

SIGN LANGUAGE RECOGNITION BY IMAGE ANALYSIS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BUKET BÜYÜKSARAÇ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

FEBRUARY 2015

Approval of the thesis:

SIGN LANGUAGE RECOGNITION BY IMAGE ANALYSIS

submitted by **BUKET BÜYÜKSARAÇ** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Gönül Turhan Sayan
Head of Department, **Electrical and Electronics Eng.**

Assoc. Prof. Dr. Mehmet Mete Bulut
Supervisor, **Electrical and Electronics Eng. Dept., METU**

Prof. Dr. Gözde Bozdağı Akar
Co-supervisor, **Electrical and Electronics Eng. Dept., METU**

Examining Committee Members:

Prof. Dr. Uğur Halıcı
Electrical and Electronics Engineering Dept., METU

Assoc. Prof. Dr. Mehmet Mete Bulut
Electrical and Electronics Engineering Dept., METU

Prof. Dr. Gözde Bozdağı Akar
Electrical and Electronics Engineering Dept., METU

Prof. Dr. Abdullah Aydın Alatan
Electrical and Electronics Engineering Dept., METU

Assoc. Prof. Dr. Sinan Kalkan
Computer Engineering Dept., METU

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : BUKET BÜYÜKSARAÇ

Signature :

ABSTRACT

SIGN LANGUAGE RECOGNITION BY IMAGE ANALYSIS

Büyüksaraç, Buket

M.S., Department of Electrical and Electronics Engineering

Supervisor : Assoc. Prof. Dr. Mehmet Mete Bulut

Co-Supervisor: Prof. Dr. Gözde Bozdağı Akar

February 2015, 82 pages

The Sign Language Recognition (SLR) Problem is a highly important research topic, because of its ability to increase the interaction between the people who are hearing-impaired or impediment in speech. However there are several limitations of the existing methods. Most applications need different necessities like making the user wear multi-colored or sensor based gloves or usage of a specific camera. We propose a simple but robust system that can be used without the need of any specific accessories. The proposed system consists of three main steps. First we apply segmentation to the face and hand region by using Fuzzy C-Means Clustering (FCM) and Thresholding. FCM is a clustering technique which employs fuzzy partitioning, in an iterative algorithm. After the face and hands are segmented, the feature vectors are extracted. The feature vectors are chosen among the low level features such as the bounding ellipse, bounding box, and center of mass coordinates, since they are known to be more robust to segmentation errors due to low resolution images. In total there are 23 features for each hand. After the feature vectors are extracted, they are used for recognition with discrete Hidden Markov Model (HMM). Recognition stage is composed of two stages, namely training and classification. The Baum Welch algorithm is used for HMM training. In classification part the likelihood of

each HMM is calculated and the HMM with the highest likelihood is chosen. In order to measure the success rate of the system, user-dependent and independent tests were conducted for 10 Turkish Sign Language gesture and the system is shown to be working with 85.8% accuracy in the user independent case and 100% in user dependent case.

Keywords: Sign Language Recognition, Machine Vision, Machine Learning, Discrete Hidden Markov Model, Fuzzy C-Means Clustering, Baum Welch

ÖZ

GÖRÜNTÜ İŞLEME TEKNİKLERİYLE İŞARET DİLİ TANIMA

Büyüksaraç, Buket
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Assoc. Prof. Dr. Mehmet Mete Bulut
Ortak Tez Yöneticisi : Prof. Dr. Gözde Bozdağı Akar

Şubat 2015, 82 sayfa

İşaret Dili Tanıma (İDT) problemi işitme ve konuşma engelli insanlar arasındaki engeli aşma yeteneğinden dolayı, çok önemli bir araştırma konusudur. Ancak, mevcut yöntemlerin çeşitli sınırlamaları vardır. Çoğu uygulama kullanıcıların renkli veya algılayıcı tabanlı eldiven kullanmaları ya da özel bir kamera kullanımı gibi farklı ihtiyaçlar gerektirmektedir. Bu tez çalışmasında herhangi bir özel aksesuara gerek kalmadan kullanılabilen, basit, ancak güçlü bir sistem öneriyoruz. Önerilen sistem üç ana adımdan oluşmaktadır. İlk olarak Bulanık Kümelerin Ortalamasıyla Bölütleme (BKO) ve Eşikleme kullanılarak yüz ve el bölgesine bölütleme uygulanır. BKO tekrarlayıcı bir algorithmada bulanık bölütleme kullanan bir kümeleme tekniğidir. Yüz ve eller bölütlendikten sonra özellik vektörleri oluşturulur. Özellik vektörleri sınırlayıcı elips, sınırlayıcı kutu ve ağırlık merkezi koordinatları gibi düşük seviye özellikler arasından seçilir, çünkü onların düşük çözünürlüklü görüntülerden kaynaklanan bölütleme hatalarına karşı daha güçlü oldukları bilinmektedir. Her bir el için toplam 23 özellik vardır. Özellik vektörleri çıkartıldıktan sonra ayrık Saklı Markov Modeli (SMM) ile tanıma için kullanılmıştır. Tanıma evresi eğitim ve sınıflandırma olarak adlandırılan iki aşamadan oluşmaktadır. Baum Welch algoritması SMM öğrenmesi için kullanılmaktadır. Sınıflandırma bölümünde her

SMM olasılığı hesaplanır ve en yüksek olasılıklı SMM seçilir. Sistemin başarı oranını ölçmek amacıyla, kullanıcı bağımlı ve bağımsız testler 10 Türk İşaret Dili jesti için yapılmıştır ve sistemin kullanıcı bağımsız durumda % 85.8 kullanıcı bağımlı durumda % 100 doğruluk ile çalıştığı görülmüştür.

Anahtar Kelimeler: İşaret Dili Tanıma, Bilgisayarla Görme, Otomatik Öğrenme, Ayrık Saklı Markov Modeli, Bulanık Kümelerin Ortalamasıyla Bölütleme, Baum Welch

To My Family

ACKNOWLEDGMENTS

I express my sincere appreciation to my thesis supervisor Assoc. Prof. Dr. Mehmet Mete Bulut and co-supervisor Prof. Dr. Gözde Bozdağı Akar for their guidance, insight and elegant attitude throughout the research.

I wish to thank my parents Hediye and Mesut Adak and my brother Bahadır Adak for their support, encouragement and confidence throughout the years of my education.

I would like to thank to my company ASELSAN and my colleagues for their understanding.

And of course I would like to express my greatest thanks to Serdar for his love, invaluable support, encouragement and patience.

TABLE OF CONTENTS

ABSTRACT	v
ÖZ.....	vii
ACKNOWLEDGMENTS.....	x
TABLE OF CONTENTS	xi
LIST OF TABLES	xiv
LIST OF FIGURES.....	xv
LIST OF SYMBOLS AND ABBREVIATIONS	xvi
CHAPTERS	
1. INTRODUCTION	1
1.1 Motivation	1
1.2 Scope	2
1.3 Outline.....	3
2. LITERATURE REVIEW	5
2.1 Data Acquisition.....	5
2.2 Feature Extraction	8
2.3 Classification.....	11
3. FACE AND HAND SEGMENTATION	15
3.1 Introduction	15
3.2 Theoretical Background on Clustering	17
3.3 Fuzzy C-Means Clustering.....	18

3.4	Post Processing of the Clustered Image	20
3.5	Results from the Recorded Videos	20
3.6	Conclusion.....	22
4.	FEATURE EXTRACTION.....	23
4.1	Introduction	23
4.2	Details of Feature Vectors	26
4.2.1	The Best Fitting Ellipse	26
4.2.2	The Bounding Box	29
4.2.3	Location Information of the CoM	30
4.3	Normalization.....	31
5.	SIGN LANGUAGE RECOGNITION	33
5.1	Introduction	33
5.2	Theoretical Background	35
5.2.1	Classification	38
5.2.2	Training	39
5.3	Left-to-Right Discrete Hidden Markov Models	41
5.4	Hidden Markov Model Application	42
6.	TEST RESULTS AND APPLICATIONS OF THE THEORY	45
6.1	Signer Independent Tests	46
6.1.1	Tests with Equally Weighted Feature Vectors	47
6.1.2	Tests with Different Weighted Feature Vectors.....	51
6.1.3	Tests with Different Number of Training Samples ...	54
6.1.4	Tests with Different Datasets	55
6.2	Signer Dependent Tests.....	58

6.3 Comparison with the Other Implementations	58
7. CONCLUSIONS	61
REFERENCES.....	65
APPENDICES	
A. GESTURES.....	73
A.1 Arm	73
A.2 Box.....	74
A.3 Chess.....	75
A.4 Early.....	76
A.5 Elbow	77
A.6 Lung	78
A.7 Sea.....	79
A.8 Swim	80
A.9 Wound.....	81
A.10 Wrist.....	82

LIST OF TABLES

TABLES

Table 1: Some of the SLR Systems.....	14
Table 2: Hand Shape Features.....	25
Table 3: Gesture Numbers and Names.....	46
Table 4: Signer Independent Test Results.....	47
Table 5: Success Rate of All Subjects in Gesture Base	48
Table 6: Classification Results and Confusion Matrix.....	49
Table 7: Signer Independent Test Results with Weighted Feature Vectors.....	51
Table 8: Subject 1 Success Rate in Gesture Base with Weighted Feature Vectors ...	52
Table 9: Subject 2 Success Rate in Gesture Base with Weighted Feature Vectors ...	52
Table 10: Subject 3 Success Rate in Gesture Base with Weighted Feature Vectors .	53
Table 11: Confusion Matrix with Weighted Feature Vectors	53
Table 12: Tests with Different Number of Training Samples.....	54
Table 13: Comparison with Other Implementations	56
Table 14: Tests with Enterface Dataset.....	56
Table 15: Success Rate of All Subjects in Gesture Base for Enterface Dataset	57
Table 16: Confusion Matrix for Enterface Dataset	57

LIST OF FIGURES

FIGURES

Figure 1: Image Segmentation Steps with Blurred Image	21
Figure 2: Image Segmentation Steps with Lightening Differences	22
Figure 3: The Best Fitting Ellipse	27
Figure 4: The Bounding Box	29
Figure 5: Location Information of the CoM	30
Figure 6: Feature Vector Before Normalization	32
Figure 7: Feature Vector After Normalization.....	32
Figure 8: Diagram of the Hidden Markov Model Parameters	37
Figure 9: 4 State Left-to-Right HMM.....	41
Figure 10: Correctly Classified Feature Vectors for Left Hand.....	50
Figure 11: Correctly Classified Feature Vectors for Right Hand	50
Figure 12: The Ten Postures of the Triesch Data Set	59
Figure 13: Arm Gesture	73
Figure 14: Box Gesture	74
Figure 15: Chess Gesture	75
Figure 16: Early Gesture	76
Figure 17: Elbow Gesture	77
Figure 18: Lung Gesture	78
Figure 19: Sea Gesture	79
Figure 20: Swim Gesture	80
Figure 21: Wound Gesture	81
Figure 22: Wrist Gesture.....	82

LIST OF SYMBOLS AND ABBREVIATIONS

ANMM	: Average Neighborhood Margin Maximization
ASL	: American Sign Language
BPN	: Back Propagation Network
BSL	: British Sign Language
CSL	: Chinese Sign Language
CSR	: Continuous Sign Recognition
FCM	: Fuzzy C-Means Clustering
FS	: Finger Spelling
GSL	: German Sign Language
HMI	: Human Machine Interaction
HMM	: Hidden Markov Model
HNN	: Hopfield Neural Network
IBL	: Instance Based Learning
SD	: Signer Dependent
SI	: Signer Independent
SL	: Sentence Level
SLR	: The Sign Language Recognition
SON	: Self Organization Network
SPT	: Sequential Pattern Trees
SVM	: Support Vector Machine
TDNN	: Time Delay Neural Network
TSL	: Taiwanese Sign Language
WL	: Word Level

CHAPTER 1

INTRODUCTION

1.1 Motivation

People express time-varying motion patterns (gestures) in order to transmit a message to a receiver. Nowadays the most popular receivers are the computers. People spend most of their times by interacting with the touchscreens of their phones or tablets or the touchpads of their laptops either for work or social purposes. What if we can eliminate all the interlayers such as touchpad, mouse, or touchscreen and the computer can detect and distinguish these motion patterns directly as it happens in the movie of Minority Report. This is the research interest of a very huge area called Human Machine Interaction (HMI).

Speech and gesture are the naturally used methods for communication between humans and it is imitated by the HMI systems. The research on speech recognition is more mature than gesture recognition. The systems which use speech as an interaction method is started to be used commercially for years. As a result of the recent development in the quality and availability of the phone and web cameras, gesture recognition has started to attract attention of many researchers, and shown to be a growing potential.

Probably the most explicit example of human gesture is sign language, which has a well-defined vocabulary and grammar. Yet, there has been little advance in the sign language recognition even though the developments in the gesture recognition is very noticeable.

Sign language gestures composed of many elements such as facial expression, body shape and hand movements. However the most important information about the performed gesture is conveyed through the hands. This is the reason why this thesis study is concentrated mainly on the hand gestures.

Most of the researches in the literature are based on the SLR which requires the multi-colored or sensor based gloves or usage of a specific camera. This reduces the user-friendliness of the system. If in the future, these systems will be used in daily life regularly, the recording system shall not include specialized cameras in the recording stage. Because of this reason, in this thesis work, the recordings were made by using the phone cameras, which almost everyone has nowadays. Also, the skin color information is used in the segmentation step in order to prevent the user to wear cumbersome gloves.

According to the [15], one of main problems in the SLR researches is the absence of the signer independent experiments and datasets. Although larger datasets are constructed and tested by some researchers, few of them allow testing them in signer independent way. Actually this is stated as one of the most urgent problems in SLR. Creating very large datasets for further usage are not in the scope of this study, however in order to test our system in user independent way a small vocabulary is created and tested in this thesis work. To sum up, this study is aimed to work and tested for user independence since it is crucial for the system to be used in real world.

1.2 Scope

In this thesis work, a generic method to recognize the isolated sign language gestures in a signer independent way is presented. Vision based techniques are used to extract the feature vectors from the prerecorded videos. Although in the dataset Turkish Sign Language gestures are used, the study could be used to train and test other sign languages or gesture sets which are based on manual gestures.

The system is composed of a phone camera for recording and a computer to evaluate the recorded videos. Since the view of the camera covers the upper body of the performer, using the phone camera provides much more mobility. Besides in future the implemented algorithms in this study could be converted to a mobile phone application and by this way, the applied theory in this thesis work, could reach many people.

Occlusion of the hands with each other and with the head is not in the scope of this thesis. Also the segmentation errors resulted from the extreme lighting conditions are not covered in this study.

1.3 Outline

In this chapter, the reasons why this study is chosen and why it is important are explained. In Chapter 2, previous studies on the SLR are explained. In Chapter 3, the method used for segmentation is explained and some test images are presented. The second step in SLR systems is feature extraction and it is explained in Chapter 4. The recognition method and its application to the presented work are examined in Chapter 5. In Chapter 6, the overall success rates of the system and test techniques to evaluate success of the system are explained. Finally in Chapter 7, the overall system is concluded and some future works to improve the system are commented.

CHAPTER 2

LITERATURE REVIEW

Most of the SLR systems can be divided into the three main steps. First, data acquisition, and tracking of the hand or the body parts must be utilized. In the second step, features describing the manual or non-manual information must be defined and extracted from the input data. In the last step by using the data gathered from the feature descriptors, the performed sign is classified.

Although there are some non-manual features which can change the context of the sign completely, the meaning of a sign language gesture is mostly expressed by using manual features. This thesis covers the manual features and the non-manual features are not in the scope of this study. However there are a few studies which combine the non-manual information with manual ones, interested readers could read the studies in [2], [6] and [58].

