2006

Implementation of image processing approach to translation of ASL finger-spelling to digital text

Divya Mandloi

Kanthi Sarella

Chance Glenn

Implementation of Image Processing Approach to Translation of ASL Finger-spelling to Digital Text

Divya Mandloi, Mani Kanthi Sarella, and Chance M. Glenn, Sr.
The Laboratory for Advanced Communications Technology
Rochester Institute of Technology
mks4500@rit.edu, dxm9088@rit.edu, cmgiee@rit.edu

Abstract

The present analysis is the phase one of a broader project, the Sign2 Project, which is focused on a complete technological approach to the translation of American Sign Language (ASL) finger-spelling to digital audio and/or text. The methodology adopted in this analysis employs a gray-scale image processing technique. We will describe the latest results as well as future enhancements to the system. This paper will also discuss the application of the approach to the telecommunication Industry.

1. Background

Sign language is the primary means of communication with and among the deaf community and is defined as a visual-gestural language [1]. American Sign Language (ASL) is the fourth most commonly used language in the United States and Canada [2]. It is one of the most widely used and most popular way of communication between the above mentioned communities in the United States of America. There had been a constant effort for the past few years to successfully convert the ASL to text through an artificial interface substituting a human interpreter. ASL is a gestural language and the entire ambit of the ASL involves various hand gestures and facial expressions like eyebrow movement and lip-mouth movement. It also utilizes the space around

the signer to specify places and people. [3] While there is ongoing work to recognize facial expression through vision based and data point based technologies [4], this paper only considers the finger spelling aspect of ASL, in an endeavor to accomplish the task mentioned. These technologies have most distinctively taken two approaches. 1) The visual based approach and 2) Instrumented gloves approach. The problem statement being considered in the present analysis is the creation of a vision-based translation system which converts American Sign Language to text in a natural environment which is unobtrusive to both the signers and the viewers.

There has been a continual resurgence of attempts to successfully recognize hand gestures over the past many years. The problem of posture and gesture recognition is chiefly solved by a consistent method of operation, which includes collection of raw data, and analysis of the data using recognition algorithms to extract meaning from the data. The data is basically collected in two distinct ways. The first approach involves input devices worn by the subject and the second method involves using a vision-based approach in which the user's gestural images are collected as raw data by one or more cameras.

The current commercially viable solutions for translation like the glove-based approach tend to be intrusive and require the subject to wear colored gloves or tracking targets. In this particular approach, the user is required to wear one or two data gloves that measure the angles which are created by the joints of the hand and a six degree of freedom tracking device that helps extract hand position and orientation data. The data gloves measure the finger movement of the wearer using sensors which are embedded in the data glove. This raw data is then relayed to a computer for further analysis. Though this approach had been commercially more viable and yielded acceptable results, it provided only an invasive and expensive solution.

In the second approach, which is a vision-based approach, one or more cameras are used to collect the images of the signer. The cameras might collect a random number of frames and subject the chosen frames to image processing algorithms to recognize the demonstrated gestures. The image processing algorithms can vary from approaches using Hidden Markov Models [5] to Gray-scale based algorithms [6]. This approach is noninvasive and has more scope for implementation in real time devices.

2. Approach

The scope of this investigation does not include a translation system which attempts to encompass the entire ASL vocabulary. Rather, it considers the alphabet of ASL and strives to convert it into English text. Another goal is to design this system so as to be able to incorporate this technology in real-time systems further down the line. The extension of this concept to create a possibility of a commercial product implementing this translation system which includes the entire ASL vocabulary is anticipated. A high accuracy also takes a great precedence in this particular project.

The main objectives of this analysis are

- To create a translation system that can successfully convert the ASL alphabet into digital English text.
- To collect the raw data required for this investigation employing one or more cameras.
- To implement a visual-based technology so as to make the translation system non-intrusive for the signer.
- To obtain a high degree of accuracy while realizing this translation system.
- To investigate the possibility of the extension of Sign2 project into real time systems.

The various technologies which were used before including the Glove-based technology have been found to be too cumbersome and intrusive to the signer. In most of the cases, these techniques tended to be unpopular as they do not allow the signer any freedom of movement, required other appurtenances which need to be setup every time the system needs to be used and the raw data collected needed to undergo a vast amount of processing adding to the run time of the systems. They also did not appear attractive as possible commercially viable devices like real time systems which could be incorporated in every day electronic devices like cell phones, smart-phones and high-tech kiosks. This has lead some of the research to channelize towards the more convenient vision-based systems which require less processing of data due to the ready availability of image compression and processing techniques. It does not include complex calculations which consequently results in less processing overhead.

