American Sign Language translation using edge detection and cross co-relation

By

Anshal Joshi

A thesis submitted in partial fulfillment of the requirements for the degree

of

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

UNIVERSITY OF PUERTO RICO

MAYAGUEZ CAMPUS

2017

Approved by:

Emmanuel Arzuaga, Ph. D. President, Graduate Committee	Date
Pedro I. Rivera Vega , Ph. D. Member, Graduate Committee	Date
Heidy Sierra, Ph.D. Member, Graduate Committee	Date
José G. Colom Ustariz, Ph.D. Department Chairperson	Date
Dr. Hilton Alers Valentn , Ph.D. Graduate Studies Representative	Date

Acknowledgement

I express my sincere appreciation to my thesis supervisor Prof. Emmanuel Arzuaga, co-supervisors Prof. Heidy Sierra and Prof. Pedro I. Rivera Vega for their guidance, insight and elegant attitude throughout the research.

I wish to thank my parents Kailash Chandra Joshi and Meena Joshi my sisters Prachi Joshi and Neha Joshi and my brother Pranav Joshi for their support, encouragement and confidence throughout the years of my education.

Abstract

According to the World Health Organization (WHO), there are approximately 360 million people worldwide that have disabling hearing loss and 70 million that are mute. Developing communication advancements is very complex and its been a challenge for many years. Currently, American Sign Language, which is expressed through the hands and face and perceived through the eyes, is the standard language of communication for the Deaf community. However, the development of better communication mechanisms for the hearing impaired is still a big challenge. Our main objective is to implement an automated translation system which can translate the American Sign Language to English text using common computing environments such as a computer and a generic webcam.

In this investigation, a real-time approach for hand gesture recognition system is presented. Two different approaches are used to translate English letters and words. In the method to recognize letters, first, the hand gesture is extracted from the main image by the image segmentation, morphological operation and edge detection technique and then processed to feature extraction stage. And for the words, a video sequence is captured then divided into frames and process them for the frame selection stage. In frame selection stage, frames are sampled and selected for feature extraction and then the gesture is extracted from all of the frames by the same using the same technique as image segmentation, morphological operation, edge detection technique and combined by Montage. In feature extraction stage the Cross-correlation coefficient is applied on the gesture to recognize it. In the result part, the proposed approach is applied on American Sign Language (ASL) database and we are able to achieve 92 - 94% accuracy in translation.

Resumen

Según la Organización Mundial de la Salud (OMS), hay alrededor de 360 millones de personas en todo el mundo que tienen pérdida auditiva discapacitante y 70 millones que son mudos. Desarrollar los avances de la comunicación es muy complejo y ha sido un desafío por muchos años. En la actualidad, el lenguaje de señas americano, que se expresa a través del uso de las manos y la cara y se percibe a través de la vision, es el lenguaje estándar de comunicación para la comunidad sorda. Sin embargo, el desarrollo de mejores mecanismos de comunicación para las personas con discapacidad auditiva sigue siendo un gran desafío. Nuestro objetivo principal es implementar un sistema automatizado de traducción que sea capaz de traducir el lenguaje de señas americano a texto en inglés utilizando entornos informáticos comunes como una computadora y una cámara web genérica.

En esta investigación presentamos un software para el reconocimiento y traducción a texto de señas de ASL en tiempo real. Se utilizan dos enfoques diferentes para traducir letras y palabras al idioma inglés. En el método de reconocimiento de letras, la seña de la mano se extrae de la imagen principal mediante el uso de técnicas de segmentación de imágenes, la operaciones morfológicas y la técnica de detección de bordes (edge detection en inglés) y luego se procesa para la etapa de extracción de la letra. La técnica de reconocimiento de palabras o frases, se utiliza una secuencia de vídeo que luego se divide en marcos (frames en inglés) y los procesa para la etapa de selección del marco. En la etapa de selección del marco, muestreamos y seleccionamos aquel marco que mejor nos permita extraer la seña y luego aplicamos la misma técnica de segmentación de imagen, operaciones morfológicas, técnica de detección de borde y combinado por Montage. En la etapa de extracción de señas, se aplica el coeficiente de correlación cruzada en la seña para reconocimiento. En la parte de resultados, el enfoque propuesto se aplica a la base de datos de lenguaje de signos americano (ASL) Y somos capaces de lograr 92 - 94 % de precisin en la traduccin.

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Motivation

Communication and community are significant parts of human life. Deaf people are isolated from the most common forms of communication in today's society such as warnings and sound alerts, or any other form of oral communication between people in regular daily activities like visiting the doctor or communicating in the street. In other words, deaf people can often feel disassociated and thus find it hard to get information or help in daily activities or even when encountering emergency situations. As a consequence, deaf people are twice as likely as hearing people to be affected by depression, anxiety and similar problems[1, 2].

A deaf person mostly relies on vision for clues to what people are communicating as well as other clues like vibrations, sense of touch in floors or around them. Often other people will change the way they act towards deaf people and can even become irritated with having to repeat statements, or feel frustrated on the lack of a mutually intelligible language.

In deaf communication, hands play the same role that the tongue plays in the hearing community. Context of a non-verbal discourse is conveyed through a series of distinct

kinematic configurations of hands within the linguistic extent of its language.

Besides the use of text messaging through tablets and smartphones or online chating using computer equipment, the state of the art in person to person communication for the deaf community are: 1) lip reading and 2) interaction with a sign language interpreter. Although some people mumble or have difficulties speaking to the point where they are practically impossible to read their lips, most people are easy to read. Additionally because lip reading depends on visual cues, lip readers must have good eyesight. Lip readers also need to have a clear light since it is almost impossible to lip read in the dark. Lip reading is much easier when it involves the lip reader's first language. For example, an English speaker will find it much easier to lip read English than to lip read a second language.

The most popular communication medium for a deaf person is to have an interpreter who converts the verbal communication into sign form.

Since sign language is used for translating the communication of a deaf person to a hearing person and vice versa, it has received special attention [3]. A significant amount of software based systems have been proposed to recognize gestures using different types of sign languages [4]. For example, the use of histogram boundary algorithms for American Sign Language (ASL) recognition, MLP neural network and dynamic programming matching [4]. Japanese sign language (JSL) has been recognized using Recurrent Neural Network,(42 alphabet and 10 words)[5]. Arabic Sign language (ArSL) has been recognized using two different types of Neural Network, Partially and Fully Recurrent neural Network [6].

Different tracking methods have been developed, presented and discussed by researchers, but they generally require specialized hardware or equipment and usually perform a very specific task. For example, research has shown that it is viable to use the Microsoft kinect [7] along a special glove [8] to track hand gestures for sign language translation. But these

methods require this hardware setup tying the user with a special environment to perform the translation. Other existing methods are also person dependent and can only recognize the sign for the person to whom it is designed. In the process of extracting signs we are searching for an unknown word sequence for which the sequence of features of interest best fits to trained data images. In this work, our system will take the video as the input from a regular webcam, such as the ones available in portable computers, smartphones and tablets and process it to extract the features and then translate it into English text. For example, a person will produce the sign and the written text will be produced by the system in real time so that a deaf person can read it from the device monitor.

This thesis is focused on the implementation of an efficient translation system which can identify all the American Sign Language (ASL) alphabets and convert them into English alphabets. We are creating an ASL database of signs with different hand samples to use as training data for our algorithms. In order to create the sign database we have implemented Graphical User Interface (GUI) Application to capture signs. The GUI will capture the image within certain boundaries, and then image segmentation and morphological filtering algorithms are applied to reduce noise and convert the captured image into binary form. As few signs are very much similar to each other and very hard to differentiate in a black and white image, we use an edge detection technique to include the edges of the fingers as new features that can help us better to discriminate patterns in the gesture. After adding these edges, the training image is saved in sign database that will be later used for the classification of real time gestures.

For the translation, we have implemented a second GUI in which we have a video frame where the camera output can be observed. This application is designed to collect video frames from the camera and extract the sign from an image frame to compare it to the image database and check which of the signs best matches the extracted sign. Finally, the application generates a display of the classified frame, which is an English

translation of the ASL sign. In our initial attempt we have implemented a generic sign translator which anyone can easily use at their home using their everyday devices like a laptop computer. The clear advantage of such approach is that the users do not have to spend extra money to buy additional or expensive hardware.

1.2 Applications

Computer recognition of hand gestures may provide a more natural-computer interface, allowing people to point, or rotate a CAD model by rotating their hands. Hand gestures can be classified in two categories: static and dynamic. A static gesture is a particular hand configuration and pose, represented by a single image. A dynamic gesture is a moving gesture, represented by a sequence of images. We will focus on the recognition of static images.

Interactive applications pose particular challenges. The response time should be very fast. The user should sense no appreciable delay between when he or she makes a gesture or motion and when the computer responds. The computer vision algorithms should be reliable and work for different people. There are also economic constraints: the vision-based interfaces will be replacing existing ones, which are often very low cost. A hand-held video game controller and a television remote control each cost about $40. Even for added functionality, consumers may not want to spend more. When additional hardware is needed the cost is considerable higher. Academic and industrial researchers have recently been focusing on analyzing images of people. While researchers are making progress, the problem is hard to solve till present day.

Creating a proper sign language (ASL at this case) dictionary is the desired result at this point. This would combine structure understanding of the system. The ASL will be used as the database since it is a tightly structured set. From that point further applications can be suited. Computer interfaces using typical input devices in conjunction with

gesture recognition approaches can be used to perceive some of the user feelings as well. Taking ASL recognition further, a full real-time dictionary could be created with the use of video. Another application is huge database annotation. It is far more efficient when properly executed by a computer, than by a human.

1.2.1 American Sign Language

American Sign Language is the language acquired by most deaf people in the United States. It is part of the deaf culture and includes its own system of puns, inside jokes, etc. However, ASL is one of the many sign languages of the world. As an English speaker would have trouble understanding someone speaking Japanese, a speaker of ASL would have trouble understanding the Sign Language of Sweden. ASL also has its own grammar that is different from English. ASL contains approximately 6000 gestures of common words with finger spelling used to communicate obscure words or proper nouns. Finger spelling uses one hand and 26 gestures to communicate the 26 letters of the alphabet. Some of the signs can be seen in See Figure 1.1.

