IMPROVED SIMILARITY MEASURES FOR AMERICAN SIGN LANGUAGE RECOGNITION USING MANUAL HANDSHAPES AND HAND APPEARANCES

by

SIDDHARTHA GOUTHAM SWAMINATHAN

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT ARLINGTON
May 2016

Acknowledgements

I would like to thank my supervisor, Dr.Vassilis Athitsos for constantly motivating me on my thesis and helping me out with his invaluable inputs throughout the course of my thesis. I wish to thank my committee members Dr. Ioannis Schizas and Dr. Michael Manry for their interest in my research and for taking time to serve on my thesis committee.

I would also like to thank Chris Conly for introducing me to American Sign Language and also for his constant support and help throughout my thesis.

Finally, I would like to express my gratitude towards my parents for their encouragement and support throughout these years. I also thank my friends who have supported and helped me throughout the course of my Masters studies.

April 18, 2016

Abstract

IMPROVED SIMILARITY MEASURES FOR AMERICAN SIGN LANGUAGE RECOGNITION USING MANUAL HANDSHAPES AND HAND APPEARANCES

Siddhartha Goutham Swaminathan, MS

The University of Texas at Arlington, 2016

Supervising Professor: Vassilis Athitsos

American sign language is a primary language for approximately 0.5 to 2 million people who are deaf or hard of hearing in the United States[22][23]. When a user encounters an English word which he does not understand, he looks up the meaning for it in a dictionary. However, when an American Sign Language (ASL) user encounters an unknown sign, looking up the meaning of that sign is not an easy task. There are many systems available to access ASL signs that require articulatory properties such as handshapes but these systems fail if there is a slight variation of what the user is looking from the actual ASL dictionary. The existing system proposes a baseline similarity measure based on dynamic time warping(DTW)[8] where feature vectors are extracted based on trajectory of the hand and DTW is applied on this time series of feature vectors to obtain the similarity measure between the query sign and the database of videos. Handshape is also one of the essential component that makes up an American sign language along with the trajectory information. However it is not easy to recognize different hand shapes. The system implemented here performs improvements on methods proposed by [3] [2] and [5] by incorporating hand shape information from the signer. The Goal is to evaluate methods based on two topics, one based on Manual handshape inputs from the user which is considered to be a best case, where the user manually spends more time specifying the hand shapes that constitute a sign for attaining better accuracy and other method based on hand appearances. we also investigate how well the method based on hand appearances fare with the near perfect accuracy method of manual hand shapes.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

Introduction

There exists many different types of gestures among which the most well-structured and defined set belong to the sign language sets [1]. Each gesture in the sign language has a specific meaning denoting it and in order to make it tractable strong rules and grammar may be applied to the gesture recognition process. American Sign Language (ASL) is the most preferred sign language among the others for most of the deaf people in the USA.

Finding an appropriate or an unknown sign from the ASL dictionary can be a cumbersome task [2]. ASL dictionaries allow looking at the signs based on their English translations. So it becomes more difficult if the user doesn't know the meaning of the sign or its direct translation in English as these are typically English to ASL dictionaries [2]. There are many systems available to access ASL signs that require articulatory properties such as handshapes but these systems fail if there is a slight variation of what the user is looking from the actual ASL dictionary. So there is need of a system which can help to look up unknown signs which in turn benefits millions of users and learners of sign language around the globe.

The system implemented based on [2] assists the users to retrieve unknown signs from the ASL. The user can perform a sign in front of a webcam recording it at the same time and submit it as a query to the system for which the system retrieves the signs which are best matched to the query from the system database(Ideally these are ASL to English translations). The user can then browse through the top matched signs from the system to find the sign of interest.

The datasets used in [2] were created by [3] which contain examples for nearly 3000 signs from the Gallaudet dictionary. The dataset consists of 2 frontal views, a side view and a view zoomed in on the signer's face[5]. Out of these we use one of the two frontal views for our system implemented here where we have 3 examples per sign class and the number of sign classes totaling 1113. As mentioned in [3] the signs are distinguished from one another by the shape of the hands their orientation and the relative position in the signing space with respect to the body and movement.

The system implemented here performs improvements on methods proposed by [3] [2] and [5] by incorporating hand shape information from the signer. In order for the system to work well the system implemented here is not a fully automatic system [2] and requires knowing the bounding box of the hands in both the query and the videos in the database. The database videos were created by [3] which includes manual annotation of the location of both the dominant (right-hand) and non-dominant hands (left-hand) and also the location of the face. The bound boxes for the dominant and non-dominant hands are annotated for each frame of the sign while the bounding box for the face is annotated only on the first frame of the sign[3].

[4] Provides a method based on temporal and spatial gesture segmentation to recognize manual gestures. A vital component of any hand gesture recognition system is the hand tracker and detector. The method proposed in [4] doesn't require the user to specify hand locations for both dominant and non-dominant hands but this might work only for a system with a small vocabulary of sufficiently dissimilar gestures with approximate estimations of hand locations. But in actual more accurate hand position information is needed for better results with a large vocabulary of complex two-handed gestures, which are typically found in sign languages [5].

[2] proposes a baseline similarity measure based on dynamic time warping(DTW)[8]. Feature vectors are extracted based on trajectory of the hand and DTW is applied on this time series of feature vectors to obtain the similarity measure

between the query sign and the database of videos. DTW allows comparison of two different feature vectors by finding an optimal match even for the stretched and compressed sections of the time series. The feature vectors are extracted for every frame of the gestured sign in the query video and compared with the database of videos to obtain a combined similarity score based on dynamic programming. The similarity measure mentioned above is also improved by incorporating hand appearances along with the DTW.

This system works on hand motion to discriminate between signs as proposed in [2][3][5] and incorporate hand shape and hand appearance information (manual, semi-automatic methods) to improve the accuracy significantly. As we had already mentioned that the system is not fully automatic and the system should work well for public use[2], it requires the user to provide the bounding box for both the dominant hand and non-dominant hand (if performed sign is two-handed) and the head as well as the start and end frames along with the query.

The manual method doesn't require the user to specify the bounding boxes for the dominant and non-dominant hands as well as the bounding box for the face but requires the user to specify the hand shape information for the start and end frames for both the dominant and non-dominant hands as well as the start and end frames for the recorded video as the query input. This minimizes the work that the user must perform to match a sign or gesture[5] . For this to accurately work the system should be capable enough to detect and track the hand locations automatically in every frame without any intervention from the user. This is achieved by the RGB-D technology of the kinect camera used in [5]. The kinect offers the functionality to automatically locate and track the user’s hands (both dominant and non-dominant) throughout the gesture, and thus eliminating few steps[5]. The process is automated by using readily available skeleton tracking algorithms to estimate hand positions in frame of the query video provided by the user[5]. The manual method thus incorporates the score obtained from the conditional probability table(a probability table computed with hand shape information among the database hand shape signs) by user specifying the hand shape information along with the DTW scores(Feature vectors for the database videos were computed offline) to improve the accuracy of the overall system.

In the semi-automatic method the hand regions which are specified as bounding box by the user in the start and end frames of the query video are matched with the database videos (which already possess the hand locations annotated by [3]) by similarity measures such as Zero Mean normalized cross correlation(ZNCC) and Histogram of oriented gradients(HOG) in combination with dynamic time warping(DTW) scores based on trajectory. The DTW scores were computed offline for the semi-automatic method for the ease of experimentation while the RGB-D skeleton tracker mentioned in the manual method can also be used to obtain the feature vectors for the query.

Chapter 2

Related Work

The system proposed here works with [2] and differs with the existing work by incorporating manual hand shapes(manual-method) scores computed from the conditional probability table along with the Dynamic Time Warping(DTW)[8] which works on hand motion to discriminate between the signs and an semi-automatic method based on hand shape appearance along with Dynamic Time Warping(DTW) to improve the accuracy.

[6] provides an approach to recognize hand gestures based on binary motion energy images which basically represents the occurrence of any movements between the images and motion history image a scalar valued image where the intensity of individual pixels is a function of motion. These temporal templates are compared with the stored models to evaluate the power of representation between them[6].[7] talks about identifying human actions based on space time shapes which are computed by combining the 2D silhouettes in every frame to get a 3D model. There are also model based methods proposed based on Hidden Markov Models(HMM) and Hidden conditional random fields. The disadvantage with these approaches is they offer very low immunity to noises in the background, accurate silhouettes extraction of hands are required even if the bounding box for the hands are known and also we do not have enough training examples for using model based methods such as HMM .

[2] and [5] proposes a baseline similarity measure based on dynamic time warping(DTW). The signs used for the experiments by this method were annotated by [3].

2.1 Dynamic Time Warping(DTW)

The DTW algorithm[8] finds optimal alignment between two time series. The cost obtained through DTW is an optimal alignment between the trajectories under test. DTW works well even if one time series may be warped non-linearly, compressed or stretched with respect to other time series.

The DTW matches two time series sequences a model X and a query Q and computes a cost D(Q,X) which is the minimum cost warping path between the model X and the query Q which is used for comparing the similarity between the query and the model sequences and thus classifying the query sequence with respect to the best matched models. Dynamic programming [9] was employed to calculate the optimal warping path and DTW cost D(Q,X) with a time complexity of $O(|Q||X|)$ where $|Q|$ and $|X|$ are the total number of frames in a query sign Q and model sign X.

2.2 Improved Accuracy with Hand Appearance along with DTW

The Dynamic time warping(DTW) discussed above is based on trajectory of hand motion for both the dominant and non-dominant hand. The hand appearance is also an important parameter for recognizing a sign along with the trajectory information.

Recognizing the actual hand shape can be a cumbersome task especially if the signed hand is in front another object which is of the same color as the hand[2]. So the hand appearance alone cannot be used to improve the accuracy of the recognition. But it does give valuable cues which can be combined with DTW to improve the similarity measure. [2] proposes a simplest possible solution which is based on Euclidean distances between the hand shapes. This has provided significant improvement to the accuracy along with DTW considering its simplicity[2]. The total distance for the hand is calculated as a sum of Euclidean distances between the start and end frame of the dominant hand and non-dominant hand if it's a two handed sign.

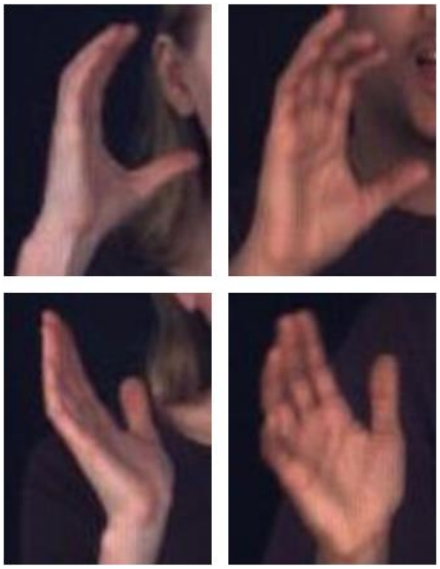


Figure 2-1 Hand appearance of 2 different signers for the same sign (sign identity : cousin, chat)

D(Hand) = ||Ds(1) - Ds(2)|| + ||De(1) - De(2)|| + ||NDs(1) - NDs(2)|| + ||NDe(1) - NDe(2)||

Ds(1) - Hand appearance in the start frame for the dominant hand of signer 1

De(1) - Hand appearance in the end frame for the dominant hand of signer 1

Ds(2) - Hand appearance in the start frame for the dominant hand of signer 2

Ds(2) - Hand appearance in the end frame for the dominant hand of signer 2

NDs(1) - Hand appearance in the start frame for the non-dominant hand of signer 1

NDe(1) - Hand appearance in the end frame for the non-dominant hand of signer 1

NDs(2) - Hand appearance in the start frame for the non-dominant hand of signer 2

NDe(2) - Hand appearance in the end frame for the non-dominant hand of signer 2

Total distance computed for the hand shapes is combined along with the DTW score and fine tuned with a weight W to improve the overall accuracy. Before the distance for the hand shape appearance is computed [2] implements a skin detection on the bounding box of the hand shapes. skin detection can be useful in finding hands in controlled environments where the background is guaranteed not to contain any objects which are of the same color as the skin. Illumination effects also has an enormous impact on the object under test in the image. changing the direction of illumination can lead to shift in the location of shadows, change in the direction of gradient . Here we implement a system based on manual hand shapes as inputs from the user to improve the accuracy especially in the signs where there is little to no hand motions along with the DTW scores. we also implement better similarity measures other than Euclidean distance such as zeros mean normalized cross correlation(ZNCC) and Histogram of oriented gradients(HOG) for the hand appearance and investigate its impact along with the DTW scores to improve the accuracy.

Chapter 3

Contribution

The existing American sign language recognition system helps users to look up unknown signs by comparing trajectories of hand motion between the query sign submitted by the user against the database of videos which were already annotated by [3] and also combining hand shape appearance information as a similarity measure between the start and end frames of both dominant and non-dominant(if two-handed sign) hands along with the trajectory of hand motion. The similarity measure used for comparing trajectories of hand motion is Dynamic time warping(DTW)[8] which was already mentioned in the previous section which finds an optimal alignment between the two trajectories of hand motion under test and Euclidean distance as similarity measure for hand shape appearance between the query and the model sequence.

The contribution for this thesis is as follows:

- We use the existing similarity measure Dynamic time warping(DTW) for comparing trajectories of hand motion.
- Computed a conditional probability table by incorporating relation between hand shape pairs in all the model classes(database sign videos).
- Obtained a similarity score from the conditional probability table based on the hand shape information provided by the user with reference to the hand shape information for the example classes.
- Combining the similarity scores obtained from the trajectory of hand motion along with the similarity scores obtained from the conditional probability table for the hand shape information(manual method) provided by the user and obtain an global similarity score which will be used to rank the sign under test.

- Incorporated hand shape appearances(semi-automatic method) by using similarity measures such as Zero mean normalized cross correlation(ZNCC) and Histogram of oriented gradients(HOG) along with DTW to obtain a global similarity score which will be used to rank the sign under test.

Additional contribution:

- Creating a GUI to visualize hand shape information obtained by annotating hand shape for the start and end frames of each sign for all the database videos.
- Creating a real time demo by integrating manual hand shape method based on similarity scores obtained from the conditional probability along with the existing American sign language demo based on DTW.

Chapter 4

System Overview

The system implemented assists the users to retrieve unknown signs from the ASL. The user can perform a sign in front of a webcam recording it at the same time and submit it as a query to the system for which the system retrieves the signs which are best matched to the query from the system database. We evaluate two methods for the described system here. System based on manual hand shapes and a system based on hand appearances.