The approach taken to cater to this study is formulated into three steps [6]. First, we established a standardized set of physical measurements for ASL finger spelling. Second, we formulated a generalized set of measurements for the broad statistical range of subjects we would encounter. Finally we correlated these measurements with statistical range of subjects for letter recognition. We process static images of the subject considered and then match them to a statistical database of pre-processed images to ultimately recognize the specific set of signed letters using the mean square error (MSE) and the peak signal-to-noise ratio (PSNR) [6]. Figure 1 shows the block diagram of the approach undertaken.

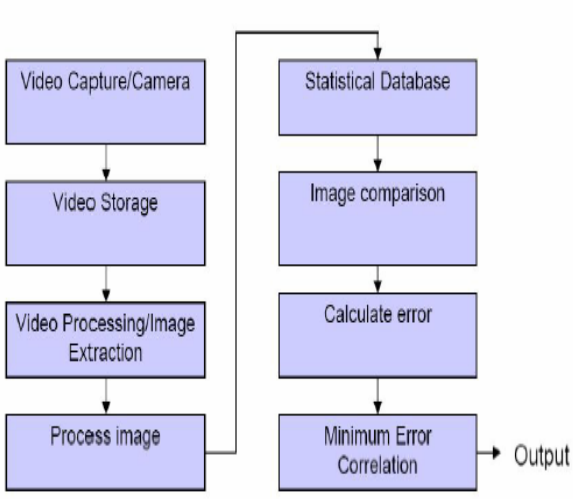


Fig.1. The Sign2 system functional block diagram

3. Implementation

3.1 Data Collection: Video Capture and Video storage

In the data collection phase, videos of various subjects spelling out the alphabet in ASL were captured. Figure 2 shows the pictorial representation of the assembly used for capturing the video. Each of the subjects was made to spell out the complete alphabet. This process was carried out multiple times to capture many videos, so as to build an extensive database.

3.2 Data Processing

Data Processing consists of three phases

1. Video Processing and Image Extraction

2. Image Processing

3. Image Storage

3.2.1 Video Processing and Image Extraction

This phase consists of post processing of the captured video. The captured video is read in MATLAB. The system allows the flexibility to specify the start and end frame of the video for which the processing is carried out. Individual frames are extracted as images which are used in the image processing phase of the project. Image extraction is not only used for the statistical database but also for the extraction of images from the test video for further comparison with the already built statistical database. Figure 2 shows the extraction of a frame from a video.

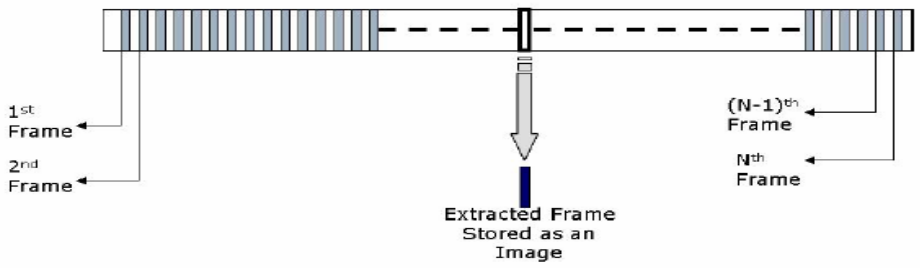


Fig.2: Frame extraction from a video clip

3.2.2 Image Processing

Image processing is used to process the extracted images. Frames extracted by the image extraction technique are converted to black and white images in a pixel by pixel fashion. The images thus obtained are color images which are in an RGB format. The red, green and blue (RGB) components of the image are extracted separately and each component is converted to double precision. Binary thresholding is performed based on the user specified threshold value, thus converting the image into black and white. This black and white image is cropped based on the detection of the edges (left, right and top edges) of the hand. Because of the varying hand sizes of the subjects, the resulting cropped images are of different sizes. To counter this problem and to provide a generalized system, the cropped images of all subjects were again resizes to a consistent size of 150 x 80.