2.1 Data Acquisition

The data acquisition methods are mainly composed of two approaches such as vision based and glove based. In vision based methods, sign gestures are captured by a camera. The captured images involve location, posture and motion features of the face, palms and fingers. After the capture phase, an image processing step is required in order to isolate the signer's face and hands from other objects in the frame and the background. Although vision based methods are more natural and useful for real time applications, they are commonly susceptible to noises from different sources such as camera input system, different lighting conditions and complex background. On the other hand in glove based approach, detection of the hand and face is eliminated by

the sensors in the glove. Features are easily extracted and become ready for analysis. The downside of this method is that those gloves are usually expensive and unnatural. Also there are color based gloves which do not include sensors or accelerometers. They are usually in different colors in order to overcome the occlusion of the hands with each other and with the face. They are segmented also by using the vision based techniques but since their colors are fixed it is an easier work to segment them than applying skin color segmentation.

In the early years of the research on SLR, data gloves and accelerometers are used as an example of the glove based systems. The Polhemus tracker [59] and PowerGlove [32] are examples of the sensor based gloves which perform the measurements such as x, y, z location, orientation etc. Vogler and Metaxas used magnetic sensors in combination with the vision based techniques in order to track the hands accurately [55]. Brashear et al. [11] designed a completely wearable system consisting of a hat mounted camera pointing downward and accelerometers on the wrists.

While segmentation and tracking phase of hand, the most common problem is segmenting and tracking accurately the hands in presence of occlusion. There are many sign language recognition systems to solve this problem by using colored gloves in order to overcome the occlusion. Usually these gloves are chosen to be different colors for each hand. In [3], Aran et al. used multi colored gloves in order to track the hands. They used blue and yellow painted gloves and they applied thresholding to the histogram bins in order to segment the hands from the image. In some cases these gloves are designed to emphasize the hand pose and finger positions better. In [63] Zhang et al. used also multi colored gloves where fingers and palms of the hand were painted differently.

Using colored gloves reduces the uneasiness of wearing sensor based gloves, but does not eliminate it completely. For the users of SLR systems the most comfortable and natural approach would be without gloves. In this approach the most common hand detection approach is using the skin color information. In these systems there are some common restrictions such as wearing long sleeves [7], [29]. Also there

could be restrictions on background. In [28], the background is restricted to be a specific color, and in [50] it is restricted to be uncluttered and static.

Using the depth information could simplify the problem. In [39] and [25], hands are segmented based on the assumption that the hands will be the closest objects to the camera. Uebersax et al. [53] also used depth camera in order to recognize the 56 different ASL words constructed with finger spelling.

In [29] Imagawa et al. showed that the segmentation could be made based on solely skin color information, and applied a Kalman filter during the tracking. Holden et al. [26] used snake tracking in order to separate the head from the hands, snake tracking also solves the problem of occlusion. Awad [7] et al. combined skin segmentation, frame differencing and predicted position in a probabilistic way to track the face and hands.

Microsoft™ released Kinect® in 2010, a motion sensing input device which provides the features such as depth-sensing, skeletal tracking and voice recognition. Also it is possible to obtain the raw data from its sensors. This device offers the researchers a short-cut to the tracked data in real time performance. In [17], Doliotis et al. use the Kinect® in order to improve the tracking results of their previously conducted study by using the skin color information for tracking. The results are improved from 20% to 95% on a dataset which consists of 10 different complex gestures. Cooper et al. also extended their previous work [14] to use the 3D tracking ability of the Kinect® sensor [46]. They show the results on two different dataset consisting of 20 Greek and 40 German sign language gestures respectively. They showed that the system is improved and could produce a solution capable of signer independent recognition. Isikligil [30] also used the same dataset obtained from the Kinect® device and used the combination of Sign Graphs and K Nearest Neighbors algorithm with a success rate of 59.3% in signer independent and 91% in signer dependent case.

Tracking the hands could be a burden on SLR systems. Because of this reason some non-tracking based methods are proposed in the literature. Cooper and Bowden [13]

proposed a system such that, instead of detecting the hands, patterns of motion are identified. They compared the study with another system in which the hands are completely detected and tracked using colored gloves. Their system was resulted with 74.3% recognition rate with a very large lexicon composed of 164 different sign.

In Table 1, the summary of the most prominent SLR systems could be seen.

2.2 Feature Extraction

Defining accurate hand shape information is one of the most crucial tasks in gesture controlled computer applications. Specifically in SLR systems, the gesture numbers are large and that makes the hand shape definition more difficult. According to Wu and Huang [60] hand motion has approximately 27 degree of freedom. If 2D images are captured by a single camera that makes extraction of the hand shapes from the video more complex, since the third dimension information is eliminated.

The properties of an efficient hand shape feature descriptor should have certain aspects such as translation, rotation and scale invariance, resistance to the noise and should be identifiable easily. Also the descriptors must be extracted in an acceptable computation time and should be stored without consuming large memory. In literature there are many different hand shape feature descriptors are proposed. Some of them are basic and address the problem of easily extractability and some of them are more complex in order to define the hand shape more accurately. In general these methods can be grouped as appearance based methods in which recording is done in 2D and model based methods in which the recording is done in 3D.

Although the human vision system apprehends the world in three dimensions, for a vision system recording as how it appear in two dimensions is more preferable. Consequently many SLR systems especially the ones which operates in real time prefer appearance based methods due to their low computation time, user friendly environment, cost effectiveness and simplicity. The features used in appearance

based methods can be grouped in two according to the descriptors they use such as region based feature descriptors and texture based feature descriptors.

Region based feature descriptors use the binary image of the segmented hand in order to extract the information such as outer contour of the hand, some geometric features such as area, bounding ellipse, center of the gravity, width and height of the bounding box, axis of least inertia, image moments. Segmentation errors and occlusions in tracking step make extracting the outer contour of the hand accurately more difficult. As a result using region based feature descriptors are encouraged especially when segmentation accuracy is high. However geometric feature descriptors are resistive to noise and by normalizing these features, translation and scale variance can be eliminated. The negative side of these features is they only provide general information about the hand shape. In order to gather more information about hand such as finger positions and orientation one could use texture based feature descriptors. Geometric features are especially used with low resolution systems [16] also could be used in combination with other information such as facial expression [5], [57].

Texture based feature descriptors contain the information about the finger's position and orientation which are especially useful to differentiate the gestures with similar silhouettes.

Orientation histograms are used as a texture based feature descriptors in recent studies [23]. By using the edge direction or distribution of local intensity gradient orientations they describe the shape of an object. The shape information by using the orientation histograms are robust to the illumination change also it gives the system translational invariance. On the other side if the hands occlude, the system could not produce meaningful information hence accurate segmentation of the hands are required. In [43], a system which does not completely rely on hand tracking and very accurate segmentation is proposed. In this study the edge information at the boundary and inside the skin colored areas are used as low level image features. Then, sign

language recognition is performed by using the distribution of pair wise relationships between these low level features.

Local Binary Pattern (LBP) is another example of texture based feature descriptors. In this method local statistical distribution of the pixel values are described by the relationship between the intensity values of neighboring pixel values. The studies which use LBP based methods are generally used to recognize face [18] or facial expressions [51]. In recent times, there is a study which uses LBP based method to describe hand shape information. According to this study LBP features could be used for hand recognition successfully [51].

Discrete Cosine Transform (DCT) is also used to describe the features of hand shape. In [9] DCT is used for American Sign Language letters and digits. Binh et al. converted the spatial representation in the image into frequency representation and calculated the distance between the DCT vectors.

In [21] Scale Invariant Feature Transform (SIFT) features are used to describe the hand shape. Repeatable characteristic feature points are extracted and the descriptors are generated which represents the texture around the feature points by SIFT algorithm from an image. One advantage of using the SIFT features is that, they are invariant to image scale and rotation. Another advantage is the robustness to noise, changes in illumination also minor changes in viewpoint as a result, SIFT features are partially robust to occlusion. This algorithm is the last example of the appearance based methods.

Model based methods are also used as an alternative to appearance based methods. In model based methods, stereo cameras or sensor equipped devices are used to obtain the 3D model of the hand. In [20] and [36] sensor gloves are used to gather the shape information and hand trajectory of each hand. Hand and finger configurations are transformed into joint angle data by the gloves and this information with the hand trajectory information is used directly to recognize the performed signs.

In [57] the SLR system collects a database with a large number of postures seen from many different angles by using the colored gloves with six differently colored visual markers. By using this database two dimensional features and three dimensional hand posture parameters are obtained in a stable and descriptive way. Postures are captured for each frame of the video and a smooth posture sequence is collected at the end of the video sequence.

Another very important feature set for describing hand gesture is the location and the trajectory of the hands in the video. In appearance based methods, hand locations could be obtained by using the center of mass information. In [1], [5], [16], [52] and [58] center of mass coordinates of the segmented hands are used to obtain the trajectory of the hands in the video. As a side effect of using appearance based features, the trajectory information could be affected by the segmentation errors because of the lighting conditions and occlusion. In order to overcome these errors the noise reduction techniques and trajectory smoothing techniques could be applied to the segmentation part. Moreover, in some applications [1], [16] and [52] first order derivative of the hand trajectory, hand velocity, is used as a feature descriptor.

In addition to the hand location, the position of the face could also be a valuable information source as a feature descriptor. It is used as a reference position for the hands in [5], [16], [56] and [58]. Also in some approaches, [19] and [20], in addition to the hand location, the hands relative position with respect to each other is used a feature descriptor.

2.3 Classification

In the previous chapter, the features which describe a gesture are explained. In this chapter, the machine learning techniques, which uses the combination of these features and classify them as a sign, will be explained.

Researches, in early years of the SLR, start with the Artificial Neural Networks (ANN) to classify the sign. Kim et al. [35] used data gloves as data acquisition tool

and 3D coordinates and angles as feature vectors. They trained a Fuzzy Min Max ANN to recognize 25 isolated signs and achieved to a success rate of 85%. Huang et al. [28] used Hopfield ANN to construct a simple isolated sign recognition system. Yang et al. [62] extracted motion trajectories of ASL gestures, and recognize them by using a Time Delay Neural Network (TDNN). Yasir Niaz et al. [42] use a three layer ANN, where the input layer contain 7 neurons and the feature data is gathered by using data glove sensors. In this system the output layer contain 26 neurons each one defines a letter from alphabet. 88% recognition success is obtained by using this system.

Dynamic Time Wrapping (DTW) is also a well-studied approach in the field of bioinformatics and also could be used as signature verification [34], [47] and speech recognition projects [48]. In [16] a small vocabulary of sign language gestures is recognized by using DTW. In this study 5 different gestures are recognized by using the DTW and the hand tracking is done by using the skin color information.

Despite the fact that there are studies which use DTW and ANN for sign language recognition, HMM based methods are shown to be more successful and applicable, on the ground that temporal form of SLR is coped with automatically by the nature of the HMMs [45]. In [52] Starner and Pentland indicated that HMM offer a strong technique for SLR and they use HMM to recognize 40 words from American Sign Language (ASL) by using a coarse description of hand shape, orientation and trajectory.

In [2] an interactive system is developed where signing of users is classified and evaluated by using HMM. In this system colored gloves are used in order to overcome occlusion problem. In addition, this system uses hand motion and shape analysis together with head motion analysis. In the end 7 gestures with 19 variations is trained and classified.

There are many HMM variations with respect to the different implementations. Parallel HMMs are the very notable ones. They are independent HMMs with

separate outputs; the probabilities of them are combined in the end. In [56] 22 different ASL gestures of the left and right hand are modeled using parallel HMMs and a 3D tracking system is used for tracking. The success rate of this system is 84.85%

Liang and OuhYoung [45] constructed a SLR which uses DataGlove as a data acquisition device, and feature vectors such as posture, position, orientation, and motion. They trained three different HMMs, and these HMMs are combined with a weighted sum. The one generating the highest probability is chosen as the winner gesture.

Zieren and Kraiss [64] also used HMM to implement a SLR system and tested their system in signer independent way. They also offer a hand tracking system consisting of multiple hypotheses, which is robust to the different background conditions.

Another popular classification technique for SLR systems is Dynamic Bayesian Network (DBN) and used in [49]. In this study mixtures of DBNs, mixed-state DBNs, and coupled HMMs are compared by using 11 gestures.

Kadous [32] suggested using instance-based learning, K-Nearest Neighbors (KNNs) and decision tree to classify isolated signs. The data was gathered by using the DataGlove. However the results are not as good as ANN or HMM based systems. Therefore instance based approach such as KNN may not be a suitable approach for SLR systems.

Uebachs et al. [53] presented a system recognizing letters and finger spelled words. The letters were classified by using Average Neighborhood Margin Maximization (ANMM), and for word classification the results from the letter classification were summed up. They achieved 87.8% success for a multi user system.

Table 1: Some of the SLR Systems

Work	Acquisition Method	Classification	Dataset	Accuracy (%)
Waldron and Kim [59] (1995)	DataGlove mounted with a Polhemus sensor	BPN SON	14 ASL	86 (BPN) 84 (SON) (Not completely SI)
Kadous [32] (1996)	PowerGlove	IBL and Decision Trees	95 Auslan	80 SD 15 SI
Vogler and Metaxas [55] (1997)	Magnetic sensors and computer vision	HMM	53 ASL	87 SD and CSR
Huang and Huang [28] (1998)	Skin segmentation	3-D modified HNN	15 TSL	91 SD
Starner and Pentland [52] (1998)	Hat Mounted Camera and accelerators	HMM	40 ASL	97 SD and CSR
Yang et al. [71] (2002)	Motion Trajectories	TDNN	40 ASL	93.42 SD
Brashear et al. [11] (2003)	Accelerometers and hat-mounted camera	HMM	5 ASL	90.48 SD
Zhang et al. [63] (2004)	Colored gloves	HMM	439 CSL	92.5 SD
Holden et al. [26] (2005)	Skin segmentation by motion cues and snake algorithm	HMM	163 Auslan	97 SD SL 99 SD WL
Zieren and Kraiss [64] (2005)	Multiple hypothesis	HMM	232 BSL 221 BSL 18 BSL	99.3 SD 44.1 SI 87.8 SI
Cooper and Bowden [13] (2007)	Skin segmentation	2-level classifier including Markov Chain	164 BSL	74.3 SD
Aran et al. [3] (2009)	Colored gloves	HMM	19 ASL	94.34 SD 75.53 SI
Kelly et al. [33] (2010)	Contour segmentation by Canny Edge Detector	SVM	10 static letter gestures	91.8 SI
Uberseax et al. [53] (2011)	Depth camera	ANMM	56 ASL	87.8 SI FS
Ong et al. [46] (2012)	Kinect [®]	SPT	40 GSL	55.4 SI
Isikligil [30] (2014)	Kinect [®]	Sign Graphs and K Nearest Neighbor Algorithm	40 GSL	59.3 SI 91 SD

CHAPTER 3

FACE AND HAND SEGMENTATION

3.1 Introduction

Face and hand segmentation is the beginning stage in many of the SLRs. In order to gather the valuable information and analyze the acquired video correctly, one has to extract the desired data in the entire set of pixels from the sampled image of the video. Since hands do not have a strict shape, and change in size for different gestures, segmentation could be a very challenging task in an SLR. We intend that our SLR system could be used by using a simple phone camera. Therefore the complicated features such as depth or joint angle data is not provided during the data acquisition step. This kind of information is provided for some systems such as [17] and [39] in their data acquisition step and they use specialized devices such as DataGlove or Kinect[®]. Detailed information about these systems and other examples of that kind of systems could be found in Chapter 2.

In our system the input data is colored images, sampled from the prerecorded videos. By using this kind of input, contour information could be used as a segmentation method. In [28] Huang and Huang use first Otsu thresholding and then find the contours. Their idea of hand tracking is based on the assumption that shape of the moving object does not change significantly between two consecutive frames. If the difference between two consecutive frames is under some threshold, they assume that the tracked object is the hand. Also in [26], contour information of the foreground object is used in combination with motion cues and snake algorithm.