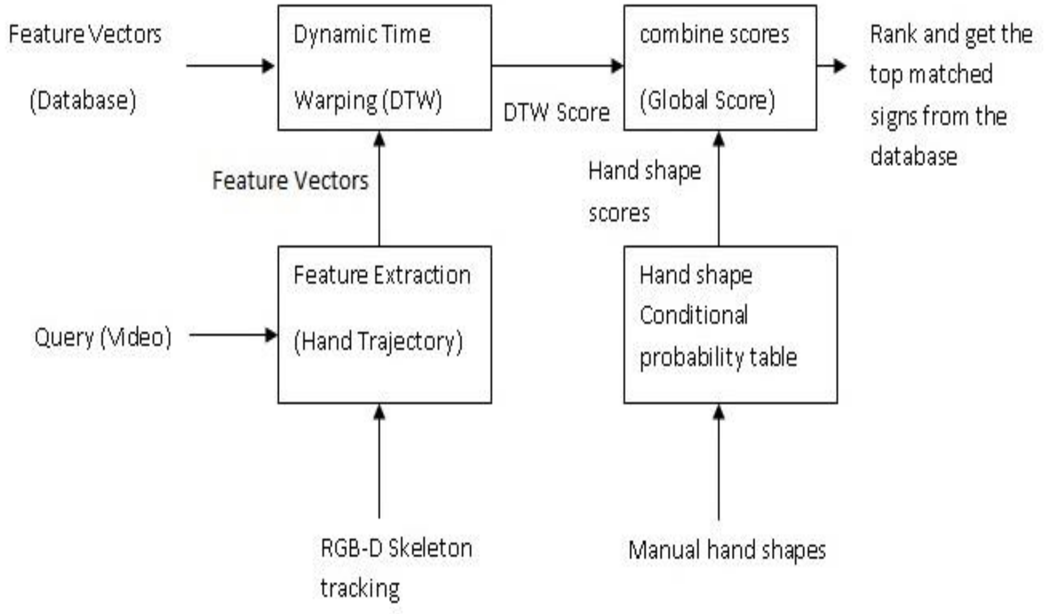


Figure 4-1 American sign language recognition flowchart based on DTW and manual hand shapes

The American sign language recognition system based on manual hand shapes as mentioned in Figure 4-1 consists of the following modules. A feature extraction module which obtains feature vectors for the hand motion with the help of RGB-D skeleton tracking of the kinect for the given query video. The feature vectors extracted from the query video for the dominant and non-dominant hands (if two-handed) are matched with the feature vectors for signs in the database videos (which are computed offline and stored) using dynamic time warping(DTW) by establishing an optimal warping path to obtain the DTW score which is based on hand motion. The user inputs manual hand shapes to the best of his knowledge for the query sign submitted through which an hand shape score is computed based on the hand shape conditional probability table which is combined with the DTW score to obtain a global score. Based on the global score the system ranks and retrieves the top matched signs from the database videos for the user to browse through and find his sign of interest.

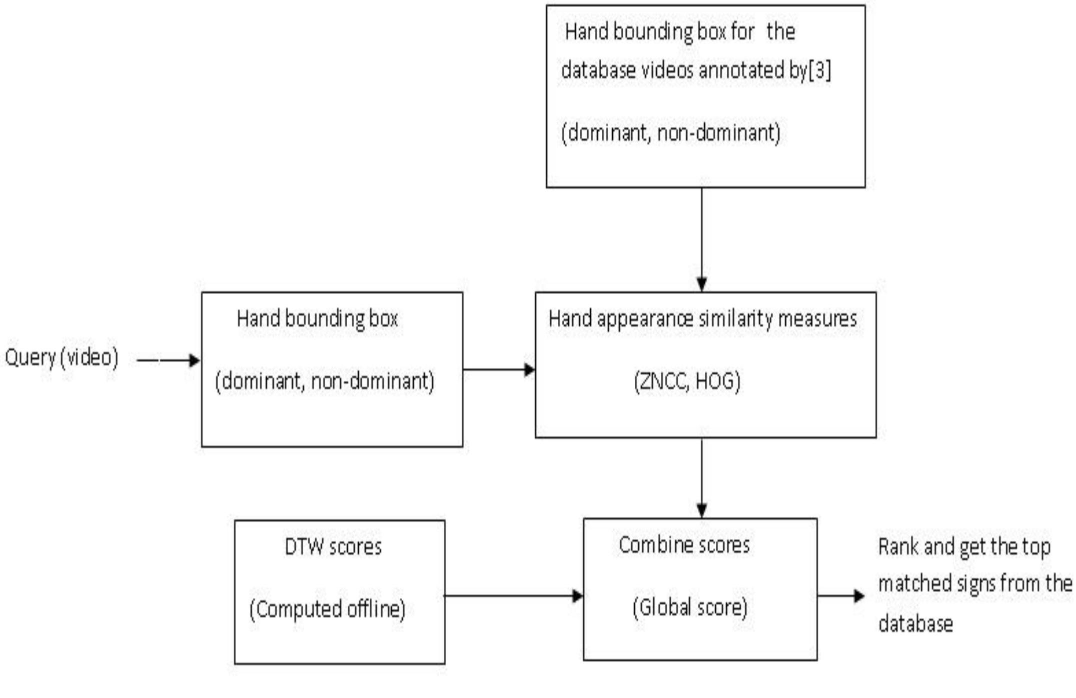


Figure 4-2 American sign language recognition flowchart based on DTW and hand appearances

For the American sign language recognition system based on hand appearances the hand bounding box for the start and end frames for both the dominant and non-dominant hands(if two-handed) are submitted along with the query video. Popular template matching methods such as Zero mean normalized cross correlation(ZNCC) and histogram of oriented gradient based template matching (HOG) are used to compute the similarity measures between the hand appearance of the query and the database videos(bounding box annotated by [3]). The similarity score obtained for the hand appearances is sum of combination of both the dominant and non-dominant(if two-handed) hand appearances in the start and end frames. The DTW score(computed offline for the ease of experimentation) and the hand appearance similarity scores are combined to obtain a global score based on which the system ranks and retrieves the top matched signs from the database videos for the user to browse through and find his sign of interest.

The accuracy of the system is calculated based on the rank assigned to the retrieved sign from the database of videos. Say if the assigned rank is 10 then the user has to browse through 10 signs before he encounters his sign of interest.

Chapter 5

Feature Extraction and DTW

The system proposed here improves upon the existing system for American sign language recognition which is based on Dynamic time warping(DTW)[8]. The cost obtained through DTW is an optimal alignment between the trajectories under test. DTW or any another method which provides trajectory information is still a vital part of the system proposed here and we need to improve upon the existing methods based on trajectory with hand shape information.

5.1 Feature Extraction

Hand detection is not a part of this system as we have the bounding box of the hand annotated by [3]. For the feature extraction centroid's of hand locations needs to be identified and tracked. For the manual method we use RGB-D skeleton tracker of kinect[5] to obtain the centroid's of both the dominant and non-dominant hands .For the semi-automatic method we use the hand bounding box annotated by[3] for ease of experimentation.

The DTW[8] matches two time series sequences a model M and a query Q and computes a cost which is used for comparing the similarity between the query and the model sequences and thus classifying the query sequence with respect to the best matched models.

Let X be the model sign sequence represented as X_1, X2,…..$X_{|x|}$ where |X| is the number of frames and X(t) be the t^{th} frame from the model sign sequence X. From the model sign sequence X we extract the following features:

- F_d(X,t) and F_{nd}(X,t) : The centroid (x,y) of the dominant(right hand) and non-dominant(left-hand) of the sign for the t^{th} frame.

- $F_\Omega(X,t)$: The relative position between dominant and non-dominant hand of the sign for the t^{th} frame,($F_\Omega(X,t) = F_d(X,t) - F_{nd}(X,t)$).
- $I_d(X,t)$ and $I_{nd}(X,t)$: The direction of motion represented by the unit vectors for the dominant and non-dominant hand respectively from F_d (X,t-1) frame to F_d (X,t+1) and F_{nd} (X,t-1) frame to F_{nd} (X,t+1)
- $I_\Omega(X,t)$: The direction of motion represented by the unit vectors between the dominant and non-dominant hands respectively from $F_\Omega(X,t-1)$ frame to $F_\Omega(X,t+1)$ frame.

This constitutes a 1×12 feature vector represented as:

[$F_d(X,t)$, $F_{nd}(X,t)$, $F_\Omega(X,t)$, $I_d(X,t)$, $I_{nd}(X,t)$, $I_\Omega(X,t)$] (Each point stores both x and y coordinates). In order to establish a reference coordinate system which account for variations in spatial scale and translation we use a face as reference coordinate[2]. Here the position of the face is annotated beforehand for the models and stored as bounding boxes . For the query we either have to provide the bounding box for the face position manually or use some face detector to detect the faces. The face size is the diagonal of the bounding box and is used to normalize both x and y positions. Thus the feature vectors $F_d(X,t)$, $F_{nd}(X,t)$, $F_\Omega(X,t)$ are all normalized with respect to face centric coordinate system[2].

The signers performing the sign in both the query and models may sign at different speeds. Since we use the trajectory information for the DTW, the DTW is biased more towards the signs which are performed for longer duration than short duration. Hence in order to account for this we normalize the duration of the sequences such that it is even for all the signs. Here normalization length of 25 is used and bicubic interpolation was used for normalizing the sequences.

5.2 Comparing trajectories with DTW

we need to compare the trajectories of both the query Q and the model X with the feature vectors obtained. Let X(t) be given by:

$$X(t) = [F_d(X,t), F_{nd}(X,t), F_{\Omega}(X,t), I_d(X,t), I_{nd}(X,t), I_{\Omega}(X,t)] \qquad (5.1)$$

where X(t) is obtained by combining the feature vectors for the t^th frame. Let |X| be total number of frames in X and |Q| be total number of frames in Q. For the given signs Q and X we need to compute the warping path W which establishes the alignment between Q and X. Let W be given by:

$$W = ((m_1,n_1),\ldots\ldots\ldots\ldots(m_{|W|},n_{|W|})) \qquad (5.2)$$

where |W| is the length of the warping path. (m_i,n_i) provides the relation between i^th frame of Q and X. The warping path W must follow the below 3 constraints adopted from [10]:

- **Boundary condition**: $m_1 = 1$, $n_1 = 1$ and $m_{|W|} = |Q|$ and $n_{|W|} = |X|$. The first element of m and n should match and the last element of m and n should match.
- **Monotonicity**: $m_{i+1} - m_i >= 0$ and $n_{i+1} - n_i >= 0$. The alignment between m and n cannot cross or cannot go backwards.
- **Continuity**: $m_{i+1} - m_i <= 1$ and $n_{i+1} - n_i <= 1$.The alignment between m and n cannot skip elements.

If the sign is one handed the hand features $F_{nd}(X,t)$, $F_{\Omega}(X,t)$, $I_{nd}(X,t)$, $I_{\Omega}(X,t)$] are set to 0.The cost C(W,Q,X) is the sum of individual local costs $C(Qm_i,Xn_i)$ which corresponds to the matching between each Qm_i and Xn_i.

$$C(W,Q,X) = \sum_{i=1}^{|W|} C(Qmi, Xni) \qquad (5.3)$$

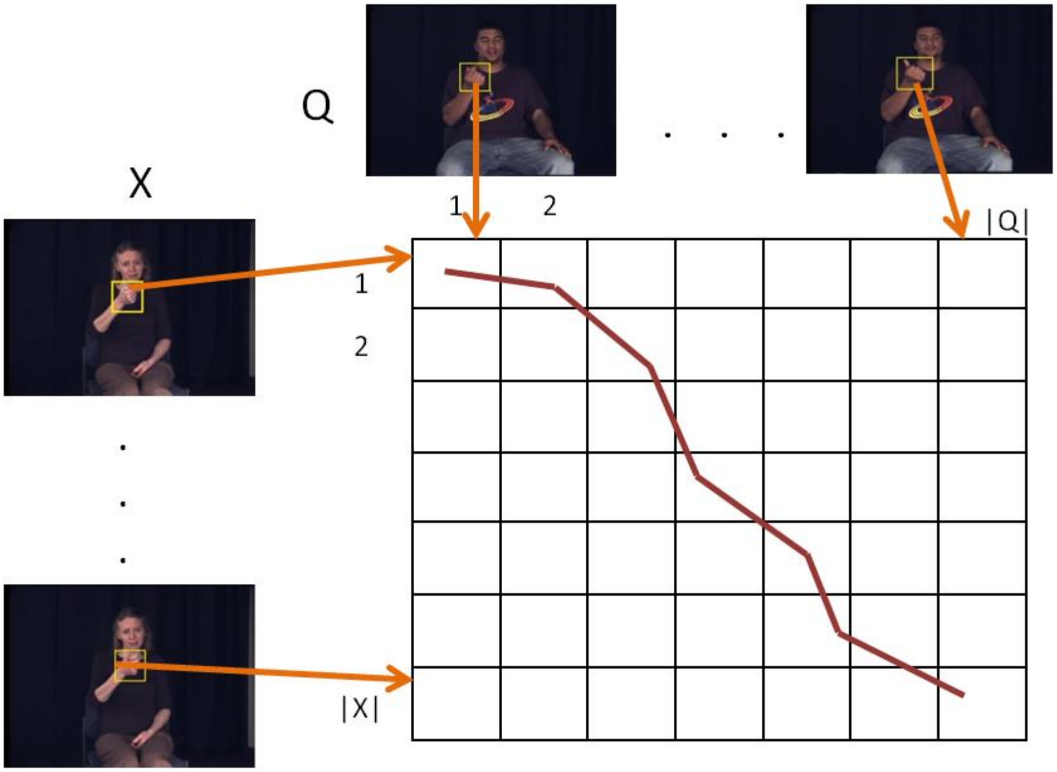


Figure 5-1 DTW warping path between Model X_i and Query Q_j

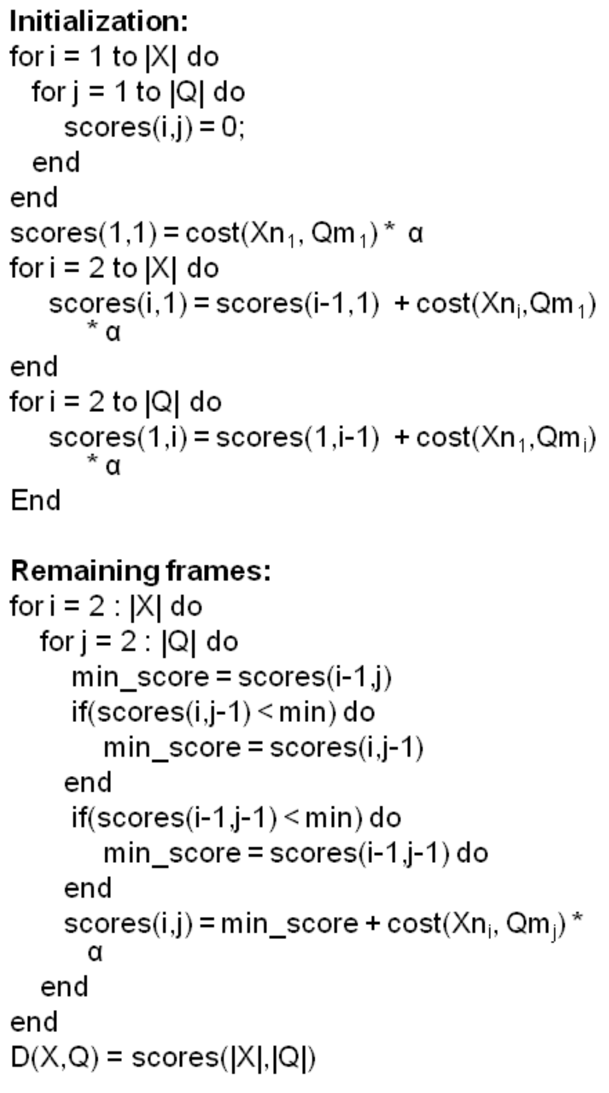
The local $C(Qm_i,Xn_i)$ is calculated as weighted linear combination of Euclidean distance of 12 feature vectors given as:

$$\begin{aligned} C(Qm_i,Xn_i) = {} & \alpha 1\|F_d(Q, mi) - F_d(X, ni)\| + \alpha 2\|F_{nd}(Q, mi) - F_{nd}(X, ni)\| + \\ & \alpha 3\| F_\Omega (Q, mi) - F_\Omega (X, ni)\| + \alpha 4\| I_d (Q, mi) - I_d (X, ni)\| + \\ & \alpha 5\| I_{nd} (Q, mi) - I_{nd} (X, ni)\| + \alpha 6\| I_\Omega (Q, mi) - I_\Omega (X, ni)\| \end{aligned} \qquad (5.4)$$

The weights α_i are chosen experimentally by testing with the training set which improves the detection accuracy. The DTW cost D(Q,X) is the minimum cost warping path.

$$D(Q,X) = min \sum_{i=1}^{|W|} C(Qmi, Xni) \qquad (5.5)$$

Dynamic programming was employed to calculate the optimal warping path and DTW cost D(Q,X) with a time complexity of O(|Q||X|). Let X be the model sequence represented by { Xn_1, Xn_2,…………$Xn_{|X|}$} and Q be the query sequence represented by { Qm_1, Qm_2,…………$Qm_{|Q|}$}.



```
Initialization:
for i = 1 to |X| do
  for j = 1 to |Q| do
     scores(i,j) = 0;
  end
end
scores(1,1) = cost($Xn_1$, $Qm_1$) *  $\alpha$
for i = 2 to |X| do
    scores(i,1) = scores(i-1,1)  + cost($Xn_i$,$Qm_1$) * $\alpha$
end
for i = 2 to |Q| do
    scores(1,i) = scores(1,i-1)  + cost($Xn_1$,$Qm_i$) * $\alpha$
End

Remaining frames:
for i = 2 : |X| do
   for j = 2 : |Q| do
      min_score = scores(i-1,j)
      if(scores(i,j-1) < min) do
         min_score = scores(i,j-1)
      end
      if(scores(i-1,j-1) < min) do
         min_score = scores(i-1,j-1) do
      end
      scores(i,j) = min_score + cost($Xn_i$, $Qm_j$) * $\alpha$
   end
end
D(X,Q) = scores(|X|,|Q|)
```

Figure 5-2 DTW Algorithm [2]

Chapter 6

Manual Handshapes with DTW

As we had mentioned the hand shape inputs from the users will significantly improve the accuracy especially in the signs where there is little to no hand movements along with the DTW scores. We need to incorporate the relation between hand shape pairs in all the example classes for each individual sign to generate a conditional probability table from which a handshape similarity score will be computed at runtime for the given handshape input from the user and combine with the DTW scores to improve the system.

Handshape is one of the essential component that makes up an American sign language sign. They can represent the characteristic of the object or how it is handled. If we change the handshape of a sign we change the meaning of the sign. The two main structures that define a handshape are the joints and the fingers. The joint structure basically signifies the disposition of the joints (closed,open,flex,spread) and the selected finger signifies what other fingers has been fore grounded.

If we are right-handed we say our dominant hand is right and if we are left-handed we say our dominant hand is left. The system worked upon here is a right-handed dominant system and all the analysis and experimentation is carried out with this assumption. For one-handed signs we use only the dominant hand and for two-handed signs we use both the dominant and the non-dominant hands.

The handshape information in the start and end frames of the sign performed along with the hand motion can identify a sign. Thus two hand shapes constitutes a one-handed sign and four hand shapes constitutes a two-handed sign. The relation among handshape pairs between different examples is a viable source of cue to improve the accuracy of the existing ASL system.

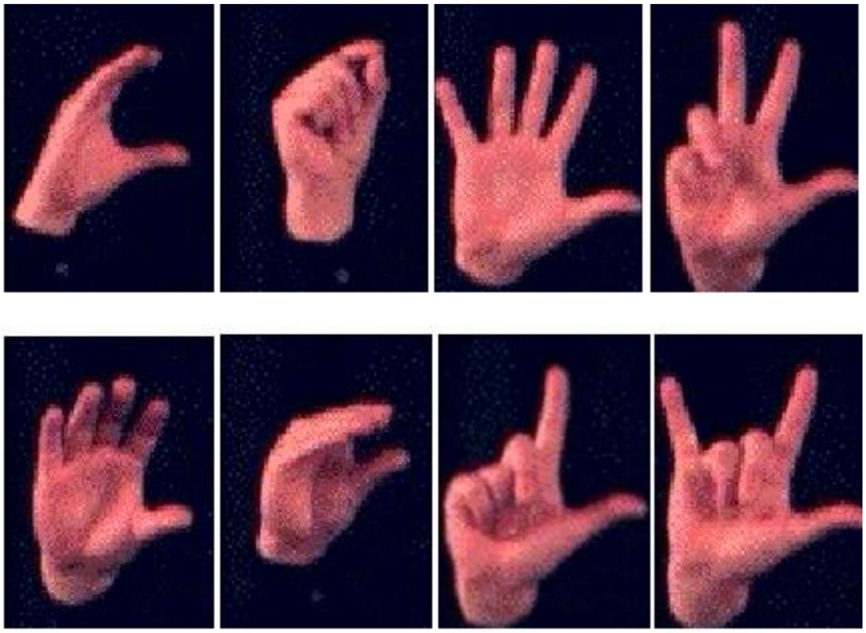


Figure 6-1 Hand shapes representing different identities for each[Row(1) - (C, baby-O,5,3) Row(2) - (crvd-sprd-B, tight-C,L, I-L-Y)]

Let W denote the set containing handshape pairs which is created from the example classes. Each sign class has handshape info for the start frame and the end frame for both the dominant and non-dominant hands respectively. For the system implemented we have considered 3 examples per sign class. There are $^3P_2$ combinations of handshape pairs possible for each start frame, end frame for both the dominant and non-dominant hands. For a one-handed sign totally $^3P_2 \times 2 = 12$ combinations of handshape pairs are possible where 2 - denotes number of hand shapes for a one-handed sign. For a two-handed sign totally $^3P_2 \times 4 = 24$ combinations of handshape pairs are possible where 4 - denotes number of hand shapes for a two-handed sign. Let M_1,M_2, and M_3 denote the 3 examples or 3 different signers.

- $M_{1,i}(DS)$ denote the handshape in dominant start frame for ith sign of signer 1.
- $M_{1,i}$ (DE) denote the handshape in dominant end frame for ith sign of signer 1.
- $M_{1,i}$ (NDS) denote the handshape in non-dominant start frame for ith sign of signer 1.
- $M_{1,i}$ (NDE) denote the handshape in non-dominant end frame for ith sign of signer 1.

Similarly we have the start frames and end frames for both dominant and non-dominant hand shapes for signers 2 and 3. |N| denotes number of signs in each example M. Thus W will have |N| × n × $^3P_2$ handshape pairs(n can be 2 or 4 - one handed/two-handed).

$$\begin{aligned} W_i \;=\; [&(M_{1,i}\,(DS),\, M_{2,i}(DS)),\ ((M_{1,i}\,(DS),\, M_{3,i}(DS)),\ ((M_{2,i}\,(DS),\, M_{3,i}\,(DS)),\\ &(M_{2,i}\,(DS),\, M_{1,i}\,(DS)),\ ((M_{3,i}\,(DS),\, M_{1,i}\,(DS)),\ ((M_{3,i}\,(DS),\, M_{2,i}\,(DS)),\\ &(M_{1,i}\,(DE),\, M_{2,i}\,(DE)),\ ((M_{1,i}\,(DE),\, M_{3,i}\,(DE)),\ ((M_{2,i}\,(DE),\, M_{3,i}\,(DE)),\\ &(M_{2,i}\,(DE),\, M_{1,i}\,(DE)),\ ((M_{3,i}\,(DE),\, M_{1,i}\,(DE)),\ ((M_{3,i}\,(DE),\, M_{2,i}\,(DE)),\\ &(M_{1,i}\,(NDS),\, M_{2,i}\,(NDS)),\ ((M_{1,i}\,(NDS),\, M_{3,i}\,(NDS)),\ ((M_{2,i}\,(NDS),\, M_{3,i}\,(NDS)),\\ &(M_{2,i}\,(NDS),\, M_{1,i}\,(NDS)),\ ((M_{3,i}\,(NDS),\, M_{1,i}\,(NDS)),\ ((M_{3,i}\,(NDS),\, M_{2,i}\,(NDS)),\\ &(M_{1,i}\,(NDE),\, M_{2,i}\,(NDE)),\ ((M_{1,i}\,(NDE),\, M_{3,i}\,(NDE)),\ ((M_{2,i}\,(NDE),\, M_{3,i}\,(NDE)),\\ &(M_{2,i}\,(NDE),\, M_{1,i}\,(NDE)),\ ((M_{3,i}\,(NDE),\, M_{1,i}\,(NDE)),\ ((M_{3,i}\,(NDE),\, M_{2,i}\,(NDE))] \end{aligned} \qquad (\,6.1\,)$$

where i varies from 1 to |N| and W = union{ W_i } where W is a set of all W_i. Let T denote the set of all handshape templates and |T| is the number of handshape templates. Totally 87 handshape templates were considered for the current system(|T| = 87) . Let $A_{1,}$ A_2,…….$A_{|T|}$ denote handshape 1 in and $B_{1,}$ B_2,…….$B_{|T|}$ denote handshape 2 in the handshape pairs W. Let $[A_i,B_j]_n$ denote the number of occurrences of the pair A_i,B_j in W. Thus we can create a 2D frequency table $F(A_i,B_j)$ as follows:

$$F(A_i,B_j) = W\left([A_i,B_j]\ n\right) \qquad (\,6.3\,)$$

where i varies from 1 to |T| and j also varies from 1 to |T|.

The frequency table F(A,B) are the joint occurrences of the handshapes A and B. Now we need to create a conditional probability table P(A|B) from the joint probabilities.

$$P(A|B) \;=\; \frac{F(A_i,B_j)}{P(B_j)} \qquad (6.4)$$

$$\text{where } \; P(B_j) \;=\; \sum\nolimits_{i=0}^{|T|} F(A_i, B_j) \qquad (6.5)$$

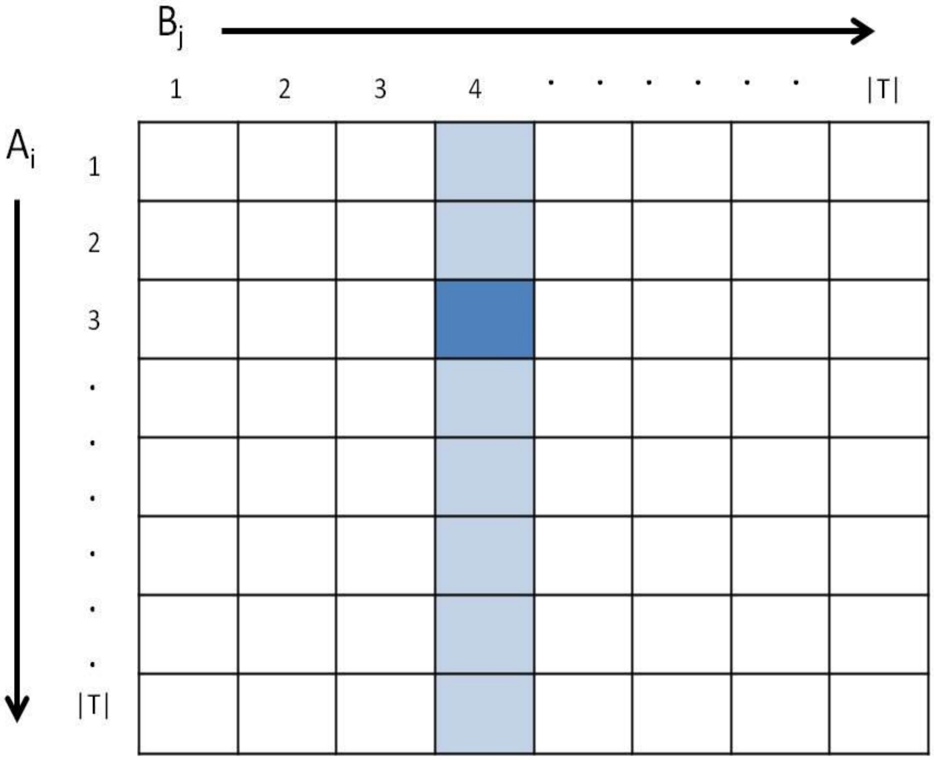
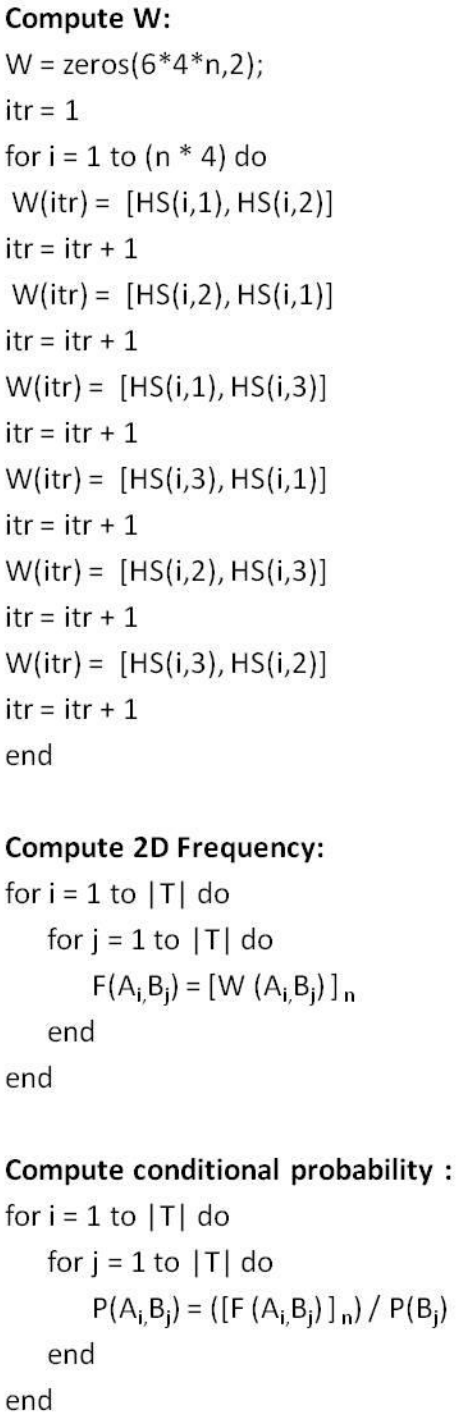