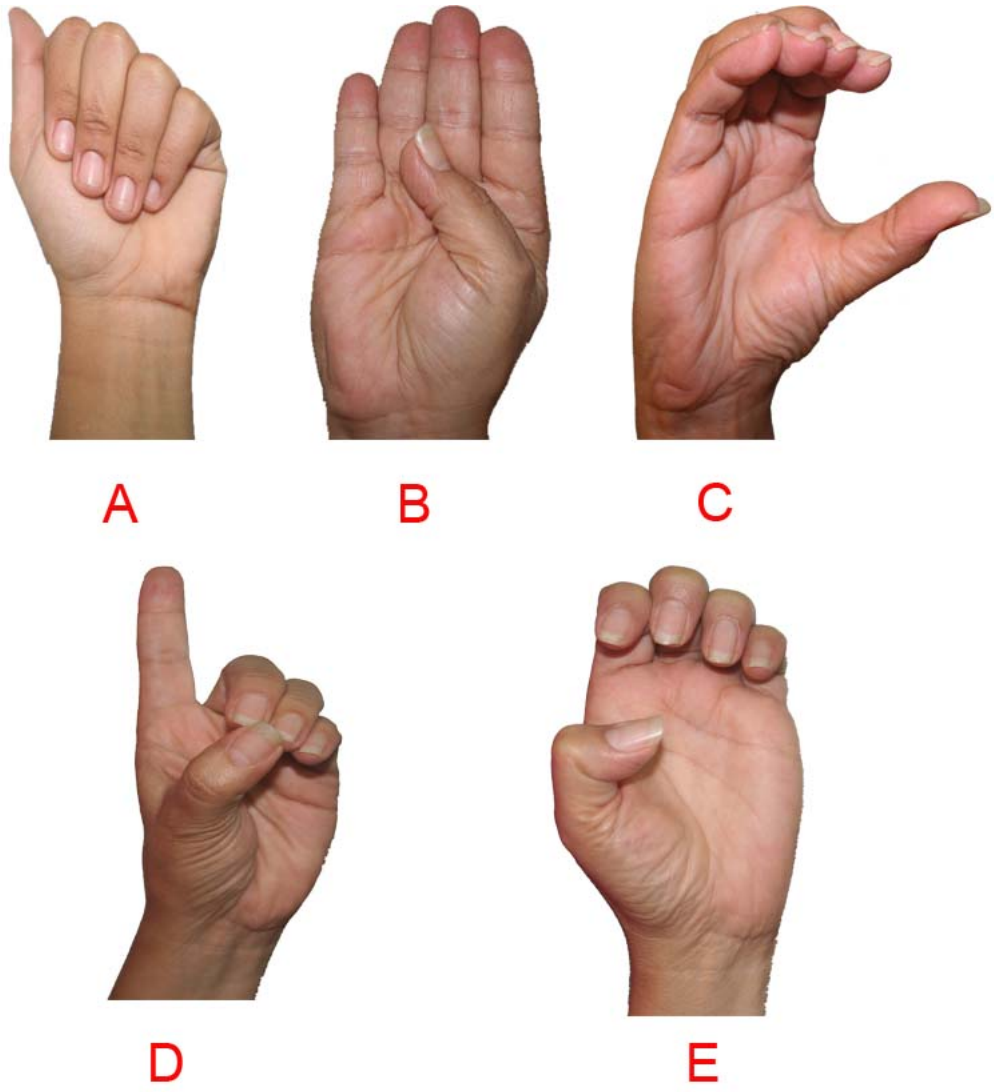


Fig 1.1: ASL examples.

ASL uses facial expressions to distinguish between statements, questions and directives. The eyebrows are raised for a question, held normal for a statement, and furrowed for a directive. There has been considerable work and research in facial feature recognition, they will not be used to aid recognition in the task addressed. This would be feasible in a full real-time ASL dictionary.

1.3 Objective

1.3.1 General Objective

- The general objective of this research work is to create an ASL sign translator to English text software that can later be easily integrated to a video chat environment or similar type of interaction.

1.3.2 Specific Contributions

The main contributions of this thesis work in particular are:

- Design of a computing application that is able to extract sign features and convert them into plain English text.

- Creation of a ASL sign database to be generally available for the general public.

Chapter 2

LITERATURE REVIEW

There is a considerable amount of work related to sign language translation and interpretation. Researches in china created a sign language translator using the Microsoft kinect motion sensing camera [7]. Kinect has been designed for Xbox gamming, in which it can read a specific movement of the human body and translate them into game control commands using its special sensors. In November 2010 after the release of Microsoft Kinect, a number of researchers almost immediately focus their interest in this area. In 2011, Microsoft released the SDK for kinect, which gave a boost to all the interested researchers. At Microsoft Research Asia, head researcher Ming Zhou proposed their work in sign language translation. They have been able to create a translation system that can capture sign convert them into written text and spoken translation in real time. The non signer is represented by an avatar which takes his spoken words and then accurately converts them into written text so that the deaf person can read it.

In April 2016, two researchers from University of Washington won the Lemelson-MIT student prize $10,000 for the development of gloves that can translate ASL signs into speech [8]. Their invention is called SignAloud, in which each glove contains sensors that captures the hand position and motion and send it to computer via Bluetooth, then the computer searches for the appropriate hand gesture through various sequential statistical

regressions, similar to a neural network. If the data match a gesture, then the associated word or phrase is spoken through a speaker.

Hasan [9] applied multivariate Gaussian distribution to recognize hand gestures using non-geometric features. The input hand image is segmented using two different methods [10]; skin color based segmentation by applying HSV color model and clustering based thresholding techniques [9]. Some operations are performed to capture the shape of the hand to extract hand feature; the modified Direction Analysis Algorithm are adopted to find a relationship between statistical parameters (variance and covariance) [9] from the data, and used to compute object (hand) slope and trend [9] by finding the direction of the hand gesture [9]. Then Gaussian distinction is applied on the segmented image, and it takes the direction of the hand.

Form the resultant Gaussian function the authors divide the image into circular regions, in other words, that regions are formed in a terrace shape to eliminate the rotation affect [9, 10]. The shape is divided into 11 terraces with a 0.1 width for each terrace [9, 10]. 9 terraces are resultant from the 0.1 width division which are; (1-0.9, 0.9-0.8, 0.8-0.7, 0.7-0.6, 0.6, 0.5, 0.5-0.4, 0.4-0.3, 0.3-0.2, 0.2-0.1), and one terrace for the terrace that has value smaller than 0.1 and the last one for the external area that extended out of the outer terrace [9, 10].

Each terrace is then divided into 8 sectors which named as the feature areas, empirically discovered that number 8 is suitable for features divisions [9], To attain best capturing of the Gaussian to fit the segmented hand, re-estimation are performed on the shape to fit capturing the hand object [9], then the Gaussian shape are matched on the segmented hand to prepare the final hand shape for extracting the features.

After capturing the hand shape, two types of features are extracted to form the feature vector [9, 10]; local feature, and global features. Local features using geometric central

moments which provide two different moments 00, 11 as shown by equation 2.1

$$\mu_{pp} = \sum_{x} \sum_{y} (x - \mu_x)^p (y - \mu_y)^P f(x, y) \tag{2.1}$$

$$\mu_{pp}^{(k)} = \sum_{y} \sum_{x} (x^{(k)} - \mu_x^{(k)})^p (y^{(k)} - \mu_y^{(k)})^p f(x^{(k)}, y^{(k)}) \tag{2.2}$$

$$\forall k \in \{1, 2, 3, ..., 88\} \& \forall p \in \{0, 1\}$$

Where μ_x and μ_y is the mean value for the input feature area [9], x and y are the coordinates, and for this, the input image is represented by 88*2 features, as explained in detail in equation 2.2. While the global features are two features the first and second moments [9, 10] that are the computed for the whole hand features area [9]. These feature areas are computed by multiplying feature area intensity plus feature areas map location [9]. In this case, any input image is represented with 178 features [9, 10]. The system carried out using 20 different gestures [10], 10 samples for each gesture, 5 samples for raining and 5 for testing, with 100% recognition percentage and it decreased when the number of gestures are more than 14 gestures [10]. In [9] 6 gestures are recognized with 10 samples for each gesture. Euclidian distance used for the classification of the feature [9, 10].

Kulkarni [11] recognize static posture of American Sign Language using neural networks algorithm. The input image are converted into HSV color model, resized into 80x64 and some image preprocessing operations are applied to segment the hand [11]from a uniform background [11], features are extracted using histogram technique and Hough algorithm. Feed forward Neural Networks with three layers are used for gesture classification. 8 samples are used for each 26 characters in sign language, for each gesture, 5 samples are used for training and 3samples for testing, the system achieved 92.78 % recognition rate using MATLAB language [11].

Hasan [12] applied scaled normalization for gesture recognition based on brightness factor matching. The input image with is segmented using thresholding technique where the background is black. Any segmented image is normalized (trimmed), and the center mass [12] of the image are determined, so that the coordinates are shifted to match the centroid of the hand object at the origin of the X and Y axis [12]. Since this method depends on the center mass of the object, the generated images have different sizes [12], for this reason a scaled normalization operation are applied to overcome this problem which maintain image dimensions and the time as well [12], where each block of the four blocks are scaling with a factor that is different from other blocks factors. Two methods are used for extraction the features; firstly by using the edge mages, and secondly by using normalized features where only the brightness values of pixels are calculated and other black pixels are neglected to reduce the length of the feature vector [12]. The database consists of 6 different gestures, 10 samples per gesture are used, 5 samples for training and 5 samples for testing. The recognition rate for the normalized feature problem achieved better performance than the normal feature method, 95% recognition rate for the former method and 84% for the latter one [12].

Wysoski et al. [2] presented rotation invariant postures using boundary histogram. Camera used for acquire the input image, filter for skin color detection has been used followed by clustering process to find the boundary for each group in the clustered image using ordinary contourtracking algorithm. The image was divided into grids and the boundaries have been normalized. The boundary was represented as chords size chain which has been used as histograms, by dividing the image into number of regions N in a radial form, according to specific angle. For classification process Neural Networks MLP and Dynamic Programming DP matching were used. Many experiments have implemented on different features format in addition to use different chords size histogram, chords size FFT. 26 static postures from American Sign Language used in the experiments.

Homogeneous background was applied in the work.

Stergiopoulou [3] suggested a new Self-Growing and Self-Organized Neural Gas (SGONG) network for hand gesture recognition. For hand region detection a color segmentation technique based on skin color filter in the YCbCr color space was used, an approximation of hand shape morphology has been detected using (SGONG) network; Three features were extracted using finger identification process which determines the number of the raised fingers and characteristics of hand shape, and Gaussian distribution model used for recognition.

In recent years, studies are mostly concentrated on Boosting and HMM [13] (which have shown a high detection rate for these cases). The most tempting side of using those methods is that they usually work with grayscale images instead of colored images and thus it eliminates the drawbacks of such color based noise issues. This innovative approach is using a well-known technique namely Adaboost classifier which was mentioned in [14]. Adaboost classifier is an effective tool to select appropriate features for face detection. This feature extraction technique does not need skin color information and have less computation time with the use of integral image concept. But the drawback of this method is that it requires a training process. This process often needs a large amount of sample images to have a high detection rate. Sample images would require thousands of positive images (include face) and thousands of negative sample images (does not include face). The training process would also have a high computation time and it might require several days to complete the training process.