Instead of using the shape info, distinctive color information of the human skin could be used as a starting point for face and hand segmentation. In [5], skin color from a training set is learned and a Gaussian Mixture Model (GMM) is created. A look-up table is computed from the probability density function given by GMM. The color of each pixel is compared with the value in the look-up table and decided whether it is skin pixel or not. The blobs are detected from the filtered image and their CoM information is used to track them.

In this thesis work the main concern is the training and classification part. Some of the feature vectors explained in Chapter 4 heavily depends on the hand shape information and in addition to the hand based feature vectors, location of the head is also used. Therefore segmenting the hand and face region accurately is very important to the success of the recognition part. We need an accurate algorithm in segmentation step which gives the shape and location information correctly. Motion based contour algorithms are eliminated since they require an additional algorithm to detect the face area. Also considering the fact that the recognition part heavily depends on training based algorithms, we chose an easy to implement but accurate, non-training based solution. In this thesis work hands and face are segmented by using Fuzzy C-Means Clustering and Thresholding method.

The images are sampled from the video in 2 frame intervals. After the image is obtained, hand recognition system executes. In order to detect the hands and face, a two steps system is designed. First Fuzzy C-Means algorithm clusters the image according to the color information. The clustering step is done according to the method proposed by [12]. Then the mean values of the clusters are calculated for each of the three color component. After that, one of the mean values is chosen by thresholding according to the possible values a skin might have. In the end of this step face and hands are segmented. Face information is used in addition to the hands information by its position values. Hand information contains many aspects such as shape, location and area. These feature vectors are explained in detail in Chapter 4.

3.2 Theoretical Background on Clustering

Clustering can be considered as the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way.

A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

The main requirements that a clustering algorithm should satisfy are:

- Scalability;
- Dealing with different types of attributes;
- Discovering clusters with arbitrary shape;
- Minimal requirements for domain knowledge to determine input parameters;
- Ability to deal with noise and outliers;
- Insensitivity to order of input records;
- High dimensionality;
- Interpretability and usability.

Clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

In the first case, the data is grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. (Example: K-Means)

On the contrary, the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value. (Example: Fuzzy C-Means)

The third type, the hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. (Example: Hierarchical clustering)

Finally, the last kind of clustering uses a completely probabilistic approach. (Example: Mixture of Gaussians) [65]

3.3 Fuzzy C-Means Clustering

Fuzzy C-Means Clustering (FCM) is a clustering technique which employs fuzzy partitioning, in an iterative algorithm. The aim of FCM is to find cluster centroids that minimize a dissimilarity function. The Fuzzy C-Means algorithm

[8] minimizes (J_m) given by Formula (1).

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2 \quad , \quad 1 \leq m < \infty \quad (1)$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j , x_i is the i^{th} of d -dimensional measured data, c_j is the d -dimension center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center. In this project Euclidian distance is used as a similarity measure.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by (2).

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad \text{where} \quad c_j = \frac{\sum_1^N u_{ij}^m \cdot x_i}{\sum_1^N u_{ij}^m} \quad (2)$$

The iteration will stop when (3) occurs. This procedure converges to a local minimum or a saddle point of J_m .

$$\|U^{k+1} - U^k\| < \xi \quad \text{and} \quad |J_m^{k+1} - J_m^k| < \xi \quad \text{where} \quad \begin{matrix} 0 < \xi < 1 \\ \text{k: Iteration Step} \end{matrix} \quad (3)$$

The algorithm is composed of the following steps:

1. Initialize $U=[u_{ij}]$ matrix, $U(0)$
2. At k -step: calculate the centers vectors $C(k)=[c_j]$ with $U(k)$ in (2)
3. Update $U(k)$, $U(k+1)$, calculate cost function by using (1) and (2).
4. When (3) occurs, STOP; otherwise return to step 2. [8]

3.4 Post Processing of the Clustered Image

Clustering only gives the different clusters in the image. In order to determine which one of the clusters is the skin color cluster, thresholding is used. Threshold values are determined a priori by examining the sampled images from different performers' videos and the same threshold values are used throughout the study. Thresholding is applied to the mean value of every cluster's hue, saturation and value components.

The binary image of the face and hands are obtained after clustering and thresholding. According to the difference in lighting conditions and clustering errors there could be gaps in the resulted binary image. Morphological operations are applied to the image by using their related Matlab functions in order to smooth the resulted binary image.

Finally, in the resulted binary image face, right and left hand are separated according to their position information. Face is assumed to be always the uppermost one and left and right hands are assumed to be leftmost and rightmost ones.

3.5 Results from the Recorded Videos

In order to obtain the most accurate feature vectors from the segmented image, videos were recorded in the optimum conditions for segmentation. The room was lightened as equally distributed as possible and the performer's clothes and the background is chosen uniform to avoid the over segmentation errors. Cluster number was given the system a priori and the user wore long sleeve clothes. The camera was focused on the upper body of the signer and the features are not invariant to viewpoint of the camera. Although the classification stage is tolerant of the small rotations that can naturally occur, the performers are required to sign the gestures by facing the camera for an accurate analysis.

Also the segmentation algorithm must be applicable to the low resolution images as well as the high resolution ones. In Figure 1, there are some low resolution captured

images from the sign video which shows us the step by step segmentation process. In this image the blurring effect occurs caused by the fast movement of the hands. It could cause segmentation errors, but in this image the segmentation algorithm overcomes this difficulty successfully. In Figure 2, if we notice the hand section, the shadows in the palm and overly brightened areas in the fingertips could easily be seen. If the segmentation algorithm was unsuccessful, the hand area would be overly segmented after the clustering. Although by looking at the properly segmented face and hand region in the resulted binary image, one could tell that the segmentation algorithm also overcomes this difficulty.

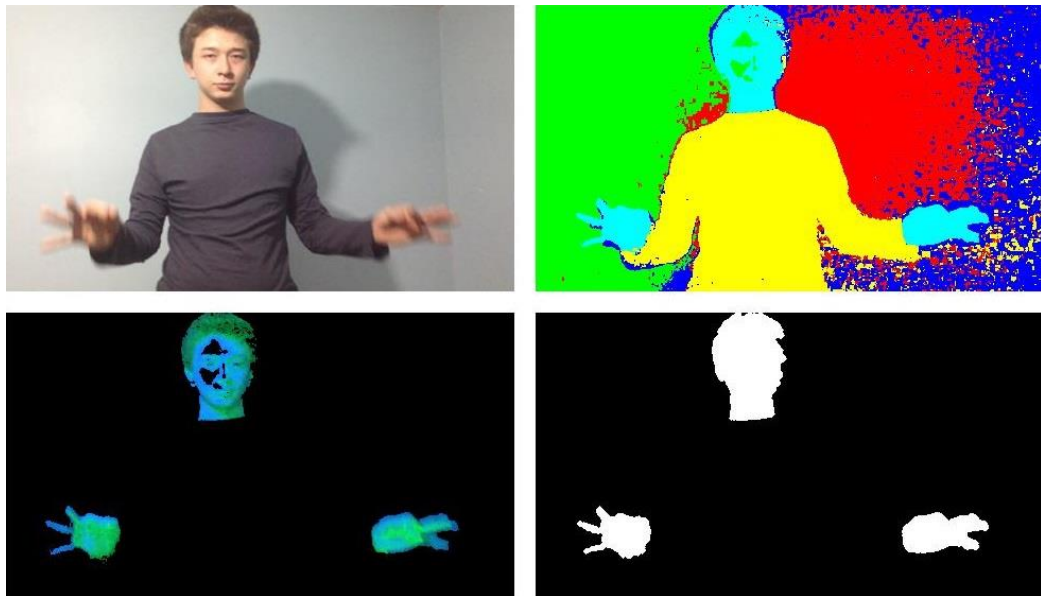


Figure 1: Image Segmentation Steps with Blurred Image

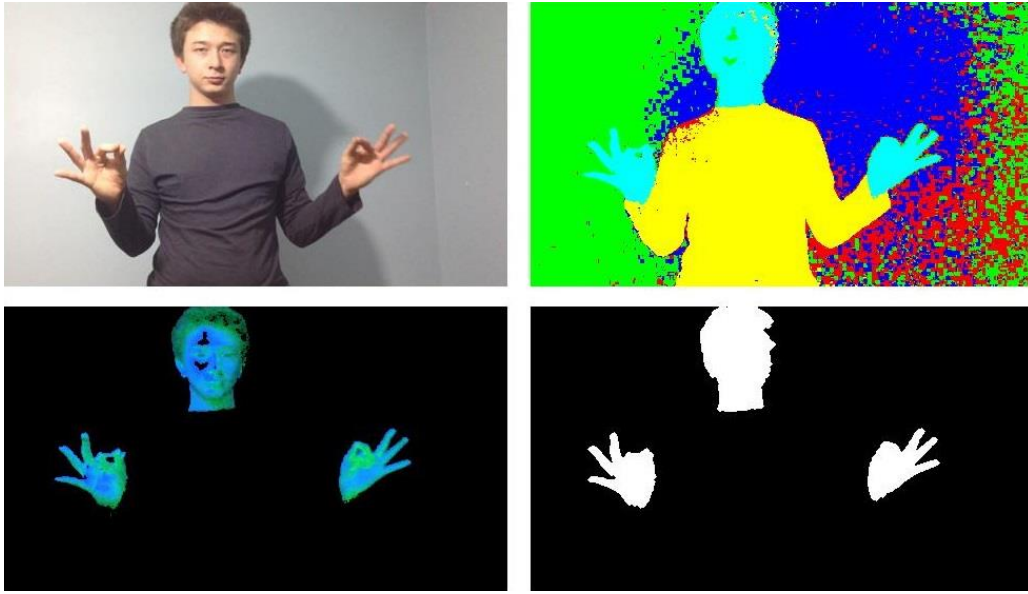


Figure 2: Image Segmentation Steps with Lightning Differences

3.6 Conclusion

The application of the algorithm is applied successfully to the captured images from the video and by using some morphological and thresholding operations; the captured image becomes ready for the next stage, feature extraction.

In conclusion the Fuzzy C-Means clustering combined with thresholding is a very effective method to segment the skin color. In some situation such as the complex background or non-uniform light distribution, it could be boosted with connected component analysis, morphological operations or color constancy algorithms. It is also shown to be an easy to implement algorithm to segment the face and hands from the videos in SLR systems.

CHAPTER 4

FEATURE EXTRACTION

4.1 Introduction

Each gesture in sign language has a specific act of the hands, hand postures and the head. The movement of the head contributes less the meaning of the sign than the hands and hand posture movements, and is not in the scope of this study. Hand shape and finger configuration contribute notably to sign recognition. Furthermore, there are signs which completely build upon the hand shape.

There exist different kinds of feature vectors proposed in the literature. Some of them use the outer contour information of the hand such as bounding ellipse, axis of least inertia, image moments etc. [16]. They could be extracted using the binary image of the segmented hand. Usually they do not provide the information about the specific shape of the hands and finger configurations. There are studies which uses the edge direction and distribution of the local intensity gradients to define the shape of the hands more specifically [23]. LBP, in which statistical distribution of the pixel values is used, is a way to define the hand shape information to extract the feature vectors [51]. Another way is to convert the spatial representation of the image into the frequency domain and to use the DCT vectors as feature descriptors [9]. SIFT features, which are known to be scale and rotation invariant, are also used in some studies as feature vectors [21]. Moreover, motion vectors of the hands and the relative positions of the hands with respect to each other or another static point also carries the information about the gesture and widely used in literature [1], [16], [36] and [56]. The detailed information about the existing feature extraction methods is given in Chapter 2.2.

The system proposed in this study is intended to work with a single phone camera whose field of view covers the upper body of the performer. The hands area is not directly focused and the images used for segmentation have 360x270 resolution. Consequently, the hand area is smaller than 40x40 pixels.


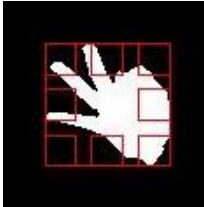
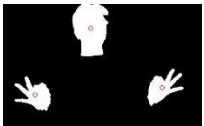
One problem with the overall system is the absence of a proper database for isolated signs without wearing gloves or without cameras equipped with specialized sensors. There exist two databases [4], [41] and both of them include 7 isolated manual signs and both of them are restricted with multi-colored gloves. There are also Kinect[®] based databases [46]. Yet none of them are proper for the presented work in this thesis. In order to create our own database, 3 people are trained for 10 different Turkish Sign Language (TSL) gestures. All of the gestures include both of the hands. Since the performers are not native signers, training the performers takes a great deal of time. Also the variations between the same gestures would be greater than the variations from the case which includes native signers.

Another problem is since the gestures include movements of the hands along a trajectory segmentation errors could happen due to the fast movement. This could reduce the accuracy of hand shape information.

By considering all these problems, the system is decided to work with only low-level features, which are known to be more robust to segmentation errors and compatible with low resolution images [3]. Accordingly simple appearance-based shape features are calculated from the hand contour. The features are selected to emphasize the characteristics of the hand shape and finger configurations. Since different users are used to train the system and hand dimensions show diversity among different people, the features have better to be scale invariant.

Features are selected from the work in [3] with small differences applied in order to make them compatible with our system. All features are listed in Table 2.

Table 2: Hand Shape Features

Feature Number	Method	Feature	Invariant	
			Scale	Rotation
1	Best Fitting Ellipse 	Best fitting ellipse width		•
2		Best fitting ellipse height		•
3		Compactness (perimeter ² /area)	•	•
4		Ratio of hand pixels outside / Total hand pixels	•	•
5		Ratio of hand pixels inside / Total number of pixels inside ellipse	•	•
6		$\sin(2 * \alpha)$ (α = angle of ellipse major axis)	•	
7		$\cos(2 * \alpha)$ (α = angle of ellipse major axis)	•	
8		Elongation (ratio of ellipse major/minor axis length)	•	•
9	Bounding Box 	Percentage of NW (north-west) area filled by hand	•	
10		Percentage of N area filled by hand	•	
11		Percentage of NE area filled by hand	•	
12		Percentage of E area filled by hand	•	
13		Percentage of SE area filled by hand	•	
14		Percentage of S area filled by hand	•	
15		Percentage of SW area filled by hand	•	
16		Percentage of W area filled by hand	•	
17		Total area (pixels)		•
18		Bounding box width		
19		Bounding box height		
20	Center of Mass (CoM) 	Horizontal location of CoM wrt. Horizontal location of head		
21		Vertical location of CoM wrt. Vertical location of head		
22		Horizontal location of CoM		
23		Vertical location of CoM wrt. Vertical location of head Vertical location of CoM		

The feature vectors can be divided into three different subsections according to the method which they are calculated. They are

- The best fitting ellipse
- The bounding box
- Location Information of the CoM

All of them are explained in the Chapter 4.2.

The feature vectors are calculated for both hands separately and after the extraction they are concatenated. After the features are extracted a normalization process is applied to the features explained in Chapter 4.3. Normalization process is the last step in feature extraction phase. After then, an HMM for every feature vector of each hand is constructed. In total 46 HMMs are built and these processes are explained in detail in the Chapter 5.

4.2 Details of Feature Vectors

4.2.1 The Best Fitting Ellipse

First seven of the features in Table 2 are formed on the best fitting ellipse to the hand contour. Ellipse is drawn by using the least squares method explained detailedly in [22]. Ellipses have two mutually perpendicular axes about which the ellipse is symmetric. These axes intersect at the center of the ellipse due to this symmetry. These two axes: the ellipse width, the ellipse height and the angle of the major axis are used as feature vectors and are shown in Figure 3. The angle of the ellipse major axis could be in the range $[0, 360]$. In order to use this 4-fold symmetry of the ellipse this angle α is assumed to be in the range $[0, 180]$ and cos and sin values are calculated according to the 2α .