Figure 6-2 Example showing conditional probability calculation for handshape pair (3,4) from the frequency occurrence table

Let n = |N| number of signs in a model class and HS denotes an array containing all the handshapes of the example classes M1 M2 and M3 and is of dimension (4 × n × 3).



```
Compute W:
W = zeros(6*4*n,2);
itr = 1
for i = 1 to (n * 4) do
 W(itr) =  [HS(i,1), HS(i,2)]
itr = itr + 1
 W(itr) =  [HS(i,2), HS(i,1)]
itr = itr + 1
W(itr) =  [HS(i,1), HS(i,3)]
itr = itr + 1
W(itr) =  [HS(i,3), HS(i,1)]
itr = itr + 1
W(itr) =  [HS(i,2), HS(i,3)]
itr = itr + 1
W(itr) =  [HS(i,3), HS(i,2)]
itr = itr + 1
end

Compute 2D Frequency:
for i = 1 to |T| do
    for j = 1 to |T| do
        $F(A_{i,}B_j) = [W\ (A_{i,}B_j)]_n$
    end
end

Compute conditional probability :
for i = 1 to |T| do
    for j = 1 to |T| do
        $P(A_{i,}B_j) = ([F\ (A_{i,}B_j)]_n) / P(B_j)$
    end
end
```

Figure 6-3 Algorithm to calculate conditional probability table for handshape pairs

Let S_{HA_Man} be the Manual hand shape score obtained by the product of individual conditional probability scores *P(A|B)* between the query and the database handshapes. Let n be the number of handshapes under comparison for calculating S_{HA_Man} and is 2 (start and end-frames for dominant hand) for a one-handed sign, 4 (start and end-frames for both dominant hand and non-dominant hand) for a two-handed sign. Let A be the handshape of query sign under test and B be the handshapes of sign in the database.

$$S_{HA\_Man} = \prod_{i=1}^{n} P(A_i\,, B_j) \qquad (6.6)$$

The values for S_{HA_Man} range from 0 to 1 as they are just the product of individual conditional probabilities. DTW scores are calculated offline before combining with the Manual hand shape scores to obtain a global score for the ease of experimentation. Let S_{DTW} be the DTW score and S_{HA_Man} be the Manual hand shape score for a particular sign and we need to combine S_{DTW} and S_{HA_Man} to obtain $S_{DTW\_HA\_Man}$ global score.

As we had already mentioned n is the number of handshapes and will vary accordingly with respect to one-handed or two-handed sign. we know that the DTW scores are measure of distance between two time series under test. The lower the DTW score better the similarity and higher the DTW score lesser the similarity. On the contrary for the hand shape scores the higher the handshape scores better the similarity and lower the handshape score lower the similarity. The global score $S_{DTW\_HA\_Man}$ calculated is still in reference to DTW score S_{DTW} and follows the same significance as the DTW score.

Hence we just need to invert the probability of Manual handshape scores S_{HA_Man} and take the product along with the DTW scores to obtain the global similarity score $S_{DTW\_HA\_Man.}$

$$S_{DTW\_HA\_Man} = S_{DTW} * (1 - S_{HA\_Man}) \qquad (6.7)$$

6.1 Hand Shape Templates

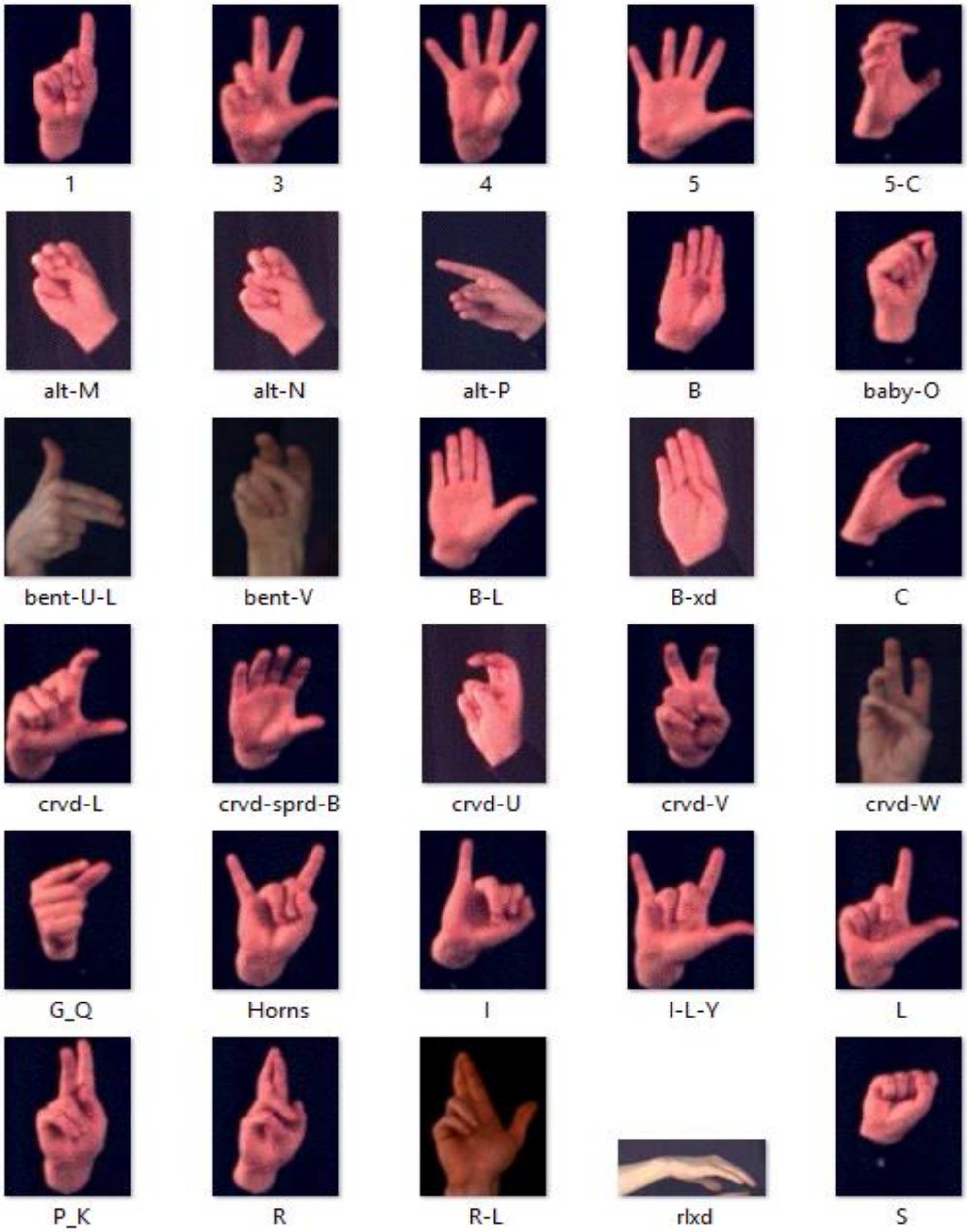


Figure 6-4 Handshape Templates(1-30)

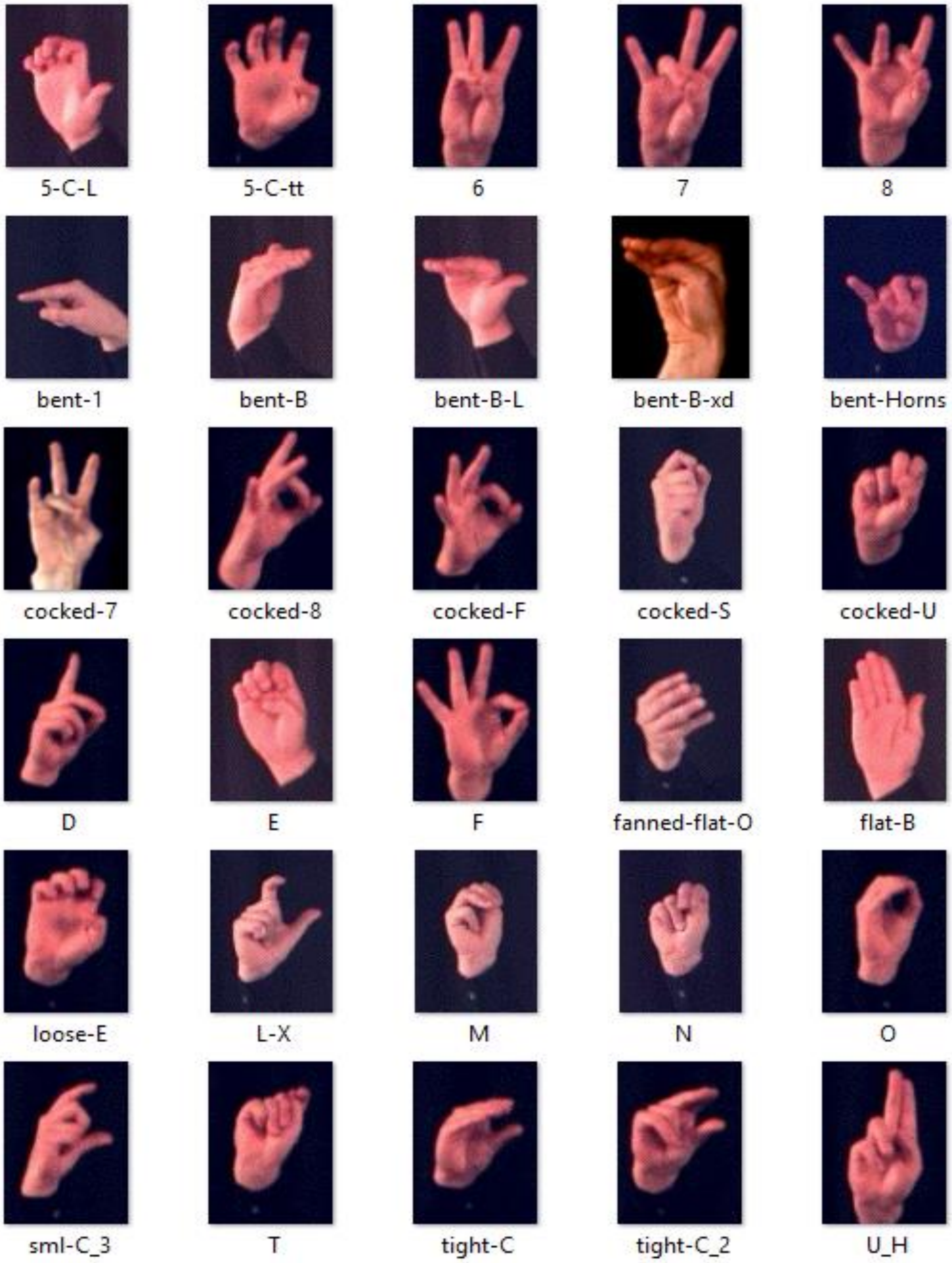


Figure 6-5 Handshape Templates(31-60)

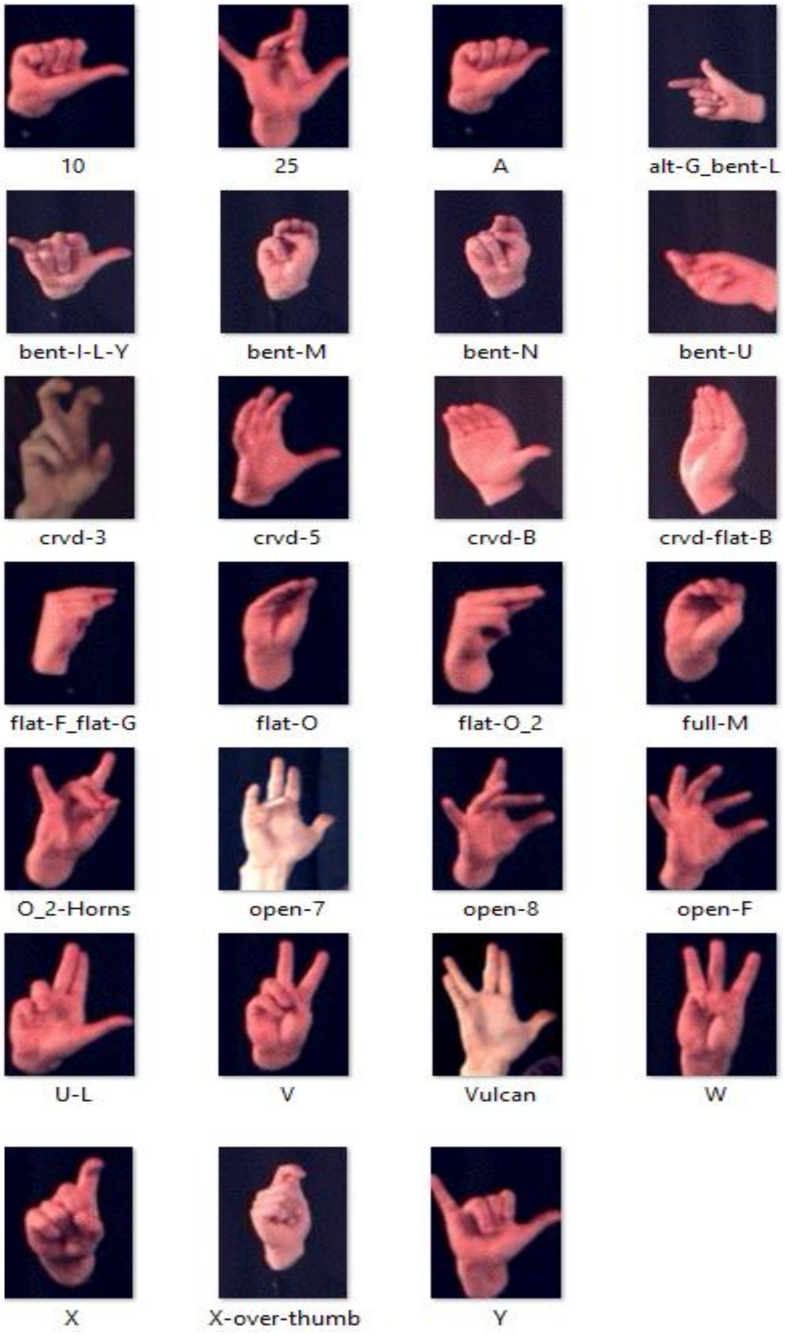


Figure 6-6 Handshape Templates(61-87)

Chapter 7

Hand Appearances

7.1 Zero mean normalized cross correlation(ZNCC) with DTW

we had already mentioned that hand appearance alone cannot be used to improve the accuracy of the recognition. But it does give valuable cues which can be combined with DTW to improve the similarity measure. Here we choose Zero mean normalized cross correlation a popular template matching method in computer vision as a similarity measure for evaluating hand shape appearances.