Finding out hands directly would not be an effective way since hands do not have a strict shape. Once the face is detected, the other skin pixel blobs can be supposed as hands. Hand gesture process is started at this point. For hand gesture recognition part of this study, HMM or Adaboost classifier type training based methods have very limited usage because of the non-strict structure of the hand. To recognize a gesture

once need to train positive images (include the defined gesture) and negative images (do not include the defined gesture). But negative images have a serious role at this point. Since many hand poses might yield similar training data, reliance to the training data would be limited. So an adaptation of a well-known hand gesture recognition method was implemented in this study. According to the proposed methods in prior studies [15] and [16] centroidal profile extraction of the hand is extracted around the center of the palm and histogram clustering is applied to the resultant data to recognize the gesture. Such an algorithm would typically count the number of fingers being shown to the camera. this can capture 6 gestures for each hand namely 1, 2, 3, 4 or 5 fingers or no fingers(punch) conditions. If the algorithm is used for two hand gesture recognition 6x6 = 36 gestures can be recognized by using the mentioned method. In this thesis, an adaptation of that method is proposed and a higher correct detection rate is provided. There is a limited amount of studies in literature for the hand gesture recognition. Recognition methods, like in the detection procedure, mainly rely on algorithms which need training or different environmental constraints. A clear summary of such algorithms are shown in Table 2.1

Reference	Primary Method OF Recognition	Number of Gesture Recognized	Background to Gesture Images	Additional Markers Required	Number of Training Images
39	Hidden Markov Models	97	General	Multicolored gloves	400
5	Entropy Analysis	6	No	No	400
41	Linear Approximation to non-linear point distribution models	26	Blue Screen	No	7441
42	Finite State Machine modeling	7	Static	Marker on Glove	10 sequence of 200 frames each
43	Fast Template Matching	46	Static	Wrist Band	100 examples per gesture

Table 2.1: Gesture Recognition Methods.

Chapter 3

METHODOLOGY

In this chapter we are going to discuss how various methods are used in the process of translating ASL into English text. The methodology consists of gesture data acquisition from the web-cam for the extraction of features used in sign language in parts of the body such as left hand, right hand, face. Once the samples of each gesture are obtained from a camera, the translation program finds a the match on each of these samples. The feature extraction stage consists of computing the sign features on the live camera. The basic concept of the our system is illustrated in the Figure 3.1.

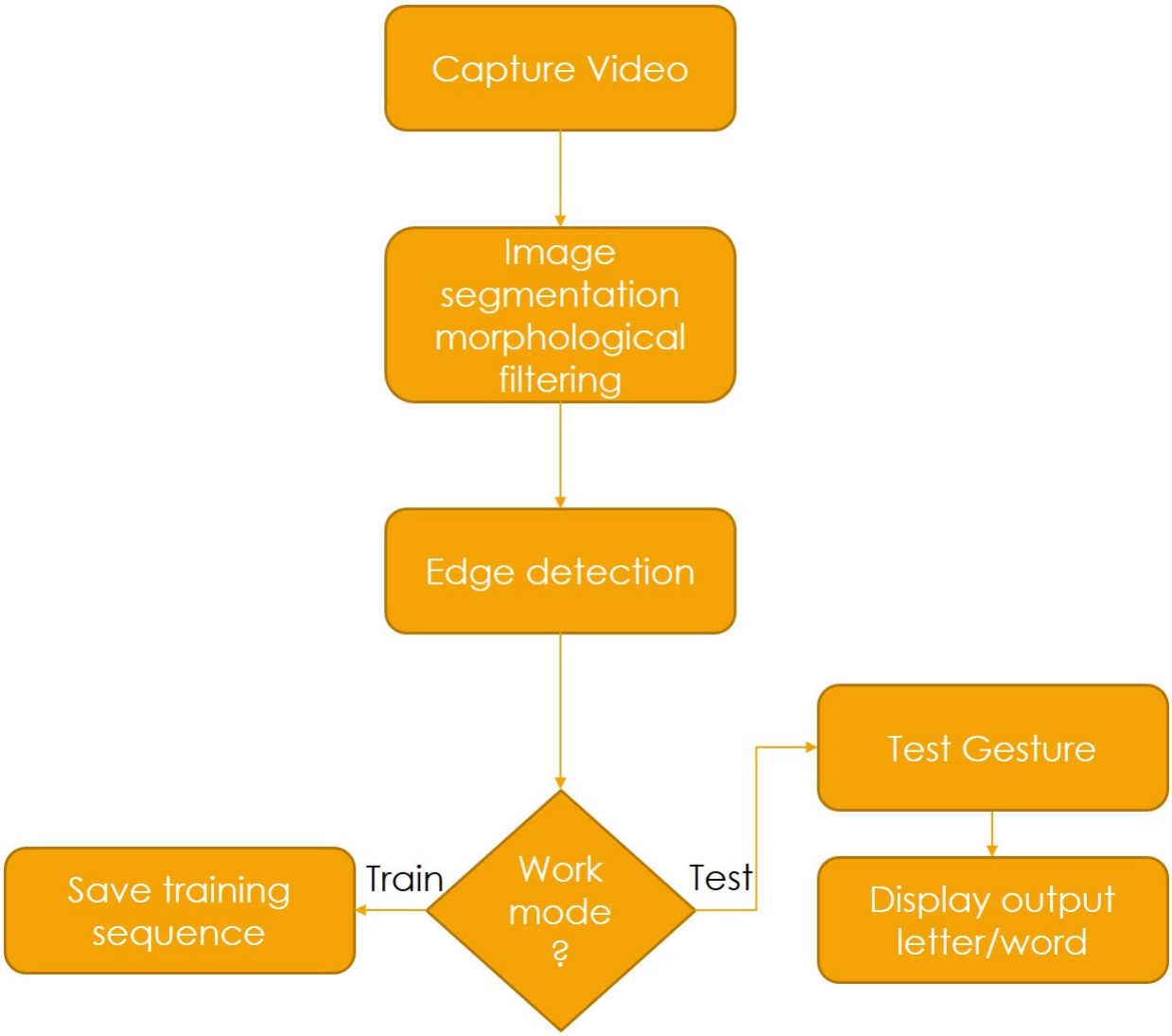


Fig 3.1: System Concept.

3.1 ASL Sign Database Creation

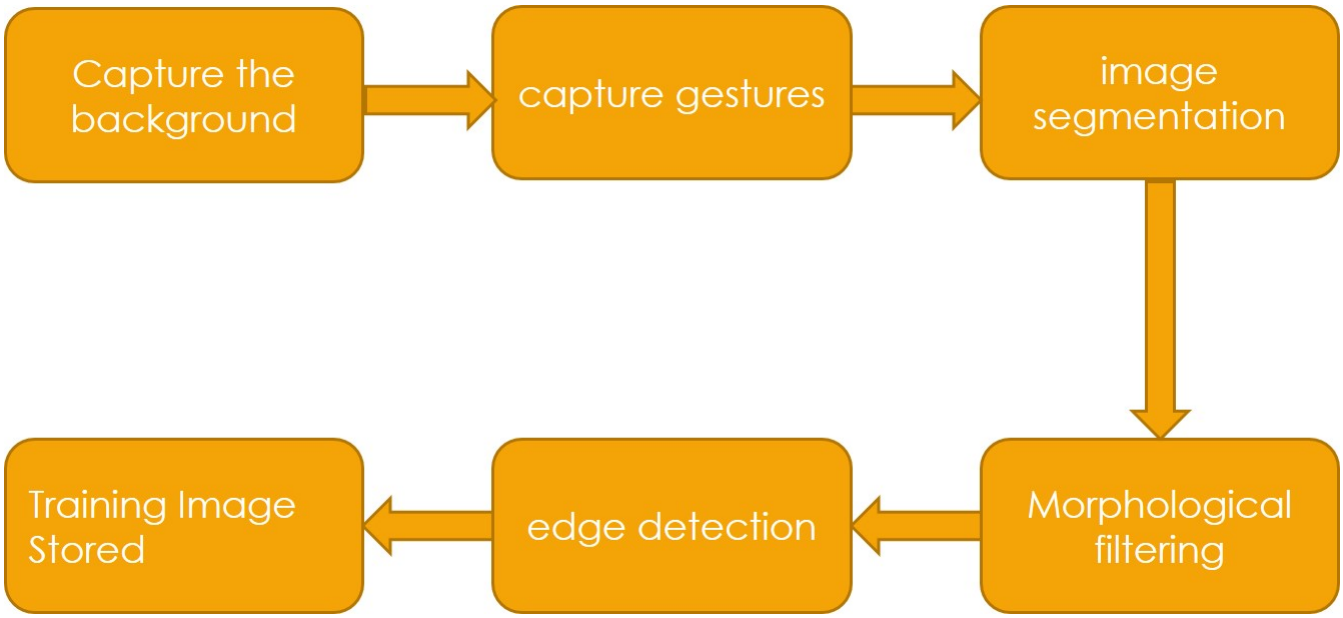


Fig 3.2: Database Training Process.

Figure 3.2 shows the ASL database training process. ASL Sign Database is a collection of different gestures used in American Sign Language. It consist of ASL letters (A to Z) and ASL words (like Good Morning, Hi, Bye). The purpose of ASL sign Database is to provide a data training set to our algorithm. We have created two systems to generate the database for English alphabets and words. Which are explained as follows:

3.1.1 Alphabet Database

In order to create the ASL sign database for alphabets, we have implemented a sign image acquisition software. This application provides a graphical user interface (GUI) environment that allows the user to collect the ASL image sings. These sign images will be used to train our system. See Figure 3.3.

