion trajectories to solve many traditional problems such as illumination changes and projective distortion. The study shows that the fused tracker is competent for accurate 3D hand tracking even under very complex scene. The motion-based tracker is not affected by nearby motionless objects, but it is susceptible to moving distractors. However, the areas near the target are not pruned, causing it difficult to distinguish whether they belong to the target or not.

Wu et al. (2016) proposed static and dynamic gesture recognition method using several algorithms and classifier. Background subtraction algorithm is used to remove

invalid gestures. Next, an adaptive square is applied to find the hand region. Once the hand region is obtained, the hand features are extracted and the static hand pose is classified using Support Vector Machine (SVM). However, the point created by the hand tracking is not persistently in the center of the hand. The system cannot utilize the stature of palm to obtain the square with its center at the hand tracking point. There are four white regions which are the convex hull comprised in the contour. The system also cannot use a fixed-size rectangle to extract the hand region because the point generated by the hand tracking is not always in the middle of the hand, so the system cannot directly use the height of palm to obtain the square with its center at the hand tracking point.

In a recent study, Kim et al. (2017) used a real-time 3D information from sensors such TOF and Kinect depth sensor for hand detection and tracking. As a 2D image based feature extraction methods such as LBP utilized texture information, they cannot be connected to depth or extend image, as it has no such information. It affects the performance of the system since the trajectory produces noise to the depth image. Therefore, the trajectory of hand tracking should be filtered so that it will not be affected by the small movement of the hand. This work cannot be used for other depth sensors except Kinect.

The summary of hand detection and tracking method using Kinect is described above is presented in Table 2.5. From the comparison of the techniques for improved the hand tracking was some limitations such as in Kinect camera work depend on depth camera for calculated the position of one or two hands front the body and also the Kinect is working depend on skeleton and joints of the human.

Table 2.5 Summary of hand detection and tracking method using Kinect

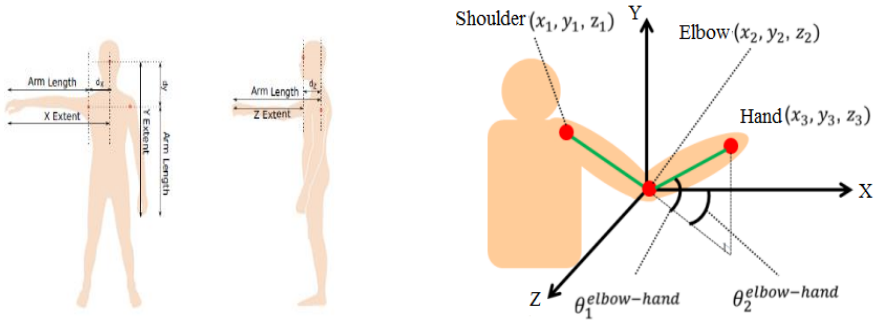
Research	Method	Feature Extraction	Limitation
Palacios et al. 2013	Six dynamic gestures	Based decision tree and the convexity defects	User must not hold her/his hand closer than a particular distance, or otherwise, the system could not return any estimation.
Czarnuch & Mihailidis 2014	COACH framework and RDF	Invalid pixels by depth value	Tracking methodologies are not appropriate for this application
Joo, et al. 2014	DAM-Shift algorithm	Depth-difference	The converged point may not fall in the hand region
Liu et al. 2015a	Based on synchronized color and depth video	Depth feature, salient hand skin feature, and motion feature	The gesture trajectory shape is not appropriate for gesture recognition.
Wu et al. 2015	Depth Mean Shift Tracker	Depth values of depth images	Areas near the target are not pruned, causing it difficult to distinguish whether they belong to the target or not
Wu et al. 2016	Background subtraction, adaptive square, and SVM	The contour of the hand	The system cannot use a fixed-size rectangle to extract the region of the hand
Kim et al. 2017	LBP	Depth images	Trajectory produces noise to the depth image

a. Feature Extraction

Feature extraction is a process that attempts to extract certain measurable inputs that can be used to classify a section of a signal. If the isolated part of an image contains what human interpret as a hand sign with a ‘thumbs up’ gesture, then it is important to extract information that would make this gesture unique compared to other gestures. The success of any classification task relies on the ability to develop unique and robust features. As in sign language context, even the same user would not be able to perform the same gesture again precisely, meaning that any gesture has certain variability and a certain degree of

uniqueness. Human have evolved more subtly to remember and understand this variability and uniqueness. To develop machine capabilities to interpret this information from an image is not trivial. Therefore, a robust feature or set of features should uniquely describe the gesture to achieve reliable recognition. In other words, different gestures should result in different real discriminable features. Shift and rotation invariant features lead to a better recognition of hand gestures even if the hand gesture is captured from different angles.

The depth image coordinates, a given x-coordinate, y-coordinate and z-coordinate and its corresponding depth value, (X_i, Y_i, Z_i, D) can be converted to signer coordinate space. In this case, the signer coordinate space refers to the real-world;' space centered at the sign origin as shown in Figure 2.4 (a). Signer coordinate space of convert image coordinates. It first converts the image triplet (X_i, Y_i, Z_i, D) to camera coordinates. Although this information is available at runtime through Microsoft Kinect SDK API's, it does not store this data as it would slow down the data capture frame rate. Instead, the clear, correctly tracked hand points, from the skeleton tracking information is used to obtain corresponding camera coordinates (X_c, Y_c, Z_c, D) for the depth image coordinates (X_i, Y_i, Z_i, D). Then, it estimates the transformation matrix and converts any given image triplet (X_i, Y_i, Z_i, D) into the camera coordinate space.



(a) (b)

Figure 2.4 Process of normalizing 3D hand coordinates using X-, Y- and Z-extents computed based on arm lengths. (a) 2D, (b) 3D.

b. Classification

American Sign languages are rich, faceted language, and their full complexity is beyond current gesture recognition technologies. The interpersonal communication problem between signer and hearing community could be resolved by building up a new communication bridge integrating components for sign. Several Significant problems specific to Automatic Sign Recognition are i) Distinguishing gestures from signs ii) Context dependency (directional, verbs, inflections, etc) iii) Basic unit modeling (how do we describe them?) iii) Transitions between signs (Movements) iv) Repetition (Cycles of movement). Classification is using an instance-based classifier can be a simple matter of locating the nearest neighbor in instance space and labelling the unknown instance with the same class label as that of the located (known) neighbor. This approach is often referred to as a neighbor classifier. Studies have also shown that deaf children who have not been exposed to sign language at an early age have lower comprehension and diminished use of sign language at a later age. The main motivation of this study is to provide ASL learning tools that enable ASL exposure to deaf children of hearing parents. Therefore, it becomes imperative that we expand our research in ASL recognition.

Classifier always tries to improve the classification rate by pushing classifiers into an optimized structure. Classifier is suitable for large-scale prediction and classification tasks on complex and incomplete datasets. Classification is used to justify the objects using new methods to get a maximum. The features are the parameters extracted from the sign frame image which are taken from a camera. In each frame, a measure of properties is taken to determine the sign in different position. They estimate the probability that a sign belongs to each of the target classes that is predetermined. In the training phase, the training set is used to decide how the parameters must be weighted and combined in order to separate the various classes of signs (Premaratne 2014).

While the application of recognition systems based on sign is limited to rather small vocabularies, systems based on subunit models are able to handle larger vocabularies. This limitation results from the training problem. In order to adequately train a set of sign

models, each sign in the vocabulary must appear several times in different contexts. For large vocabularies, this implies a prohibitively large training set. Moreover, the recognition vocabulary may contain signs which have not appeared before. Consequently, some form of sign models compositions technique is required to generate models for those signs which have not been seen sufficiently during training. Many researchers built datasets and use a different method to split the training and testing data for classification. Some researchers randomly split the data by subject into training (70%) and testing (30%) sets such as (Molchanov et al. 2016) or training (80%) and testing (20%) set (Suthaharan 2015). Other researchers used 10% of labeled data and 90% of unlabeled data (Tencer et al. 2017). Table 2.6 summarizes the data split for classification in previous sign language recognition (SLR) research.