The correlation between a template and a matching region might be less than the correlation between a template with a brighter spot. Correlation between two images might fail if the image energy is varying. The camera sensor will have different response characteristics for the same image at different instances and also the illumination cannot assumed to be constant throughout. Zero mean normalized cross correlation overcome these shortcomings by normalizing the pixels in the image by subtracting the mean and dividing by the standard deviation thus achieving intensity normalization[11].

The hand regions are extracted offline by hand bounding box information obtained from the dataset annotated by[3] and this is performed for both the query and the database videos for ease of experimentation. we only need the bounding box for start and end frames of both the dominant and non-dominant hand regions for calculating similarity scores between the query image and the database videos. The bounding box info annotated by[3] was for uncompressed dataset with image resolution 480 * 640 but the system implemented here uses compressed dataset whose resolution is lesser than the uncompressed videos. Hence we rescale the bounding box info for hand regions according to the compressed scale before extracting the hand regions from the database videos. The extracted images for hand regions are converted to gray scale images before proceeding with the template matching. For a two-handed sign we extract 4 hand regions from the start and end frames of both the query and the database videos and for a one-handed sign we only extract 2 hand regions from the start and end frames of both the query and the database videos. The bounding box values slightly vary for every hand region extracted and we need to maintain a constant image size while matching the templates and hence we resize all the extracted hand images to 50 * 50 resolution. The Zero mean normalized cross correlation(ZNCC) for two images img1 and img2 of resolution p * q is given as:

$$ZNCC(img1,\ img2) = \frac{\sum_{i,j}^{p,q}(img1(i,j) - \overline{img1}(i,j)) * (img2(i,j) - \overline{img2}(i,j))}{\sqrt{\sum_{i,j}^{p,q}(img1(i,j) - \overline{img1}(i,j))^2} * \sqrt{\sum_{i,j}^{p,q}(img2(i,j) - \overline{img2}(i,j))^2}} \quad (7.1)$$

where $\overline{img1}$ is the mean of img1 and $\overline{img2}$ is the mean of img2.

Let S_{HA} be the hand shape appearance score obtained by the sum of individual ZNCC scores between the query and the database handshapes. Let n be the number of handshapes under comparison for calculating S_{HA} and is 2 (start and end-frames for dominant hand) for a one-handed sign, 4 (start and end-frames for both dominant hand and non-dominant hand) for a two-handed sign. Let Q be the query sign under test and M be the sign in the database.

$$S_{HA} = \sum_{i=1}^{n} ZNCC(Qi, Mi) \quad (7.3)$$

The values for ZNCC range from -1 to 1 as we get a cosine like correlation coefficient since we normalized the image and feature vectors to unit length. -1 signifies the two images under test are least correlated and +1 signifies highest correlation or similarity between the images under test.

The hand appearance scores and DTW scores are calculated offline before combining to obtain a global score for the ease of experimentation. Let S_{DTW} be the DTW score and S_{HA} be the hand appearance score for a particular sign and we need to combine S_{DTW} and S_{HA} to obtain S_{DTW_HA} global score.

$$S_{DTW\_HA} = S_{DTW} + \alpha * (n - S_{HA}) \qquad (7.4)$$

As we had already mentioned n is the number of handshapes and will vary accordingly with respect to one-handed or two-handed sign. we know the DTW scores are measure of distance between two time series under test. The lower the DTW score better the similarity the higher the DTW score lesser the similarity. On the contrary for the hand appearance scores the higher the handshape scores better the similarity and lower the handshape score the lower the similarity. The global score S_{DTW_HA} calculated is still in reference to DTW score S_{DTW} and follows the same significance as the DTW score. For a two handed sign the maximum value that can be attained for S_{HA} is 4 and for a one handed sign the maximum value that can be attained by S_{HA} is 2. Hence we subtract S_{HA} score from n as a measure of deviation and multiply it with a factor α which is chosen experimentally to maximize the accuracy.

7.2 Histogram of Oriented gradients(HOG) with DTW

The edge or gradient information for the hand appearances are an important aspect to be used in the recognition. The above mentioned template matching using ZNCC does not offer translational invariance and acceptable level of illumination invariance. Histogram of oriented gradients(HOG) characterizes the object appearance and shape by well distributed local intensity gradients even without the actual knowledge about the edge positions[12]. The orientation of the gradients offers robustness to an acceptable degree of illumination changes and the part of histogram binning offers translational invariance[13].

The hand region extracted using the hand bounding box information obtained from the dataset annotated by[3] were decomposed into cell size of 8 * 8 blocks for our experimentation and computed the histogram of oriented gradients for each cell and contrast normalized the local responses for better illumination invariance. The HOG descriptors were calculated using VLFeat a cross platform open source collection of vision algorithms.

The bounding box values slightly vary for every hand region extracted and we need to maintain a constant image size while matching the templates and hence we resize all the extracted hand images to 50 * 50 resolution. The HOG descriptors were extracted for the hand images of resolution (50 * 50) thus leading to descriptor size of 6*6*n for a cell size of 8 * 8 where the first 2 dimensions represents the number of rows and column blocks(overlapping) and the third dimension "n" represents the feature length for that particular block. The cell size has to be chosen appropriately to capture the relevant large scale details or small scale details as there is a trade-off between the two and for our experimentation cell size of 8 * 8 was found to be optimal. We converted the HOG descriptors to natural image space by using inverse HOG code by [14].

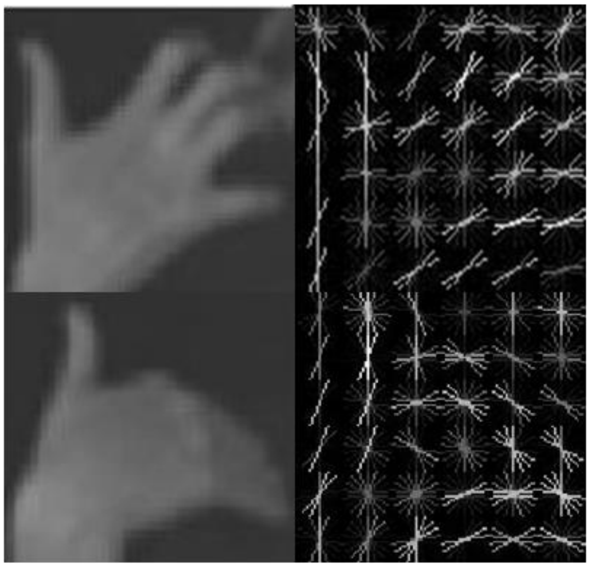


Figure 7-1 Gray scale image of extracted hand region and its HOG descriptors (cell size 8*8)

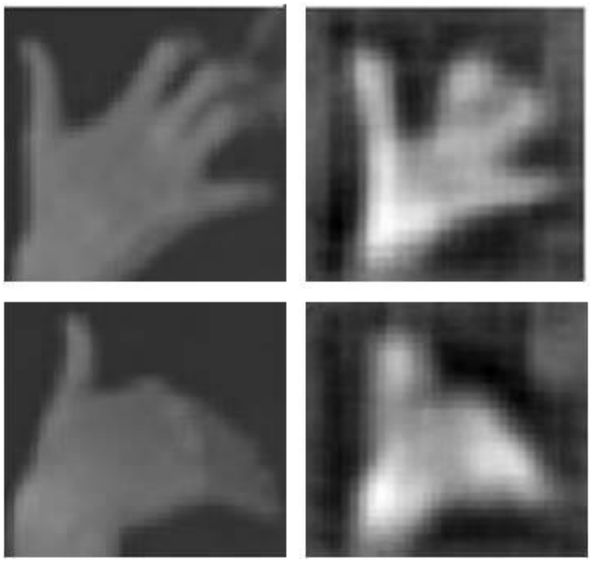


Figure 7-2 Gray scale image and its inverse HOG image as proposed by [14]

The HOG features were calculated offline by using the bounding box information for the hand regions as annotated by[3] for both the query and the database videos for the ease of experimentation. we only need the bounding box for start and end frames of both the dominant and non-dominant hand regions for calculating similarity scores between the query image and the database videos. The bounding box info annotated by[3] was for uncompressed dataset with image resolution 480 * 640 but the system implemented here uses compressed dataset whose resolution is lesser than the uncompressed videos. Hence we rescale the bounding box info for hand regions according to the compressed scale before extracting the hand regions from the database videos. The extracted images for hand regions are converted to gray scale images before proceeding with the HOG feature extraction and its conversion to inverse HOG image.

We evaluate the similarity between the two images as the Euclidean distance between the two HOG feature space of the images under test[24]. Let S_{HA} be the hand shape appearance score obtained by the sum of individual Euclidean distances of the HOG feature descriptor between the query and the database handshapes. Let n be the number of handshapes under comparison for calculating S_{HA} and is 2 (start and end-frames for dominant hand) for a one-handed sign, 4 (start and end-frames for both dominant hand and non-dominant hand) for a two-handed sign. Let Q be the query sign under test and M be the sign in the database.

$$S_{HA} = \sum_{i=1}^{n} ||HOG(Qi) - HOG(Mi)|| \qquad (7.5)$$

The hand appearance scores and DTW scores are calculated offline before combining to obtain a global score for the ease of experimentation. Let S_{DTW} be the DTW score and S_{HA} be the hand appearance score for a particular sign and we need to combine S_{DTW} and S_{HA} to obtain S_{DTW_HA} global score.

$$S_{DTW\_HA} = S_{DTW} + \alpha * S_{HA} \tag{7.6}$$

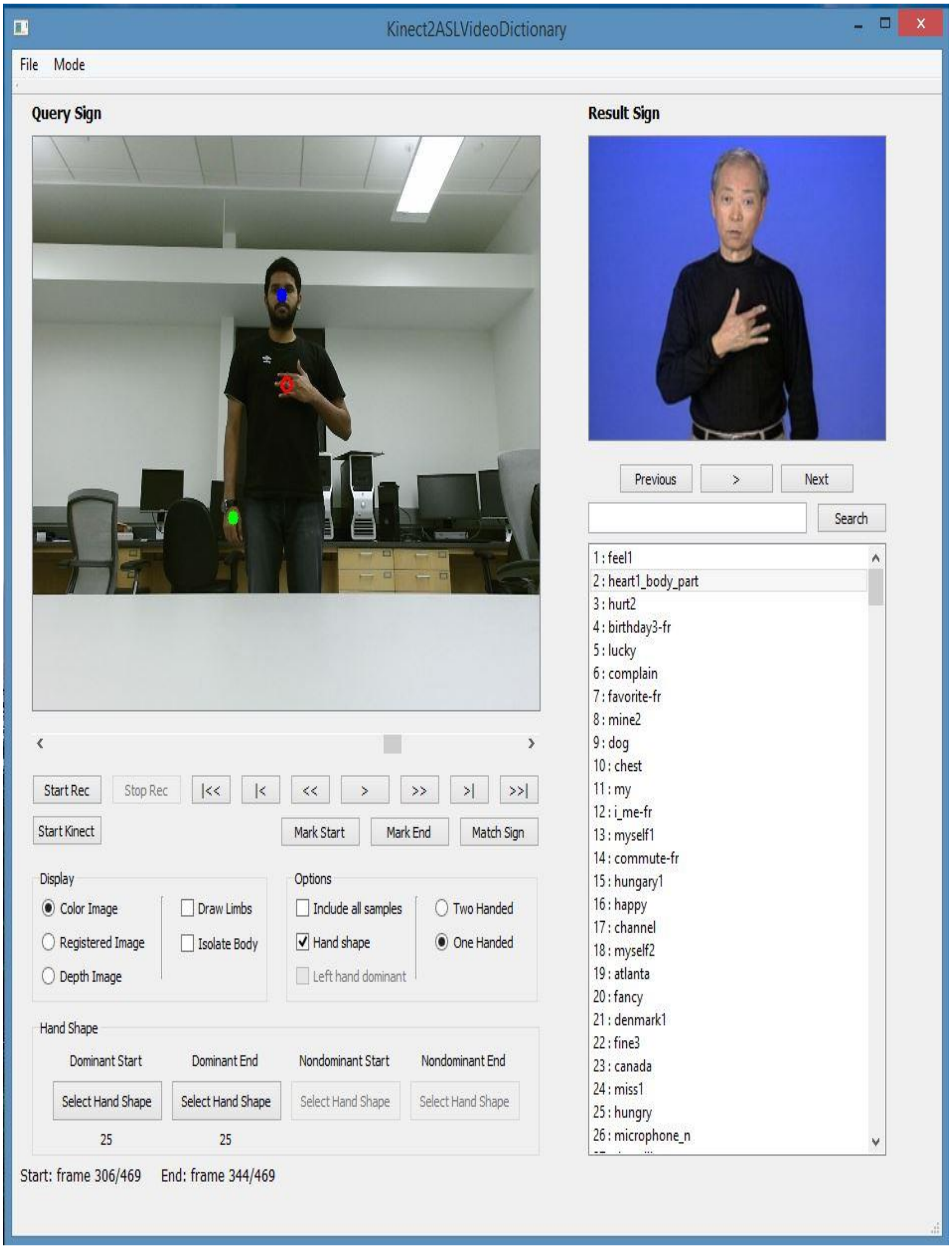
The hand shape appearance score S_{HA} is combined along with the dtw scores as a function of parameter α which is chosen experimentally to maximize the accuracy.

