

A TUTOR FOR THE HEARING IMPAIRED (DEVELOPED USING AUTOMATIC GESTURE RECOGNITION)

Ila Agarwal¹, Swati Johar² and Dr. Jayashree Santhosh³

¹Software Developer, Volant Trading, New York, USA.

ila.agarwal@gmail.com

²Computer Scientist, Defence Institute of Psychological Research, Defence Research and Development Organization, New Delhi, India.

joharswat@gmail.com

³Faculty, Computer Services Centre, Indian Institute of Technology, New Delhi, India.

jayashree@cc.iitd.ac.in

ABSTRACT

Automatic gesture recognition could be used to provide feedback in a computer-based learning environment to practice sign language skills for the deaf. In the present work the gestures made by the deaf as a communication aid, as per the Indian Sign Language were captured and a method was evolved to recognize them as English alphabets. After segmenting and removing the noise, the gestures were recognized in two stages. In the first stage, a common feature vector was calculated for all the alphabets, which captured the properties of overall shape of the gesture. If the gesture passed the first stage, then in the second stage a feature vector was calculated for that part of the gesture that actually represented the alphabet. The approach was evaluated by training the module on data of five users and testing on data of another five users. The dataset was characterized by varied user characteristics including clothing and skin tones. The module is user-independent and could be used for education as well as for entertainment purposes.

KEYWORDS

Gesture Recognition, Hearing Impaired, Indian Sign Language (ISL), Color Segmentation, Shape Descriptors

1. INTRODUCTION

Gestures can originate from any bodily motion or state. Gestures can be made with your hands, face, arms, shoulders, legs, feet or a combination of these but hand gestures are probably the most common. Gesture Recognition is a topic in computer science with the goal of interpreting human gestures via mathematical algorithms. McNeill and Levy [1] note that gestures have a preparatory phase, an actual gesture phase, and a retraction phase. The present work uses the actual gesture phase which conveys maximum information out of the three phases.

Sign Language is a communication system using gestures that are interpreted visually. Many people in deaf communities around the world use sign language as their primary means of communication. These communities include both deaf and hearing people who converse in sign language. But for many deaf people, sign language serves as their primary language, creating a strong sense of social and cultural identity.

Linguists have found that sign languages and spoken languages share many features [2]. Like spoken languages, which use units of sounds to produce words, sign languages use units of form. These units are composed of four basic hand forms: hand shape, such as an open hand or closed fist; hand location, such as on the middle of the forehead or in front of the chest; hand movement,

such as upward or downward; and hand orientation, such as palm facing up or down. Indian Sign Language (ISL) or Indo-Pakistani Sign Language is possibly the predominant sign language variety in South Asia used by at least several hundred thousand deaf signers (2003) [3][4]. It uses both hands to form alphabets of the English language.

It has been estimated that in India, the deaf population is approximately 3 million and ninety percent of deaf children are born to hearing parents who may not know sign language or have low levels of proficiency with the sign language. In line with oralist philosophy, deaf schools attempt early intervention with hearing aids etc, but these are largely dysfunctional in an impoverished society.

To our knowledge, no software is currently available to make the children practice the skills of ISL. Moreover, the software available would prompt children to mimic signs but do not provide a feedback to the child. The present work aims to give an evaluation to the child about the correctness of the sign performed, hence making the module a better and interactive tutor.

In earlier work on gesture recognition, Bowden et al. [5] proposed the use of a 'two stage classification' that does not involve the use of Hidden Markov Models. This approach can be used to extract a high level description of hand shape and motion which is based upon sign linguistics and describes actions at a conceptual level easily understood by humans. Pavlovic [6] developed a systematic analysis of different aspects of gesture interaction. According to this system a gesture model is the one whose parameters belong to a parameter space which contains the position of all hand segment joints and fingertips in a 3-D space. Orientation histogram was used as a feature vector for gesture classification and interpolation by Freeman and Roth in American Sign Language (ASL) [7]. The recognition was relatively robust to changes in lighting and based on a pattern recognition technique developed by McConnell [8]. Cristina et al. [9] presented a real-time algorithm to track and recognize single hand gestures for interacting with a videogame. A system developed by Shahzad [10] presented the implementation and analysis of a real-time stereo vision single hand tracking system that can be used for interaction purposes. In real time, the system can track the 3D position and 2D orientation of the thumb and index finger of each hand without the use of special markers or gloves, resulting in up to 8 degrees of freedom for each hand. The approaches mentioned above present recognition of gestures made using one hand and are useful in recognizing a particular gesture, i.e. differentiating between gestures of different alphabets. However, to develop a tutor providing feedback one needs to develop a scheme that moves one step ahead from recognizing, to verifying the correctness of the gestures. Hence the present study used a two-stage process as explained in rest of the paper.