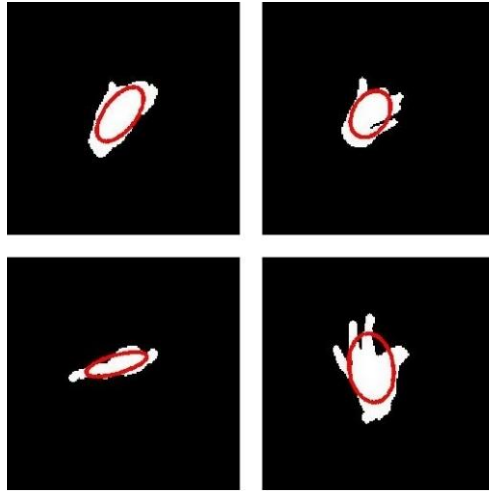


Figure 3: The Best Fitting Ellipse

In addition to the basic ellipse parameters the ratio of the hand pixels outside of the ellipse total hand pixel number and the ratio of the hand pixels inside of ellipse to total number of pixels inside the ellipse are used to contribute the information of the hand shape. The pixels are determined whether they are inside or outside of the ellipse according formula (6).

First pixel coordinates are translated and rotated to align with the ellipse (4) and (5):

x = x coordinate of the pixel

y = y coordinate of the pixel

center x = x coordinate of the center of the ellipse

center y = y coordinate of the center of the ellipse

α = angle of ellipse major axis

X = translated x coordinate

Y = translated y coordinate

$$X = (x - centerx) * cosa + (y - centery) * sina \quad (4)$$

$$Y = -(x - centerx) * sina + (y - centery) * cosa \quad (5)$$

Then if the (6) less than 1 the point is in the ellipse outside it is outside of the ellipse.

$$\frac{X^2}{centerx} + \frac{Y^2}{centery} < 1 \quad (6)$$

Another feature vector which is determined according to the ellipse parameters is compactness. It is invariant to the scale and rotation and calculated according to the following formula (7).

perimeter = perimeter of the best fitting ellipse

area = Hand area

$$Compactness = \frac{perimeter^2}{area} \quad (7)$$

The last feature vector gathered from the best fitting ellipse is elongation. It is calculated according to the formula (8).

$$Elongation = \frac{Ellipse\ major\ axis\ length}{Ellipse\ minor\ axis\ length} \quad (8)$$

4.2.2 The Bounding Box

In order to specify the area parameters more specifically the bounding box is used. Features #9 to #16 in the Table 2 are used to determine finger configurations more accurately. They calculate the percentage of the hand pixels inside the given orientation of the box. The area of the bounding box is divided into eight regions. In Figure 4 the division could be seen. Also bounding box width and height carry information about the hands.

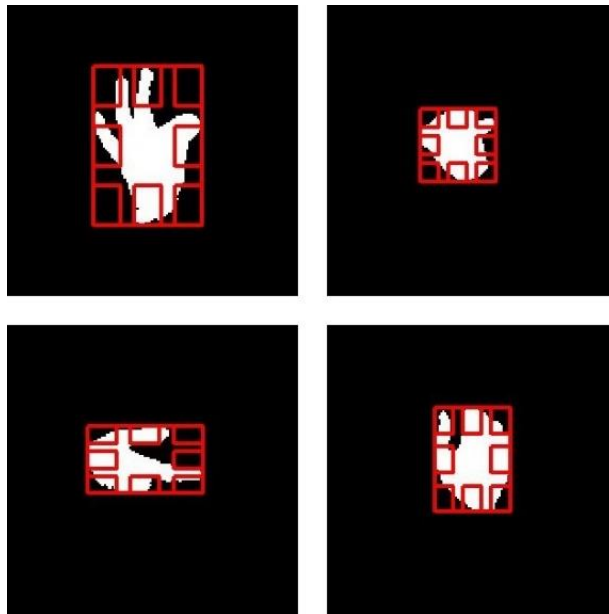


Figure 4: The Bounding Box

4.2.3 Location Information of the CoM

The motion of the hand is processed by tracking the Center of Mass (CoM) of head and hands. Hand location individually and with respect to the head are used as feature vectors. In Figure 5 the trajectory of the hands and CoM indicator could be seen. In addition to the x and y coordinates of the CoM, the relative distances to the head also used as feature vectors in this system.

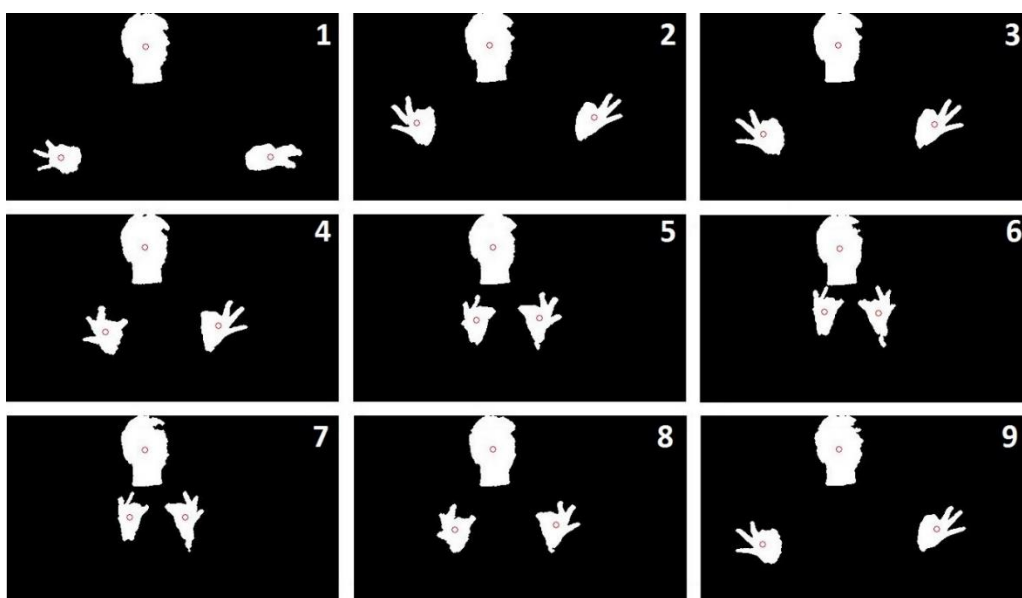


Figure 5: Location Information of the CoM

4.3 Normalization

In this study discrete HMMs are used. All 23 feature vectors should be normalized for the HMM algorithm. Discretization value in the algorithm is 5 and the range of the discretization value is chosen as [1 5] discrete interval in this study. The normalization is done according to the formula (9).

$$F_n = \frac{F - \min}{\max - \min} * \text{Discretization value} \quad (9)$$

F_n is casted to the nearest and lowest integer value. If its value is 0, it is casted to 1. Any value exceeding [1 5] interval is truncated. In Figure 6 and Figure 7 the feature vector which belongs to the x location of the box gesture could be seen before and after normalization respectively.

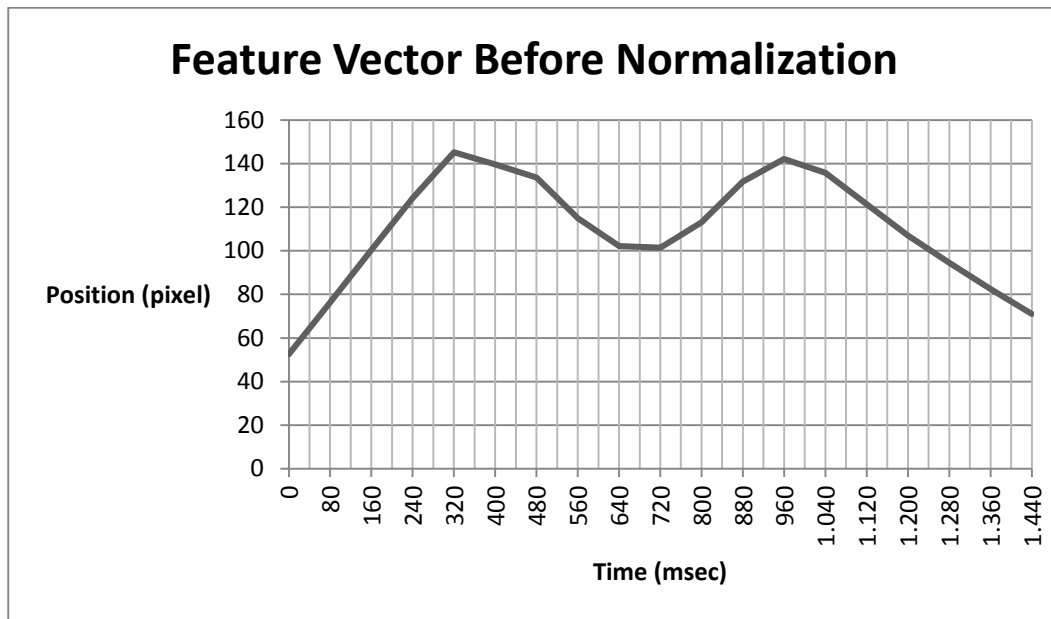


Figure 6: Feature Vector Before Normalization

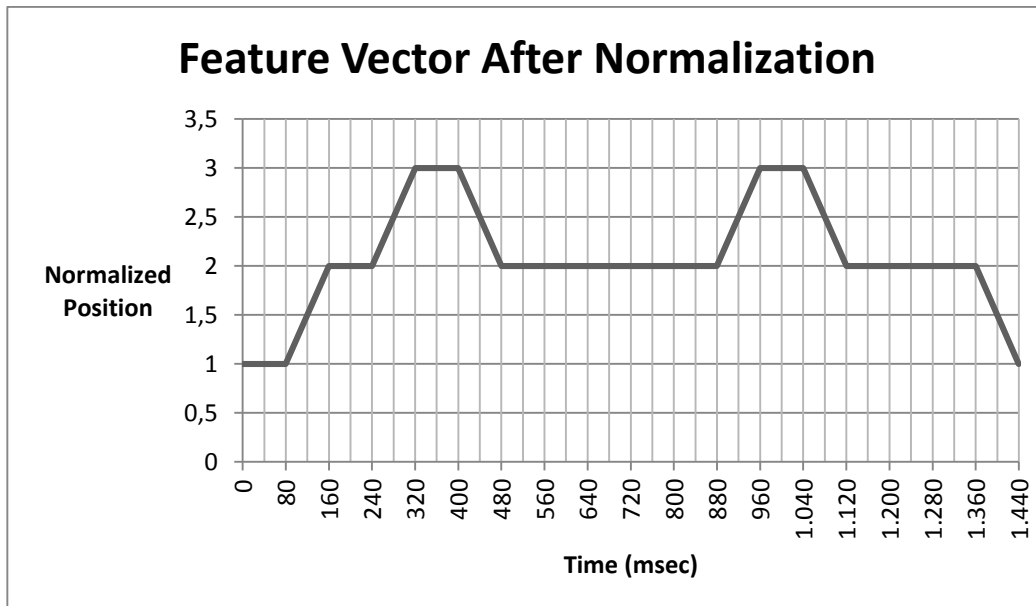


Figure 7: Feature Vector After Normalization

CHAPTER 5

SIGN LANGUAGE RECOGNITION

5.1 Introduction

Recognition is the main step in an SLR system. Previously gathered feature vectors are transformed into the meaningful signs in this step.

If we take a look at the different algorithms used for recognition, the machine learning based ones are more dominant in the literature [15]. Although the first studies in SLR started with ANN [59], HMM based methods are more widely used lately [3], [13]. After the HMMs are applied to the speech recognition successfully, they became widely used for gesture recognition due to the similar nature of two pattern recognition problems. Although the most widely used methods are HMM and ANN, there exist other methods used to classify the sign gestures. Such that, Kelly et al. used SVM in order to recognize 10 different static sign gestures with eigenspace size function and HU moments features [33]. In [46], Ong et al. used SPT with the robust features based on hand trajectories. Different sign recognition algorithms are explained particularly in Chapter 2.3.

In this thesis work, to recognize the gestures HMMs are used. Dealing with the temporal feature of SLR becomes easier due to the nature of the HMMs [45]. HMM is a statistical Markov Model in which the system being modeled is assumed to be a Markov Process with unobserved (hidden) states. A Markov Process is described for the systems which have Markov Property. A process satisfies the Markov Property, if the predictions for the future of the process can be made by looking only on its present state, as well as by looking the process's entire history. They are largely used

in gesture recognition systems. In an HMM there are three parameters which are used to model the gesture, namely state transit probability matrix, symbol output probability matrix, and initial state probability matrix. These parameters are explained in detail in Theoretical Background section. In this thesis study, left-to-right discrete HMMs are used in order to train different gestures. For every gesture, 46 different HMMs are constructed (for each feature vector described in Chapter 4, there exist one HMM) and every one of these HMMs have those three parameters. In order to classify the test video, constructing the HMMs and calculating the HMM parameters properly are very important.

The Baum Welch algorithm is used to find the unknown parameters of an HMM. It makes use of the forward-backward algorithm. The Baum Welch algorithm uses Expectation Maximization (EM) algorithm to find the maximum likelihood estimate of the parameters of a Hidden Markov Model given a set of observed feature vectors.

At first the program randomly estimates the initial values for prior, transition and observation probabilities. After that it computes the expected sufficient statistics for a discrete HMM. While calculating these values, the program uses the forward backward algorithm. Forward backward algorithm returns likelihood value and every time of this iteration the code checks whether the code is converged or not converged by looking into the likelihood and previous likelihood. The algorithm converges if the slope of the likelihood function falls below a threshold. If the values are converged or exceed the max number of iteration the expectation maximization loop finishes with the final values of HMM parameters.

In classification part the observed data is classified with the forward part of the forward backward algorithm with the trained parameters of different gestures. There exist 10 different HMM sets and each set contain 46 different HMMs. The gesture which has the highest value of the likelihood in most of these 46 HMMs is the winner gesture of the classification part.

5.2 Theoretical Background

HMM is a statistical model capable of modeling spatio-temporal time series. An HMM has a finite set of states governed by a set of transition probabilities. In a particular state, an outcome or observation can be generated according to an associated probability distribution. HMM is used in robot movement, bioinformatics, speech and gesture recognition. They are usually chosen for their capability to grant an efficient way of handling the temporal variability among sequences and missing data [50].

Using HMM for sign recognition is motivated by the successful application of the techniques of Hidden Markov Model to speech recognition issues. The similar points between speech and sign suggest that effective techniques for one problem may be effective for the other as well. First, like spoken languages, gestures vary according to position, social factors, and time. Second, the movements of body, like the sounds in speech, transmit certain meanings. Third, signs regularities performances while speaking are similar to syntactic rules. Therefore, the methods elaborate by linguistic may be used in sign recognition. Sign recognition has its own characteristics and issues. Because sign is an expressive motion, it is natural to describe such a motion through a sequential model. Based on these criterions, Hidden Markov Model is appropriate for sign recognition. A multi-dimensional HMM is able to deal with multi-path signs, which are general cases of sign recognition.

An HMM is composed of a number of states each of them has a probability of transition from one state to another. With time, state transitions occur probabilistically. States at any time depend only on the states at the preceding time as a property of being a Markov Model. As a result of being a Hidden Markov Model, states are not directly observable, and can be predicted only through a sequence of observed symbols.

To describe a discrete HMM the following definitions are made:

T = length of the observation sequence.

$Q = \{q_1, q_2, \dots, q_N\}$: set of states.

N = number of states in the model.

$V = \{v_1, v_2, \dots, v_M\}$: set of possible output symbols.

M = number of observation symbols.

$A = \{a_{ij} | a_{ij} = \Pr(s_{t+1} = q_j | s_t = q_i)\}$: state transit probability, where a_{ij} is the probability of transiting from state q_i to state q_j .

$B = \{b_j(k) | b_j(k) = \Pr(v_k | s_t = q_j)\}$: Symbol output probability where $b_j(k)$ is the probability of output symbol v_k at state q_j .