Chapter 8

GUI for Real Time ASL with Manual Handshapes

The GUI developed here for the real time demo with manual handshapes improves upon the existing American sign language recognition by combining the manual handshape similarity scores along with the DTW similarity scores. The GUI allows the user to record the video and specify/mark the start and end frames for the recorded video to signify the performed sign from the recorded video as a input query for the sign match. There are two displays one for displaying the recorded query sign and the other for the resulting sign obtained by match from the database videos for the submitted query. The recorded video can be displayed as a color frame, color registered depth frame or only the depth frame and has provisions for selecting any one.

By default when the GUI starts, the manual handshapes mode is disabled and the demo works only with the existing DTW which works on hand motion. The system is capable enough to detect and track the hand locations automatically in every frame without any intervention from the user and this is achieved by the RGB-D technology of the kinect camera used in [5]. The user has to make sure that the hands and faces are detected by the kinect RGB-D tracker before recording the sign. Before submitting the query sign the user has to specify whether it is a one-handed or a two-handed sign under the options menu as the submitted query will be matched only with one-handed or two-handed signs from the database videos.



Display

● Color Image
○ Registered Image
○ Depth Image

□ Draw Limbs
□ Isolate Body

Options

□ Include all samples
☑ Hand shape
□ Left hand dominant

○ Two Handed
● One Handed

Dominant Start	Dominant End	Nondominant Start	Nondominant End
25	25		

Start: frame 306/469 End: frame 344/469

1 : feel1
2 : heart1_body_part
3 : hurt2
4 : birthday3-fr
5 : lucky
6 : complain
7 : favorite-fr
8 : mine2
9 : dog
10 : chest
11 : my
12 : i_me-fr
13 : myself1
14 : commute-fr
15 : hungary1
16 : happy
17 : channel
18 : myself2
19 : atlanta
20 : fancy
21 : denmark1
22 : fine3
23 : canada
24 : miss1
25 : hungry
26 : microphone_n

Figure 8-1 GUI for real time ASL demo with Manual Handshapes

The Manual handshapes mode is enabled by selecting Hand Shape under the options menu. Once this option is selected the buttons under the Handshape menu is enabled and the user can select the handshapes for the start and end frames for both the dominant and non-dominant hands. The handshape buttons for non-dominant hands be disabled or enabled based on whether the user selects the option for one-handed or a two-handed sign. Once all the options are chosen accordingly the user can click the match sign button for proceeding the matching process with the database videos and the top matches are displayed as a resulting sign where the user can search his sign of interest and play the appropriate sign.

Chapter 9

Experiments

9.1 Dataset

For all the experiments in the existing system we use the ASLLVD dataset. The signs included in ASLLVD dataset are restricted to class of signs which are referred as lexical signs[3]. The ASLLVD dataset consists of total 1113 sign vocabulary recorded with 3 different signers[5] thus making three examples for each sign. The aim is to extend the total set to 3000 signs as found in the Gallaudet dictionary and have more signers perform the sign thus increasing the number of examples per sign as each signer performs the sign with some variations[5].

The video sequences for this dataset are captured simultaneously from four different cameras, providing a side view, two frontal views, and a view zoomed in on the face of the signer[3]. For our system we use the one of the frontal views. The videos are captured at 60 frames per second with a resolution of 480 * 640[3].The datasets are available in both lossless ad lossy formats and we use the lossy formats for our processing.

The annotations for the video sequences contains the lexicons, the start and end frames, type of sign(one-handed, two-handed), manual annotations for the location of two hands and face . For the hands the bounding boxes are marked at each frame of the sign while for the face the bounding box is marked only at the first/start frame[3].

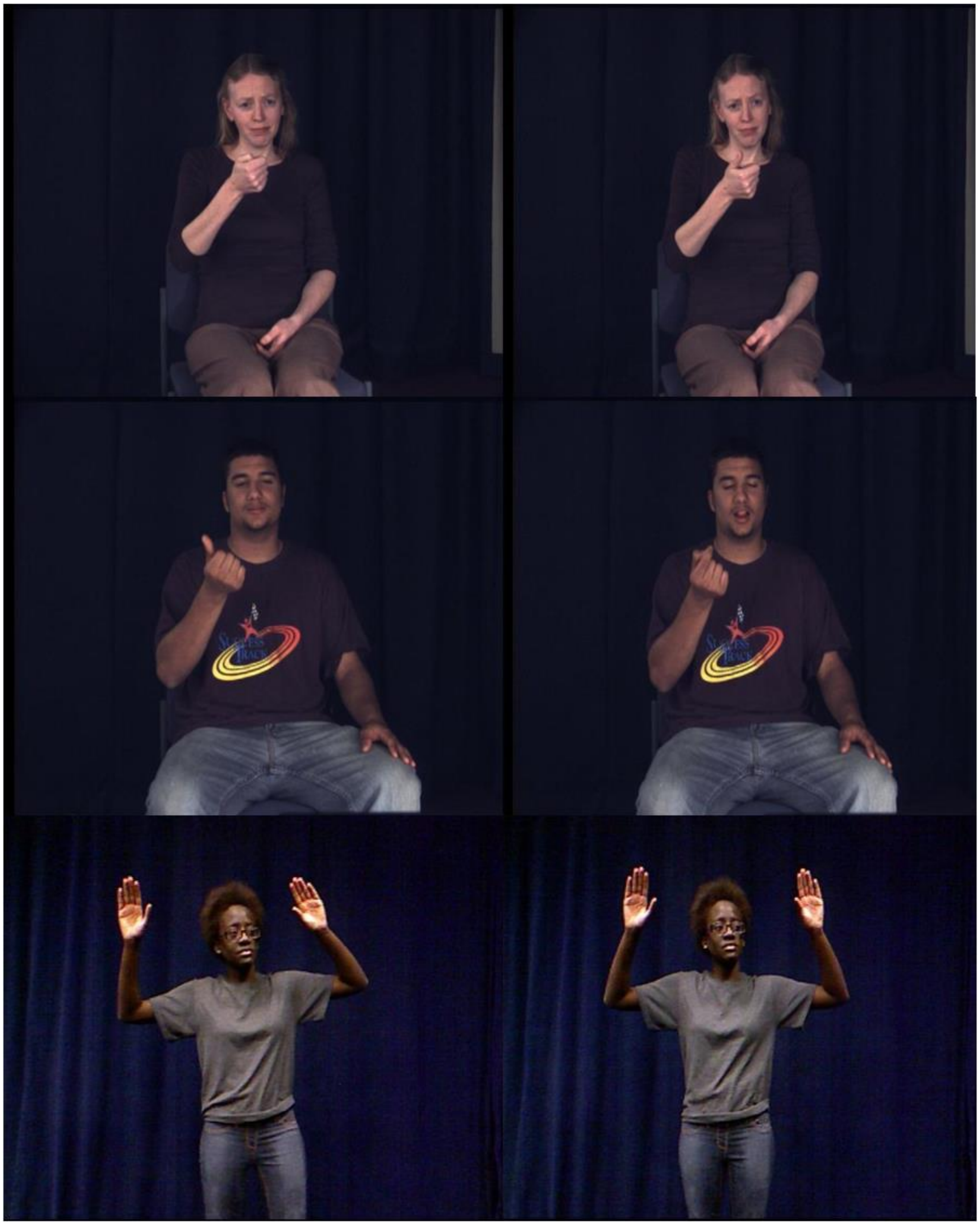
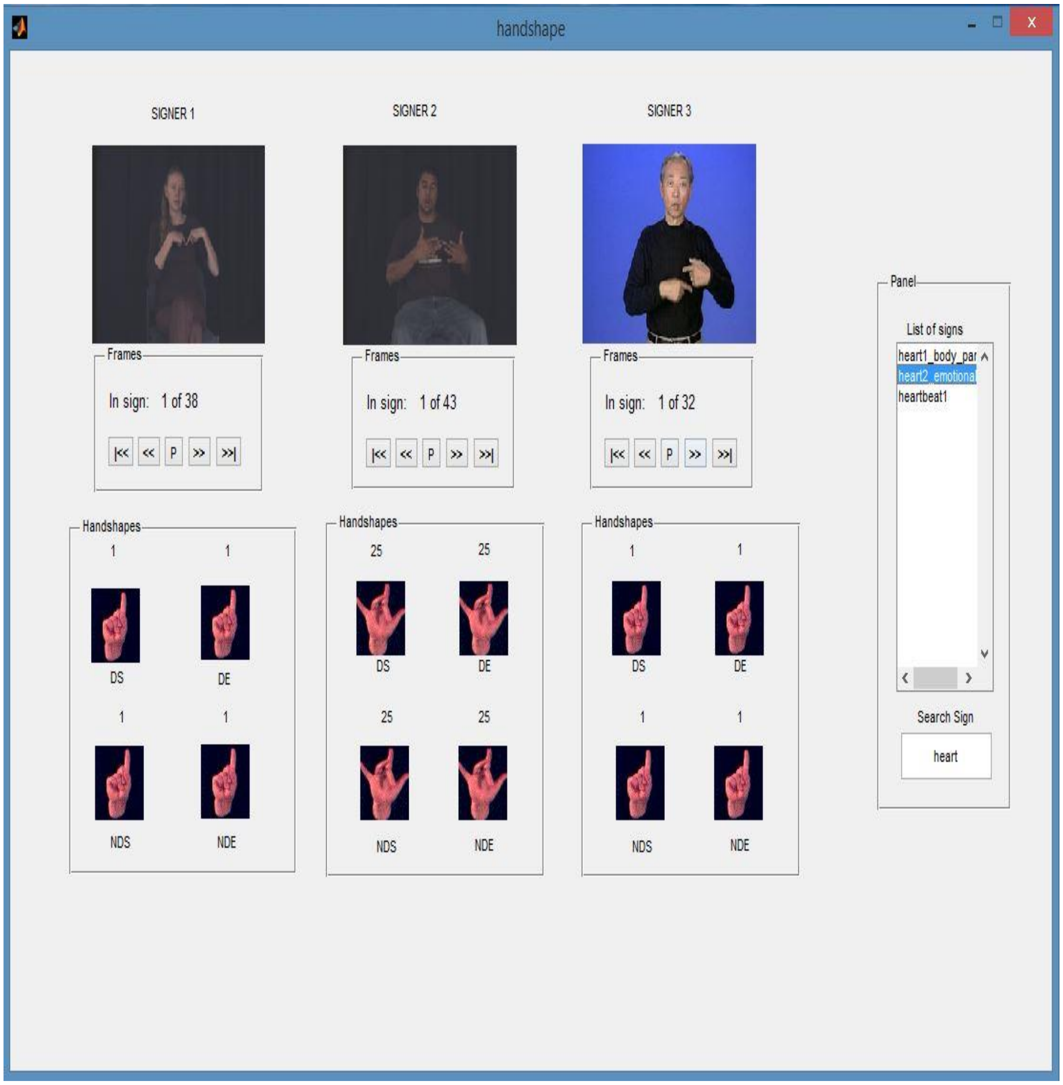


Figure 9-1 Example of a one-handed sign(start and end frames) performed by three different signers which is included in ASLLVD

9.2 GUI to visualize ASLLVD signs with manual handshapes



In sign: 1 of 38

In sign: 1 of 43

In sign: 1 of 32

heart1_body_par
heart2_emotional
heartbeat1

Search Sign: heart

Figure 9-2 GUI to search and visualize a particular sign of interest along with the handshapes that contribute to it

The GUI was developed to visualize the handshapes for a particular sign from the ASLLVD dataset. The GUI was initially developed to manually annotate the handshapes for the complete set of ASLLVD dataset from a list of handshape templates which totals to 87 by just viewing the start and end frames of a particular sign by using the hand bounding box info annotated by[3]. The annotated handshapes were integrated back to be part of the GUI where the user will be able to view and play his sign of interest as well as view the handshapes that contribute to the sign. There is a search box to search a particular sign of interest ,which will retrieve a list of all possible matches for the sign text inputted. The user can then choose his intended sign of interest from the list box and view the different signers who are signing the chosen sign on the display axes and also view the handshapes for both the dominant and non-dominant hands which occurs in the start and end frames for a particular sign .The user will be able to view three signers(three examples) for a particular sign.

9.3 Measures of Accuracy

Here we evaluate the system for sign language recognition based upon the retrieval of query sign Q from the database of signs M . M has a total of 1113 distinct sign classes and 3 examples per sign class. The system is said to have performed its job correctly if the intended query sign Q was chosen correctly from the available database of signs else the system is said to have failed[2]. The system is said to have failed in its job if the intended match for the query sign Q was not at all present in the database of signs(we don't consider this failure case as a part of our accuracy calculation) or the user could not find the intended match for the query sign Q among the best T matches from the database of signs where T is an integer and user chosen parameter value[2].

For a query sign Q we define the accuracy as a measure of rank R(Q) (same as [5]) where the rank R for the query sign Q is assigned by the system for every correct match among the database of signs. Given an integer T we define the correctness of the system as C(Q,T) by assigning a value 1 if $R(Q) <= T$ else assigning a value 0 if $R(Q) > T$[2]. So we measure the accuracy C(T) over the total query signs |Q| as average of the correctness of the system as C(Q,T) over |Q| signs.

As we have 3 examples per sign class M_{ij} where i varies from 1 to 1113 and j varies from 1 to 3 while being evaluated with the query sign Q_k where k varies from 1 to |Q|. we obtain the similarity scores S_{kji} for Q_k against M_{ij}. The similarity scores can be calculated by any means of similarity measures but care has to be taken based on how the scores represent a relation between the query sign Q and the model sign M. i.e., S > better(higher the score) or S < better(lesser the score).