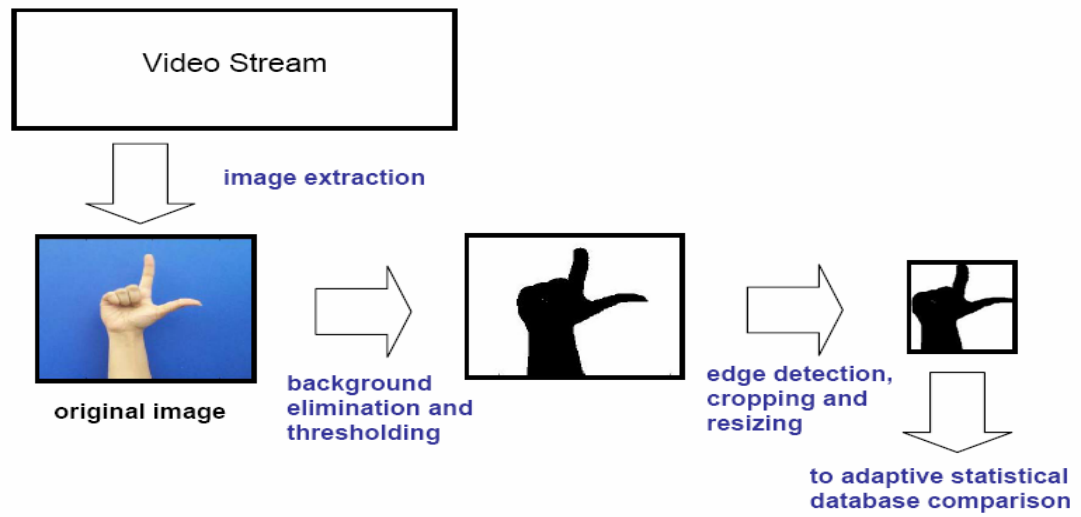


Fig.3: Processing procedure

The following Figure 4 shows the graphical user interface developed to post process the captured video in order to extract the frames from it. This GUI has 3 main modules: first module is display module to show the post-processing of the video, the second module shows the black and

white cropped image of the last frame extracted and third module shows the error graph between the extracted frames. This GUI allows a user to specify the frame difference (module 6) between the two consecutive frames extracted in the text box specified as the Frame Difference. Module 5 specifies the start as well as the end of frame of the video. In the example shown in the figure, every third frame will be extracted and processed as defined in the frame difference module. The graph shown in the GUI depicts the error between the original frames extracted as well as the processed frame.

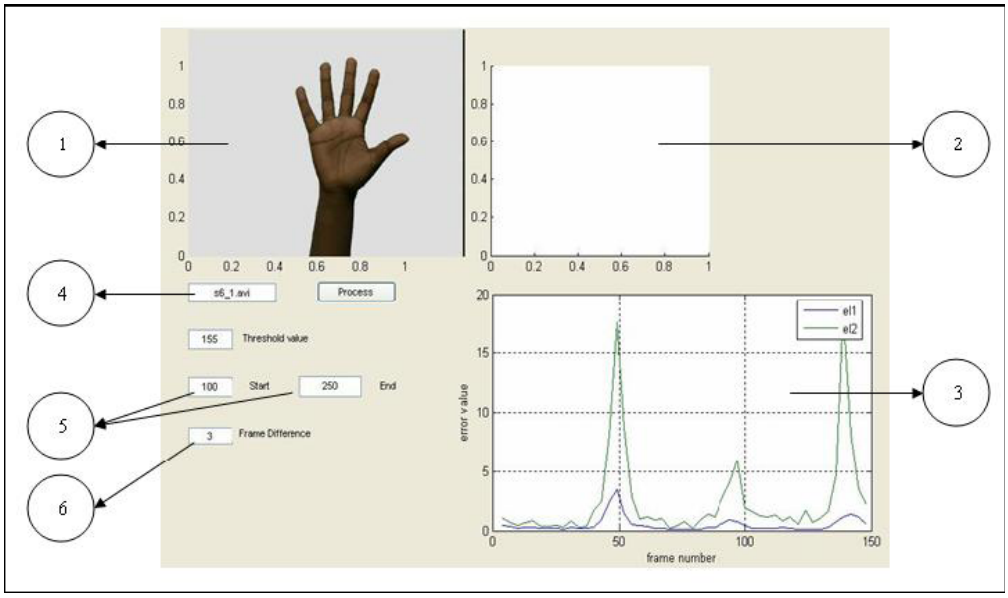


Fig.4: Frame Extraction GUI

3.2.3 Image Storage (Statistical Database)

The heart of the study lies in the statistical database, which determines the reliability of the results. It becomes crucial to capture as many hand structures as possible for the generalization of this methodology. To improve the reliability, each subject was asked to sign the alphabet from A to Y repetitively. The repetitive technique is used to record the changes in the hand gestures of

the signs at different instances. The database consists of the black and white resized images extracted using the image processing and cropping techniques.