$\pi = \{\pi_i | \pi_i = \Pr(s_1 = q_i)\}$: Initial state probability.

$\lambda = \{A, B, \pi\}$: Complete parameter set of the model

Using this model, transitions are described as follows:

$S = \{s_t\}, t = 1, 2, \dots, T$: State s_t is the t^{th} state (unobservable)

$O = O_1, O_2, \dots, O_T$: Observed symbol sequence (length = T)

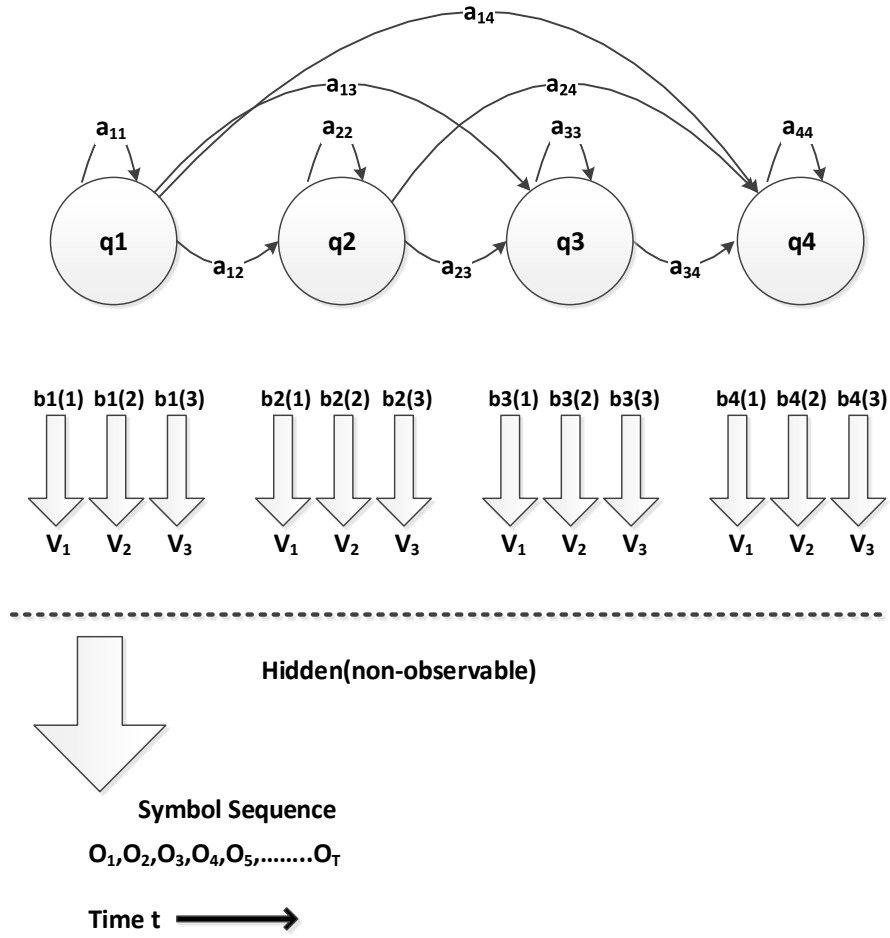


Figure 8: Diagram of the Hidden Markov Model Parameters

As can be seen from Figure 8, the concept of HMM is illustrated with a transition graph. The states are represented by circles and each line shows the transition from one state to another also there are transition probabilities indicated by the character alongside the line. There are also transition paths from states to themselves. These paths allow the HMM to stay in the same state for any duration. This property of HMM is important because the system is time-scale invariant due to these transitions. Each state of the HMM probabilistically outputs a symbol. In state q_j , symbol v_k is the output with a probability of $b_j(k)$. If there are M kinds of symbols, $b_j(k)$

becomes an $N \times M$ matrix. The HMM outputs the symbol sequence $O = O_1, O_2, \dots, O_T$ from time 1 to T. The HMM states are unobservable, only the symbol sequence outputs are observable. The initial state of the HMM is also determined stochastically by the initial state probability π . An HMM is characterized by three matrices: state transit probability matrix A, symbol output probability matrix B, and initial state probability matrix π . These parameters are determined during the training process; one HMM is constructed for each category to be recognized. In classification part the system determines which HMM could produce the observed symbol sequence. The training and classification parts are explained in detail in the following sections.

5.2.1 Classification

In order to classify the observed symbol sequences, one HMM is created for each classification category. Let's assume that there are C categories, the model which best fits the observed data is chosen amongst C HMMs $\lambda_i = \{A_i, B_i, \pi_i\}$, $i = 1 \dots C$. This means that when an observation data of unknown category is given, $\Pr(\lambda_i|O)$ value is calculated for each HMM λ_i , and λ_{c^*} is selected, where (10).

$$c^* = \operatorname{argmax}_i (P(\lambda_i|O)) \quad (10)$$

If the observation sequence is $O = O_1, O_2, \dots, O_T$ and HMM is λ_i , $P(\lambda_i|O)$ is calculated by using the forward algorithm [27].

The forward algorithm is defined in (11).

$$\alpha_t(i) \equiv P(O_1, O_2, \dots, O_t, s_t = q_i | \lambda_i) \quad (11)$$

$\alpha_t(i)$ is called the forward variable and can be calculated recursively as follows:

$$\alpha_t(j) = \left\{ \sum_i \alpha_{t-1}(i) a_{ij} \right\} b_j(O_t) \quad (12)$$

$$\alpha_1 = \pi_i \cdot b_i(O_1) \quad (13)$$

Then

$$P(O|\lambda) = \sum_{i \in S_T} \alpha_T(i) \cdot \lambda_{c^*}, \quad \text{where } c^* = \operatorname{argmax}_i (\Pr(\lambda_i|O)) \quad (14)$$

In classification part the likelihood of each HMM is calculated by using the above formula and the HMM with the highest likelihood is chosen. Because of the fact that the likelihood is calculated by using the entire pattern length as described above, time scale variance, time shifts and some failure in vector quantization have little effect on the accuracy of the final likelihood. This factor brings the advantage of HMMs for time-sequential pattern recognition: robustness to time-scale variance and shift.

5.2.2 Training

In the training phase each category of HMMs are trained so that it gives the most likely parameter set for its category. In other words, training an HMM means optimizing the model parameters (A, B, π) in order to maximize the probability of the observation sequence $\Pr(O|\lambda)$. To find the unknown parameters of HMM and related estimations Baum-Welch algorithm, which makes use of the forward backward algorithm, is used.

Define:

$$\beta_t(i) \equiv P(O_{t+1}, \dots, O_T | s_t = q_i, \lambda) \quad (15)$$

$\beta_t(i)$ is called the backward variable and can be solved recursively in a similar way to the solution of $\alpha_t(i)$ as explained in classification part.

$$\gamma_t(i) \equiv P(s_t = q_i | O_1, O_2, \dots, O_t, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} \quad (16)$$

$$\varepsilon_t(i, j) \equiv P(s_t = q_i, s_{t+1} = q_j | O_1, O_2, \dots, O_t, \lambda) = \frac{\alpha_t(i)\alpha_{ij} b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} \quad (17)$$

Using these equations, HMM parameter set λ can be improved to $\bar{\lambda}$ by using the Expectation Maximization algorithm. The reestimation equations from $\lambda = (A, B, \pi)$ to $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$ are:

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \varepsilon_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (18)$$

$$\bar{b}_i(k) = \frac{\sum_{t \in \{t | O_t = v_k\}} \gamma_t(i)}{\sum_{t=1}^T \gamma_t(i)} \quad (19)$$

$$\bar{\pi}_i = \gamma_1(i) \quad (20)$$

Training phase converges if $(\bar{\lambda} = \lambda)$. The Baum-Welch algorithm does not always find the global maximum, but it finds the local maximum of $(P(O|\lambda))$. In practice according to the experiments, it is not a significant problem.

5.3 Left-to-Right Discrete Hidden Markov Models

In this thesis work left-to-right discrete Hidden Markov Model is used in the recognition step. In left-to-right HMMs, the transition to the other states is blocked, as a result the states are ordered in time. Transition is only permitted when it is made to a state with an index that is greater than or equal to the index of the current state. In this thesis work 4 state HMM is used to model the gestures. The diagram of the 4 state left-to-right HMM could be seen in Figure 9.

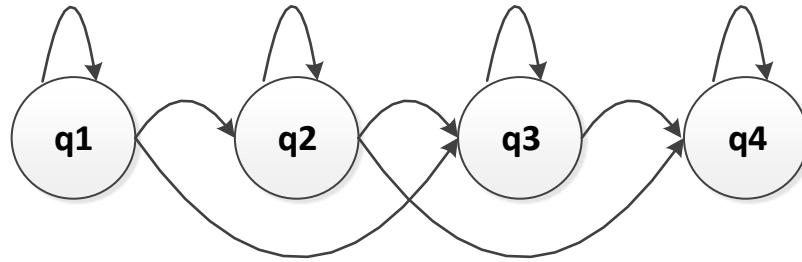


Figure 9: 4 State Left-to-Right HMM

In the HMMs considered above, the state space of hidden variables is discrete, while the observations themselves can be discrete or continuous. The way to model the observations identifies whether the model is continuous HMM or discrete HMM. In this thesis work the feature vectors are discretized in [1 5] interval.

5.4 Hidden Markov Model Application

After the segmentation and feature extraction steps feature vectors are ready for recognition phase. For every different gesture there exists 5 different train videos used in order to train one gesture for only one performer. In total 3 different performers are trained to perform the selected 10 different signs. However the number of the videos depends on the user and can be increased. Not all the frames are processed in the video; instead frames are picked in two frame intervals. This decrease the segmentation time and one frame in two frame intervals are more than enough for training and classification parts.

There are 23 feature vectors for each hand in every frame. These feature vectors are explained specifically in Chapter 4. 46 feature vectors (23 for each hand) are recorded in different times. There are 5 different feature vector set for every gesture and for each performer. In training stage 2 of the performers data is combined and 10 feature vector for each gesture is obtained. For testing one performer's data is used and 5 different test feature vector could be used for testing in user independent case. For user dependent classification 4 feature vector set of every gesture and for each performer is used and the remaining 1 feature vector set is used for classification.

While constructing the HMMs, parallel training approach is accepted. In parallel HMMs N different channel with N independent outputs modeled. In our case the N channel represents the 23 different feature vectors for each hand and there exist 46 different channels in total. The state probabilities only affect each other if they are in the same channel. Parallel HMMs are based on the assumption that the outputs of the different channels are statistically independent from each other. In classification part the probabilities for each different channel of HMMs are calculated separately and the one which obtains the maximum likelihood values from most of the channels is the winner gesture. As shown in Chapter 6.1.2 not all of the feature vectors affect the system in the same way. This voting approach allows us to assign the feature vectors, which carry more importance, higher weight values. Also in literature the parallel

approach in HMMs is shown to be a more suitable and successful way to model SLR systems [31], [56].

Feature vectors are quantized in [1 5] discrete interval before the training stage begins. Finding the proper discretization interval is one of the challenging parts of the training process. Although at first it seems that using larger discretization intervals such as [1 50] would increase the recognition rate, it actually decreases. Because the likelihood parameters are decreased between different videos which belongs to the same gesture when large intervals are used. Even the performer is the same person for all gestures; it is not possible that she/he performs the gesture exactly the same way in all different training and test videos. Three different intervals are used during implementation process of the Hidden Markov Model. In [1 50] interval the system hardly recognizes at most 3 gestures in 5, in [1 10] interval the success rate increase significantly such as at most 9 different gesture in every 10 different gesture. Whereas in [1 5] interval the system could recognize 10 gestures.

CHAPTER 6

TEST RESULTS AND APPLICATIONS OF THE THEORY

In this thesis the accuracy of the recognition performance is evaluated by using a dataset which contains 10 different gestures from Turkish Sign Language. Table 3 lists the signs used in this study. All these signs include both of the hands. Occlusion of the hands with each other and with the head is not in the scope of the study and hands are chosen such that they are never required to occlude. For each sign, five repetitions from three subjects are recorded. In total there are 150 videos used in training and test stages of this system. Subjects do not have pre-knowledge on sign language performance and learned the signs for the first time for this study.

The worthiness of a recognition system comes from its ability to generalize a learned concept to the new instances. In sign language there are two instances for a particular sign. In the first one the performer whose videos are in the training set, performs the same sign in a different time and it is used as test case. These tests are called as signer-dependent tests. In the second one the performers whose videos are used for training and testing are completely different people. These tests are called the signer independent tests. Both of these test cases are applied to the designed system in order to measure its performance in an accurate way. Although there are many studies which publish only the signer dependent results, in order to correctly show the success of a SLR system, measurements with the signer independent method gives the actual accuracy.

Table 3: Gesture Numbers and Names

Gesture Number	Gesture Name
1	Lung
2	Wrist
3	Box
4	Sea
5	Elbow
6	Early
7	Arm
8	Chess
9	Wound
10	Swim

6.1 Signer Independent Tests

The success of signer independent experiments is the main concern of this study since SLR systems are designed to overcome the communication problem of different signers in real world. In these tests videos gathered from one signer are retained for the test set and eliminated from the remaining training set. The same procedure is applied to all three signers and by this way three-fold cross validation is obtained.

For each sign there are 60 training operations and in total 600 training operations are performed. In each training instance, each gesture is tested with 5 different videos including the same sign and recorded in different time instances. In the overall test system, 3000 classification test is performed for all of the three performers.

6.1.1 Tests with Equally Weighted Feature Vectors

In this test case the weights of the feature vectors are equal. After the training step of recognition is performed, there exist 46 HMM for every feature vector. The test video is classified for every feature vector. There exist 46 different classification result for one test video. After the feature vector classification ends, the voting stage begins. The gesture which wins the most of the 46 feature vector is the winner gesture. In these tests, the weights of the feature vectors are equal in voting.

The success rate of these experiments is given in Table 4 and Table 5.

Table 4: Signer Independent Test Results

Subject Name Used for Test	Subject 1	Subject 2	Subject 3	In Total
Trial Number	1000	1000	1000	3000
Correct Classification Number	805	983	707	2495
Success rate of the Classification (%)	80.5	98.3	70.7	83.17

As can be seen from the results the success rates are very depended on the performer. Subject 2 has a very high recognition rate whereas Subject 3 has the lowest. The success rate would be higher, with the native signers' performance.

In addition to the overall success of the system for every gesture the recognition rates are calculated in subject base. They are given in Table 5 for each subject. According

to the results Box and Sea are recognized by every signer with a 100% success rate. In contrast wound and elbow have the lowest recognition rate among all the gestures.

Table 5: Success Rate of All Subjects in Gesture Base

	Subjects	Lung	Wrist	Box	Sea	Elbow	Early	Arm	Chess	Wound	Swim
Accuracy (%)	S1	99	45	100	100	77	59	89	100	41	95
	S2	98	100	100	100	95	100	100	98	94	98
	S3	76	90	100	100	10	82	94	89	32	44

The confusion matrix for 3000 trial is shown in Table 6. According to these results, most confused gesture is wound. The rough definition of the gestures can be seen from Appendices. Wound is mostly confused with swim. If we take 14 wrongly classified wound sample, in 10 of them, 27th feature vector, (ratio of hand pixels outside / total hand pixels of the right hand) is confused with the swim gesture. Other most misclassified wound feature vectors are 16th, 26th and 29th (percentage of west area filled by hand of right hand, compactness of the left hand and sin value of the left hand respectively). They are confused 7 times out of 14 with swim gesture.