Before calculating the rank $R(Q_k)$ we need to pick the best example among the 3 examples against whom the query sign Q_k is being matched. The best example is chosen such that, S_{ki} = max[S_{kji}] (along the j dimension which varies from 1 to 3) if S represents a similarity relation such the higher the score S the better it is (or) S_{ki} = min[S_{kji}] (along the j dimension which varies from 1 to 3) if S represents a similarity relation such the lower the score S the better it is. If the scores are tied we choose the median of j. S_{ki} = return{median(j)} if S_{k1i} = S_{k2i} = S_{k3i} where j varies from 1 to 3. We rearrange the scores S_{ki} either in ascending or descending order based on the similarity relation it establishes with the query sign Q_k under test. S_{ki} = ascending[S_{ki}] such that $S_{ki} < S_{k(i+1)}$ is better (or) S_{ki} = descending[S_{ki}] such that $S_{ki} > S_{k(i+1)}$ is better.

Now we start to assign the ranks $R(Q_k)$ by comparing the index positions from the calculated similarity scores S_{ki} with Q_k. $R(Q_k)$ - The rank that the system assigns to the query sign for every correct match S_{ki} such that $R(Q_k)$ = i (index at which signid(Q_k) == signid(S_{ki})) where k varies from 1 to |Q| and i varies from 1 to (total number of one-handed signs or two-handed signs). In case of a tie i.e., if the similarity scores are similar for the k^{th} sign with other signs among S_{ki} take the median rank ' i ' (index at which signid(Q_k) == signid(S_{ki})) in whichever order the similarity scores S_{ki} were arranged. The correctness of the system is given by:

$$C(Q_k,T) = 1 \text{ if } R(Q_k) <= T \qquad (7.7)$$

$$C(Q_k,T) = 0 \text{ if } R(Q_k) > T \qquad (7.8)$$

$$C(T) = mean(C(Q_k,T)) = \frac{\sum_{k=1}^{|Q|}(C(Qk,T)==1)}{|Q|} \qquad (7.9)$$

Where Q_k is a test sign from |Q| is the accuracy measure. In other words the user has to see at most best T sign matches from the database of signs for every query sign Q where Q ∊ |Q| . The user has to view T signs until the correct sign was found which includes the T^{th} sign among the database of signs. The user doesn't have to finish viewing until the T^{th} sign although he is willing to view but can stop immediately once he encounters the best sign match among the database of signs.

9.4 Results

The experiments are performed with four different signers M1,M2,M3 and M4. M1,M2 and M3 each has 1113 signs and M4 has 834 signs. M1,M2 and M3 can be used as database set as well as individual query set but M4 can be used only as a query set and not as a database set. We have performed our evaluation for five different cases:

- Query - M1 and Model - M2,M3
- Query - M2 and Model - M1,M3
- Query - M3 and Model - M1,M2
- Query - M4 and Model - M1,M2,M3
- Average of (M1,M2,M3,M4)

Each of the case has been evaluated for five different types of similarity measures and compared based on their measures of accuracy. The first similarity measure is based on DTW which works on hand motion. The second similarity measure is based on combining hand appearance using ZNCC along with DTW. The third similarity measure is based on combining hand appearance using HOG along with DTW. The fourth similarity measure is based on manual handshape inputs alone without DTW and the final similarity measure is based on combining manual handshape similarity scores with DTW.

For the first case from Figure 9-3 we see that the lowest performing similarity measure is DTW and the highest performing similarity measure is manual handshapes with DTW. The DTW performs well for longer signs as it is more biased towards it and performs poorly for signs which are static with less hand motion. By combining manual handshape similarity scores which is computed according to the user inputs with DTW the accuracy is significantly improved especially for all the static signs. The hand appearance similarity measures using ZNCC and HOG offer better recognition accuracy compared to DTW as they incorporate valuable cues from hand appearances which improves the recognition accuracy for both static signs and signs with hand motion. The HOG slightly outperforms ZNCC as they incorporate edge or gradient information for the hand appearances which ZNCC fails as the similarity purely intensity based. The same behavior of accuracies are reciprocated for the remaining cases and we plot and tabulate the accuracy results below.

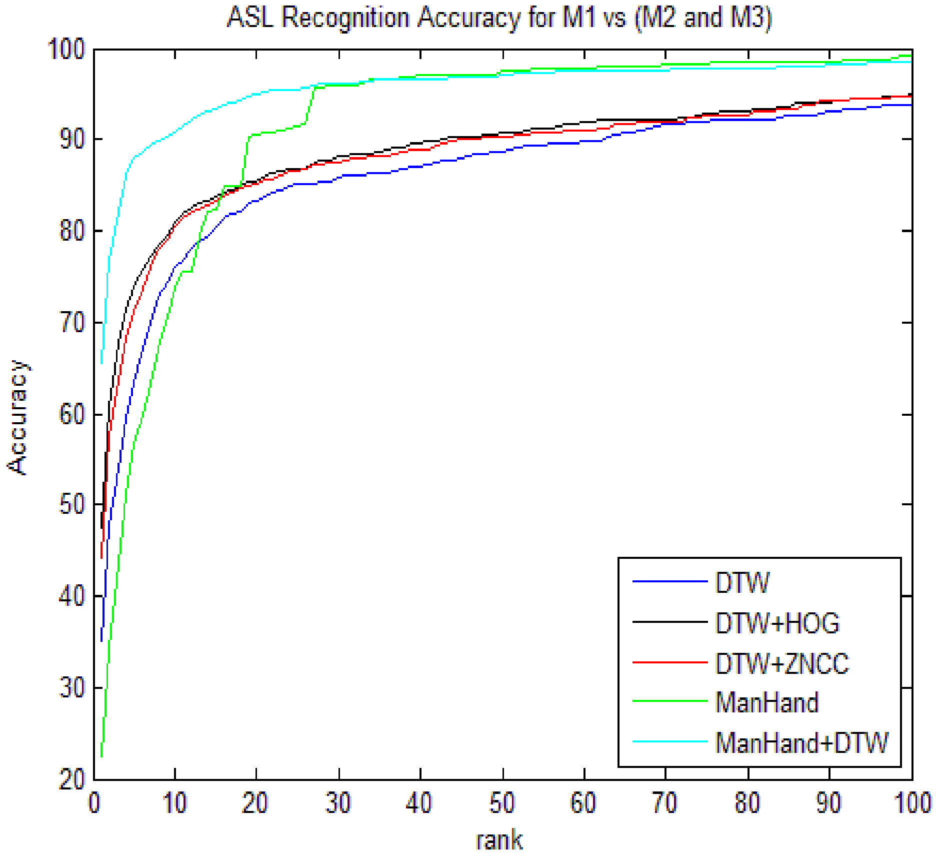


Figure 9-3 ASL Recognition Accuracy for M1 vs. (M2 and M3)

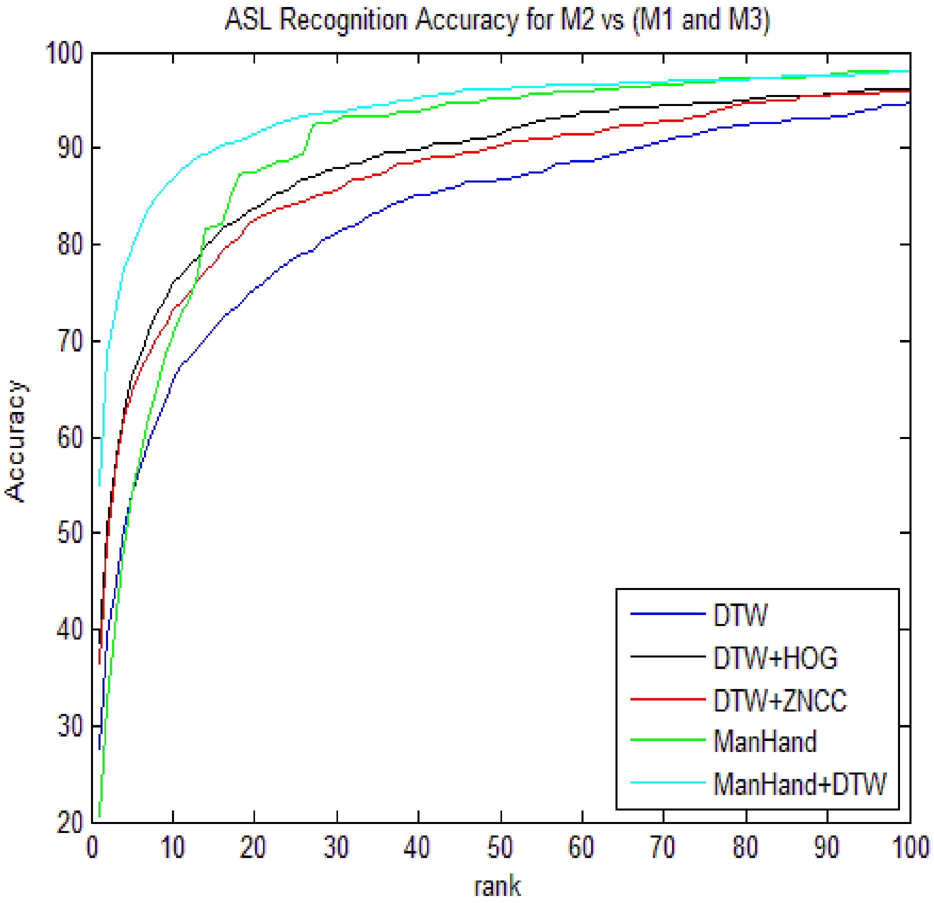


Figure 9-4 ASL Recognition Accuracy for M2 vs. (M1 and M3)

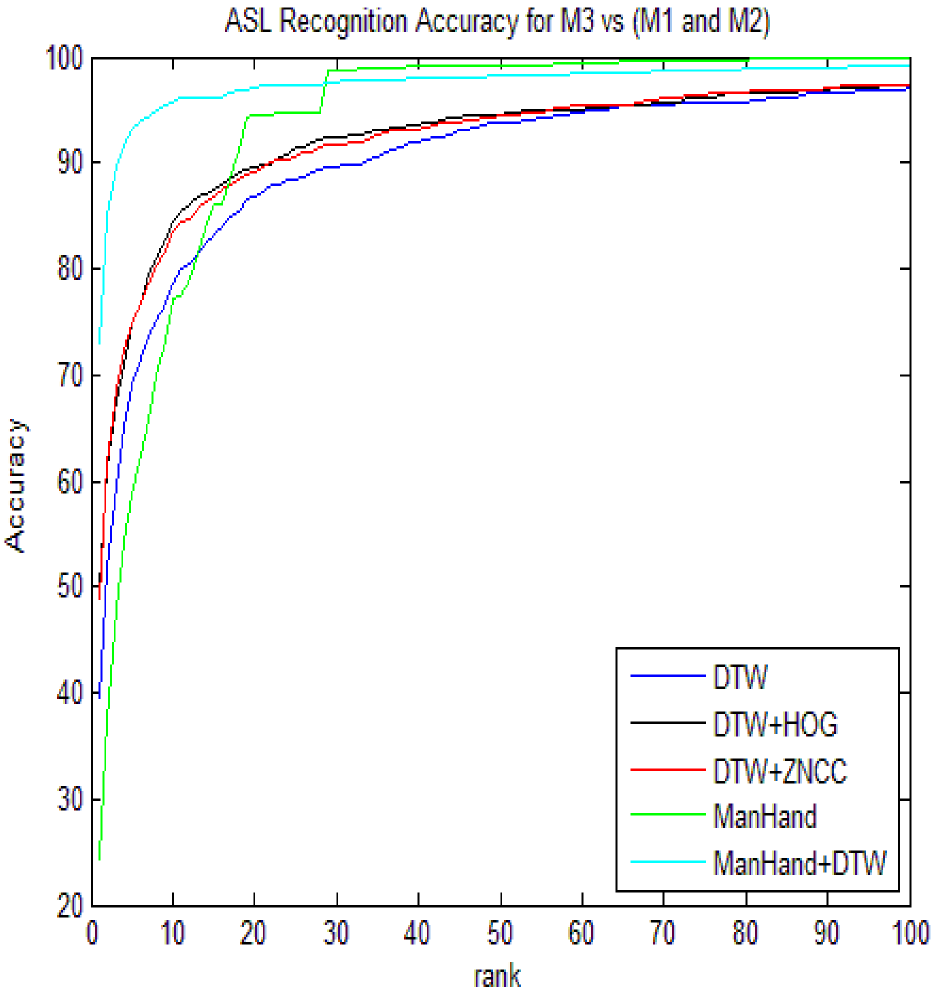


Figure 9-5 ASL Recognition Accuracy for M3 vs. (M1 and M2)

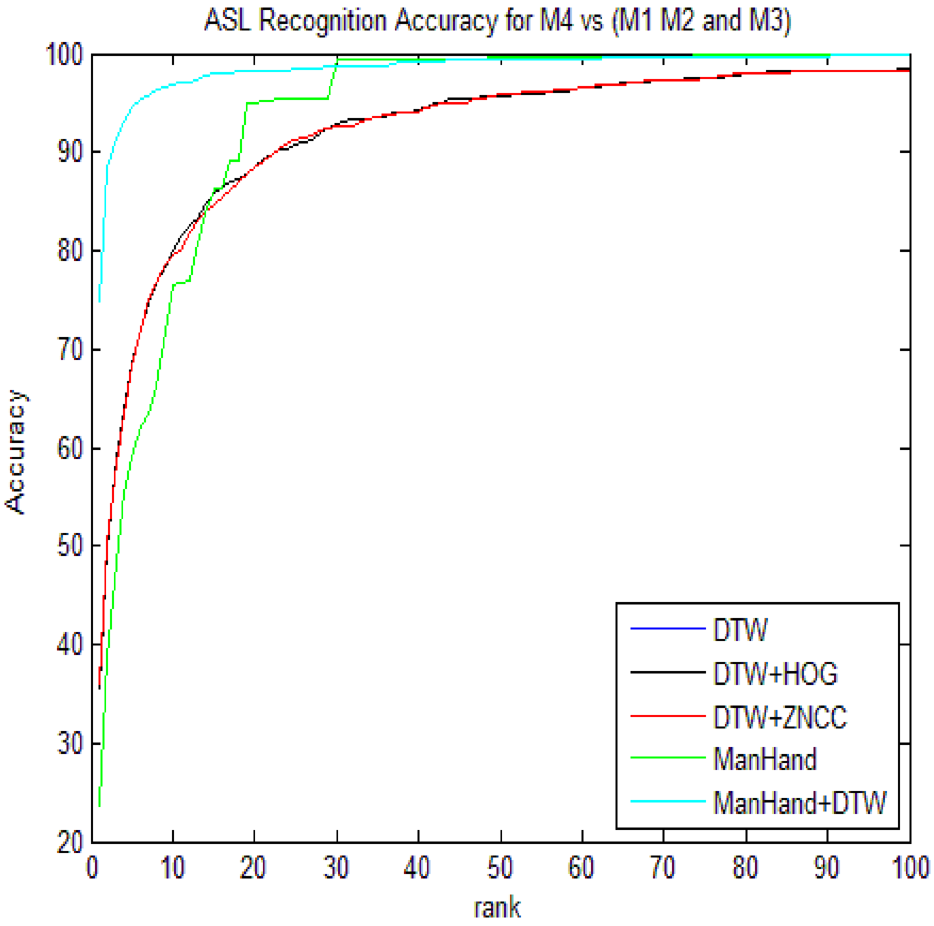


Figure 9-6 ASL Recognition Accuracy for M4 vs. (M1 ,M2 and M3)