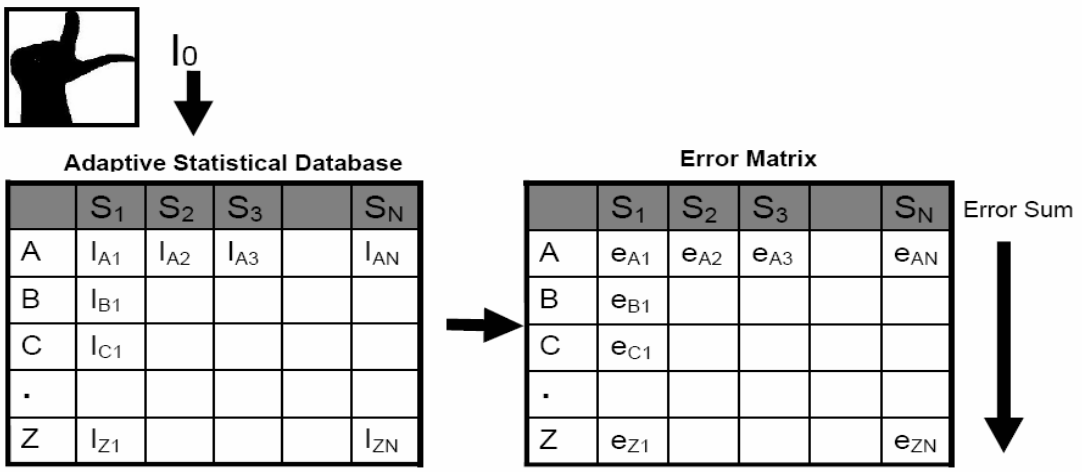


Figure 4. Diagram of the statistical database showing images for each letter and each subject resulting in an error matrix

Fig.5: Diagram of the statistical database showing images for each letter and each subject resulting in an error matrix

The cropped and resized test image is compared with all the images in the statistical database images based on the definition of **Mean Square Error (MSE)** and **Peak signal to Noise Ratio (PSNR).**

The **Mean Square Error** is given by

$$MSE = \frac{1}{LW}\sum_{l=1}^{L}\sum_{w=1}^{W}(I(l,w) - I'(l,w))^2.$$

The **Peak Signal to Noise Ratio** is given by

$$PSNR = 20log_{10}\left[\frac{255}{\sqrt{MSE}}\right]$$

An error matrix is thus created having error values for each of the corresponding images in the database. The set of image in the database that corresponds to a given letter and has the lowest cumulative error, revels the highest priority of the correct letter being returned.

To further understand the letter recognition phase of this investigation, a video where the word LAY is being spelled is considered as an example. It is assumed that the subject holds each signed letter for a small period of time before signing the next letter. The differential error between the consecutive frames remains almost constant. This marks the presence of a letter being spelled. Rapid fluctuations in the differential errors between frames imply a transition from one letter to another. Figure 6 indicates the error between the consecutive processed frames. A window size of 10 frames falling below the error threshold e_{thresh} is considered to be a letter being spelled. e_{thresh} is given by

$$e_{thresh} = \frac{3}{10}(e_{max} - e_{min})$$

where e_{max} is the maximum error recorded and e_{min} is the minimum error recorded.

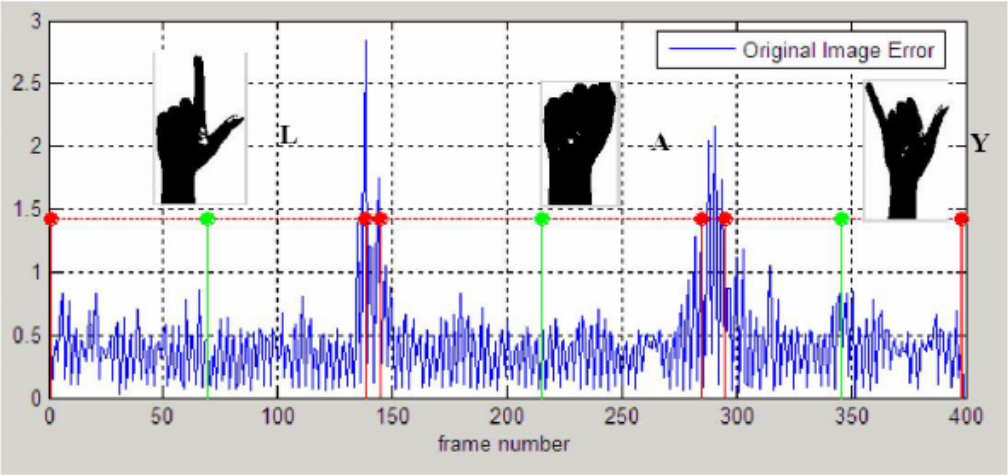


Fig.6: Frame to Frame error calculation and letter recognition

In the above example, the letters ‘L’, ‘A’, ‘Y’ are recognized where 10 consecutive frame falling below the threshold are detected. The frames that are not constant for 10 consecutive frames are discarded.