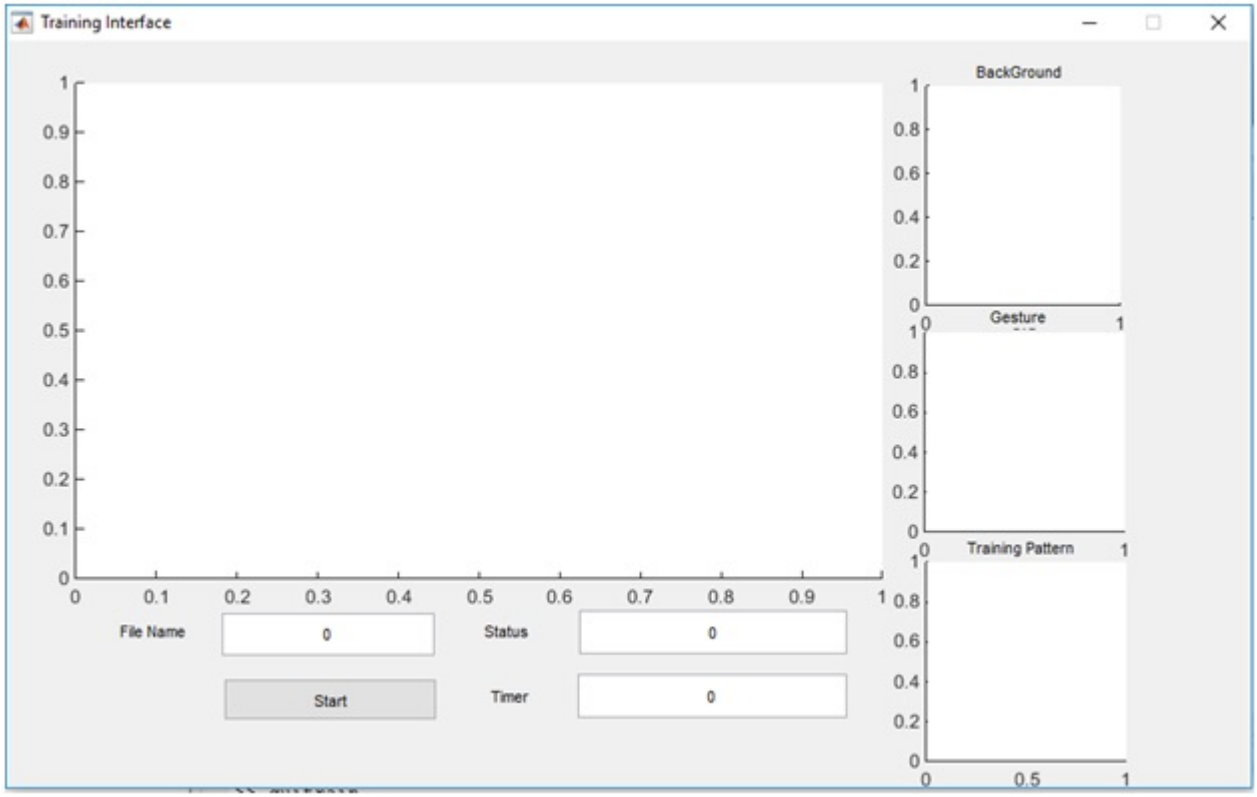


Fig 3.3: Training GUI.

To start collecting samples live, a user needs to write the file name in the File Name field and press start. When start button is clicked the training function is called and it initiates the camera view. During the training process the system captures an image of the background. Once the background is captured, system will move to capturing gestures. Once both background and gesture are captured, the camera view will stop. The system will start the image segmentation. It is a process in which we convert a RGB image or gray scale image into binary (Black and White) image. This simplifies our classification algorithm to discriminate two objects i.e. black (background) and white (hand) in our image. To obtain best result we have to choose best possible threshold value and segmentation can be done according to that value. Otsu algorithm is used to convert image into binary [17]. Suppose there are two classes of pixels with a0 as background pixel and b1 as pixel (hand). a0 shows the pixels with intensity level [1,2.....K] and b1

shows the pixels with intensity level [K+1.....L], from these classes we get the threshold value K* which is in between value of K and K+1 and now hand pixel is assigned value 1 and background pixel is assigned value 0 and we get our desired binary image See Figure 3.4 and Figure 3.5.

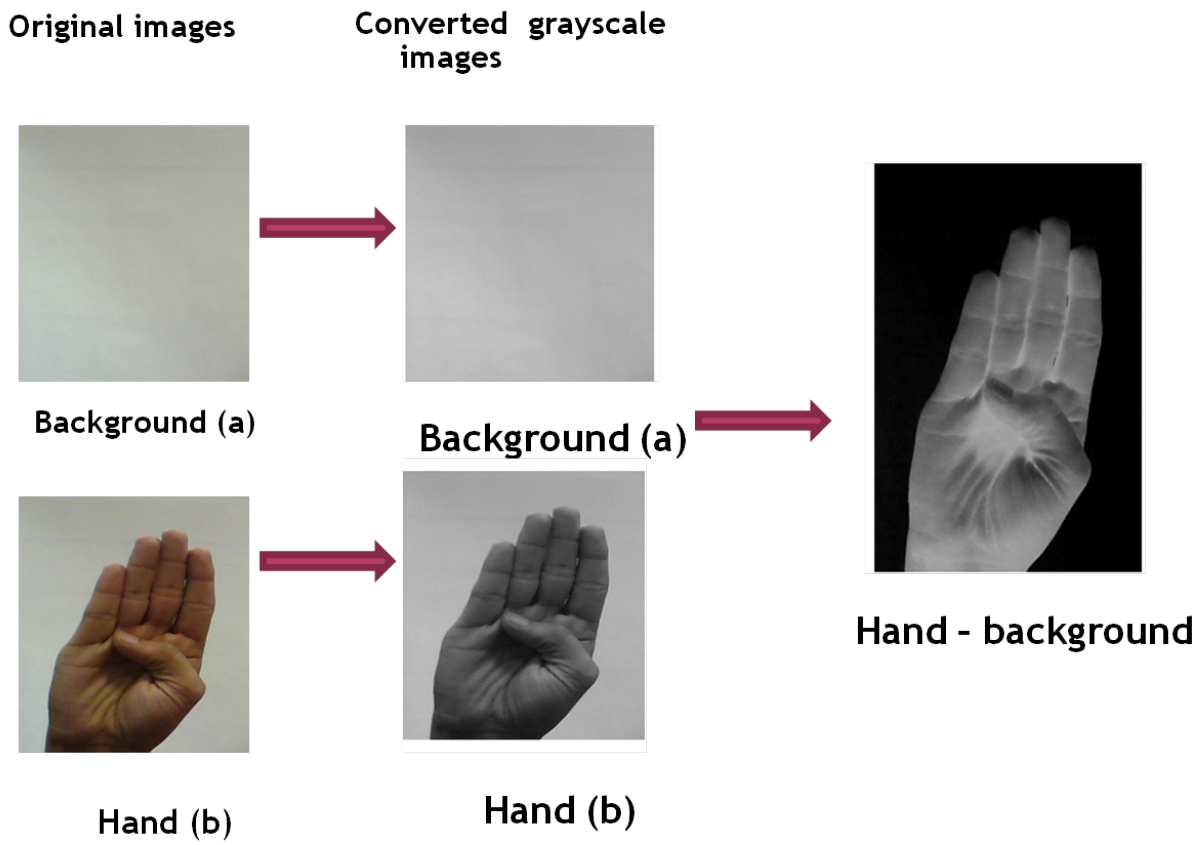


Fig 3.4: Image Segmentation part 1.

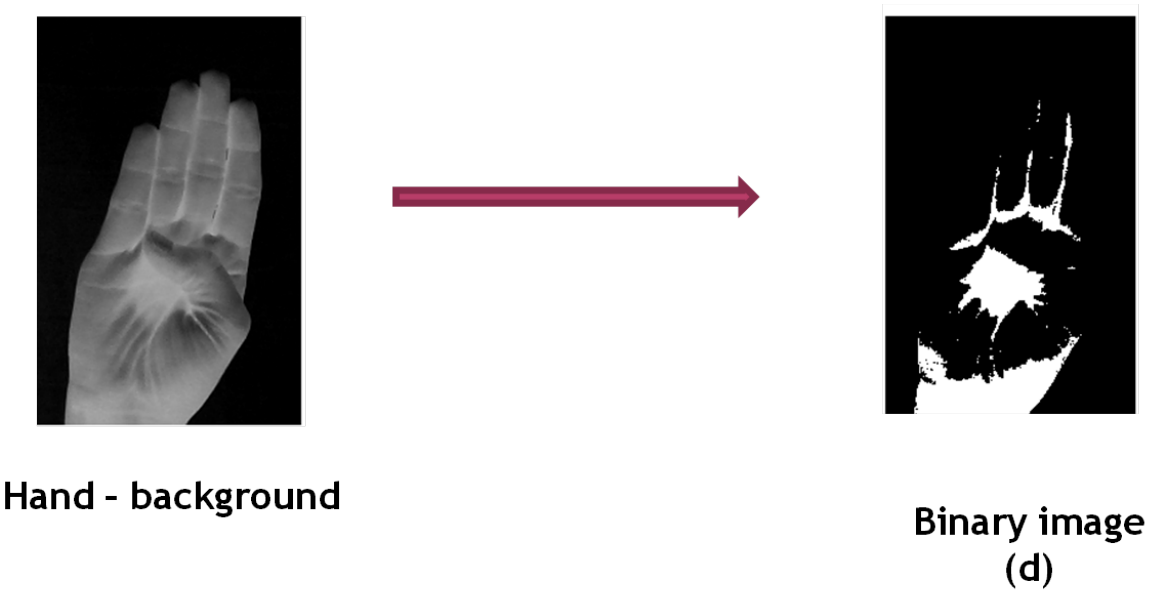


Fig 3.5: Image Segmentation part 2.

The segmented images by the Otsu algorithm still need more processing to remove unwanted data and errors. For example there might still remain some background parts containing 1s and some hand parts which denote 0s. In order to remove that noise we have to apply morphological filtering techniques on those segmented images. Dilation' Erosion' Opening' and Closing is the basic operator that work in morphological filtering [9]. See Figure 3.6.