Table 2.6 Summary of data split for classification in SLR research

Research	Sign Language	Research Field	Test	Sample Size	Train	Acquisition Tool
Matthes et al. 2012	ASL	SLR	N/A	68	16-18	Video Camera
Forster et al. 2013	German SL	SLR	911	1980	7	Video Camera
Schembri et al. 2013	British SL	Linguistic	N/A	4000	249	Video Camera
Escalera et al. 2014	Italian SL	SLR	20	13858	27	Kinect v1
Escudeiro et al. 2015	Portuguese SL	PLR	100	1680	40	Kinect v1
Gutierrez et al. 2016	Spanish SL	Linguistic	2400	2400	2	Video Camera
Yang et al. 2016	Chinese SL	SLR	N/A	100	5	Kinect v2
Kumar et al. 2017a	India SL	SLR	50	7500	N/A	Kinect v1

Classifiers are a powerful tool that offers huge expressive power to signers, allowing them to represent complex concepts briefly. Classifiers in SLR system usually use a replacement of the hand shape and provide information about their position movement and appearance. The spatial construction of classifiers takes advantage of the fact that visual-spatial information is one of the most reliable human sensing modality. Studies have shown that deaf signers have enhanced visual-spatial processing capabilities in tasks such as mental rotation, visual attention and face recognition.

In the field of ASL recognition, much attention has been given to recognition of linguistic structures such as fingerspelling, basic signs, and interrogative words, while recognition of ASL sentences and classifier-based grammatical structures is still comparatively underexplored. The complex spatial constructions in ASL classifiers make recognition challenging with input from a limited source of information such as a regular RGB camera. However, the addition of a depth camera can potentially provide ample information that would be sufficient to recognize classifier constructions in ASL sentences successfully. It is assumed that an additional work can extend its ability to understand sign language by leveraging depth maps to develop a technique using improved hand detection and hand shape classification in order to recognize selected classifier-based grammatical structures of ASL (Kong & Ranganath 2014).

A sentence is a set of sequential feature vectors where successive vectors of similar labels constitute a word belonging to that sentence. The outcome of the classifier is a series of predicted labels associated with a test sentence. As such, the corresponding actual labels of each predicted test sentence are known. Accordingly, the class recognition rate is defined as the ratio of correctly classified feature vectors to the total number of test feature vectors (Yeh et al. 2013). Classification typically used a statistical algorithm to find the recognition accuracy. This section focuses on statistical classification methods suited for isolated and continuous sign language recognition. Statistical classification requires that, for each sign of the vocabulary to be recognized, a reference model must be built beforehand. Depending on the linguistic concept, a reference model represents a single sign either as a whole or as a composition of smaller subunits similar to phonemes in spoken languages (Von Agris et al. 2008).

i. Hidden Markov Model (HMM)

Hidden Markov Model (HMM) is a collection of finite states connected by transitions. Two sets of probabilities characterize each state: i) transition probability (the probability of taking a particular transition to a particular state) and ii) discrete/continuous output probability (the probability of emitting a particular output symbol from a given state),

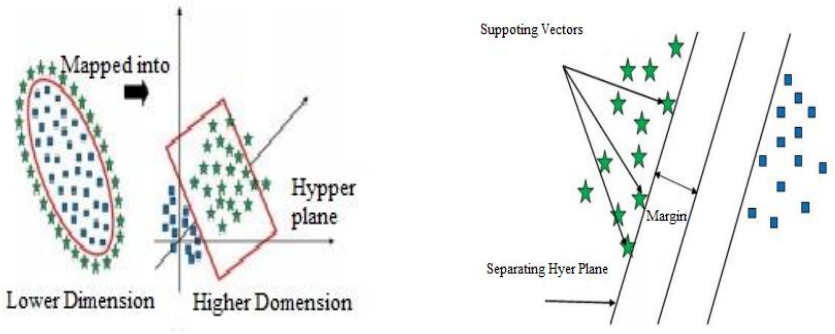
which given the state, defines the conditional probability of emitting each output symbol from a finite alphabet or a continuous random vector (Wu et al. 2017). It uses the compact notation $\lambda = \{A, B, \pi\}$ to indicate the complete parameter set of the model where C is a matrix storing transitions probabilities and a_{ij} denotes the probability of making a transition between states s_i and. D is a matrix storing output probabilities for each state and π is a vector storing initial state probabilities.

The use of HMM for hand gesture recognition stems from its ability to model speech recognition successfully. The similarities between gesture and speech propose that HMM might work on gesture recognition as well. Hand gestures, like spoken language, vary according to the person, location or background, time, gender, age and social factors. There is a common ground between speech and gestures which are known to have syntactic rules (Premaratne 2014). Since gesture is an expressive motion, it is natural to describe such a motion through the sequential model. HMM can be used for object classification, when given several models; it is possible to determine the model which will produce a given sequence of observations with the highest probability. Thus, if for each class there is a model with the states, transitions, and probabilities set appropriately, the Baum-Welch algorithm and Viterbi search algorithm can be used to calculate the model that most probably resulted in the sequence of observations (Petropoulos et al. 2017). It can be extended to a weak temporal classification system where a single model is constructed for each class, and given an unlabeled instance; the probability of each HMM generated that sequence of output symbols is calculated. The model with the highest probability is identified as the predicted class. The task of selecting the correct model and appropriate states and transitions is complicated, and trial and error method is usually applied to identify the number of states and transitions. Knowledge of the problem domain and experience would play a significant role in the success of its application.

ii. Support Super Vector Machines (SVM)

Support Vector Machine (SVM) is a promising algorithm that can classify both linear and nonlinear data. A Linear mapping is used to search for an optimal linear line to isolate hyperplanes, while nonlinear mapping is utilized to convert the training data inside a higher measurement into an unused measurement. The SVM learner supports multiple-class problems by computing the hyperplane between each class and the rest. It also supports some different kernels such as hyper tangent, polynomial, and radial basis function (Campbell et al. 2006).

SVM is a discriminative classifier defined by a separating hyperplane (Hearst et al. 1998; Smola & Schölkopf 2004). Classification decision boundaries are identified by maximizing a margin distance. The optimal separation hyperplane maximizes the margin of training data. The training example closest to the optimal hyperplane is called support vectors. Figure 2.5 illustrates the example of linear and nonlinear problem in SVM. A common problem for SVM is that the number of support vectors grows linearly with the size of the training sets (López et al. 2014).



(a) (b)

Figure 2.5 Example of (a) non-linear problem, and (b) linear problem Support Vector Machine (Hearst et al. 1998; Smola & Sch€olkopf 2004)

There are several of the research using SVM classifier for gesture recognition in the literature (Almeida et al. 2014; Marin et al. 2015) and some research combined other

classification methods to improve the performance of gesture classification (Suau et al. 2014; Singha et al. 2016).

iii. Hidden Conditional Random Fields (HCRF)

Conditional Random Fields (CRF) is a discriminative graphical model that does not accept that perceptions are independent, given the values of the covered up factors. CRF attempts to model the conditional probability $p(x \setminus y)$ directly. In general definition, one can consider a factor graph partitioned into a set of clique templates $C = \{C_1, C_2, ..., C_P\}$. Each group templates C_P specify a set of sufficient statistics $\{f_{pk}(x_p, y_p)\}$ and parameters, such that the general model for a CRF can be written as Equation (2.1).

$$p(y \setminus x) = \frac{1}{Z(X)} \prod_{C_P \in C} \prod_{\Psi_c \in C_p} \Psi_c(x_c, y_c; \theta_p) \tag{2.1}$$

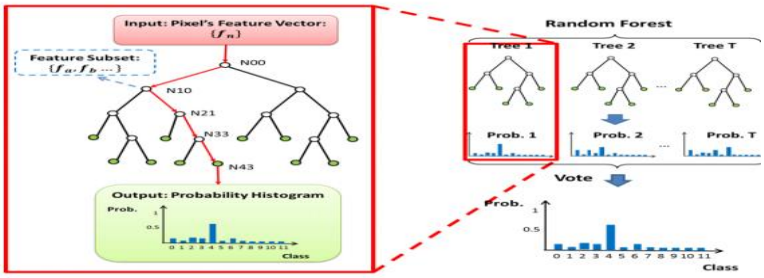
where $\Psi_c(x_c, y_c; \theta_p) = \exp\left\{ \sum_{k=1}^{k(p)} \theta_{pk} f_{pk}(x_c, y_c) \right\}$ and in which $Z(x)$ is a normalization.