4. Results

4.1 Alphabet and word recognition

Results are recorded based on the data gathered by capturing the videos of different subjects spelling out English alphabet and words. The videos consist of letters from A to Y and three, four or five letter words.

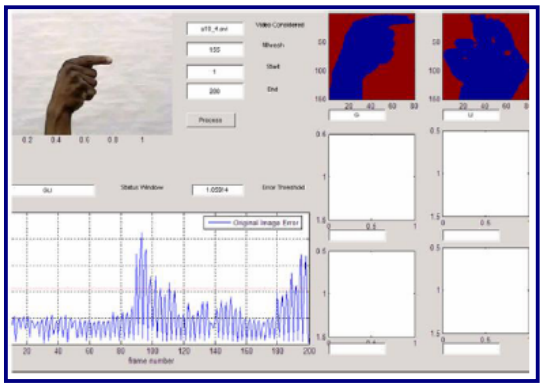


Fig.7: Sign2 graphical user interface for analysis.

Figure 7 shows the graphical user interface (GUI) that is used for analysis of the video frames as they are being extracted, processed and compared with the statistical database.

First phase of the analysis is based on the calculation of Recognition Ratio (γ) of English alphabet. Whenever a letter is recognized correctly, it is given a value 1 for each iteration and if it is not recognized correctly it is assigned a value 0.

We define the recognition ratio as, $\gamma = \frac{\sum_{i=1}^{n} \alpha_i}{n}$, where α is the is total number of successful recognitions of an alphabet.

Alphabets	Total Number of Iterations	Successful Recognition	Recognition Ratio γ	Recognition Percentage $\gamma\%$
A	14	11	0.78571	78.571
B	12	11	0.91666	91.666
C	15	6	0.40	40
D	12	11	0.91666	91.666
E	15	11	0.73333	73.333
F	11	7	0.63636	63.636
G	14	13	0.92857	92.857
H	14	13	0.92857	92.57
I	15	14	0.93333	93.333
J	11	8	0.72727	72.727
K	10	8	0.80	80
L	10	10	1	100
M	13	6	0.46153	46.153
N	13	7	0.53846	53.846
O	13	8	0.61538	61.538
P	11	8	0.72727	72.727
Q	13	12	0.92307	92.307
R	8	3	0.375	37.5
S	11	6	0.54545	54.545
T	9	4	0.44444	44.444
U	5	6	.80	80
V	9	9	1	100
W	7	6	0.85714	85.714
X	4	3	0.75	75
Y	8	8	1	100

Reliability ratio for alphabet

We also define the estimated recognition ratio of a word as, $\omega = \frac{\sum_{j=1}^{c} \gamma_j}{c} = \frac{\sum_{j=1}^{c}\sum_{i}^{n} \alpha_{ij}}{nc}$, where c is the number of characters in a word. The experimental results for a chosen set of words generated by different signing subjects are shown below.

Word	Experimental Recognition Ratio ω_{exp}	Experimental Recognition Percentage $\omega_{exp}\%$	Estimated Recognition Ratio ω	Estimated Recognition Percentage $\omega\%$
ASL	0.66666	66.666	0.77705	77.705
HAT	0.44444	44.444	0.71957	71.957
LET	0.49999	49.999	0.72592	72.592
LAY	0.73333	73.333	0.92957	92.957
$CARD$	0.50000	50	0.619345	61.9345
$WORD$	0.66667	66.667	0.69104	69.104
$DEAF$	0.41667	41.667	0.76801	76.801
$BABY$	0.5625	56.25	0.90476	90.476
$THINK$	0.67142	67.142	0.72896	72.896
$SHARE$	0.54000	54	0.673614	67.3614
$PAPER$	0.56667	56.667	0.66971	66.971
$PRINT$	0.63333	63.333	0.60370	60.370

Estimated and Experimental reliability ratio of words.

The above table presents contrast between experimental recognition percentage and estimated recognition percentage for each word under consideration. For the word “WORD”, the difference between estimated recognition ratio and experimental recognition ratio is the lowest (2.437%), whereas this difference is highest for word “DEAF” (35.134%).