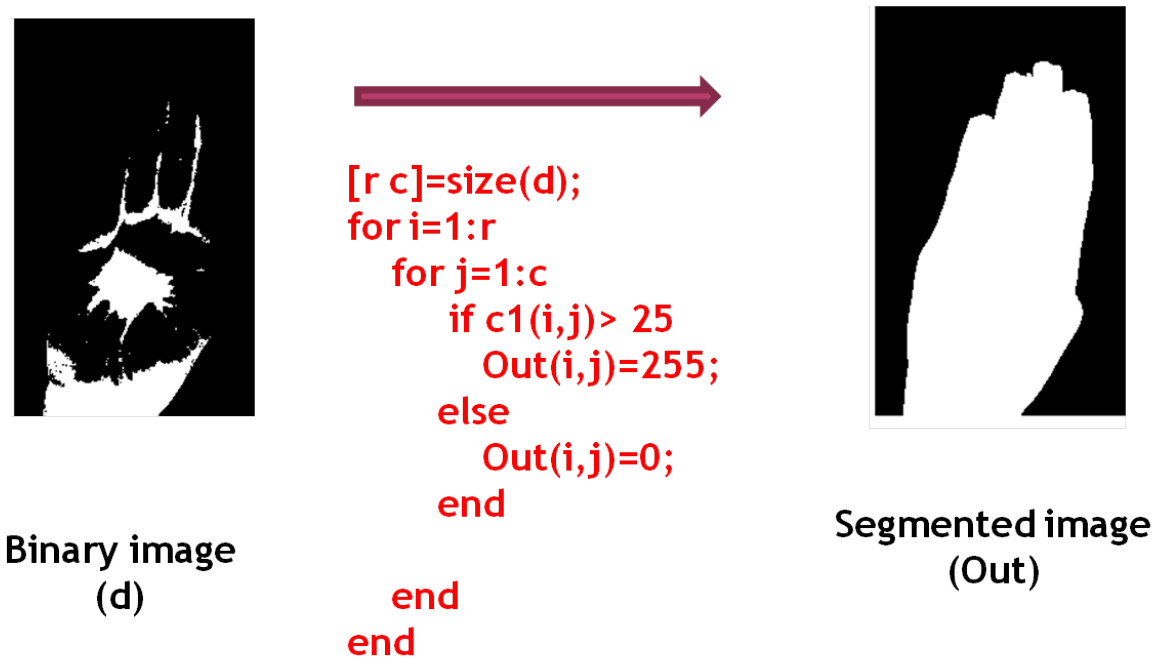
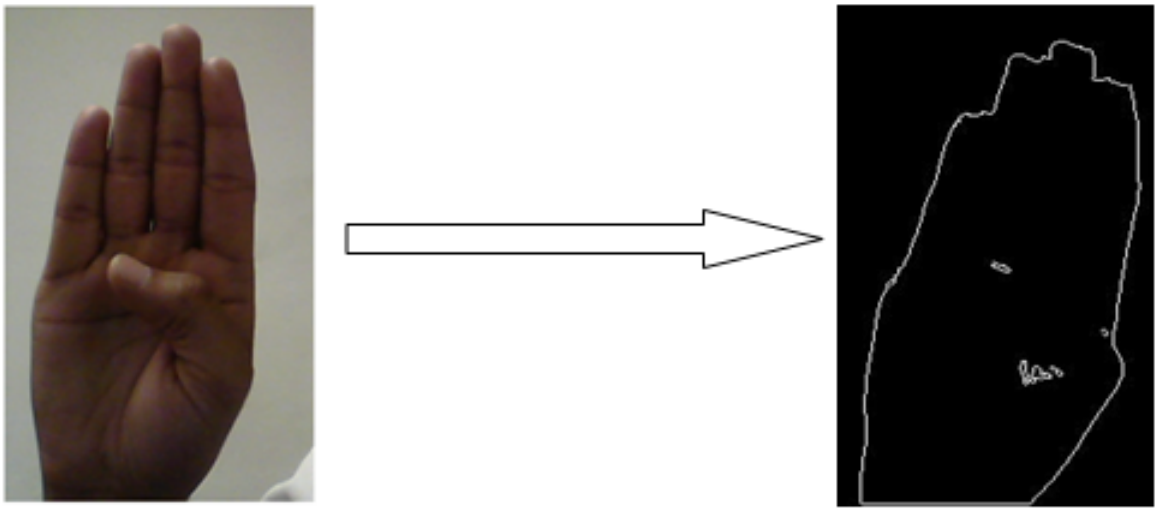


Fig 3.6: Morphological Filtering.

Now we have the morphed image for the sign but its a binary image with very less details and in binary form few of the signs look similar to each other and thus they very difficult to identify. We apply edges of the original image using edge detection algorithm [18]. Figure 3.7 depicts this process.



```
edge(hand_image,'canny',0.3,0.8);
```

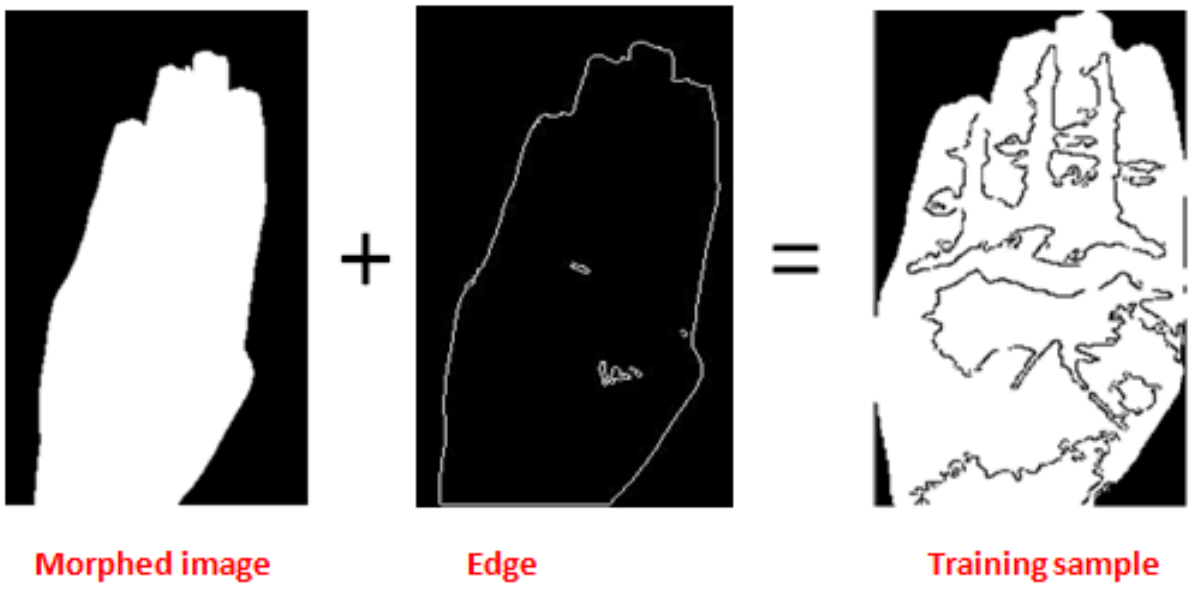


Fig 3.7: Edge Detection and Final Training Sample.

For each alphabet there are 10 different samples.

3.1.2 Words Database

In order to create words database, we first need to know that words in sign language consist of several hand movements. That means a word consists of more then one gesture in it. Therefore we need a sequence of images that can accumulately describe an English word. In this work we use a frame selection method that captures sequence images. In this method the frame rate is 35 frame/sec with a sampling factor of 4. The number of frames (nFrames) is calculated by :
nFrames = floor(NumberOfFrameInVideo / sampling_factor);

Then we apply the frame selection method to select the frames from the video we captured from live cam for further processing.
for i = 1 :nFrames
IMG = read(VideoObj, (k-1)*sampling_factor+1);
end

After all the frames are selected, each frame passes through the image segmentation, morphological filtering and edge detection phase mentioned in the above section. Then our system selects all the images and display them using montage. Montage(I) displays all the frames of a multiframe image array I in a single image object. I can be a sequence of binary, grayscale, or truecolor images. A binary or grayscale image sequence must be an M-by-N-by-1-by-K array. A truecolor image sequence must be an M-by-N-by-3-by-K array. This montage image is our words database sample for a word. see Figure 3.8

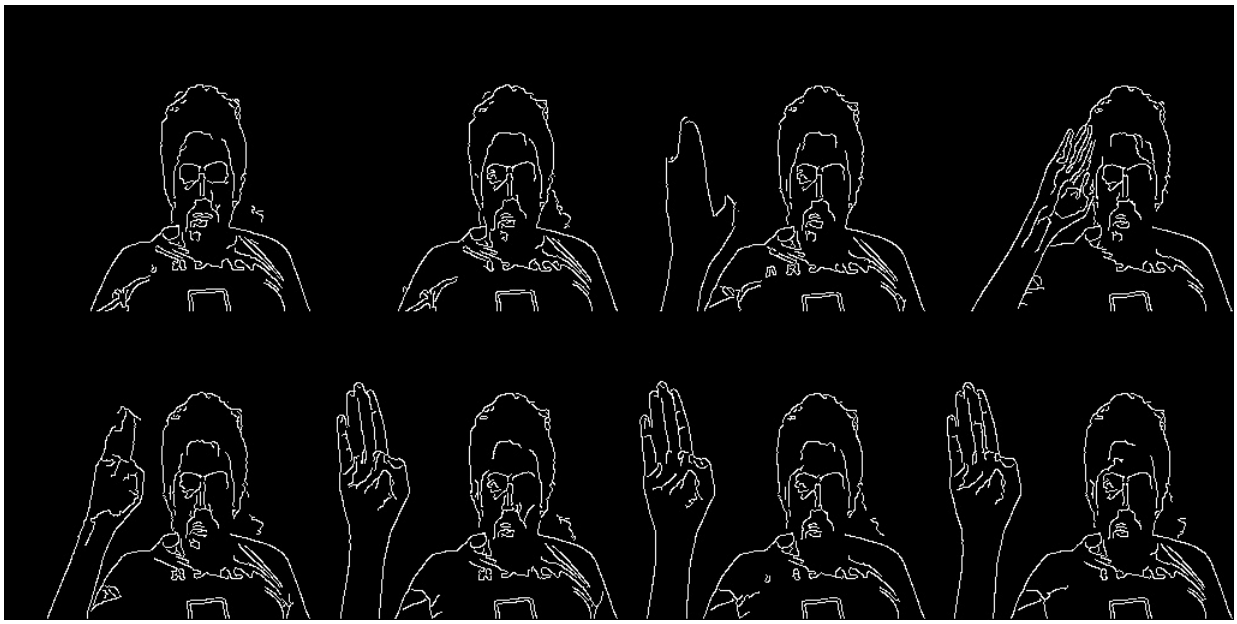


Fig 3.8: Words Database Image Sample.

3.2 Sign Translation

Sign Translation begins with the extraction of feature for gesture recognition. Feature extraction and matching is performed using the image Cross-correlation Coefficient. In signal processing, cross-correlation is a measure of similarity of two waveforms as a function of a time-lag applied to one of them. We use this function for matching of hand gesture. Cross correlation is usually applied to find the offset between two similar but time-shifted functions.If a and b are two discrete-time sequences,Cross-correlation measures the similarity between a and shifted (lagged) copies of b as a function of the lag. If a and b have different lengths, the function appends zeros at the end of the shorter vector so it has the same length, N, as the other. The cross correlation coefficient is defined in Equation: 3.1.

$$\gamma(x,y) = \frac{\sum_s \sum_t \delta_{(x+s,y+t)\delta} \delta_T(s,t)}{\sum_s \sum_t \delta^2_{(x+s,y+t)\delta^2} \delta_T(s,t)} \tag{3.1}$$

$$Where \quad \delta_{(x+s,y+t)=I(x+s,y+t)-I\acute{}(x,y)}$$

$$\delta_T(s,t) = T(s,t) - T\acute{},$$

$$s \in \{1,2,3,...,p\},$$

$$t \in \{1,2,3,...,q\},$$

$$x \in \{1,2,3,..,m-n+1\},$$

$$y \in \{1,2,3,...,n-q+1\},$$

$$I\acute{}(x,y) = \frac{1}{pq}\sum_s\sum_t I(x+s,y_t)$$

$$I\acute{=}\frac{1}{pq}\sum_s\sum_t T(s,t)$$

The value of cross-correlation coefficient γ ranges from [-1 to +1] corresponds completely not matched and completely matched respectively. For template matching the template, T slides over I and gamma is calculated for each coordinate (x, y). After calculation, the point which exhibits maximum gamma is referred to as the match point. The following step is used for matching of hand gesture:

Step 1: A hand gesture template of size m n is taken.