Hidden Conditional Random Field (HCRF) utilizes hidden variables to demonstrate the inactive structure of the input signals by characterizing a joint dispersion over the class label and hidden state names conditioned on the perceptions (Liu et al. 2015a). HCRF model can be utilized either as a gesture class finder, where a single class is discriminatively prepared against all other gestures or as a multi-way gesture classifier, where discriminative models for multiple gestures are at the same time trained (Escalera et al. 2016). Some HCRF-based methods for an arm and head gesture recognition have been proposed in previous studies (Choudhury et al. 2015; Song et al. 2015).

iv. Randomized Decision Forests (RDF)

Randomized Decision Forests (RDF) have been applied in gestures classification. The RDF

classifier comprises a set of autonomous decision trees, where at each leaf hub, the prediction is known as a set of classification probabilities $P(c \setminus f(I,x))$ for each class c (Figure 2.6a). The final prediction of the forest is obtained by a voting process of all trees (Figure 2.6b) in the training of each tree, the dataset is randomly separated into two subsets. Approximately 2/3 of the data is used for training while the rest are used for error estimation. The training subset is used to collect a statistical histogram of classifications at each leaf node. The histogram can therefore be used to estimate the classification probabilities for data samples which reach the leaf node. The thresholds are optimized to find the best split that can minimize the errors in error estimation (Deng et al. 2014).



(a) (b)

Figure 2.6 Illustration of the RDF classifier. (a) Single-tree classification: The pixel that has feature vector $\{f_n\}$ is pushed down the tree from N_{00} to a feature subset (e.g., $\{f_a, f_b ...\}$) is used to determine the split at each split node $(e.g., N_{10})$. At each leaf node $(e.g., N_{43})$, a histogram illustrates the probabilities of all classes. (b) Multi-tree voting: The output probability histogram is determined by a voting process that involves all of the trees

RDF has been proven effective for human body segmentation using depth contrast features. It is able to dodge over-fitting circumstances in the multiclass task and profoundly proficient in handling large database (Fletcher& Islam 2017). Microsoft Kinect body tracker uses the comparison of depths of pixels instead of comparing pixel intensities. At each non-leaf node, the depth of the current pixel is compared to the depth of a pixel with a certain offset from the current pixel. The node tests at a given pixel *u* for the Microsoft body tracker can be described as illustrated in Equation (2.2):

$$f(u \setminus \varphi) = z(u + \frac{\delta_1}{z(u)} - z(u + \frac{\delta_z}{z(u)}) \quad (2.2)$$

Where $\varphi(\delta_1, \delta_z)$ describe the 2D pixel offset and function $z(u)$ looks up the depth at the pixel$_u$. Each body part will have different distributions of the probability of what depth their surrounding pixels will have. For example, a pixel where there is a significant difference in depth between the current pixel and the pixel above, the current pixel would get a higher positive response to be a head pixel than, for instance, a torso pixel. Before training a random forest model, it is essential to collect features that provide an excellent description of the gestures.

Shaikhina et al. (2017) utilized Hough forests to perform activity recognition, where a set of randomized trees is prepared to perform a mapping from a densely-sampled d-dimensional space into comparing votes in Hough space. Meanwhile, (Joshi et al. 2017) utilized a transient random forest to recognize transient occasions and (Zhang et al. 2017) used WEKA to generate tree classifier to classified data using a decision tree.

c. Gesture Sign Language

Gestures sign language performed by different people usually possess significant variations in speed, range, movement, etc. These characteristics require an adaption of flexible metrics to identify each word and to delimit the words in a sentence. It is necessary to understand what constitutes a word in ASL clearly, what features may characterize it and how they can be modeled. In general, a word corresponds to a gesture in ASL gesture is considered comprises a sequence of configurations from the dominant hand, each associated with possibly a configuration of the support hand, motion and orientation of both hands. Each element of the sequence is defined as an atom of the gesture. The beginning of a gesture is marked by the adoption of a configuration by the dominant hand. The end of the gesture is characterized either by the return of the dominant hand to a neutral position or by a configuration change.

Hands can perform a complex enunciated structure comprising of numerous associated links and joints, enabling them to play a dynamic role in sign language. Generally, there are two types of hand gesture: i) static hand gesture and, ii) dynamic hand gesture. In hand pose, no developments are involved, while hand gesture is a sequence of hand poses associated with movement over a period (Ghotkar et al. 2014). In dynamic hand gesture, there are two important aspects need to be given extra attention: i) local finger motion with no changes of hand position or orientation, and ii) global hand movement with changes of hand position or orientation.

2.3.3 Hand Gesture Recognition

Hand gesture recognition based on computer vision technology has received great interests recently due to its natural human-computer interaction characteristics. Hand gestures are composed of different hand postures and motions. However, the human hand is an articulated object with over 20 degrees of freedom (DOF), and many self-occlusions would occur in its projection results. Moreover, hand motion is often too fast and complicated for current computer image processing speed. Thus, real-time hand posture estimation is still a challenging research topic with multidisciplinary research area including pattern recognition, image processing, computer vision, artificial intelligence and machine learning (Zhou et al. 2016).

Human-computer interaction with hand gestures plays a significant role in these modalities because people often rely on their hands for communication or to interact with their environment. Therefore, hand-gesture-based methods stand out from other approaches as it provides a natural way of interaction and communication (Kılıboz & Güdükbay 2015). Dynamic hand gesture involves the movement of the hand in the space. A proper interpretation of these gestures is essential so that the machine correctly defines the dynamic configurations of the hand. It has been observed from the existing research that the hand gesture databases available online consist of gestures gesticulated in a reference pattern set by them, and the users are allowed to gesticulate with this reference design (Bhuyan et al. 2014).

Gesture recognition is closely related to hand tracking issues. Gesture recognition itself is a vast topic, and therefore, only a short presentation covering the main principles can be given in this section. Gesture recognition can be described as a receiver (usually a computer in this context) can recognize a gesture performed by the user. Gestures are very familiar in the daily lives, and they perform it either consciously and unconsciously. Hand gesture recognition task is one of the essential and elemental tasks in computer vision, endorsed by the recent advancement in information technology and media (Liu et al. 2015b).

Automated human interaction systems which include hand processing tasks like hand detection, hand recognition, and hand tracking have been developed over the past decades. Once the hands are segmented, a variety of hand features can be extracted for hand gesture recognition. The hand features can roughly be classified into three categories, i.e., depth-based, color-image-based and shape-based features (Wang et al. 2017). Many hand gestures recognition methods applied skin color-based detection and used geometrical features for hand modeling. In recent years, 3D sensors, such as binocular cameras, Kinect and leap motion, have been applied for hand gesture recognition (Zhou et al. 2016).

Chen & Wang (2013) presented a gesture recognition method based on a four-channel wireless sEMG system. The application of three feature types (Hudgins time domain features (TD), autocorrelation and Across-Correlation Coefficients (ACCC) and Spectral Power Magnitudes (SPM)). Four well-known classification algorithms (k-nearest neighbor (k-NN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and support vector machine (SVM)) in offline recognition has been investigated in the study. The experimental results show that all combinations, except k-NN or LDA using ACCC features, achieved above 91% average recognition accuracy.

Lee (2013) introduced an effective approach to real time hand tracking in dynamic environments of the virtual interactive interview system. Considering simplicity and practical feasibility, we perform hand tracking in the 2D image plane. An interviewer's hand motion recognition algorithm based on the particle filters is applied in this system. The existing virtual interview system is not an automatically interactive system between

interviewee and its system.

Dinh et al. (2014) presented a novel system of hand gesture interface using depth imaging sensor for appliances control in smart home environments. The hand parts in a depth hand silhouette are recognized to control appliances with hand gestures and control commands are generated according to the implemented control interface. The system is tested on a set of 500 hand depth silhouettes containing various poses of four hand poses without body part.

Kılıboz & Güdükbay (2015) proposed an approach to recognize trajectory-based dynamic hand gestures in real time for human-computer interaction. A fast learning mechanism that does not require extensive training data to feed gestures to the system is introduced. Online gesture recognition is achieved without needing to specify the gesture start and end positions. The proposed method is very promising in terms of gesture detection and recognition performance (73% accuracy) in a stream of motion. The collected trajectory data may still contain events that are not part of the gesture due to user reaction error during the initial and final moments of the recording.

Singha et al. (2016) presented a hand gesture recognition system to address the effect of variations in gesture pattern during gesticulation. Different gestures can be gesticulated in various patterns which increase the difficulties in recognition phase. A comparative study has been done with the proposed features and the state-of-art features such as orientation, the combination of location, orientation and velocity, and the combination of ellipse and position features. The system performance is evaluated using the proposed set of features for different classifiers such as ANN, SVM, k-NN, Naïve Bayes and ELM.