5. Future Work

Two approaches have been formulated to increase the accuracy of the results obtained following the gray scale approach discussed above. The minor challenges faced to accurately recognize letters in the above approach can be attributed to varying factors like different letter formulations speeds, window sizes considered to recognize the existence of a frame, inadequacy of the data available, etc.

The two approaches being proposed so as to obtain better results are the dual-camera approach and the feature extraction method. These methods are being pursued in an endeavor to sharpen the accuracy of the results while recognizing the signed letters.

5.1 Dual Camera Approach

The following section gives a brief description of the dual-camera approach. The setup includes two cameras placed at a fixed distance of ‘D’ from each other. The distance between the subject and the line of which the cameras stand should be a fixed ‘L’. The distances ‘L’ and ‘D’ should be maintained constant. The dual camera approach follows the same methodology as that of the image processing technique illustrated in the previous sections. In the dual-camera approach we employ two cameras which result in obtaining two set of databases and hence creating twice the amount of data availability. The setup for the dual-camera approach is shown in the Fig 9.

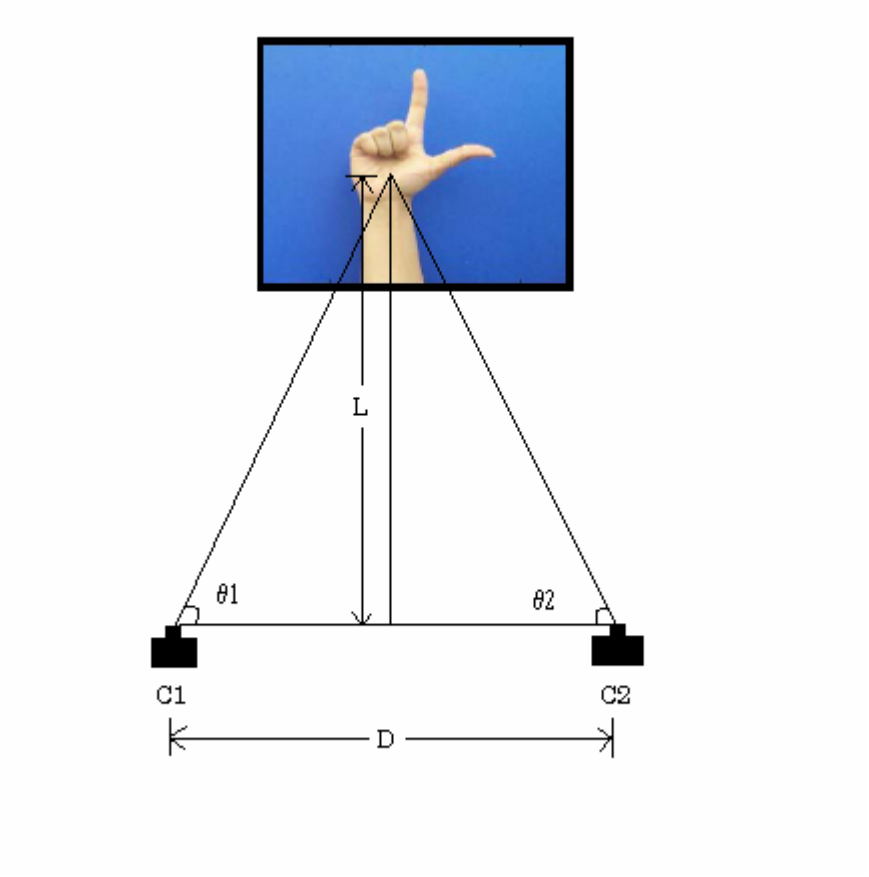


Fig.9: Dual camera approach setup

Shots taken from Camera 1 and Camera 2 are shown in the following figures.

Image taken from Camera C1

Fig.10

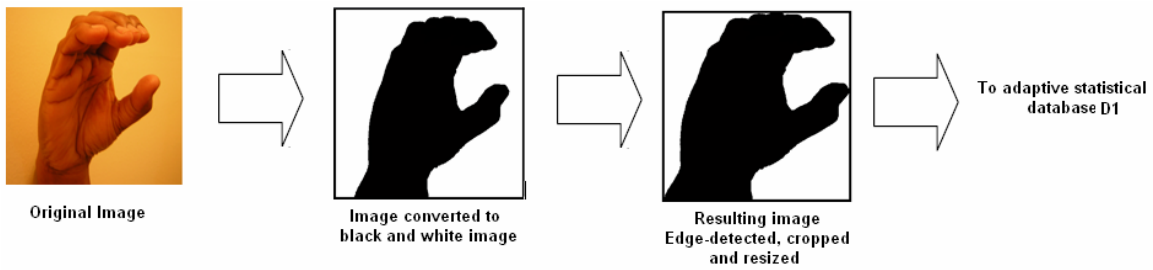
Image taken from Camera C2

Fig.11

A database D1 is formed which consists of the resultant processed images of each of the images shot from Camera 1 (C1) and another database D2 is formed consisting of all the resultant