Step 2: The normalized 2-D auto-correlation of hand gesture template is found out.

Step 3: The normalized 2-D cross-correlation of hand gesture template with various template is calculated.

Step 4: The mean squared error (MSE) of auto correlation and cross-correlation of different sample are found out. The minimum MSE is found out and stored.

Step 5: The corresponding minimum MSE represent the recognized gesture.

We have a GUI which consist of two different mechanism to translate alphabets and the words. see Figure 3.9

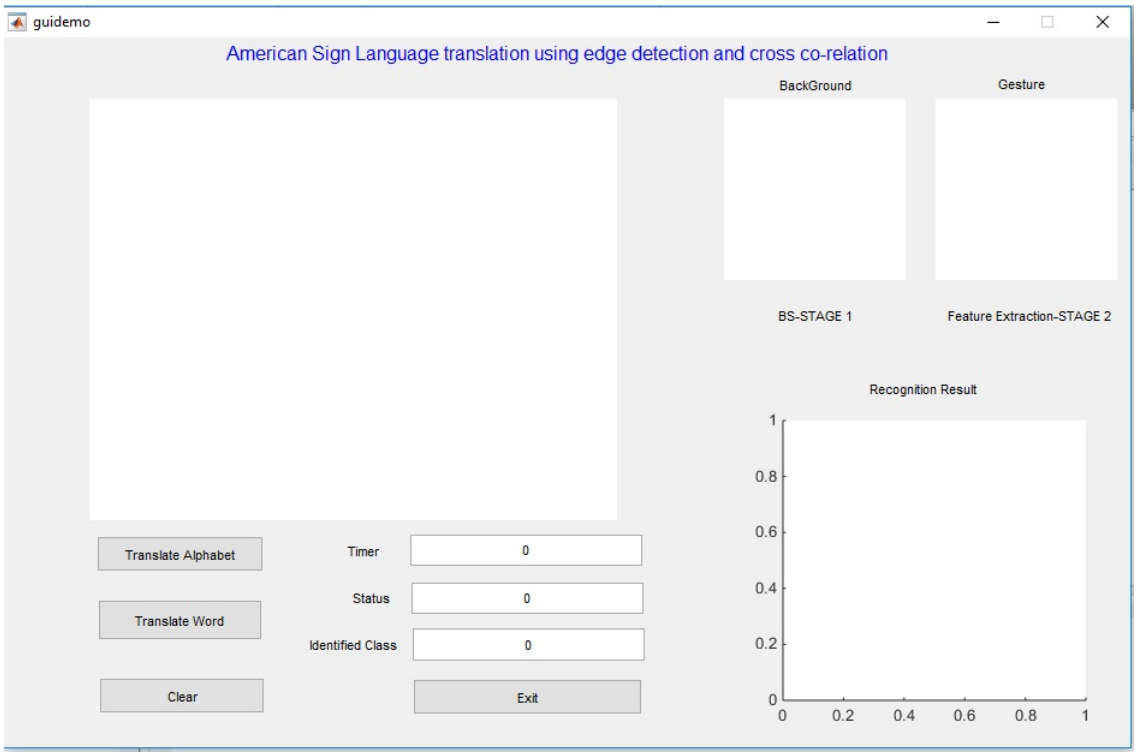


Fig 3.9: Translation GUI.

3.2.1 Alphabet Translation

Translate Alphabet button on translation GUI open a camera view in the window. It will follow the same steps of capturing background and gesture and then do image segmentation, morphological filtering and edge detection as explained in Section 3.1.1. The processed image is now compared with the alphabet database using the cross-correlation explained in Section 3.2. After this comparison the result will be displayed in the image box named Recognition Result and the alphabet in the text box named Identified Class see Figure 3.9.

3.2.2 Word Translation

The Translate word button on Figure 3.9, also opens a camera view in the application window and then capture and select the image sequences. Each frame is processed through

image segmentation, morphological filtering and edge detection. After processing the frame sequence, these processed frames converted into montage as mentioned in Section 3.1.2. At the end the montage image is compared with the database and the resultant image is displayed in the image box Recognition Result and the word is displayed in Identified Class in the translation GUI.

3.2.3 Materials

The materials that have been used in this thesis allowed us to build a prototype system capable of translating basic ASL signs to text. These are available at the Laboratory for Applied Remote Sensing and Image Processing of the Univeristy of Puerto Rico at Mayaguez. In this work we used:

1. A computer workstation with MATLAB R2015a
2. A Computer Webcam

Chapter 4

RESULTS

In this chapter, classification accuracy for ASL translation is presented. Different test data image is compared against each sample data images several times in order to find the accuracy of the system. Each English letter has 10 different sample images and each word has a sample montage (which is a sequence of image frames stored in database).

Our suggestive method have been done on Intel Core i3-2330M CPU, 2.20 GHz with 4 GB RAM under Matlab R2015a environment. Figure 4.1 shows the face of worked systems.

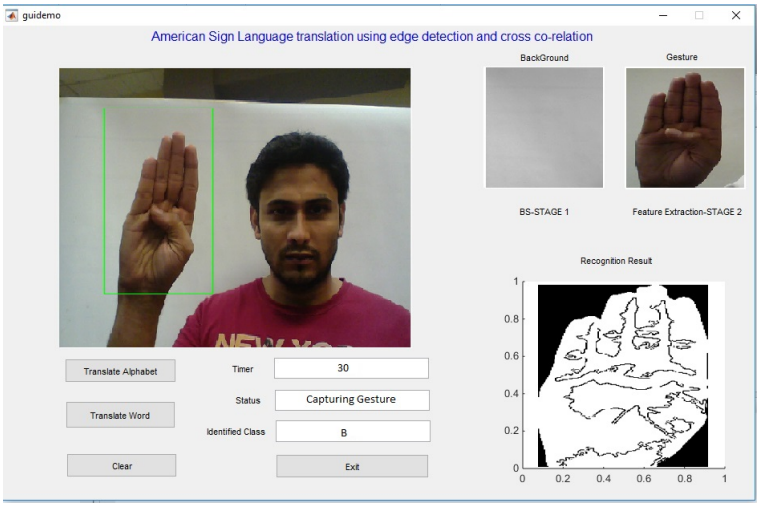


Fig 4.1: ASL Sign Language Translation System.

.

In this study, for experimental analysis, we had applied the above mention technique on our database of American Sign Language which consists of 260 images i.e. 10 images per character and we was able to recognize 26 characters out of 26 characters from sign language. Table 4.1 shows the accuracy rate for each hand gesture.

English Alphabet	Database Images	Recognized Images	Performance
A	10	8	80 %
B	10	10	100 %
C	10	10	100 %
D	10	10	100 %
E	10	7	70 %
F	10	10	100 %
G	10	10	100 %
H	10	10	100 %
I	10	10	100 %
J	10	7	70 %
K	10	10	100 %
L	10	10	100 %
M	10	7	70 %
N	10	10	100 %
O	10	10	100 %
P	10	10	100 %
Q	10	10	100 %
R	10	10	100 %
S	10	8	80 %
T	10	10	100 %
U	10	10	100 %
V	10	10	100 %
W	10	10	100 %
X	10	10	100 %
Y	10	10	100 %
Z	10	8	80 %
Total	260	245	94 %

Table 4.1: Performance of each Sign group of Alphabets.

In our experiment (with the 94.23% accuracy for alphabets), we observed confusion

hand gesture in the recognition phase between some signs. The major confusions were amongst A, S and E, M. The confusion occurs because the letters are similar to each other.

The sign for letter A (Figure 4.2a) and S (Figure 4.2b) are different from each other when you are watching someone producing the signs but when you take a 2d picture of the same sign then the confusion between the two sign is noticeable. We calculated the cross-correlation coefficient of both letters and the result was in the range of 0.7 0.8. But when we find the cross-correlation coefficient of letter A and S with other letters (B (figure 4.3a), C (figure 4.3b), D (figure 4.3c), F (figure 4.3d) etc.) it is in the range of 0.3 - 0.4, that clearly defines the difference between letter A and S from Others 24 letters. The resultant coefficient is so close to each other, that is the reason why sometimes the system recognizes them as identical.

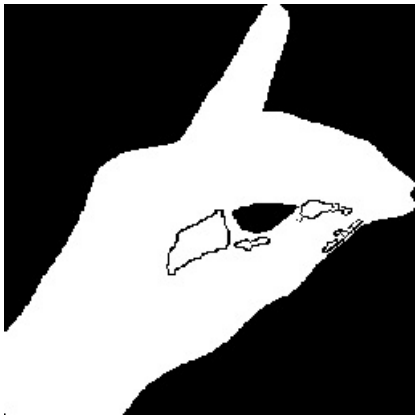
(a) Letter A

(b) Letter S

Fig 4.2: Letter A and S

(a) Letter B

(b) Letter C



(c) Letter D

(d) Letter F

The same goes for letter E (Figure 4.5c) and M (Figure 4.5d), both have similar 2d plot which makes it difficult for the system to indentify the letter from each other. The cross-correlation coefficient for letter E and M is 0.6 0.7.

(a) Letter E (b) Letter M

Fig 4.4: Letter E and M

The most difficult letter to identify is letter Z. The reason behind it is the letter itself cause the sign of Z is quite confusing as you can see on Figure. As you can see figure a, b, c, and d all are signs for the same letter Z but their cross-correlation coefficient says they are different and the range of the coefficient is inconsistent. It varies from 0.5 to 0.7 if we compare all the training set for letter Z. The same sign is produced differently like in figure b the fingers are facing the producer but in figure a its the other way around and both ways are correct.

(a) (b)

Fig 4.5: Different Variations of Letter Z

(c) (d)

Fig 4.5: Different Variations of Letter Z

For the words we choose some everyday words and tested then with our system and we got 92% accuracy see Table 4.2.