Table 6: Classification Results and Confusion Matrix

	Lung	Wrist	Box	Sea	Elbow	Early	Arm	Chess	Wound	Swim
Lung	273	24	1						1	1
Wrist	55	235	1		4				5	
Box			300							
Sea				300						
Elbow		16	6	6	182	51	29	2	8	
Early	1	1			3	241	2	48		4
Arm		1		3	4	7	283	2		
Chess						7	3	287	3	
Wound	2	12	1	2	1	1		23	157	101
Swim	3			50					10	237

As explained in Chapter 5 in detail, for every feature vector, an HMM is created. In classification step, these HMMs are evaluated and the winner gesture is the gesture which gains most of the feature vector HMMs. Correctly classified feature vector HMMs are examined and showed in a graphic form in Figure 10 and Figure 11 for left and right hand. Although these graphics are changeable according to the chosen gesture set, location feature vectors between 20 and 23 seems to be more influential to classify the gesture correctly. The feature vectors which correspond to the number in x direction of the graphics could be matched by using Table 2.

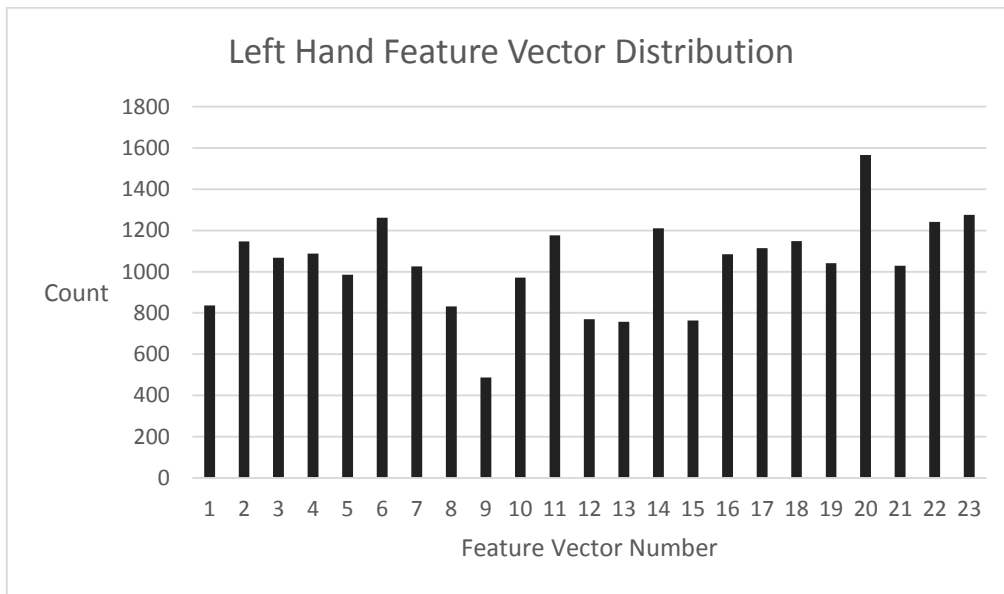


Figure 10: Correctly Classified Feature Vectors for Left Hand

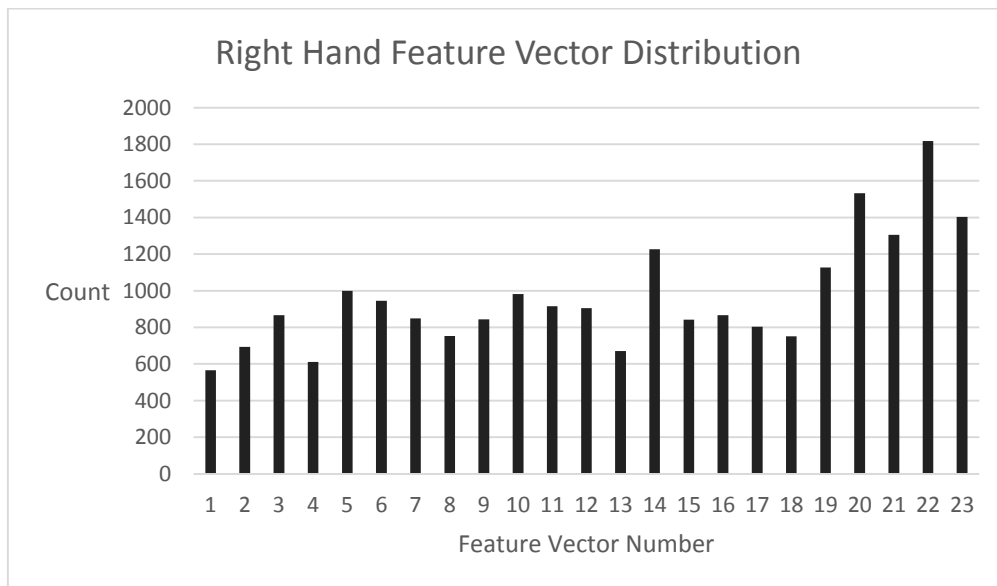


Figure 11: Correctly Classified Feature Vectors for Right Hand

6.1.2 Tests with Different Weighted Feature Vectors

As can be seen from Figure 10 and Figure 5, the feature vectors are not equal in contribution to the success rate. Some of them gives more accurate results and contribute more to the gesture to be classified correctly. For that reason, in this test step the weights of the feature vectors in the voting stage are changed according to the success rate in Figure 10 and Figure 5. The more successful feature vectors are weighted with a higher value. For example in Figure 10 the 23th feature vector is weighted with 1818, while first feature vector is weighted with 566.

The overall result of these tests is increased from 83.17% to 85.8% as can be seen in Table 7 with weighted feature vectors. Also if we examine the signer performances individually, we can see the increase in success rate. In Table 8, Table 9 and Table 10, subjects' success rates in gesture base are presented. If we look at the confusion matrix in Table 11, it can be seen that the mostly confused gestures are less confused but their names are not changed by the change in feature vectors' weights.

Table 7: Signer Independent Test Results with Weighted Feature Vectors

Subject Name Used for Test	Subject 1	Subject 2	Subject 3	In Total
Trial Number	1000	1000	1000	3000
Correct Classification Number	852	988	733	2573
Success rate of the Classification (%)	85.2	98.8	73.3	85.8
Previous success rate of the Classification (%)	80.5	98.3	70.7	83.17

Table 8: Subject 1 Success Rate in Gesture Base with Weighted Feature Vectors

Gesture	Lung	Wrist	Box	Sea	Elbow	Early	Arm	Chess	Wound	Swim
Accuracy (%)	96	73	100	100	83	65	95	100	45	95
Previous Accuracy (%)	99	45	100	100	77	59	89	100	41	95

Table 9: Subject 2 Success Rate in Gesture Base with Weighted Feature Vectors

Gesture	Lung	Wrist	Box	Sea	Elbow	Early	Arm	Chess	Wound	Swim
Accuracy (%)	99	100	100	100	94	100	100	99	98	98
Previous Accuracy (%)	98	100	100	100	95	100	100	98	94	98

Table 10: Subject 3 Success Rate in Gesture Base with Weighted Feature Vectors

Gesture	Lung	Wrist	Box	Sea	Elbow	Early	Arm	Chess	Wound	Swim
Accuracy (%)	74	93	100	100	14	83	99	92	28	50
Previous Accuracy (%)	76	90	100	100	10	82	94	89	32	44

Table 11: Confusion Matrix with Weighted Feature Vectors

	Lung	Wrist	Box	Sea	Elbow	Early	Arm	Chess	Wound	Swim
Lung	269	25	1					1		4
Wrist	26	266			1				6	1
Box			300							
Sea				300						
Elbow		15	1	3	191	30	44	8	8	
Early	1				2	248	1	44		4
Arm		1			2	2	294	1		
Chess					1	1	1	291	6	
Wound		19	1	1	1			19	171	88
Swim	3			45					9	243

6.1.3 Tests with Different Number of Training Samples

In machine learning systems increasing the number of the training samples usually increase the success rate of the system. In order to measure the dependence on training samples different tests are conducted with 4, 6 and 10 different training samples for each gesture gathered from 2 different signers and a completely different test signer. According to the results in Table 12, as the training samples increased from 4 to 10 the total success rate increases from 76.3% to 85.8%. This is a very promising result because the other HMM based SLR systems require more training samples than the one implemented in this thesis study. For example in [3] 5 samples gathered from 7 different performers, in total 35 training samples are used in order to test the system. If the training samples can be increased by adding the different signers' performances, the result of the implemented system can be improved.

Table 12: Tests with Different Number of Training Samples

Subject Name Used for Test	Subject 1	Subject 2	Subject 3	In Total
Trial Number	1000	1000	1000	3000
Success rate of the trial with 4 samples (%)	78.4	86.2	64.4	76.3
Success rate of the trial with 6 samples (%)	79.7	91.8	70.3	80.6
Success rate of the trial with 10 samples (%)	85.2	98.8	73.3	85.8

6.1.4 Tests with Different Datasets

In this section the test is done with eNTERFACE dataset [66]. The dataset belongs to the work done in [3]. In this dataset the performers use multi-colored gloves. The segmentation algorithm in this thesis work is modified in order to recognize the multi-colored gloves. There exist 8 different ASL gestures with 19 variations in meaning. Since the head motion is not in the scope of this thesis, variations are excluded from the gesture set, only base signs are considered. Therefore the tests are conducted by using the 8 base signs. There exist 8 different signers who perform the same gestures. Tests are performed in the same way as explained in the previous chapters. One signer's performance is chosen as test set and excluded from the train test and this procedure is applied to the all eight signers.

For each sign there are 80 training operations and in total 640 training operations are performed. In each training instance, each gesture is tested with 5 different videos including the same sign and recorded in different time instances. In the overall test system, 3200 classification test is performed for all of the three performers.

The success rate of the overall system is given in Table 14. 94.19% success rate is achieved in 3200 trials. A gesture is assumed to be successfully classified, if it is the exact equivalent of the performed sign amongst 8 gestures. In [3], the test is conducted by using the same dataset but they broaden the 8 gesture base dataset with head and hand motion differences to 19 gestures and assume a gesture is classified correctly if the classified sign is in the correct base sign. They achieved 99.74% accuracy. They also conducted the tests in which a sign is classified correctly if it is the exact equivalent of the same sign amongst 19 gestures. In these tests they achieved 75.53% success rate. The comparison of the two systems is given in Table 13.

Subject dependent results in gesture base and confusion matrix are shown in Table 15 and Table 16 accordingly. According to these results the most confused gesture is

fast and it is mostly confused with look at. If we look at the two gestures they are more similar than the other gestures in the dataset.

Table 13: Comparison with Other Implementations

	[3] HMM / base sign (%)	[3] HMM overall (%)	Our Work (%)
S1	100	73.68	99.75
S2	100	91.58	96.25
S3	100	71.58	100
S4	100	81.05	99
S5	96.84	62.11	87
S6	100	81.05	96.25
S7	100	65.26	97
S8	100	77.89	78.25
Total	99.61	75.53	94.19

Table 14: Tests with Enterface Dataset

	S1	S2	S3	S4	S5	S6	S7	S8	Total
Trial	400	400	400	400	400	400	400	400	3200
Correct	399	385	400	396	348	385	388	313	3014
Success Rate (%)	99.75	96.25	100	99	87	96.25	97	78.25	94.19

Table 15: Success Rate of All Subjects in Gesture Base for Enterface Dataset

	Subjects	Afraid	Clean	Door (noun)	Drink (noun)	Fast	Here	Look at	Study
Accuracy (%)	S1	100	100	100	100	100	100	98	100
	S2	100	100	100	86	100	100	84	100
	S3	100	100	100	100	100	100	100	100
	S4	100	100	92	100	100	100	100	100
	S5	100	100	100	66	100	100	42	88
	S6	100	100	100	72	100	100	98	100
	S7	80	100	100	100	100	100	100	96
	S8	100	90	100	100	6	66	100	64

Table 16: Confusion Matrix for Enterface Dataset

	Afraid	Clean	Door (noun)	Drink (noun)	Fast	Here	Look at	Study
Afraid	390							10
Clean	4	395					1	
Door (noun)			396				4	
Drink (noun)	4	17		362	1	16		
Fast	13	1		7	353		20	6
Here				17		383		
Look at	7	3	13	8	9		359	1
Study	6	15					3	376

6.2 Signer Dependent Tests

In signer dependent tests the performer in the test set and the training set are the same person, videos of the same sign are recorded in different times. Four video of the same signer are used for training and the remaining 1 video is used for testing purpose for every gesture. For each gesture and for every performer the system is trained 100 times and tested 100 times. The system achieved 100% recognition rate in this test case, there is no misclassified sign in all of the 3000 tests.

6.3 Comparison with the Other Implementations

The scope of this thesis is vision based approach in sign language recognition. As a result the comparisons are mainly performed among this kind of studies. Comparison with the systems which uses specialized capturing devices are excluded in the comparisons however other vision based systems are used to evaluate the success of the system.

In signer dependent experiments very high recognition rates are very common. As an example in [26], the system recognized 22 Auslan words with a 95% success rate with HMM (in 2005), but the sign representation mainly deals with the global motion in space with limited information on local motion of the hands. Therefore, the proposed HMM system and the chosen feature vectors cannot deal with the hand shape and orientation differences between different signs. Even so, this work segments the hands using the skin color information using the bare hands and presents valuable results to compare the work done in this thesis.

In [3], there are experiments for both signer dependent and signer independent results. The hand gestures are chosen from both one handed and two handed gestures, also non-manual information is included, colored gloves are used in the recording process. The feature vector set used in the referenced study forms the base of the feature vector set used in this thesis study with some differences. Therefore

also the study in [3] composes a very good comparison base. There are total 19 ASL gestures used to train and 5 videos from 8 signers are used for training and classification phases. 8 fold cross validation tests are applied. The achieved recognition rate in this system is 75.53%. In this study left-to-right continuous HMM is used in the training stage (in 2009).

The last example of vision based SLR system which also includes the signer independent recognition results is in [33]. In this system 10 different static hand postures are used for evaluation and SVM is used as the classification algorithm in (2010). This system in the referred paper uses only the static images gathered from benchmark database called the Jochen Triesch static hand posture database. No video is processed in this system, only the images are used to recognize the ten gestures which are shown in Figure 12.



Figure 12: The Ten Postures of the Triesch Data Set

Two different experiments are performed by using this mentioned database. In the first one 3 different signers are used for training and the remaining 21 signers are used for testing. The overall success rate of this system is 85.1%, 418 trial is performed and 356 of them are correct. In the second one 8 different signers are used for training stage and the data from 16 signer is used for testing stage. For this case the success rate is 91.8. 320 trials are performed and 294 of them are classified correctly.

Another dataset which is used in signer independent experiments is a GSL dataset consisting of 40 gestures. This dataset is recorded by using Kinect[®] device, developed by Microsoft[™]. In total there are 13 signers and dataset includes joint paths of the hands and the elbows. This dataset does not include the hand shape information, which could completely change the meaning of the sign. Ong et al. [46] used this dataset with Sequential Pattern Trees and achieved 55.4% success rate.

These four systems are the very best examples of the vision based SLR systems; three of them also give the information about the signer independent success rate of the overall system. According to the signer dependent results our system's success rate is very high and outperforms the other results in the literature. Although the number of gestures in the dataset is limited to ten, they are chosen as to include all kind of motions such as global, local and also different hand shapes. Feature vectors are chosen to be able to deal with all of these mentioned variations. Also the signer independent success rate of the system is very high by looking the other successful system in the literature. Although in [3] the dataset differs with including also non-manual components and in [33] the dataset excluding the dynamic motion only considering only static hand images, they establish a comparison base for our system. The success rate of the system in this thesis work stands in a very good place among them.

CHAPTER 7

CONCLUSIONS

In this study an SLR system which recognizes the hand based signs are proposed. The system mentioned here does not need colored or sensor based gloves or specialized camera system. It is aimed to work with the videos recorded by the user's phone, because of the mobility and availability of the phone cameras. Since the system requires that the camera angle should focus on the upper body of the performer, using phone cameras would be easier for the end users.

Although the proposed method does not establish a complete system which can alternate a sign language interpreter, it can be used as the base of such a system with the explained future work in this chapter.