Meanwhile, Raheja et al. (2016) proposed a novel technique for hand gesture recognition using Principal Component Analysis (PCA) implemented in FPGA simulator. The simulation is done using co-simulation tool of Simulink with Xilinx System Generator (XSG). The square root operation on FPGA is quite complex, causing a long delay in the

system. It is possible to work with the squared Euclidean distance (SED) to infer matches, thereby eliminating the cumbersome square root operation. The results stored in the initial cycles are not always accurate as the ED of all the gesture image is not considered in the calculation.

In recent study, Wang et al. (2017) proposed a system based on canonical superpixel-graph earth mover's distance (CSG-EMD) metric to alleviate the problem in gesture recognition. Scattered superpixels are organized in the form of canonical superpixel-graph, which can factor out non-standard finger poses, resulting from a well-structured finger-pose-neutral shape representation for hand gestures. Meanwhile, (Yang et al. 2017) presented a continuous hand gesture recognition method based on trajectory shape information that simultaneously performs segmentation and recognition process. Each component of the framework applies shape-based information to ensure robust performance for gestures with significant temporal variation. Misclassification at the current keyframe suggests that the CSG may not have functioned properly.

In the literature, there are numbers of research applied statistical or machine learning techniques for the development of either isolated or continuous sign language recognition system. Most of these techniques required a large number of training samples and are mostly dependent on the signer. In reality, independent signer method is more practical and desirable. The complexity of sign language recognition increases from alphabet recognition to word recognition and continuous sentence recognition. Table 2.7 shows the comparison of hand gesture recognition methods using various classifiers found in the literature. Some of researchers used cluttered background and other used simple background. The limitation of researcher works show above and it was observed many gaps in many research work also have many problems open research still under process need to solve.

Table 2.7 Comparison of hand gesture recognition methods

Research	Background	Feature Vector Representation	Classifier/ Recognition	Accuracy (%)
Chen & Wang (2013)	Simple background	upper arm and gross hand movement	support vector machine (SVM)	97.93
Lee. 2013	Simple background	texture of hands, for hand tracking	particles are moved according to a deterministic motion	75
Dinh et al. 2014	Cluttered	depth imaging sensor	random forests (RFs)	98.50
Chen et al. 2014	Cluttered	Segmented palm and finger	Predict the labels of hand gestures	96.6
Ghotkar et al. 2014	Extract background	Segmented hand image	K-mean, Fuzzy k-mean, minimum spanning tree	90
Kılıboz & Güdükbay. 2015	Cluttered background	six-degrees-of-freedom position	finite-state machine sequence recognizers	73
Plouffe & Cretu, 2016	Cluttered background	Identify the first pixel on the hand contour	DTW	92.4
Xu & Dai 2017	Cluttered	Kalman filter to estimate the hand position	CNN	95

2.3.4 Sign Language Recognition System

Research on hand gesture and sign recognition consist of two main measurements: i) isolated recognition, and ii) continuous recognition. Isolated recognition focuses on a single hand gesture that is performed by user and endeavors to recognize it. Meanwhile, in continuous recognition, the user is anticipated to perform gestures one after another, and the aim is to recognize each gesture performed by the user. Continuous recognition issue is slightly different from hand gesture recognition and sign language recognition. In hand gesture controlled environments, the issue can be considered as a motion spotting problem, where the task is to distinguish the meaningful gestures from the meaningless ones. Meanwhile, in sign language recognition, continuous recognition issue incorporates the co-articulation issue. The preceding sign influences the succeeding one, which complicates the recognition task as the transitions between the signs ought to be unequivocally modeled and joined into the recognition framework.

Moreover, dynamic hand gestures also prone to the motion representation problem, which includes two interrelated issues: i) how to represent the features for a single frame, and ii) how to describe and model the complete motion trajectory. Most of the studies in

the literature employed generative models with the ability to handle variable-length data, as the dynamical gestures produce variable-length sequences. Designing novel generative models and integrating discriminative approaches to existing models which can better represent the inherent dynamics of the gestures while increasing system performance is the next promising step in hand gesture recognition. The existing SLR methods were developed using different techniques and designed using different sign language while incorporating a Kinect sensor. Table 2.8 summarizes the proposed SLR systems incorporating a Kinect sensor. The limitation of sign language recognition system based on different sign language by used different datasets.

Table 2.8 Summary of SLR system incorporating Kinect sensor

Research	Method	Dataset	Accuracy (%)
Ren et al. 2013	Distance metric using template matching	10 gestures	93.20
Dominio et al. 2014	Distance, curvature and palm area features using SVM	10 gestures ASL signs	100 97.60
Pedersoli et al. 2014	Gabor features using SVM	24 ASL alphabets	70
Almeida et al. 2014	Shape, movement and position of hands using SVM	34 BSL sign	80
Kirac et al. 2014	3D hand points, regression forest, and dynamic programming	12 ASL sign	57.60
Mehrotra et al. 2015	3D skeleton points, angular and distance features using SVM	37 ISL sign	86.16
Inoue et al. 2015	HOG+SVM	JFSL	88
Chansri et al. 2016	ANN	Thai Sign Language	84.05
Yang et al. 2016	HMM	Chinese Sign Language	93.14
Nai et al. 2017	Random forest	ASL signs	81.1

a. Automatic Sign Language Recognition (ASLR)

Automatic sign language recognition (ASLR) is an open issue that has received a great deal of interest recently due to its significance to signers, as well as to other users and numerous applications. Although hearing impaired people can communicate among themselves using sign language, they face difficulties integrating into educational, social and work environment because the vast amount of people do not have any signing skills (Escudeiro et al. 2014). Moreover, most public services are not translated into sign due to expensive human translators and lack of translation tools. ASLR aims to reduce, if not, eliminate these hurdles so hearing impaired people can lead better lives.

Automatic sign language recognition and interpretation is a subset of gesture recognition. It requires a high level of expertise in different areas including electronics, machine vision, natural language processing, computer science, linguistics, artificial intelligence, machine learning, and statistics. In recent years, there has been an increasing interest in developing ASLR system to enhance communication between normal hearing and deaf people. The system mostly focuses on the manual aspect of signs and recognition of hand configurations such as hand shape, position, orientation, and movement. Although many researchers have been proposing different methods, the proposed methodology remain functional in the environments they were designed for, as the majority of the proposed methods are mainly focusing on the improvement of recognition capability of ASLR by increasing the number of words in their vocabularies.

Forster et al. (2013) introduced a large-vocabulary ASLR framework to recognize sentences in continuous sign language. It extracted features from standard single-view video cameras without utilizing additional equipment. It works on two publicly available large vocabulary databases representing lab-data (25 signers, 455 signs vocabulary, 19k sentences) and unconstrained real-time sign language (1 signer, 266 signs vocabulary, 351 sentences), but did not convey the meaning in synchronous information streams/modalities perfectly. This approach is not conceivable in ASLR since the definition and extraction of subunits are still unclear. The recognition accuracy is 80%.

Pigou et al. (2014) presented a recognition system using Microsoft Kinect, Convolutional Neural Networks (CNNs) and GPU acceleration. Instead of constructing complex handcrafted features, CNNs are able to automate the process of feature construction. The predictive model can be generalized to users and non-occurring surroundings during training with a cross-validation accuracy of 91.7%. The feature extraction is automated by utilizing CNNs. An artificial neural network (ANN) is then used for the classification task. In this study, 10-20 gestures, including noise movements that are not part of the 20 Italian motions are used.

Inoue et al. (2015) proposed an automatic hand region extraction method with a depth sensor and time-series curve for Japanese Finger-spelled Sign Language (JFSL) recognition. For a precise hand region extraction, signers are required to perform signs at a particular distance from the depth sensor, and their hands have to be at the front-most from it. The experimental result of this study shows that the accuracy of the proposed method is 88%. The proposed ASLR methods explained above are summarized in Table 2.9.

Zadghorban & Nahvi (2016) proposed automatic sign language recognition used f two main phases of sign words extraction and their classification. Several stages, including tracking and separating the sign words, are conducted in the sign word extraction phase. This algorithm decomposes sign language video into the sign words using motion and hand shape features. In the classification phase, separated words are classified and recognized using hidden Markov model and hybrid KNN-DTW algorithm, the same sign and the boundary between them is not correct. Thus, this boundary is ignored and the next boundary will be processed.