ENGLISH WORD	Database Images	Recognised Images	Performance
HI	10	8	80%
BYE BYE	10	9	90%
DAD	10	9	90%
MOM	10	10	100%
GOOD NIGHT	10	10	100%
GOOD MORNING	10	9	90%
HOME	10	10	100%
SCHOOL	10	10	100%
THANK YOU	10	9	90%
HELP	10	8	80%
TOTAL	100	92	92%

Table 4.2: Performance of each Sign group of Words.

The confusion while recognizing the words is the way of producing the sign because a word is collection of many different sequential gestures and the way it is produced once doesnt assure that it is going to be the same as previous. And selecting frame is also challenge because the frame capturing window is about 35 frames per trigger, from those frame picking the right set of frames is complex task. Not everyone does signs exactly the same way he/she did before and each person has different hands in terms of hand size, finger size, shape and thickness.

As we can see in the Table , word Hi (see figure 4.6) and help (see figure 4.7) have 80% of accuracy because both signs have very common gestures which can be easily create confusion while translating.

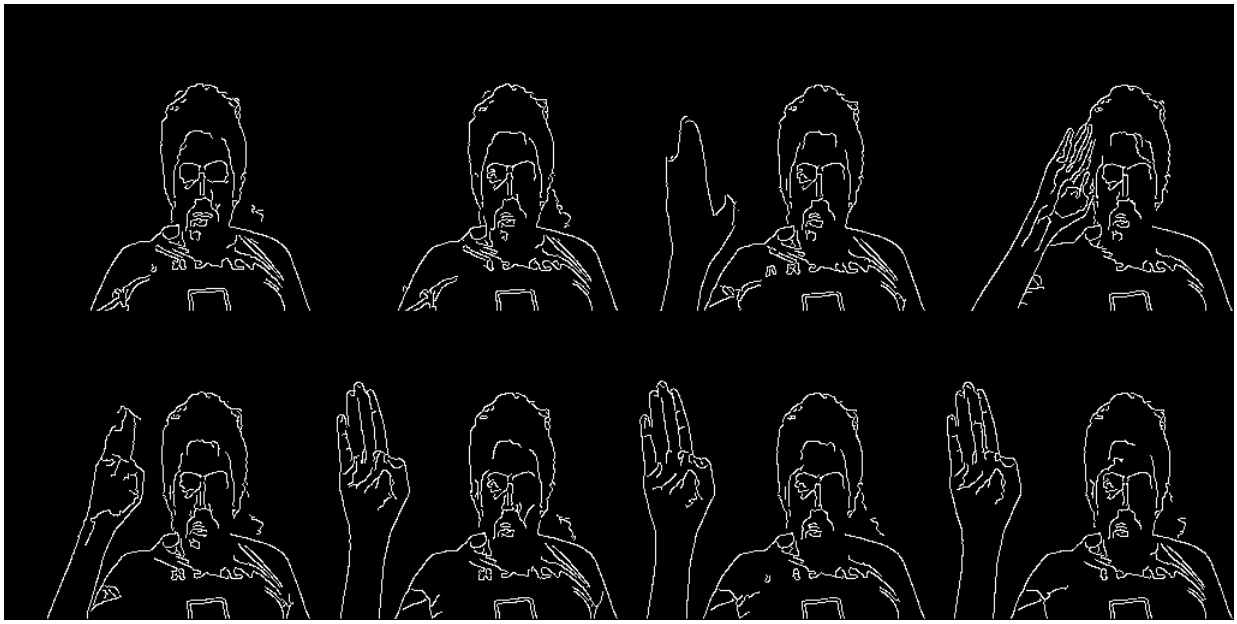


Fig 4.6: Word HI

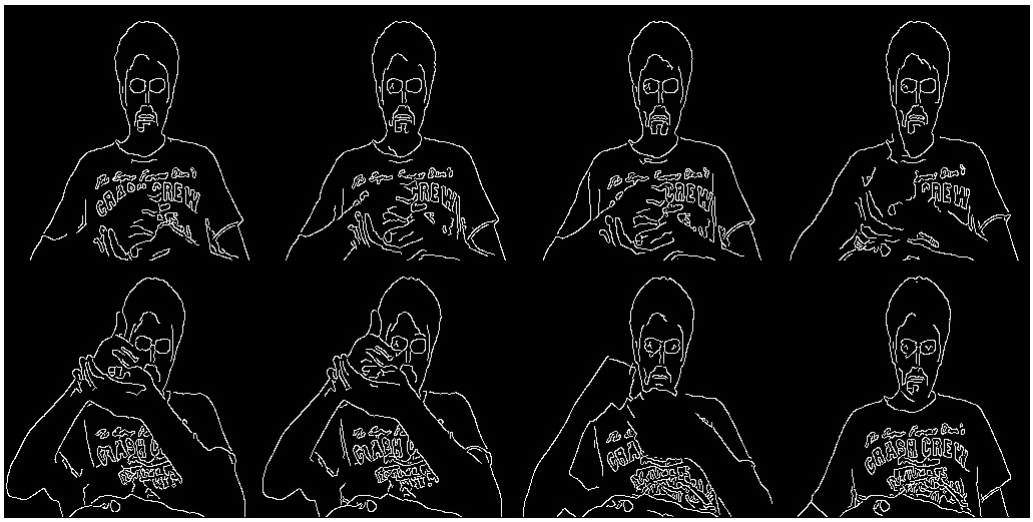


Fig 4.7: Word HELP

Fig help The accuracy for words BYE BYE (see figure 4.8), DAD (see figure 4.9), GOOD MORNING (see figure 4.10), THANK YOU (see figure 4.11) is 90%. The reason behind it is the word BYE BYE and DAD both have similar waving hand gesture which sometimes difficult to identify. Word GOOD MORNING and help both have almost identical hand raise gesture. See Figure (see figure 4.10) and (see figure 4.11)