Although the researches in the literature are mainly concerned on the colored gloves or markers, the future of the SLR systems are predicted to be user friendly systems which do not use such restriction. Although there are some restrictions on the video recording environment in this thesis study, only skin color information is used in the segmentation step and this makes the system more convenient. Since the first step of the segmentation algorithm based on clustering, complex backgrounds should be avoided in order to prevent the segmentation errors. Also extreme lighting conditions should also be avoided in order to gather maximum performance from the segmentation phase. Since this thesis main focus is on the recognition part, segmentation algorithm is chosen to be easily implementable yet effective. By looking at the segmentation results with nearly zero segmentation error, one can say that this purpose is achieved. Designing a hand segmentation algorithm which is compatible with the extreme lighting conditions and complex backgrounds are not in

the scope of this thesis work. Yet, in a preset recording environment and equally distributed lighting conditions, the hand segmentation algorithm did hardly fail to segment the hands from the sampled image data.

Fuzzy C-Means algorithm is used to cluster the sampled image from the video, and by looking the mean values of the clusters and thresholding them, the one which includes the skin color is selected.

Another focus of this thesis study is extracting the feature vector set, which best model the hand properties. Modeling the hand shape accurately is very crucial in this study, since some of the chosen signs differ mainly in shape of the hand. In order to achieve this purpose, a bounding box is chosen outside of the segmented hand and divided into eight region. Area filters are used to determine percentage of the hand area inside every one of the divided eight regions. Best fitting ellipse is also calculated and its parameters are used to determine the shape and the orientation information of the hand. In addition to the hand shape and orientation location of the hands also carries valuable data to classify the hand gesture. Center of the mass information of the hands and the head are used for this purpose.

The classification method is the main concern of this thesis study. Discrete left-to-right HMMs are used for classification stage. In addition to HMM the Baum-Welch algorithm is used in accordance with the HMM to model the unknown parameters, and reduces the computation load significantly. A different HMM is created for each one of the feature vectors and for every gesture. The test gestures are classified by looking at the highest probability values for every feature vector and after that a voting system determines the winner gesture.

The most important problem encountered in this stage is the absence of the training data set in the literature. There are example datasets which are used in researches but all of them require the gloves or specialized cameras. In order to gather the train and test videos, 3 subjects are trained for the chosen 10 signs. The training and recording process takes a good amount of effort since none of the subjects are familiar with the

sign language and the signs are not too different from each other. User dependent and independent test cases are conducted. Although the user dependent tests resulted with a very high success rate 100%, the main concern of this thesis is user independent cases since it is aimed to be used in the real world. 85.8% success is achieved in user independent case. If we look at the individual success rates of the performers, we can see that results are differ significantly in user base. The most successful performer achieve 98.7% success rate whereas the least successful subject achieves 70.7%. This shows us the user dependence of the overall success rate. In real world application since the users will be the well trained native signers, who perform the given gestures for years the success rate of the algorithm would be much higher. For all that 85.8% recognition rate is a very high rate according to the literature by considering the fact that, it is achieved with a very limited training data set.

As a future work the proposed system could be transferred to a mobile platform since the system is intended to work with phone cameras. By doing this the system could reach much more people since nearly everyone has a phone with a decent camera lately. Also the gesture set could be expanded as to include more gestures.

REFERENCES

- [1] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 9, pp. 1685-1699, 2009.
- [2] O. Aran, "Vision based sign language recognition: modeling and recognizing isolated signs with manual and non-manual components", Ph.D. dissertation, Dept. Comput. Eng., Bogazici University, 2008.
- [3] O. Aran, I. Ari, L. Akarun, B. Sankur, A. Benoit, A. Caplier, P. Campr, A. H. Carrillo, F. Fanard, "SignTutor: An Interactive System for Sign Language Tutoring" *MultiMedia, IEEE* , Vol. 16, No. 1, pp. 81-93, January-March 2009.
- [4] O. Aran, I. Ari, P. Campr, E. Dikici, M. Hruz, D. Kahramaner, S. Parlak, L. Akarun and M. Saraclar, "Speech and Sliding Text Aided Sign Retrieval from Hearing Impaired Sign News Videos", *eNTERFACE'07 The Summer Workshop on Multimodal Interfaces*, Istanbul, Turkey, 2007.
- [5] O. Aran, I. Ari, P. Campr, E. Dikici, M. Hruz, S. Parlak, L. Akarun and M. Saraclar, "Speech and sliding text aided sign retrieval from hearing impaired sign news videos", *Journal on Multimodal User Interfaces*, Vol. 2, No. 2, pp. 117-131, 2008.
- [6] I. Ari, "Facial Feature Tracking and Expression Recognition for Sign Language", M.S. thesis, Dept. Comput. Eng., Bogazici University, 2008.
- [7] G. Awad, J. Han, and A. Sutherland, "A Unified System for Segmentation and Tracking of Face and Hands in Sign Language Recognition", *18th International Conference on Pattern Recognition*, 2006 (ICPR), Vol. 1, pp. 239-242, IEEE, 2006.
- [8] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York, Pergamon Press, 1984

- [9] N. D. Binh, E. Shuichi and T. Ejima, "Real-Time Hand Tracking and Gesture Recognition System", *Proceedings of International Conference on Graphics, Vision and Image Processing (GVIP-05)*, Vol., No., pp. 362-368, 2005.
- [10] M. Brand, "Coupled Hidden Markov Models for Modeling Interacting Processes", *Technical report*, MIT Media Lab Perceptual Computing, June 1997.
- [11] H. Brashear, T. Starner, P. Lukowicz and H. Junker, "Using multiple sensors for mobile sign language recognition", *Proceedings Of IEEE International Symposium On Wearable Computing*, pp. 45-52, 2003.
- [12] Y. Chahir and A. Elmoataz, "Skin-color detection using fuzzy clustering", *IEEE-EURASIP International Symposium on Control, Communications, and Signal Processing*, March 2006.
- [13] H. Cooper and R. Bowden, "Large lexicon detection of sign language", in *Human-Computer Interaction*, London, Springer, 2007, pp. 88-97.
- [14] H. Cooper and R. Bowden, "Sign language recognition using linguistically derived sub-units", *Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Languages Technologies*, Valetta, Malta, May 2010.
- [15] H. Cooper, B. Holt and R. Bowden, "Sign language recognition", in *Visual Analysis of Humans*, London, Springer, 2011, pp. 539-562.
- [16] A. Corradini, "Dynamic Time Warping for Off-line Recognition of a Small Gesture Vocabulary", *Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS'01)*, pp. 82-89, 2001.
- [17] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard and V. Athitsos, "Comparing gesture recognition accuracy using color and depth information", *Proceedings of the 4th International Conference on Pervasive Technologies Related to Assistive Environments*, pp. 20:1-20:7, New York, ACM, 2011.

- [18] H. K. Ekenel, M. Fischer, E. Tekeli, R. Stiefelbogen and A. Ercil, "Local binary pattern domain local appearance face recognition", *IEEE 16th Signal Processing, Communication and Applications Conference (SIU08)*, pp. 1-4, 2008.
- [19] G. Fang and W. Gao, "A SRN/HMM System for Signer-independent Continuous Sign Language Recognition", *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.
- [20] G. Fang, W. Geo and D. Zhao, "Large-Vocabulary Continuous Sign Language Recognition Based in Transition-Movement Models", *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans*, Vol. 37, No. 1, pp. 1-9, January 2007.
- [21] A. Farhadi and D. Forsyth, "Aligning ASL for Statistical Translation Using a Discriminative Word Model", *Computer Vision and Pattern Recognition*, Vol. 2, pp. 1471-1476, 2006.
- [22] A. Fitzgibbon, M. Pilu, R. B. Fisher, "Direct least square fitting of ellipses", *Pattern Analysis and Machine Intelligence, IEEE Transactions*, Vol. 21, No. 5, pp. 476-480, May 1999.
- [23] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition", *In International Workshop on Automatic Face and Gesture Recognition*, pp. 296-301, 1995.
- [24] C. Güler and G. D. Thyne, "Delineation of hydrochemical facies distribution in a regional groundwater system by means of fuzzy *c*-means clustering", *Water Resour. Res.*, Vol. 40, No., pp., 2004.
- [25] S. Hadfield and R. Bowden, "Generalized pose estimation using depth", in *Trends and Topics in Computer Vision*, London, Springer, 2012, pp. 312-325.
- [26] E.-J. Holden, G. Lee, R. Owens, "Automatic Recognition of Colloquial Australian Sign Language," *Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops*, Vol. 2, No. , pp. 183-188, January 2005.

- [27] X. D. Huang, Y. Ariki and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh Univ. Press, 1990.
- [28] C. -L. Huang and W.-Y. Huang, "Sign language Recognition Using Model-based Tracking and a 3D Hopfield Neural Network" *Machine Vision and Applications*, Vol. 10, No. 5-6, pp. 292-307, April 1998.
- [29] K. Imagawa, S. Lu, S. Igi, "Color-based hands tracking system for sign language recognition," *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference*, Vol., No., pp. 462-467, April 1998.
- [30] E. Isikligil, "A method for isolated sign recognition with Kinect", M.S. thesis, Dept. Comput. Eng., Middle East Technical University, Ankara, Turkey, 2014.
- [31] M. Jebali, P. Dalle and M. Jemni, "Hmm-based method to overcome spatiotemporal sign language recognition issues" *Electrical Engineering and Software Applications (ICEESA), 2013 International Conference*, Vol., No., pp. 1-6, March 2013.
- [32] M. W. Kadous, "Machine recognition of Auslan Signs using powergloves: Towards large-lexicon recognition of sign language", *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, pp. 165-174, 1996.
- [33] D. Kelly, J. McDonald and C. Markham, "A person independent system for Recognition of hand postures used in sign language", *Pattern Recognition Letters*, Vol. 31, No. 11, pp. 1359-1368, August 2010.
- [34] A. Kholmatov, "Biometric Identity Verification Using On-Line & Off-Line Signature Verification", M.S. thesis, Dept. Comput. Eng., Sabanci University, 2003.
- [35] J.-S. Kim, W. Jang and Z. Bien, "A dynamic gesture recognition system for the Korean sign language (KSL)" *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions*, Vol. 26, No. 2, pp. 354-359, April 1996.
- [36] W. Kong and S. Ranganath, "Signing Exact English (SEE): Modeling and Recognition", *Pattern Recognition*, Vol. 41, No. 5, pp. 1655-1669, May 2008.

- [37] H. K. Lee and J. H. Kim, "An HMM-Based Threshold Model Approach for Gesture Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 10, pp. 961-973, October 1999.
- [38] J. F. Lichtenauer, E. A. Hendriks and M. J. Reinders, "Sign Language Recognition by Combining Statistical DTW and Independent Classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 11, pp. 2040-2046, November 2008.
- [39] X. Liu and K. Fujimura, "Hand and gesture recognition using depth data" *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference*, Vol., No., pp. 529-534, May 2004.
- [40] N. Liu, B. C. Lovell, P. J. Kootsookos and R. I. Davis, "Model structure selection and training algorithms for an HMM gesture recognition system", *IWFHR'04: Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition*, pp. 100-105, 2004.
- [41] S. Marcel and A. Just, "IDIAP Two Handed Gesture Dataset", IDIAP Research Institute, Switzerland, [Online], Available: <http://www.idiap.ch/resource/twohanded>, [Accessed: Feb. 25, 2015].
- [42] S. A. Mehdi, Y. N. Khan, "Sign language recognition using sensor gloves" *Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference*, Vol. 5, No., pp. 2204-2206, November 2002.
- [43] S. Nayak, S. Sarkar, and B. Loeding, "Distribution-based dimensionality reduction applied to articulated motion recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 5, pp. 795-810, 2009.
- [44] S. Nayak, S. Sarkar and B. Loeding, "Unsupervised Modeling of Signs Embedded in Continuous Sentences", *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pp. 81-88, 2005.

- [45] M. Ouhyoung and R. Liang, "A sign language recognition system using hidden markov model and context sensitive search" *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, Vol., No., pp. 59-66, 1996.
- [46] E. -J. Ong, H. Cooper, N. Pugeault and R. Bowden, "Sign Language Recognition using Sequential Pattern Trees," *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference*, Vol., No., pp. 2200-2207, June 2012.
- [47] M. Parizeau and R. Plamondon, "A Comparative Analysis of Regional Correlation, Dynamic Time Warping, and Skeletal Tree Matching for Signature Verification", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 12, No. 7, pp. 710-717, 1990.
- [48] A. S. Park and J. R. Glass, "Unsupervised Pattern Discovery in Speech", *IEEE Transactions on Audio, Speech, And Language Processing*, Vol. 16, No. 1, pp. 186-197, January 2008.
- [49] V. I. Pavlovic, "Dynamic Bayesian Networks for Information Fusion with Applications to Human-Computer Interfaces", Ph.D. dissertation, Dept. Elect. Eng., University of Illinois, Urbana-Champaign, 1999.
- [50] L. R. Rabiner and B. Juang, "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine*, Vol. 3, No. 1, pp. 4-16, January 1986.
- [51] C. Shan, S. Gong and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study", *Image and Vision Computing*, Vol. 27, No. 6, pp. 803-816, 2009.
- [52] T. Starner, A. Pentland and J. Weaver, "Real-time American Sign Language recognition using desk and wearable computer based video", *IEEE Transactions on Pattern Analysis And Machine Intelligence*, Vol. 20, No. 12, pp. 1371-1375, December 1998.
- [53] D. Uebachs, J. Gall, M. Van den Bergh and L. Van Gool, "Real-time sign language letter and word recognition from depth data" *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference*, Vol., No., pp. 383-390, November 2011.

- [54] V. Vezhnevets, V. Sazonov and A. Andreeva, "A Survey on Pixel-Based Skin Color Detection Techniques" *International Conference Graphicon*, 2003.
- [55] C. Vogler and D. Metaxas, "Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods," *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation, 1997 IEEE International Conference*, Vol. 1, No., pp. 156-161, October 1997.
- [56] C. Vogler and D. Metaxas, "Parallel hidden Markov models for American sign language recognition", *International Conference on Computer Vision*, Vol. 1, pp. 116-122, 1999.
- [57] U. von Agris, M. Knorr and K. -F. Kraiss, "The significance of facial features for automatic sign language recognition", *8th IEEE International Conference on Automatic Face and Gesture Recognition*, September 2008.
- [58] U. von Agris, J. Zieren, U. Canzler, B. Bauer and K.-F. Kraiss, "Recent developments in visual sign language recognition", *Universal Access in the Information Society*, Vol. 6, No. 4, pp. 323-362, 2008.
- [59] M. B. Waldron and S. Kim. "Isolated ASL sign recognition system for deaf persons", *Rehabilitation Engineering, IEEE Transactions*, Vol. 3, No. 3, pp. 261-271, September 1995.
- [60] Y. Wu and T. S. Huang, "Hand modeling analysis and recognition for vision-based human computer interaction", *IEEE Signal Processing Mag. – Special issue on Immersive Interactive Technology*, Vol.18, No.3, pp. 51-60, May 2001
- [61] J. Yamato, J. Ohya and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference*, June 1992.
- [62] M. –H. Yang, N. Ahuja, M. Tabb, "Extraction of 2D motion trajectories and its application to hand gesture recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, Vol. 24, No. 8, pp. 1061-1074, August 2002.

[63] L.-G. Zhang, Y. Chen, G. Fang, X. Chen and W. Gao, "A Vision-based Sign Language Recognition System Using Tied-mixture Density HMM", *Proceedings of the 6th International Conference on Multimodal Interfaces*, pp. 198-204, New York, ACM, 2004.

[64] J. Zieren and K.-F. Kraiss, "Robust person-independent visual sign language recognition", in *Pattern recognition and image analysis*, London, Springer, 2005, pp. 520-528.