Table 2.9 Summary of ASLR methods

Research	Approach	Dataset	Results (%)	Limitation
Forster et al. 2013	HMM	25 signers, 455 signs vocabulary, 19k sentences	80	Does not convey the meaning in synchronous information streams/modalities correctly
Pigou et al. 2014	CNNs	30 words Italian	91.7	Using 2D convolutions resulted in a better validation accuracy than 3D convolutions
Inoue et al. 2015	HOG + SVM	41 JFSL words	88	Lack of quantitative evaluation
Zadghorban & Nahvi 2016	HMM and KNN-DTW	20 Sentences and 46 sign words(Persian SL)	93.73	The sign ignore next boundary

b. Isolated Sign Language Recognition (ISLR)

As mentioned above, isolated sign language recognition (ISLR) focuses on a single hand gesture that is performed by the user and aim to recognize it. In this context, a hand gesture is a sequence of fixed starting/ending points as a sign boundary.

Anjo et al. (2012) presented a framework called Gesture User Interface (Gesture UI) to recognize static gestures in Brazilian Sign Language (Libras). It is developed to deal with small sets of gestures, (A, E, I, O, U) and (B, C, F, L, V). The framework comprises of two main steps: segmentation and classification. The accuracy of the method is 75.4% to 100% for both gesture sets. Unfortunately, the method is not able to make a proper decision when an unknown gesture is displayed due to eccentric mapping interior of the statistical model design.

Zarrouk et al. (2015) presented multiple hybrid systems for the recognition of isolated words from a multi-dialect Arabic vocabulary. It applied two machine learning methods which are Hidden Markov Models (HMM) and Dynamic Bayesian Networks (DBN), and Multi-Layer Perceptron (MLP) is used as an estimator of emission probabilities in HMM. The study also used Support Vector Machine (SVM) as an estimator of posterior probabilities due to its high predictive power and discrimination ability. The

experimental results show that the recognition accuracy for SVM and HMM is 87.67% and 83.01% respectively.

Wang et al. (2016) proposed a covariance matrix–based representation to fuse information from multimodal sources. The drawback induced by the commonly used Riemannian metric is that the proximity of covariance matrices is measured on the Grassmann manifold. Consequently, two signs cannot be separated without an accurate skeleton estimation. The experimental results show that the proposed method works in general cases of recognition.

In another research, Lim et al. (2016b) presented a feature covariance matrix-based serial particle filter for ISLR. At the pre-processing stage, the fusion of the median and mode filters is employed to extract the foreground and thereby enhances hand detection. The proposed method serially track the hands of the signer, instead of tracking both hands at the same time, in order to reduce the misdirection of target objects. In this work, the hand size is not included because the distance of the signer to the fixed camera is the same, and hence, the hand size of the signer does not change very much from frame to frame.

Su et al. (2016) proposed sign dialect recognition system by applying an improved decision tree as the base classifier of random forests. The system is tested on Chinese sign language subwords, which are acquired from a convenient device, a combination of accelerometers (ACC) and surface electromyography (sEMG) sensor, worn on signer's both arms. The system achieved 90% recognition accuracy.

De Edson et al. (2016) proposed a new method for static gesture recognition system using Eigen hands approach and Microsoft Kinect. The method computed the lighting normalization in the gesture image which is acquired through Kinect display. The preliminary results show that the recognition accuracy is reaching up to 89% in the static gesture recognition of Libras.

Wang et al. (2016) proposed a covariance matrix–based representation to naturally fuse information from multimodal sources. To tackle the drawback induced by the commonly used Riemannian metric, the proximity of covariance matrices is measured on the Grassmann manifold by used an isolated sign sequence. The two signs cannot be separated without a more accurate skeleton estimation.

Table 2.10 summarizes the proposed ISLR methods discussed above. There are several methods of isolated sign language recognition by used different approaches and various datasets but still have some limitation for different researcher still open research.

Table 2.10 Summary of ISLR methods

Research	Approach	Dataset	Accuracy (%)	Limitation
Anjo et al. 2012	ANNs	Numbers and alphabet	96.7	Unable to make a right decision for an unknown gesture
Zarrouk et al. 2015	SVM HMM	3,622 Arabic Words	87.67 83.01	Several of the research tried to solve some problems related to ASR using HMM by many extensions of DBN models
Wang et al. 2016	SVM	100 CSL Words	96	Two signs cannot be separated without an accurate skeleton estimation
Su et al. 2016	RDF	121 CSL subwords	90	Signer has to slowly move in the front of the chest to avoid losing the gesture
De Edson et al. 2016	PCA	480 images alphabet	89	Different lighting conditions are not considered

c. Continuous Sign Language Recognition (CSLR)

In comparison to ISLR, the nature of continuous sign language recognition (CSLR) is more complicated, as both hands are used to perform the gesture, and it can also be a combination of dynamic gesture. Continuous sign language is a sequence of gestures that generate a meaningful sentence, and a sign may begin or end anywhere in a given observation sequence. Therefore, it is challenging to recognize the signs in a real-world environment.

Kong & Ranganath (2014) proposed a segment-based sign recognition approach that is robust to variation. The sentences are segmented into signs or ME sub-segments using Bayesian network fusing the outputs of CRF and SVM. With eight signers, a recall

rate of 95.7% and a precision of 96.6% is achieved. However, in continuous signing, direct normalization based on the entire signed sentence can lead to errors as there may be variations from one sign to the next. It is hard to define sub-segment features that are similar to letter capitalization that can be consistently found to merge sub-segments into a segment.

Zhang et al. (2014) introduced an approach for CSLR which consists of offline training and online recognition. In this work, offline training is done on a threshold matrix for coarse segmentation and rate thresholds for fine segmentation. Recognition for sign sentences is obtained from the optimal path by the Viterbi algorithm. The approach is tested on a dataset consists of 510 sign samples over 34 isolated signs from 5 signers and 180 sentences samples over three different sentences from 3 signers. However, the system faces some challenges in finding the end of the sign in the sentence directly.

Koller et al. (2015) developed a real-time CSLR system using HMM-based visual model and class language model to complete the recognition task. The constrained maximum likelihood linear regression (CMLLR) is applied to deal with signer-dependency. Modeling non-gesture is a challenging task because it is not clear which parts of the temporal signal are not part of a sign. The transition between signs depends on both enclosing signs. If the tracked position of the object deviates too far from the real position, the corresponding hand-patch does not match the trained model of the recognizer.

Yang et al. (2016) proposed a method for CSLR system with a coarse segmentation approach to providing the maximal level number. HMM is used to compute the similarity between the sign model and testing sequence, and a fast method to compute the likelihood of HMM is proposed to reduce the computation complexity. Furthermore, grammar constraint and sign length constraint are employed to improve the recognition using a Kinect dataset of Chinese sign language containing 100 sentences composed of each five signs. However, the discriminative information between different signs is not considered in this study, meaning that the previous results in Level Building are not stored, causing the LB-HMM to run repeatedly and slowly.

In a recent of study, Rao et al. (2017) presented a method to bring sign language closer to the real-time application, such as mobile phone. Selfie captured sign language video is prepared by constraining the computing control of a smartphone. Pre-filtering, segmentation, and extraction on video outlines create a feature space. Least Distance and ANN classifier are then used to train the sign feature space and tested them iteratively. The head position does not change much in any of the frames captured. Hence head contour DCT remains relatively constant throughout the video sequence. Problems arise if the test dataset frame number does not match the train data set. The summary of the CSLR methods is shown Table 2.11.

Table 2.11 Summary of CSLR methods

Research	Accuracy (%)	Feature	Classifier	Dataset	Disadvantage
Zhang & Feng 2013	90	Density distribution feature	Hausdorff Distance	HCI	The time cost is not efficient
Masood et al. 2014	91	Joint of interest	DTW	Five signers, and 180 sentences	The method is only capable of detecting the gestures which don't involve specific finger detection
Zhang et al. 2014	66.5	Motion trajectory	HMM	Ten signers, 100 sentences	It is usually difficult to find the end of the sign in the sentence directly
Koller et al. 2015	55.6	Visual modelling	ANN	25 signers, 455 sign vocabulary, and 19k sentences	The corresponding hand-patch does not match the trained model
Nasri et al. 2015	90.4	SIFT feature	Template matching	CBSs of disjointed gestures	Viewing angle of camera, lighting condition, and different user clothing cause error
Yang et al. 2016	97.77	Motion trajectory	HMM	100 sentences composed of 5 signs	System will run slowly when using too high maximal level number
Li et al. 2016	87.4%	Continuous technique	HMM	510 words, 1,024 testing sentences	They adopt different features for the two hands mainly because the left hand and right hand are not equally important
Rao et al. 2017	90	Hand and head portions	ANN	Selfie sign language	Problems arise if the test dataset frame number does not match the train data set

2.4 CONCLUSION

The development of mankind is heavily dependent on the communication system, though there are some shortcomings in present communication system. One of the limitations is the communication gap between hearing impaired and normal people. Deaf and mute people use sign language to communicate with others. Unfortunately, a lot of normal hearing people find it difficult to understand them. On the other hand, normal hearing people use oral language, which is inconsolable to deaf and mute people. The social aspect of humanity can increase in individual mind by involving physically impaired people in our day to day life. Ideally, interpretation systems would make it possible to communicate with deaf people. However, some systems have little accuracy or recognition rate, even fail to interpret, due to some obvious reasons, i.e., signs vary in time and space, even if the same user repeat a sign. Slight changes of speed and position of hands will occur, lack of depth information, occlusions, the enormity of the number of gestures available and their similarity to many others, and the user's inability to sign them makes the machines unreliable with the present feature extraction and classification approaches.