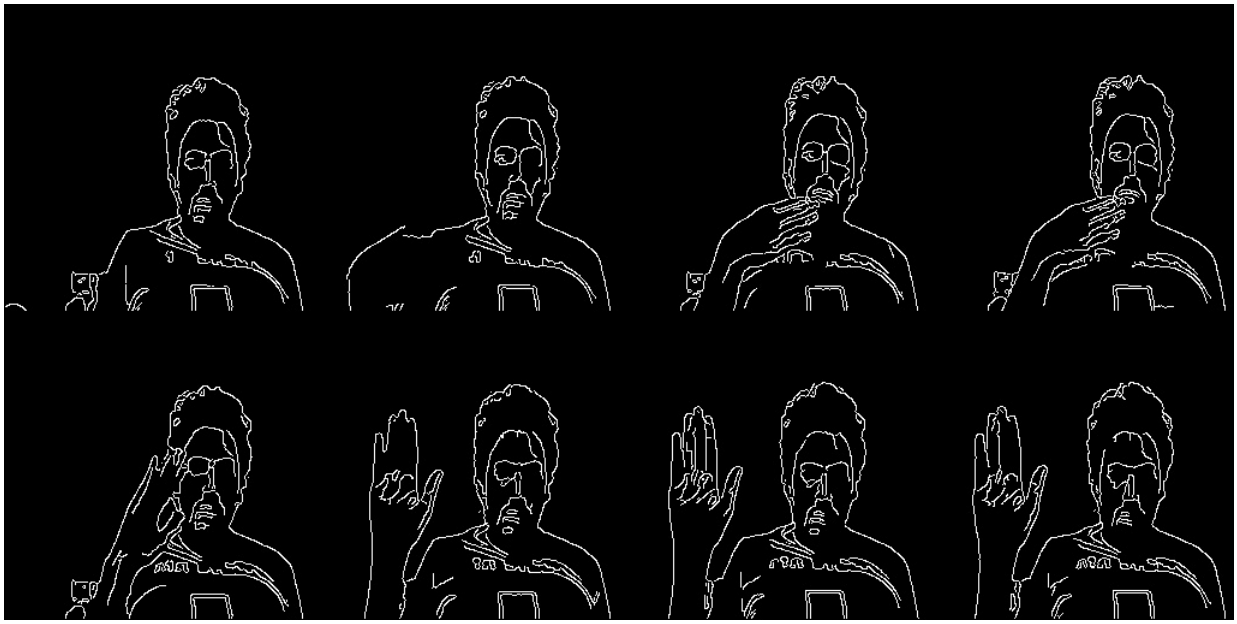


Fig 4.8: Word BYE BYE

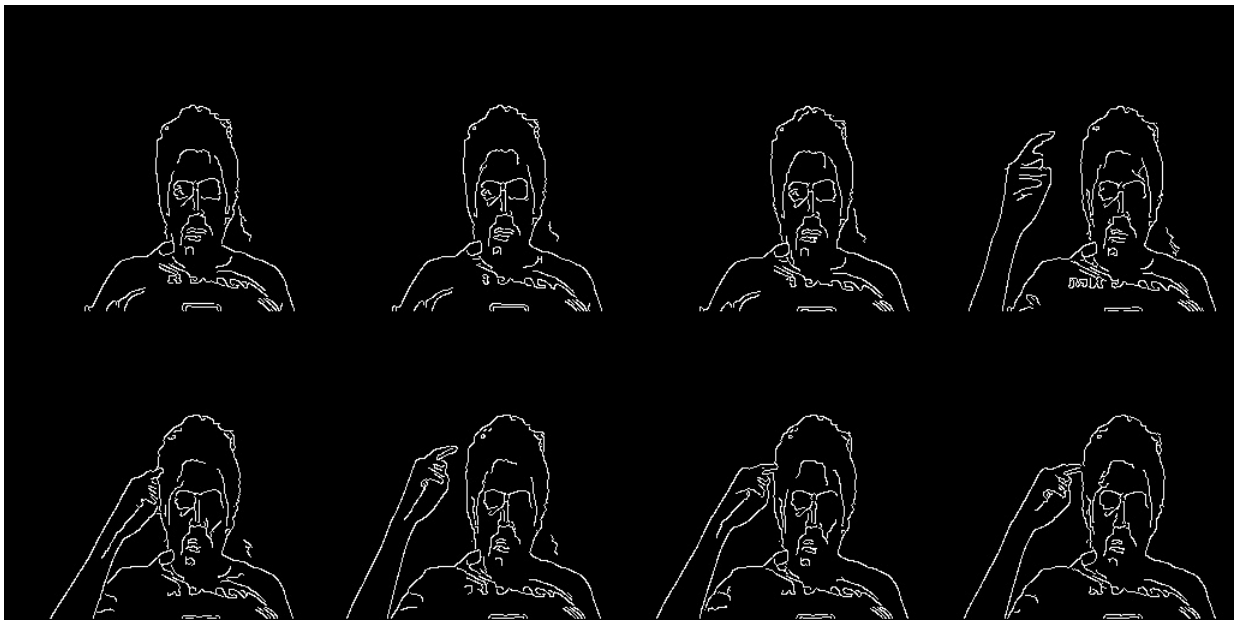


Fig 4.9: Word DAD

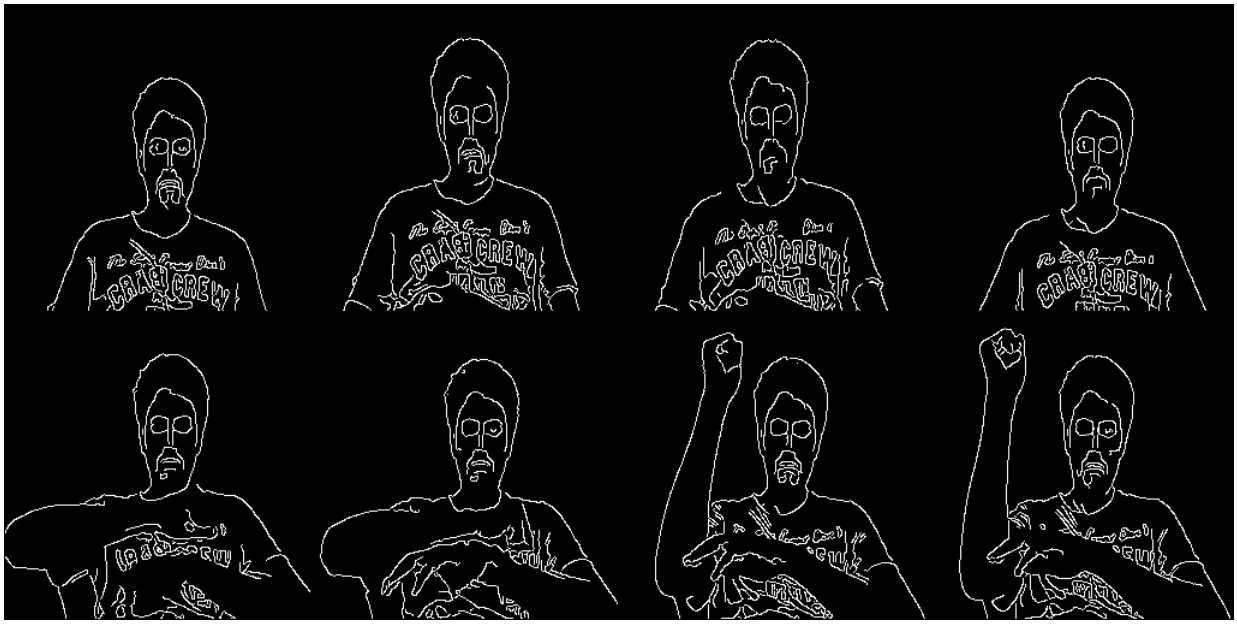


Fig 4.10: Word GOOD MORNING

Fig 4.11: Word THANK YOU

Chapter 5

CONCLUSION

5.1 Contributions

The goal of this work was to develop a practical real-life application to help remove barriers in communicating via ASL. A key aspect of our work was to develop a system that works effectively in real-time, and that requires minimum equipment such as a generic web-camera and a computer, laptop or similar. Previous work in this area is often restricted by the need for bulky hardware, probes, or cameras. The camera on a computer, acts as the input source for our classifier. Using the Cross-correlation and edge detection approach we are able to use the input image stream to identify the start and end of ASL gestures. To facilitate this identification process, we require the user to produce a sign in front of the camera mounted on the computer.

One limitation of the presented approach is its sensitiveness to the background scene. The background must be uniform because it can introduce additional structures during the edge-detection and binary image transformation stages that affect the classification if the background has many objects and glare, resulting in a miss classification of the sign. However, our experimental evaluation shows that the developed system is capable of achieving a classification accuracy of 92 to 94 % when identifying ASL gestures using real

and synthetic data. Improvement of overall classification accuracy measurements while increasing the number of ASL gestures recognizable by our system can be extensions of this work. The following section summarizes the suggested research path to extend this work.

5.2 Future Work

We have presented a methodology to translate ASL signs to text and provided a prototype implementation of such system capable of translating basic ASL signs. Broadening the scope of this work, we present five alternatives of extending this thesis. Each of these are described in the following sections.

5.2.1 Tracking of Primary Body Locations

As was determined, the primary body locations need to be tracked dynamically. The primary body locations for ASL sign communication comprise the head, torso and upper limbs. There are two means by which they could be tracked, the first by optical position trackers and the second by using a camera with depth sensing and IR (infrared) sensor. In the first instance, a minimal set of optical position trackers could be placed on the head, torso and recessive arm. From each set, a reference position and orientation could be determined. This approach could be coupled with current system as these optical sensors will provide location data to our system through custom APIs. The second approach would require infrared sensor and depth sensor integrated on the camera and they can provide the 3d image and the IR can provide infrared activity this way we dont have to depend on perfect lights where the system is used.

5.2.2 Learning Algorithms

By increasing size and resolution of the data, complexity of its management is also increased. As we discussed earlier the more training set for each letter or word we have the more accurate our results will be. A more flexible structure needs to be adopted, one that since no human movement is alike, must to be able to adapt. Such a management system is afforded by learning algorithms like neural networking.

5.2.3 Translation of Signs into Speech

Although the signs convey in most cases a literal meaning, they are not performed in the order in which English is articulated, thus a first requirement would be the translation of signs into English text. As mentioned, the expressions recognised would be instrumental in the shaping of signs into phrases. Once the signs have been translated into text, since speech synthesis has been developed quite completely, an off the shelf text to speech synthesiser would then be adopted to finally complete the translation.

5.2.4 Development of an English-to-ASL System

This system will include the entire processing architecture of the English to ASL translator. In which the dictionary words will be translated into signs. In this we will to use a speech recognition tool to convert voice to text and then those words will be represented by sign images.

5.2.5 Expression Identification

Recognition of the signs being conveyed gives the building blocks of the language, however the structuring of a phrase relies heavily on expression to shape it. To facilitate recognition of expression, the application of sensors to the face is not a practical option. A visual based technique would have to be adopted, using cameras with sensors and feature extraction

techniques. The camera would need to be located on the portable system, giving rise to a non-oblique angle from which to apply the technique, thus requiring further processing of the acquired data such that an oblique perspective could be generated.

5.2.6 Video Calling with Sign translator

One of the biggest future plan is to implement a cloud based video calling interface where all the database storage and data processing will be done on cloud virtual machines. Also provide a desktop and mobile app (which you can install in your Android/IOS) so anyone can communicate with a deaf person without the help of a human interpreter.

The proposed system can also be integrated with current video chat environments like Skype or Google Hangouts. This can be achieved using the video calling API provided by Skype or Google Hangouts. Using the api we can get the video in and out stream and we can use them as our input data.

Bibliography

[1] Landsberger SA, Diaz DR. *Inpatient psychiatric treatment of deaf adults: demographic and diagnostic comparisons with hearing inpatients.* Psychiatr Serv 2010; 61:19699.

[2] Baines D, Patterson N, Austen S. *An investigation into the length of hospital stay for deaf mental health service users.* J Deaf Stud Deaf Educ 2010; 15:17984.

[3] Mitra S., Acharya T. *Gesture Recognition* IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 2007; 37: 311-324.

[4] Wysoski S.G., Lamar M.V., Kuroyanagi S., Iwata A. *A Rotation Invariant Approach On Static-Gesture Recognition Using Boundary Histograms And Neural Networks* Proceedings of the 4th International Conference on Neural Information Processing (ICONIPOZ) 2003: 981-04-7524-1.

[5] Kouichi M., Hitomi T. *Gesture Recognition using Recurrent Neural Networks* ACM conference on Human factors in computing systems: Reaching through technology 1999: pp. 237-242. doi: 10.1145/108844.108900.

[6] Maraqa M., Abu-Zaiter R. *Recognition of Arabic Sign Language (ArSL) Using Recurrent Neural Networks* IEEE First International Conference on the Applications of Digital Information and Web Technologies, (ICADIWT) 2008, pp. 478-48. doi:10.1109/ICADIWT.2008.4664396.

[7] Microsoft Kinect Sign Language Translator, https://www.microsoft.com/en-us/research/kinect-sign-language-translator-part-1/

[8] The next web, Microsoft Research Kinect Translator, http://thenextweb.com/microsoft/2013/10/30/microsoft-research-uses-kinect-translate-spoken-sign-languages-real-time/#gref

[9] Hasan M.M., Mishra P.K. *Features Fitting using Multivariate Gaussian Distribution for Hand Gesture Recognition* International Journal of Computer Science Emerging Technologies IJCSET 2012: Vol. 3(2).

[10] Hasan M.M., Mishra P.K. *Robust Gesture Recognition Using Gaussian Distribution for Features Fitting* International Journal of Machine Learning and Computing 2012: Vol.2(3).

[11] Kulkarni V.S., Lokhande S.D. *Appearance Based Recognition of American Sign Language Using Gesture Segmentation* International Journal on Computer Science and Engineering (IJCSE) 2010:Vol. 2(3), pp. 560-565.

[12] Hasan M.M., Mishra P.K. *Brightness Factor Matching For Gesture Recognition System Using Scaled Normalization* International Journal of Computer Science Information Technology (IJCSIT) 2011: Vol. 3(2).

[13] Viola P., Jones M. *Rapid object detection using a boosted cascade of simple features* Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society 2001: doi: 10.1109/CVPR.2001.990517.

[14] Freund Y., Schapire R. *A decision-theoritic generalization of on-line learning and an application to boosting* journal of computer and system sciences 1997:55, 119 139.

[15] Storring M., Moeslund T., Yong L., Granum E. *Computer Vision-Based Gesture Recognition For an Augmented Reality Interface* 4th IASTED International Conference on Visualization Imaging, and Image Processing, pages 766-771, 2004.

[16] Shin J., Lee J., Kil S., Shen D., Ryu J., Lee E., Min H., Hong S. *Hand Region Extraction and Gesture Recognition Using Entropy Analysis* IJCSNS International Journal of Computer Science and Network Security 2006: VOL.6 No.2A.

[17] N. Otsu *A Threshold Selection Method from Gray Level Histograms* IEEE Transactions on Systems Man and Cybernetics, vol. SMC-9, NO. 1, 1979.

[18] The MathWorks Inc, Edge Detection Algorithms, `https://www.mathworks.com/discovery/edge-detection.html`

[19] Wachs J.P., Klsch M., Stern H., Edan Y. *Vision-Based Hand-Gesture Applications* Communications of the ACM 2011: Volume 54 Issue 2, Pages 60-71