[65] Clustering - Introduction, A Tutorial on Clustering Algorithms, [Online], Available: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html, [Accessed: Feb. 25, 2015].

[66] eINTERFACE'07 The Summer Workshop on Multimodal Interfaces, [Online], Available: http://www.enterface.net/enterface06/docs/results/eINTERFACE06_proceedings.pdf, [Accessed: Feb. 25, 2015].

APPENDIX A

GESTURES

A.1 Arm



Figure 13: Arm Gesture

A.2 Box

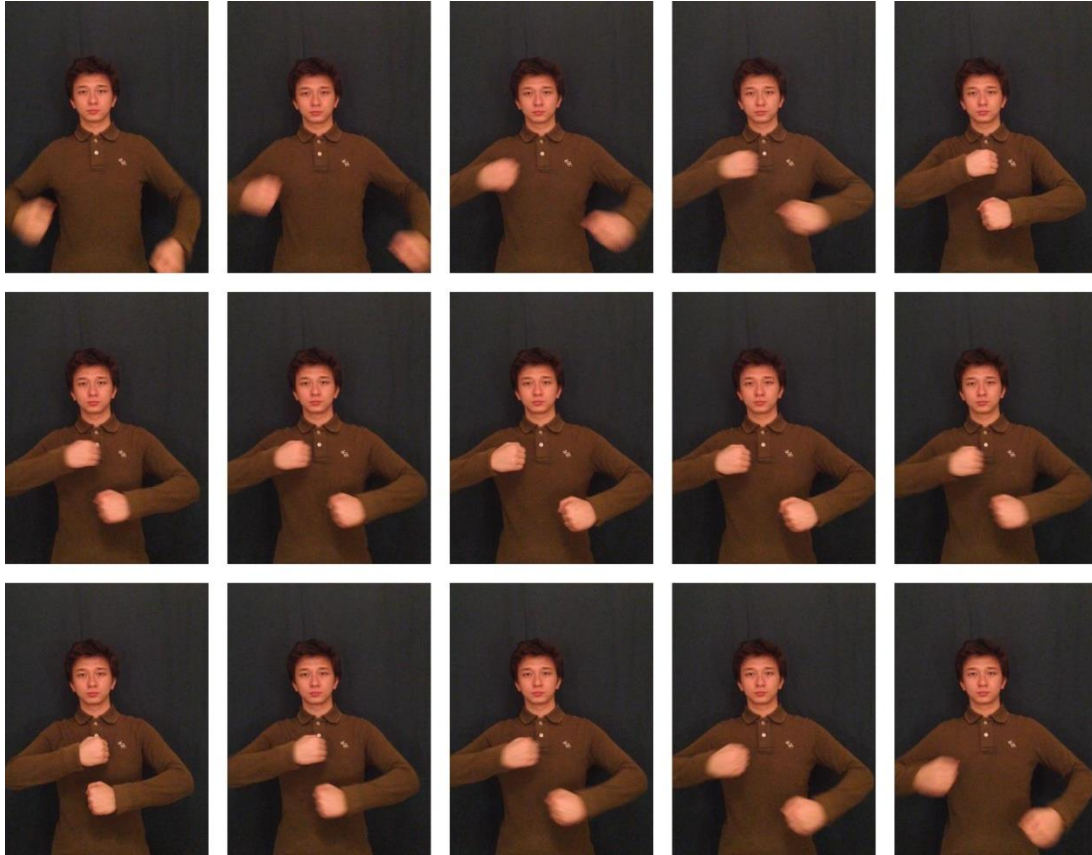


Figure 14: Box Gesture

A.3 Chess



Figure 15: Chess Gesture

A.4 Early



Figure 16: Early Gesture

A.5 Elbow



Figure 17: Elbow Gesture

A.6 Lung

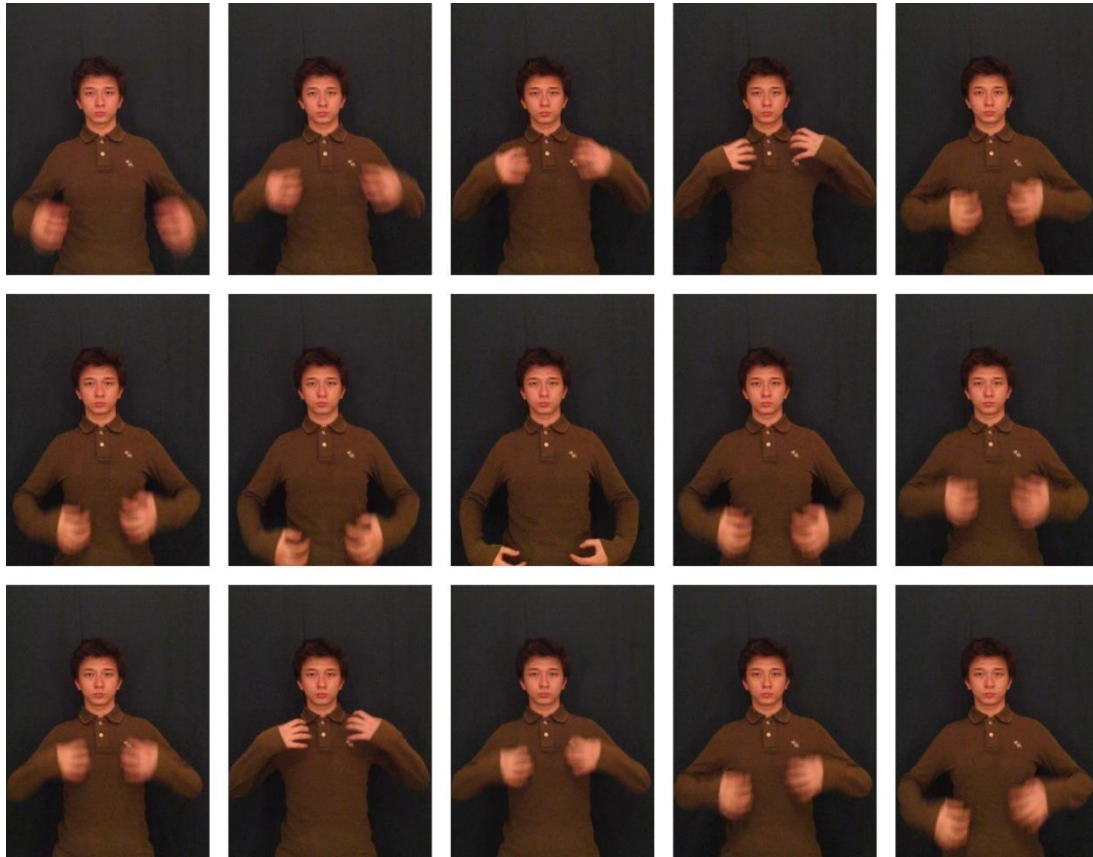


Figure 18: Lung Gesture

A.7 Sea



Figure 19: Sea Gesture

A.8 Swim

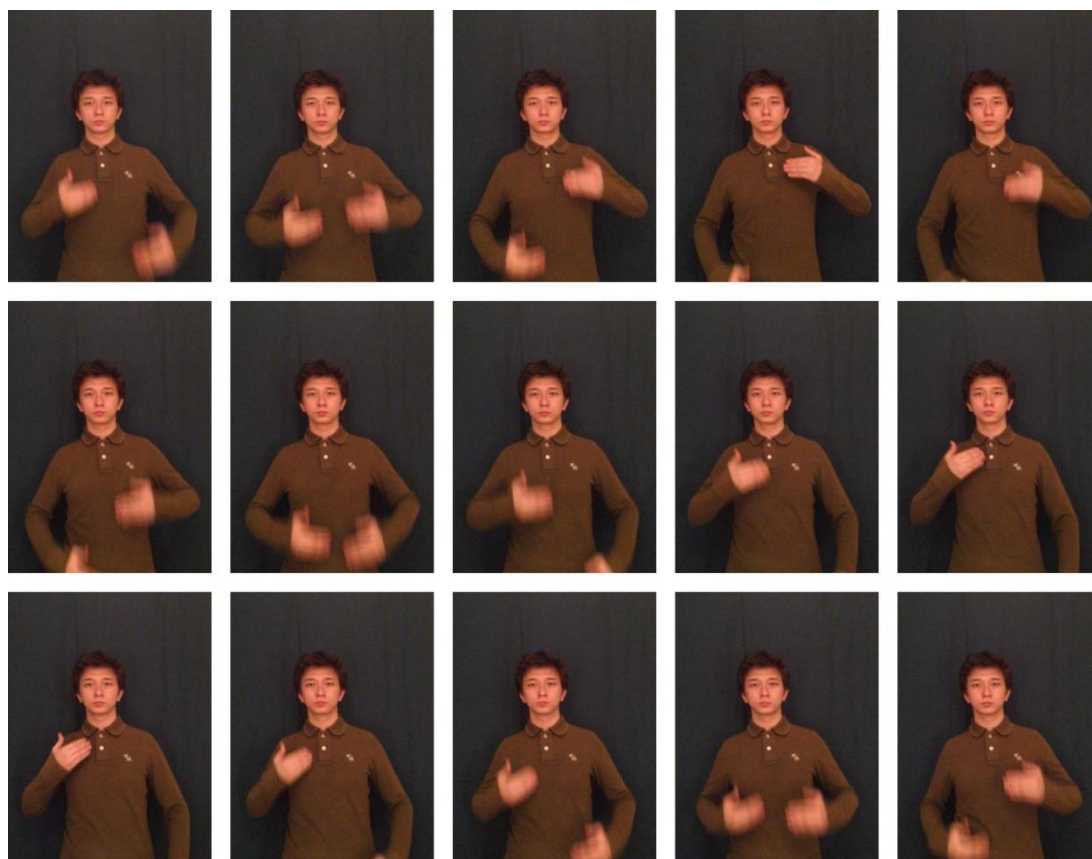


Figure 20: Swim Gesture

A.9 Wound

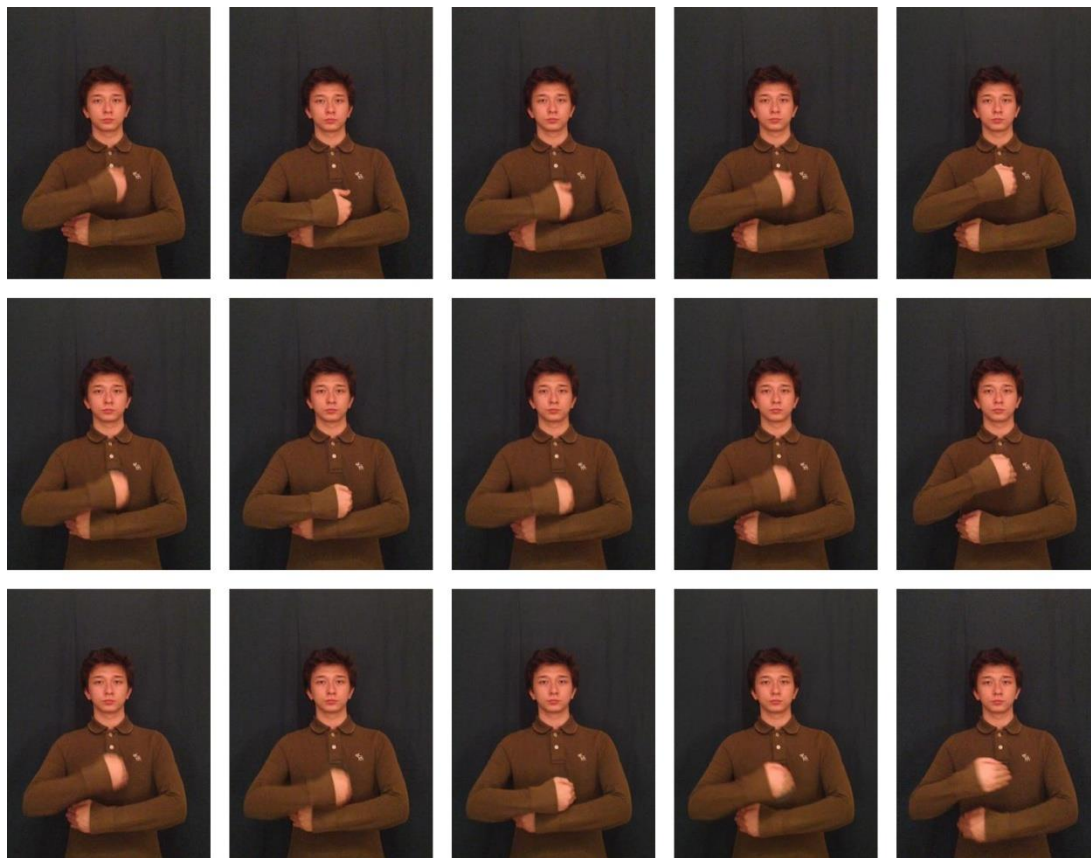


Figure 21: Wound Gesture

A.10 Wrist

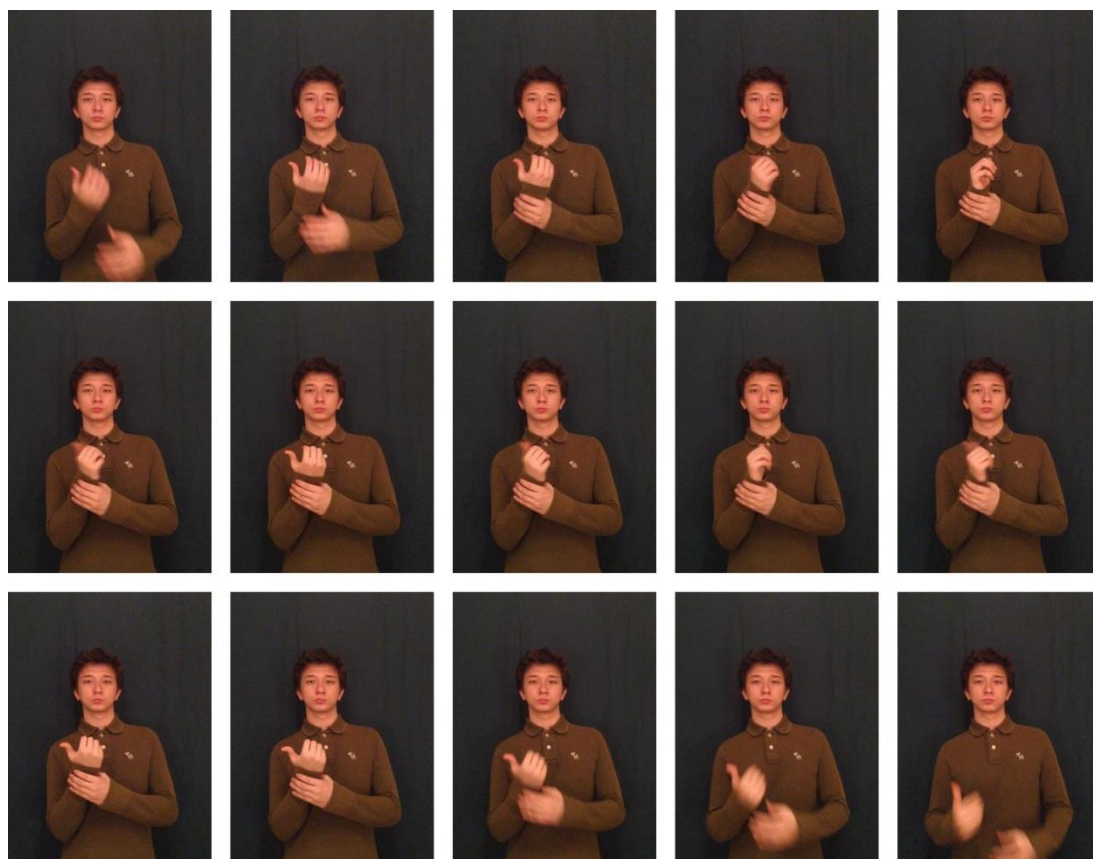


Figure 22: Wrist Gesture