This chapter consists of review for 2D static sign language and 3D dynamic sign language recognition have been presented in details then explain approximately to the device which includes a description of Kinect sensor v2 for Windows. Hand detection and tracking, SLR approach and review of previous studies in the literature have been presented in details. In general, sign language recognition can be categorized into isolated sign language recognition and continuous sign language recognition. Isolated sign language recognition aims to recognize a single sign gesture at a time. On the other hand, continuous sign language recognition aims to recognize the complete sign gesture sentences. Identification and definition of the problems related to the theoretical study will be discussed in section 3.2 of the following chapter.

The next chapter presents the Methodology which describes the steps planned and implemented for achieving the stated objectives in Chapter I. The descriptions will include the employed existing and newly developed dataset based on American Sign Language such as number, alphabet, words, and sentences.

CHAPTER III

METHODOLOGY

3.1 INTRODUCTION

Automatic sign language recognition (ASLR) is an open problem that has received a lot of attention recently due to its usefulness and its numerous potential applications. Sign languages are used worldwide, mostly by the hearing impaired communities and people associated with them such as their family members and friends. In computer vision field, several of research and development for ASLR approach have been proposed, though they are limited. In particular, previous studies on ASLR approach for American Sign Language (ASL) are available in the literature, however, to the best of our knowledge, an effective ASLR approach is still lacking. Today, an alternative solution to address this problem is possible with the advent and wide spreading of powerful devices such as Kinect depth sensor.

This chapter presents the methodology applied in this study. The methodology involves many tasks which include data collection, data processing, classification, and evaluation. This chapter is organized as follows: Section 3.2 discusses the theoretical study which includes related issues and challenges, Section 3.3 explains the research architecture, Section 3.4 discusses the experimental design and analysis, and Section 3.5 discusses the experimental finding and analysis. Finally, this chapter is summarized in Section 3.6. Research methodology is proposed in order to fulfill the objectives and answer the research questions of the study. Figure 3.1 summarizes the research methodology of this study.

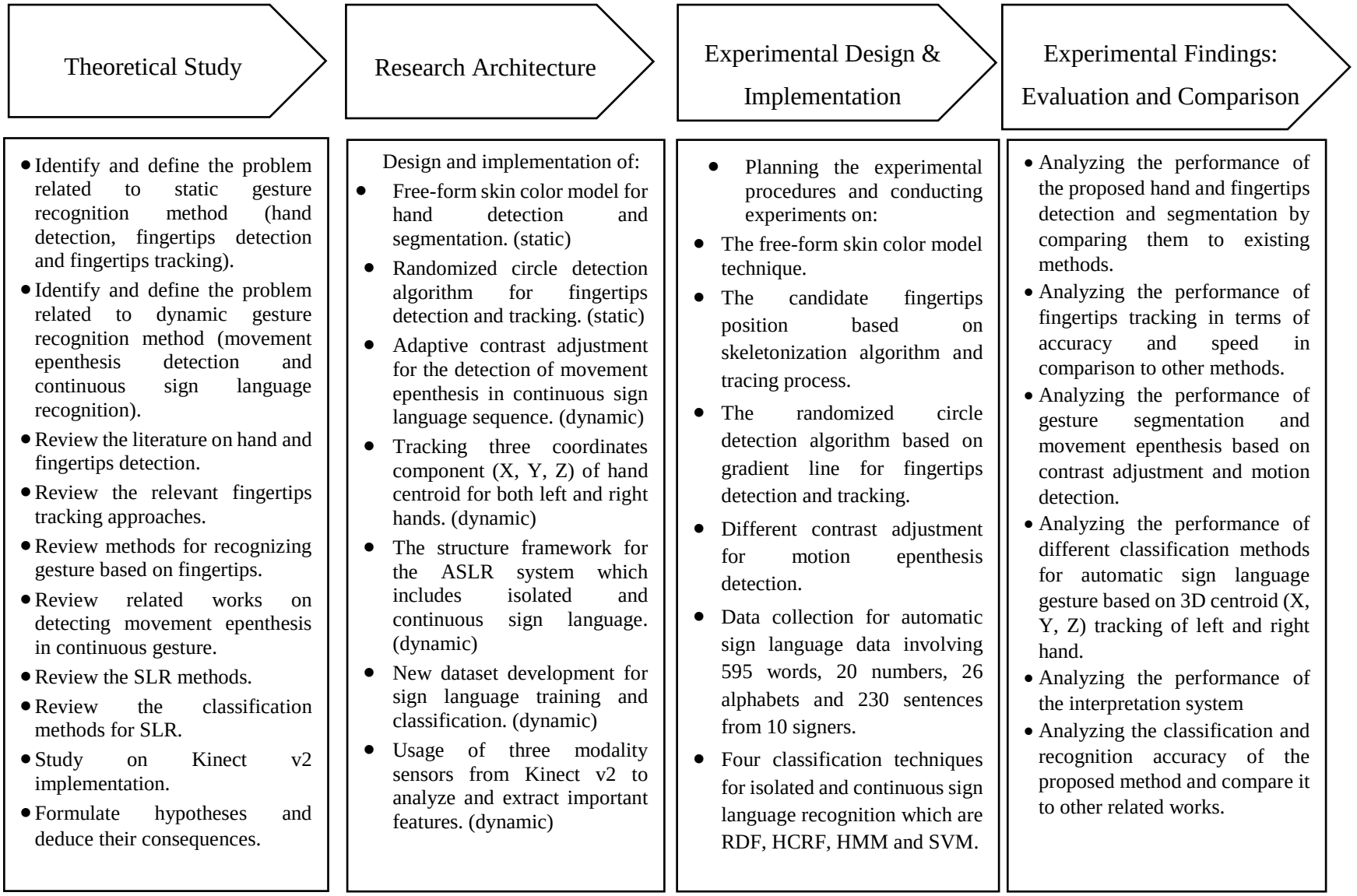


Figure 3.1 Research Methodology

3.2 THEORETICAL STUDY

Theoretical study provides a comprehensive review of automatic sign language recognition (ASLR) research in the literature, which can be found in journal articles, conference proceedings, books, and thesis. It includes two subjects of interest which are static sign language recognition and dynamic sign language recognition methods. Static hand gesture for sign language, which mainly focuses on the recognition of well-defined signs based on a hand posture, is reviewed. Previous works on hand segmentation and fingertips detection are discussed. It is found that most works suffer from poor hand detection due to complex background and different lighting, and the same problem faced in fingertips detection. In order to reduce the processing time, a bounding box is created around the fingertips before the randomized circle detection algorithm is applied. In most sign languages, alphabets and numbers are performed with one or two hands static gesture, and few alphabets require motion. Alphabet recognition mostly requires fingerspelling of the proper noun, and in some cases, it may be part of a word.

Gesture segmentation is the first and the most critical step in continuous sign language recognition. It is the task of interpreting and spotting meaningful hand gestures from continuous hand gesture sequences. It also can be defined as a task of finding the start and the end of a communicative gesture from continuous data. Several methods have been explored to deal with the problem of gesture segmentation, some of them simply based on the definition of a threshold value for gesture data, while others are more complex. The proposed ensemble framework will be discussed in detail in Chapter IV. A movement segment occasionally needs to be inserted between two successive signs to shift the hands from the end of one sign to the beginning of the following sign during the phonological process of sign language.

Many methods have been proposed in previous studies for hand position, rotation, scale and shape tracking, and motion. In real time application, methods such as skin detection, palm centers detection and fingertips detection have been suggested. Classically, image-based tracking solutions are categorized into single-camera and multi-camera. Single-camera proposals obtain impressive results given the lack of

available information. Several previous studies extracted human pose from the single color camera.