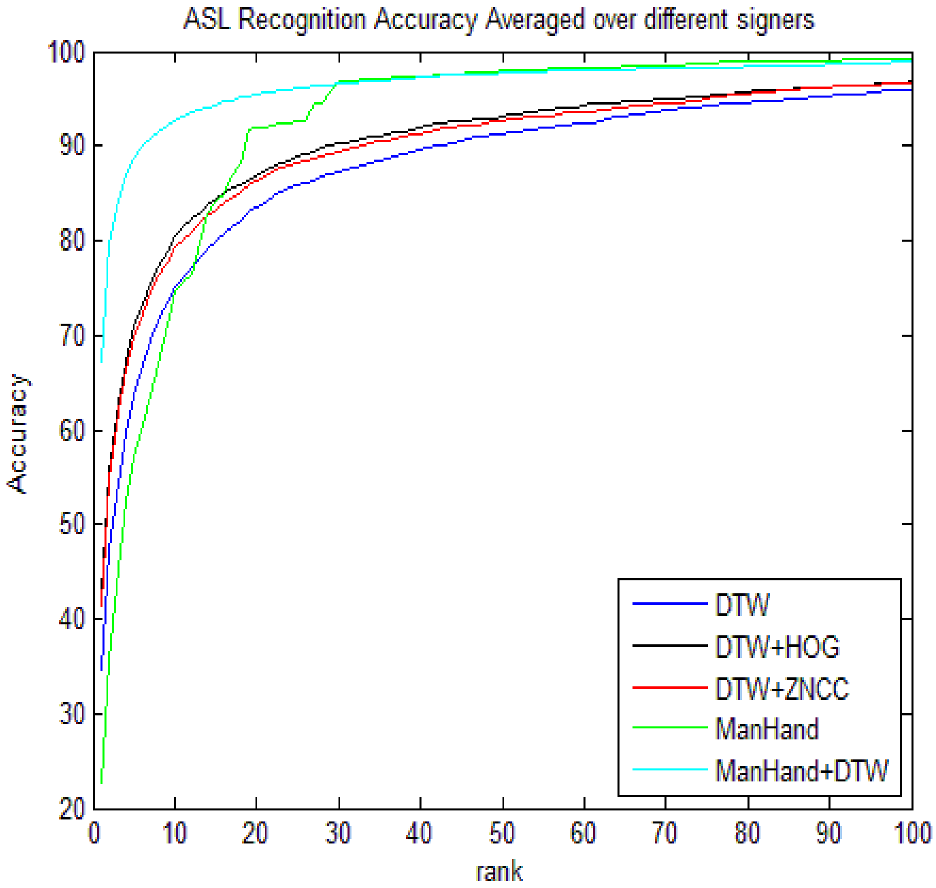


Figure 9-7 ASL Recognition Accuracy averaged over different signers(M1,M2,M3,M4)

Rank	**DTW**	**DTW+ZNCC**	**DTW+HOG**	**Manual Hand**	**Manual hand+DTW**
1	35.0404	44.2947	47.4394	22.4618	65.4088
5	63.4322	71.1590	73.9443	56.6936	87.8706
10	76.0108	80.5031	81.0422	73.7646	90.8356
20	83.3783	85.1752	85.4447	90.4762	94.9686
30	85.8041	87.6011	88.1402	95.8670	96.0467
50	88.7691	90.2066	90.6559	97.5741	97.0350
100	93.8005	94.7889	94.9686	99.1015	98.3827

Table 9-1 Recognition accuracies for ranks 1,5,10,20,30,50 and 100 evaluated for the case M1 vs.(M2 and M3) for the five different similarity measures

Rank	**DTW**	**DTW+ZNCC**	**DTW+HOG**	**Manual Hand**	**Manual hand+DTW**
1	27.5831	36.4780	38.6343	20.4852	54.9865
5	53.9982	64.5103	66.3073	54.0881	79.6047
10	65.8580	73.3154	76.0108	70.6199	86.7925
20	75.3819	82.4798	83.7376	87.6011	91.4645
30	81.1321	85.6244	87.8706	92.9021	93.8904
50	86.7026	90.3863	91.5544	95.2381	96.1366
100	94.7889	95.9569	96.2264	97.9335	98.0234

Table 9-2 Recognition accuracies for ranks 1,5,10,20,30,50 and 100 evaluated for the case M2 vs.(M1 and M3) for the five different similarity measures

Rank	**DTW**	**DTW+ZNCC**	**DTW+HOG**	**Manual Hand**	**Manual hand+DTW**
1	39.5328	48.9668	49.7754	24.3486	72.9560
5	68.9128	74.8428	74.9326	58.6703	93.0818
10	78.6164	83.4681	84.5463	77.2686	95.7772
20	86.8823	89.2183	89.6676	94.4295	97.0350
30	89.5777	91.6442	92.4528	98.7421	97.5741
50	93.8005	94.5193	94.6092	99.2812	98.2031
100	97.0350	97.3944	97.3944	99.8203	99.1914

Table 9-3 Recognition accuracies for ranks 1,5,10,20,30,50 and 100 evaluated for the case M3 vs.(M1 and M2) for the five different similarity measures

Rank	DTW	DTW+ZNCC	DTW+HOG	Manual Hand	Manual hand+DTW
1	36.0911	36.0911	35.6115	23.6211	74.8201
5	68.2254	68.2254	68.4652	58.8729	94.4844
10	79.4964	79.4964	79.9760	76.4988	96.8825
20	88.3693	88.3693	88.4892	95.0839	98.3213
30	92.5659	92.5659	92.8058	99.4005	98.6811
50	95.8034	95.8034	95.5635	99.6403	99.4005
100	98.3213	98.3213	98.4412	100.0000	99.8801

Table 9-4 Recognition accuracies for ranks 1,5,10,20,30,50 and 100 evaluated for the case M4 vs.(M1 M2 and M3) for the five different similarity measures

Rank	**DTW**	**DTW+ZNCC**	**DTW+HOG**	**Manual Hand**	**Manual hand+DTW**
1	34.5619	41.4576	42.8651	22.7292	67.0429
5	63.6422	69.6844	70.9124	57.0812	88.7604
10	74.9954	79.1958	80.3938	74.5380	92.5719
20	83.5029	86.3107	86.8348	91.8977	95.4474
30	87.2700	89.3589	90.3173	96.7279	96.5481
50	91.2689	92.7289	93.0957	97.9334	97.6938
100	95.9864	96.6154	96.7577	99.2138	98.8694

Table 9-5 Recognition accuracies for ranks 1,5,10,20,30,50 and 100 evaluated from the average of different signers(M1,M2,M3,M4) for the five different similarity measures

Chapter 10

Discussion and Future Work

We devised a system that works on hand motion to discriminate between signs as proposed in [2][3][5]and incorporate hand shape and hand appearance information (manual ,semi-automatic methods)to improve the accuracy significantly. We used the existing similarity measure Dynamic time warping(DTW)[8] for comparing trajectories of hand motion. For the manual method we computed a conditional probability table by incorporating relation between hand shape pairs in all the model classes(database sign videos) from which a similarity score was computed runtime based on the hand shape information provided by the user. The similarity scores obtained from the trajectory of hand motion and the similarity scores obtained from the conditional probability table for the hand shape information provided by the user are combined to obtain a global similarity score which is used to rank the sign under test. For the semi-automatic method the hand regions which are specified as bounding box by the user in the start and end frames of the query video are matched with the database videos (which already possess the hand locations annotated by [3]) by similarity measures such as Zero Mean normalized cross correlation(ZNCC) and Histogram of oriented gradients(HOG) in combination with dynamic time warping(DTW) DTW to obtain a global similarity score which is used to rank the sign under test. we also created a GUI to visualize hand shape information obtained by annotating hand shape for the start and end frames of each sign for all the database videos and a real time demo by integrating manual hand shape method based on similarity scores obtained from the conditional probability along with the existing American sign language demo based on DTW.

From the evaluation we conclude that DTW is the lowest performing similarity measure and the highest performing similarity measure is manual handshapes with DTW. The DTW performs well for longer signs as it is more biased towards it and performs poorly for signs which are static with less hand motion. By combining manual handshape similarity scores which is computed according to the user inputs with DTW the accuracy is significantly improved especially for all the static signs. The hand appearance similarity measures using ZNCC and HOG offer better recognition accuracy compared to DTW as they incorporate valuable cues from hand appearances which improves the recognition accuracy for both static signs and signs with hand motion. The HOG slightly outperforms ZNCC as they incorporate edge or gradient information for the hand appearances which ZNCC fails as the similarity purely intensity based. The same behavior of accuracies are reciprocated for the remaining cases .

For hand appearance matching we can use image features obtained from scale invariant feature transform(SIFT) which are robust to rotation, brightness and scale to improve the similarity measures. We can also extend the similarity measures from 2D to 3D space by using 3D hand trajectories which can be obtained by placing cameras at different positions and angles. We can definitely use better skeleton trackers and hand locating methods as the existing RGB-D kinect tracker is not up to the expected accuracy. In future we intend to integrate better trajectory measures for hand motion and more robust hand appearance measures into the existing demo to improve the recognition accuracy.

References

[1] T. Starner and A. Pentland. Real-time American Sign Language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[2] H. Wang, A. Stefan, S. Moradi, V. Athitsos, C. Neidle, , and F.Kamangar., “A System for Large Vocabulary Sign Search," *in Workshop on Sign, Gesture and Activity (SGA)*, September 2010, pp. 1-12.

[3] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, A. Thangali, H. Wang, and Q. Yuan, “Large Lexicon Project: American Sign Language Video Corpus and Sign Language Indexing/Retrieval Algorithms," *in 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign LanguageTechnologies*, 2010, pp. 1-4.

[4] Jonathan Alon, Vassilis Athitsos, Quan Yuan, and Stan Sclaroff, “A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 31(9),* pages 1685-1699, September 2009.

[5] Christopher Conly, Zhong Zhang, and Vassilis Athitsos ”An Evaluation of RGB-D Skeleton Tracking for Use in Large Vocabulary Complex Gesture Recognition”, *Pervasive Technologies Related to Assistive Environments (PETRA)*, May 2014.

[6] A. Bobick and J. Davis, “The recognition of human movement using temporal templates,” Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257–267, March 2001.

[7] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2247–2253, 2007.

[8] Sakoe, H. and Chiba, S., Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1) pp. 43- 49, 1978, ISSN: 0096-3518

[9] Kruskal, J.B., Liberman, M., “The symmetric time warping algorithm: From continuous to discrete,” *in Time Warps. Addison-Wesley*, 1983

[10] Michalis Potamias and Vassilis Athitsos, “Nearest Neighbor Search Methods for Handshape Recognition,” *Conference on Pervasive Technologies Related to Assistive Environments (PETRA)*, July 2008.

[11] J.P. Lewis, Fast Template Matching, Vision Interface 95, Canadian Image Processing and Pattern Recognition Society, Quebec City, Canada, May 15-19, 1995

[12] Navneet Dalal, Bill Triggs "Histograms of oriented gradients for human detection" Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference

[13] William T. Freeman, Michal Roth "Orientation Histograms for Hand Gesture Recognition" IEEE Intl. Wkshp. on Automatic Face and Gesture Recognition, Zurich, June, 1995

[14] Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, Antonio Torralba ,MIT "HOGgles: Visualizing Object Detection Features" Jan 2013.

[15] Dreuw, P., Deselaers, T., Keysers, D., Ney, H.: Modeling image variability in appearance-based gesture recognition. In: ECCV Workshop on Statistical Methods in Multi-Image and Video Processing. (2006) 7–18

[16] Cui, Y., Weng, J.: Appearance-based hand sign recognition from intensity image sequences. Computer Vision and Image Understanding 78 (2000) 157–176

[17] Stefan, A., Wang, H., Athitsos, V.: Towards automated large vocabulary gesture search. In: Conference on Pervasive Technologies Related to Assistive Environments (PETRA). (2008)

[18] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, Peter Presti "American sign language recognition with the kinect". In: ICMI Proceedings of the 13th international conference on multimodal interfaces, Nov 2011.

[19 P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney, “Tracking using dynamic programming for appearance-based sign language recognition,” in Proc. IEEE Int. Conf. Autom. Face Gesture Recog., 2006, pp. 293–298.

[20] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, “Hidden conditional random fields for gesture recognition,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2006, vol. 2, pp. 1521–1527.

[21] Paul Doliotis, Alexandra Stefan, Chris Mcmurrough, David Eckhard, and Vassilis Athitsos. “Comparing Gesture Recognition Accuracy Using Color and Depth Information,” *Conference on Pervasive Technologies Related to Assistive Environments (PETRA)*, May 2011.

[22] Lane, H., Hoffmeister, R.J., Bahan, B.: A Journey into the Deaf-World. DawnSign Press, San Diego, CA (1996)

[23] Schein, J.: At home among strangers. Gallaudet U. Press, Washington, DC (1989)

[24] Pat Jangyodsuk, Christopher Conly, and Vassilis Athitsos "Sign Language Recognition using Dynamic Time Warping and Hand Shape Distance Based on Histogram of Oriented Gradient Features", *Pervasive Technologies Related to Assistive Environments (PETRA)*, May 2014.

Biographical Information

Siddhartha Goutham Swaminathan was born in Chennai, India in 1989. He completed his Bachelor of Engineering in Electronics and Communication from Anna University, India in 2010. He obtained his Master of Science degree in Electrical Engineering from University of Texas at Arlington in May 2016. After completing his Bachelor's degree he joined Larsen and Toubro Ltd as a Software Engineer and worked on developing algorithms in the field of signal and image processing. He also worked for Qualcomm Inc. as an Interim Engineering intern for their Multimedia Research and Development Team from May 2015 - August 2015. His current research interests are in the area of Computer Vision, Signal processing, Image processing, Virtual and Augmented Reality.