processed images of each of the images shot from Camera 2 (C2). When a fresh image is captured, it is processed and compared to those of the images in D1 and D2. This would result in giving a more exact result. ‘A1’ corresponds to the signed letter ‘A’ as shot from the camera C1 and ‘A2’ correspond to the signed letter ‘A’ as shot from the camera C2. 9 iterations are considered for each letter. ‘A11’ would be the first iteration if letter ‘A’ shot from camera C1. The databases shown below are considered for a single subject. When multiple subjects are considered, the database takes a third dimension with the Z-axis representing the number of subjects.

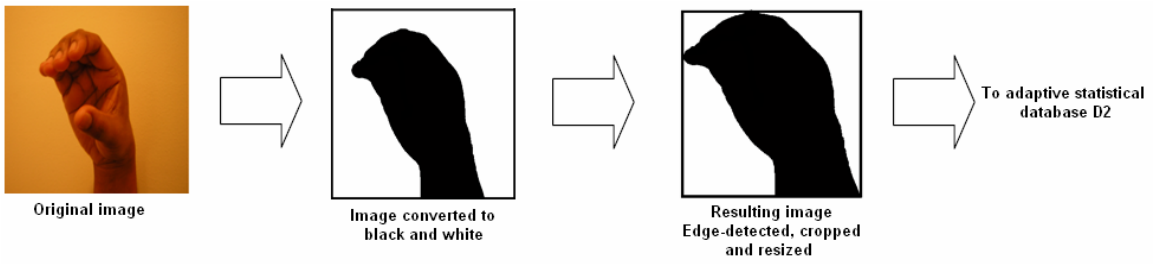
Process:



Database 1

D1	**1**	**2**	**3**	**…**	**…**	**…**	**9**
A1	**A11**	**A12**	**A13**	**…**	**…**	**…**	**A19**
B1	**B11**						
…	**…**						
…	**…**						
Z1	**Z11**						

Fig.12: Image extraction process for images shot from C1 and the corresponding statistical database.



Database 2

D2	1	2	3	…	…	…	9
A2	A21	A22	A23	…	…	…	A29
B2	B21						
…	…						
…	…						
Z2	Z21						

Fig.13: Image extraction process for images shot from C2 and the corresponding statistical database.

The Mean Square Error is calculated for each of the images and stored in the error matrix. The set of images in the database that correspond to a given letter and have the lowest cumulative error reveals the highest probability of the correct letter being returned.

This approach presents a sturdier database which provides more information and hence increases the probability of recognizing a correct letter for a signed hand signal.

5.2 Feature Extraction Method

The second approach that is intended to be pursued is the feature extraction method. In this approach, the images are converted into gray scale images as opposed to the black and white images in the previous methods. The aim of this approach is to retain the features of the image and to work on them as they provide more data than the black and white images. The features

that are extracted in this approach are the gray scale intensities of several pixels picked at the intended coordinates. The coordinates are chosen in a way to cover all the crucial areas of the image where there is the highest likeliness of change in features for each letter.