A detailed review on hand tracking based on depth sensor has also been presented in this study. Based on the reported literature, there are only two components that are often used for tracking namely the X and Y component. The Z component is often abandoned. However, the Z component might serve some benefits, and this motivates us to include it in our proposed method. The center of the palm coordinates is calculated based on the moment of hand contour feature, which can be tracked based on the depth image information obtained by the Kinect sensor.

Finally, gestures in sign language recognition from continuous data are reviewed. Most of the existing works in sign language assume that the training signs are already available and often signs used in training are set to manually selected frames from continuous sentences. Given these time series trajectories, linguistic research on isolated and continuous sign language allows challenging questions related to the appearance of dialectic variances, differences in pronunciation and sentence structures. While continuous sign language is a sequence of gestures that generate a meaningful sentence, the ability to recognize isolated signs does not guarantee the recognition of signs in continuous sentences. Each sentence is represented as a trajectory in a small dimensional space called the space of relational distributions isolated sign language, which consists of single gesture having a single word.

3.3 RESEARCH ARCHITECTURE

As mentioned in Chapter I, the main objective of this study is to develop automatic sign language recognition (ASLR) approach to interpret the words and sentences in sign language efficiently. Our proposed solution has been specified, designed, tested and preliminarily evaluated for the ASLR problems, in particular for American Sign Language (ASL). In the context of this study, computer vision techniques have been adopted along with the depth sensor application. A proper SLR approach has been proposed, where methods for feature extraction, representation, storing and classification have been suggested. Two novel algorithms to solve the problem of real-

time recognition of isolated static signs have been specified, developed, tested and evaluated accordingly. Feature selection method is also suggested to address the issues observed in related works. In short, the proposed system intends to improve the recognition accuracy of static and dynamic sign language recognition.

An ensemble framework for static and dynamic sign language for ASLR is proposed in this study. ASLR Method for static sign language is developed based on free-form skin color model for hand segmentation (Dawod et al. 2010a). The pixel value of a signer's hand is captured and represented in color model. Then, a method for fingertips detection is proposed based on skeleton algorithm and region of interest (ROI). The fingertips location is calculated based on skeletonization method and tracing process. The method begins with the conversion of an image in RGB color space into $YCbCr$ color space. The chrominance Cb and Cr are extracted from the color space. A thresholding technique, which is based on a pre-defined range for Cb and Cr components representing the skin color value, is used to extract the hand from the background to create a binary image.

A method to automatically detect and track fingertips position between frames using randomized circle detection is presented. During tracking, in order to reduce the processing time, a bounding box is created around the fingertips before the algorithm is applied. The randomized circle detection algorithm is utilized based on the fact that for a circle, the gradient vector of the edge pixels will continuously point to the center of the circle. The center of the circle can be determined by choosing the point that has the most number of gradient line intersections. This notion can be modified to be applied to semi-circle shape like fingers. The proposed ensemble system is explained in detail in Chapter IV.

So far, there are no raw datasets for Kinect v2 available in the literature. Hence, this research aims to build a new dataset for SLR based on Kinect v2 camera, which consists of sign language alphabets (A-Z), numbers (1-20), words and sentences. The dataset is collected from Community Service Centre for the Deaf (CSCD) in Malaysia. Then, it is tested with four machine learning approaches SVM, RDF, HMM, and HCRF. A good classification result has motivated us to further the work to dynamic sign

language recognition. For hand tracking, skeleton algorithm and mean-shift algorithm are applied to shift the coordinate data on Kinect v2 after the joint of interest found from normalization. Normalization method is preferred as it does not cause the sign data to be invariant to shifting while the location is normalized.

However, normalization causes sign data to be invariant to scaling. Although it handles inter-signer differences in terms of movement size, it causes sign data to lose sequences of locations, which is another discriminative feature of signs (Isikligil 2014). As an example, think about two signs which have the same type of movement but different origin and end position for the active hand. Creating data of signs which are invariant to scaling induces the locations of the active hands to have the same value during the performance of the signs and causes these two signs to no longer have a discriminative feature. Normalized features are extracted from the signing space and divided into regions of varying density based on the domain knowledge of ASL, with respect to hand locations during a signing for both hands.Then, six labels are created; for each hand possesses three labels (X, Y, and Z). The pre-processing module handles the depth data, joint information collection process, and other pre-processing tasks before passing the data to the feature extraction module.

After normalization, each of the sign is represented as a set of joint paths which keeps track of the joint trajectories during the production of that sign. The joint paths of a sign are obtained from the frames of a single skeleton. The number of frames (number of elements in a joint path sequence) is the same for all joint paths belonging to a single sign production, though it might have different values for distinct production of the same sign by the same signer. As a result, training and testing data extracted from skeleton will possibly have some nonlinear variations in time. Therefore, statistical methods are applied in this study to classify the trajectories of joints in the manner that it can evaluate joint paths independent of nonlinear variations of the data characteristics. The hand positions are obtained using the Microsoft Kinect Skeleton Tracker (MKST). They are first transformed into signer coordinate space by subtracting the sign origin from the hand coordinates, and after that, training and testing are done on the dataset for classification purpose. The result of the classification is then used for hand gesture recognition.

Processing with prediction is easier than bottom-up processing (detection, tracking, and characterization) because if errors occur in one of the steps, the remaining steps will be affected. Besides, the measurements for prediction task are more straightforward and easy to check with the prediction processing. The system can predict the hand position according to the hand speed and movement. It can also figure out if the global shape of the hand has changed or not, instead of validation, which is time-consuming. Interpretation is one of the attributes of human-machine interactions, where the machine takes control of it. Therefore, there are expectations on the response and right to exploit it in order to develop a practical system. In each interaction for interpretation, all candidate signs are predicted to be recognized. Then, the performance of the sign language interpretation is evaluated to record the accuracy of the system.

Our proposed method combines static and dynamic sign language recognition to recognize isolated and continuous sign language solely using manual components. Figure 3.2 illustrates the framework of our proposed method for automatic sign language recognition ASLR. The system include three levels of sign language recognition fingerspelling (alphabets) based on one hand , Isolated sign language recognition aims to recognize a single sign gesture at a time (words) which can used one or two hand . On the other hand, continuous sign language recognition aims to recognize the complete sign gesture are performed by two hands, thus, it is a challenging task to recognize it in the real world environment.