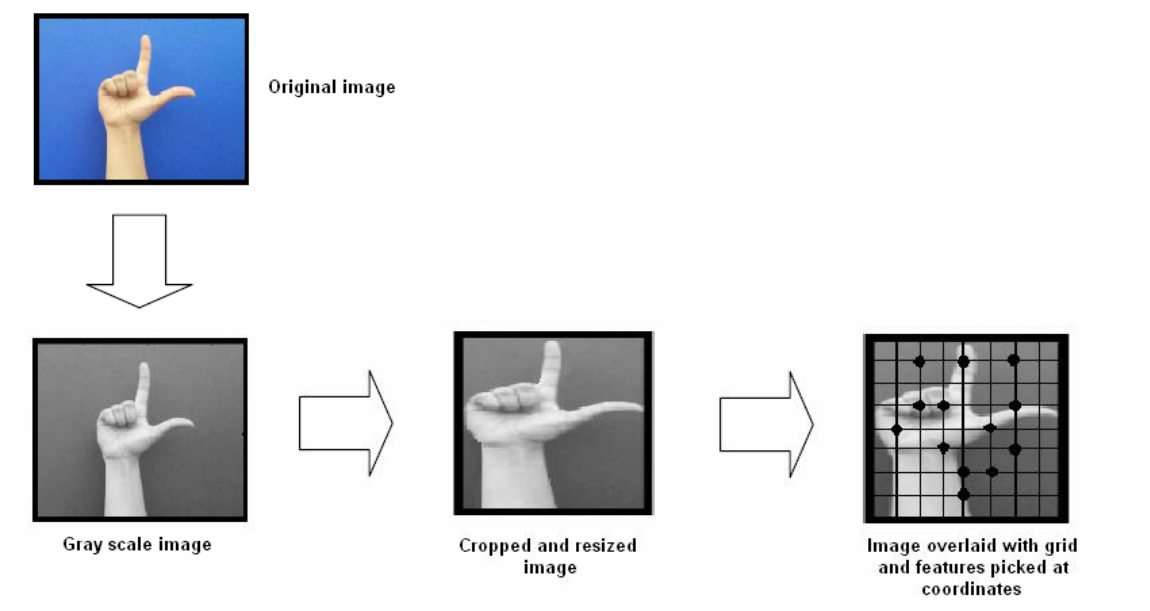


Fig.14: Feature extraction process of an image

The gray scale intensities thus extracted are converted into a statistical representation of the intensities per pixel and fed into a database for later reference.

The algorithm for recognition of newly fed images follows the same algorithm used for the feature extraction of the original images. The gray scale intensity information extracted from the fresh input image is compared to the information in the database to obtain an appropriate recognition of the letter signed. Neural networks are trained to recognize the pattern formed by each of the signed letters. Employment of the neural network helps in normalizing the slight differences obtained while recognizing the letter and smoothens them to the appropriate weighted pattern of each of the letters hence

6. Conclusion

The primary focus of this study was to examine image processing as a tool for the conversion of American Sign Language in to digital text. Further this study promises to be used in the real time application to fully recognize American Sign Language. This can be further developed into a system which can be integrated in to the upcoming telecommunication devices with cameras to bridge the communication gap between the hearing and deaf/hard of hearing communities. System can be enhanced in terms of increase in the data processing speed and data storage by using the compression techniques and feature extraction techniques. The results observed after conducting the experiments strengthens the approach being considered as an effective means to tackle the problem.

This analysis presents the results of the phase one of the Sign2 process and proposes future enhancements to the system using some of existing and upcoming technologies. It also highlights the application in the telecommunication industry so as to achieve the basic motivation of bridging the communication gap between the hearing community and deaf and hard of hearing community.

References

1. D. C. Charlotte Baker-Shenk, *American Sign Language: A Teacher's Resource Text on Grammar and Culture,* Washington, D.C.: Clerc Books, Gallaudet University Press, 1980.
2. D. Burgett, Assistive *technology: Communication devices for the deaf*, 2004
3. http://www.deaflibrary.org/asl.html
4. I. Essa, T. Darrel. and A. Pentland. Tracking facial motion. IEEEWorkshop on Nonrigid and articulated motion.
5. Thad Starner and Alex Pentland. Visual Recognition of American Sign Language Using Hidden Markov Models.
6. C. M. Glenn, D. Mandloi, K. Sarella, and M. Lonon, *An Image Processing Technique for the Translation of ASL Finger-Spelling to Digital Audio or Text,* NTID International Symposium Instructional Technology and Education of the Deaf," Conference Proceedings, June 2005
7. M. Kadous, *Grasp: Recognition of australian sign language using instrumented gloves,* 1995.
8. D. Yarowsky, *Gesture recognition using recurrent neural networks*, pp. 237–242, Journal of the ACM, January 1991.
9. *ASL Fingerspelling Conversion*, Where.com (www.where.com).
10. Alkoby, K., and Sedgwick, E., *Using a Computer to Fingerspell*. DeafExpo 99, San Diego, CA, November 19-22, 1999.
11. Furst, J., et.al, *Database Design for American Sign Language*. Proceedings of the ISCA 15th International Conference on Computers and Their Applications (CATA-2000). 427-430.
12. K. Fukunaga and W. Koontz, *Application of the karhunen-loeve expansion to feature selection and ordering*, IEEE Trans. Comput., vol. C-19, no. 4, pp. 311–318, April 1970.
13. C. M. Glenn, M. Eastman, and G. Paliwal, *A new digital image compression algorithm base on nonlinear dynamical system*, IADAT International Conference on Multimedia, Image Processing and Computer Vision, Conference Proceedings, March 2005.
14. Divya Mandloi, *Implementation of Image Processing Approach to Translation of ASL Finger-spelling to Digital Text*, January 2006.