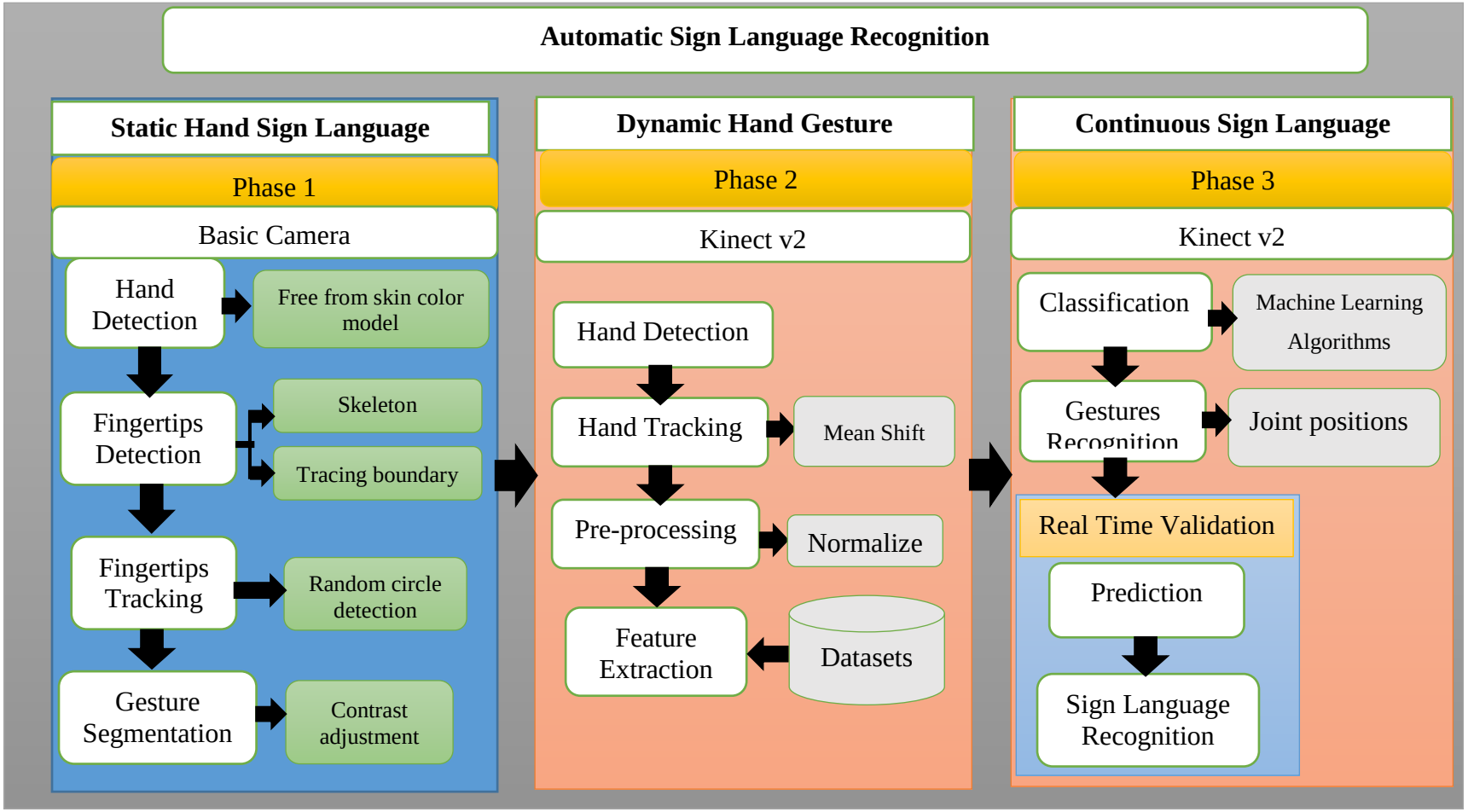


Figure 3.2 Framework of ASLR method is proposed

The research architecture involves the design and implementation of proposed framework and algorithms. Throughout this thesis, main contributions have been presented, which include static gesture for sign language, dynamic gesture for sign language, continuous sign language recognition, and development of novel dataset. Figure 3.3 shows the categorization of hand gestures based on temporal nature.

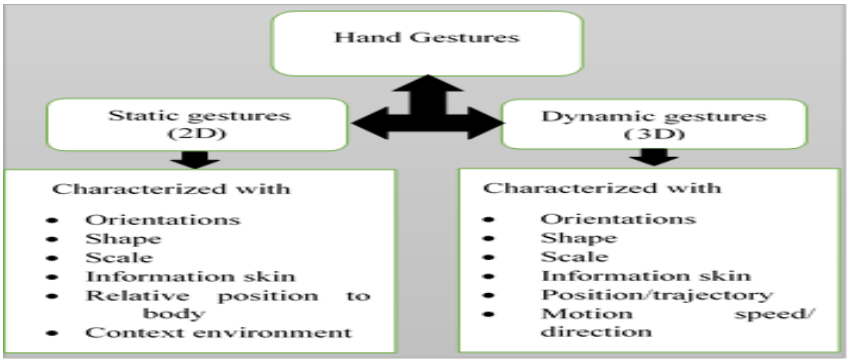


Figure 3.3 Hand gestures based on Static gestures and dynamic gestures

3.3.1 2D Static Gesture for Sign Language

As can be seen in Figure 3.2, 2D static gesture for sign language consist of 4 steps: i) hand segmentation and hand detection, ii) fingertips detection, iii) fingertips tracking, and iv) gesture segmentation.

i. Hand Segmentation and Detection

A new method for hand segmentation and detection is proposed using free-form skin color model. The pixel value of a person's hand is captured and represented in *YCbCr* color model. Problems such as skin color detection, cluttered background removal and changes of lighting are addressed in this phase.

ii. Fingertips Detection

A new method for fingertips detection is proposed based on skeleton algorithm and region of interest. The process consists of five steps which are image acquisition, segmentation, morphological operation, calculation of the hand centroid, and fingertips detection.

iii. Fingertips Tracking

The proposed method modifies the randomized circle detection technique. The randomized circle detection method is chosen based on the fact that for a circle, the gradient vector of the edge pixels will always point to the center of the circle. Therefore, if the gradient vector (which is parallel with the gradient line) for all the edge pixels around the circle is drawn, they will intersect at the center of the circle. Hence, the center of the circle can be determined by choosing the point that has the most number of gradient lines intersect. The intersection plane is considered as an integer plane. Otherwise, it is challenging to get similar values for the intersection points. Based on this principle, since fingertips shape can be assumed to have a semi-circular shape, the algorithm can be employed with some modifications that suit the fingertip shape.

iv. Gesture Segmentation

Gesture segmentation is one of the most critical steps in continuous sign language recognition. It is the task of interpreting and spotting meaningful hand gestures from continuous hand gesture sequences with non-sign transitional hand movements.

3.3.2 3D Dynamic Gesture for Sign Language

In 3D dynamic gesture for sign language, for hand gesture recognition, there are four steps involved which are hand detection and tracking, feature extraction, classification, and hand gesture recognition. Meanwhile, for sign language recognition, the steps include prediction and interpretation.

a. Gesture Recognition

The proposed gesture recognition method is divided into four stages which include hand tracking, feature extraction, classification and hand gesture recognition. Computer vision based techniques use real-time data collected by Kinect cameras. This data is used as an input for feature extraction, which then transformed into domain-based features vectors to be employed in the classification module. A classification module estimates the class these features belong to, before undergoing the recognition phase.

i. Hand Detection and Tracking

Development and use of hand tracking methods are common in $ASLR$. Hand tracking for SLR is a difficult task since the sign gestures are not designed for automatic recognition purposes, and they consist of quick hand movements and occluded body parts. Therefore, various methods are proposed for this non-trivial task in the scope of SLR.

Microsoft Kinect is used for the purpose of hand tracking. It provides different types of frames such as skeleton frame, color frame, and depth frame. Signer performs the gestures in front of the Kinect sensor. The Kinect depth camera has a skeleton tracking capability to track the approximate hand position using the skeleton data. It can track 25 main skeleton joints and the skeleton located closest to Kinect is called primary skeleton. Table 3.1 records the list of the skeleton joints and its index number.

Table 3.1 The joints name and their respective index number for 25 main skeleton joints

Description	Value	Description	Value	Description	Value
Base of the spine	0	Right wrist	10	Right foot	19
Middle of the spine	1	Right hand	11	Spine at the shoulder	20
Head	3	Left hip	12	Tip of the left hand	21
Left shoulder	4	Left knee	13	Left thumb	22
Left elbow	5	Left ankle	14	Tip of the right hand	23
Left wrist	6	Left foot	15	Right thumb	24
Left hand	7	Right hip	16		
Right shoulder	8	Right knee	17		
Right elbow	9	Right ankle	18		

ii. Feature Extraction

Feature extraction provides salient information for real-time automatic gesture recognition and provides a training sample set, a critical data source for the gesture recognition process, which will be stored in a dataset. The skeleton tracker can be

configured to return only the joints in the upper body. The shoulder, elbow and hand positions for the right and left side of the body are obtained. The body pose feature is built using the joint angles of the shoulder and the elbow, the unit vectors of the elbow with respect to the shoulder and the hands with respect to the elbows. Then two joints for two hands; three labels for the left hand $\{L_x, L_y, L_z\}$ and three labels for right hand $\{R_x, R_y, R_z\}$ are considered. There are three parameters for a manual sign which are rotation, transformation and hand shape.

iii. Classification

An overview of SLR approach is given based on employed data acquisition and feature extraction methods. It is obvious that extracting necessary data is a crucial part in SLR However, the success of recognition process depends not only on the performance of the feature extraction method but also on the performance of the classification method. SLR Approaches classify signs after features are extracted. The automatic classification of hand shapes, which are used in sign language communication after feature extraction, based on statistical method is developed for gesture recognition. For isolated and continuous gesture, these methods are found to be suitable for recognition of 3D hand gestures (x, y, z), specifically for ASL.

There are two main components in this phase: Feature extractor and classifier. Feature extractor processes the raw data of the gesture to extract a vector of characteristic features, such as length of the gesture. Once these feature vectors are extracted from the gesture, it is passed to a linear classifier. For each possible output class, the classifier contains a set of weights, which are multiplied by the input vector to yield a measure of the likelihood of that class is correct. In this study, while SVM used feature vectors, other methods depend on matric validation (HCRF, FDR, HMM). The weights of this classifier are found through iterative training on a set of simple gestures. The details of the sample training are discussed in Chapter VI.

In machine learning, one often needs to divide two different datasets, namely training and testing dataset. Essentially, one can use the “sample” command to select