Present work recognized 23 alphabets of English language leaving out H, I, J. Students wore red colored gloves, which had black color on the wrists. In this method, first the videos of the users forming the gestures were captured and from these videos stationary frames representing the gesture were manually extracted. These frames formed the database of gesture samples in the form of *png* images. Color segmentation used the RGB space and after noise removal the gestures were recognized and results were evaluated by testing on five users.

The paper is organized as follows. Section II of the paper presents the system description. Section III describes image acquisition and database. Section IV discusses the algorithm for segmentation of the hands and noise removal. Section V and VI give the feature extraction and recognition. Finally, results are given in section VII and section VIII describes the conclusion and future work.

2. SYSTEM DESCRIPTION

Fig.1 shows the block diagram of the system representing the steps involved in processing of the gestures. On a screen user was shown english alphabets from A-Z . User performed the gesture of same alphabet in ISL. The first step was *Image Acquisition*. A camera was setup facing the user to collect data. A video was taken in which subject mimiced the gestures. Video was manually processed to extract static images. These static images were fed into the *Segmentation and Noise Removal module*. In this stage only those pixels of the image that represented red gloves forming the gesture were extracted. It was observed that segmentation often left some red color blobs along hand region considered as noise. Thus additional step to remove that noise was required. These segmented images were passed to the *Feature extraction and Recognition Stage I and II*.

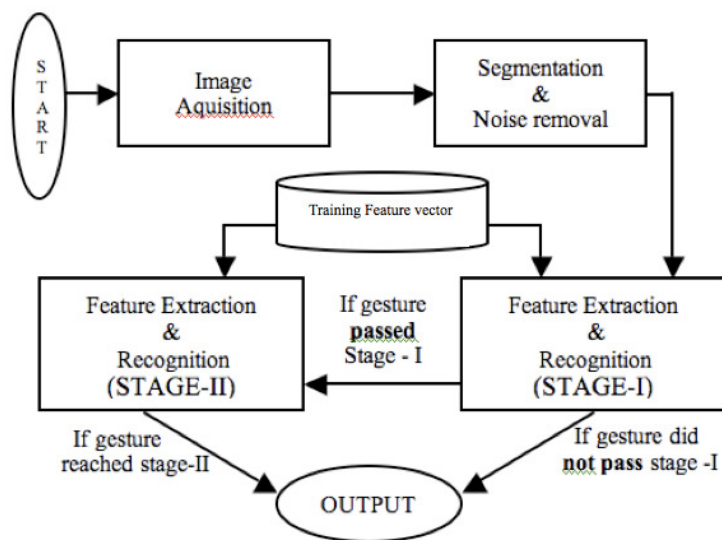


Figure 1. Block diagram of the system

In stage-I, gestures with high degree of errors were removed and stage-II was used to recognize computationally intensive features. In stage-I the feature vector describing the overall shape was calculated and recognition was done by comparing against the thresholds predetermined by the *training feature vectors (TFV)*. In stage-II, recognition criteria was more rigid and feature vector had binary coefficients. Finally, an output was given classifying the gesture as correct or incorrect along with the description of the mistake made by the user. Before the system could be used for tutoring training feature vectors were required. A set of 5 trained ISL users were asked to use the system with constraints of recognition. The feature vectors calculated in stage I and II comprised the training feature vector set.

3. IMAGE ACQUISITION AND DATABASE

Fig. 2 shows the setup created for image acquisition. The set-up was prepared with the digital camera, canon 7.1 megapixel, placed at a distance of 45 cm from the subject facing the subject. The subject was required to wear the red color gloves. Latex gloves were used as they were thin and stuck along the skin surface and took the shape of the hand. Therefore, not interfering or altering the boundaries of the gesture formed. The gloves were made complete red on the palm and black on the wrist to aid segmentation of the gesture. On the screen (Fig 2) subject was shown english alphabets from A-Z and performed gestures for those alphabets in Indian Sign Language. A continuous video of the subject was taken while he performed the gestures. Zoom in

of camera was set to normal to get best image quality. Illumination was kept uniform in the room to avoid reflection into the camera from the gloves surface due to point light sources. Resolution was set to 72 and size of each frame was 1000 X 720 . Static colored images were segmented from the video by manual processing. One frame was chosen per alphabet depicting the actual gesture phase [1]. Images extracted were saved in PNG format to avoid data loss due to image compression. Each image was resized by a magnification factor of 0.5 to reduce the computation time using bilinear transformation.

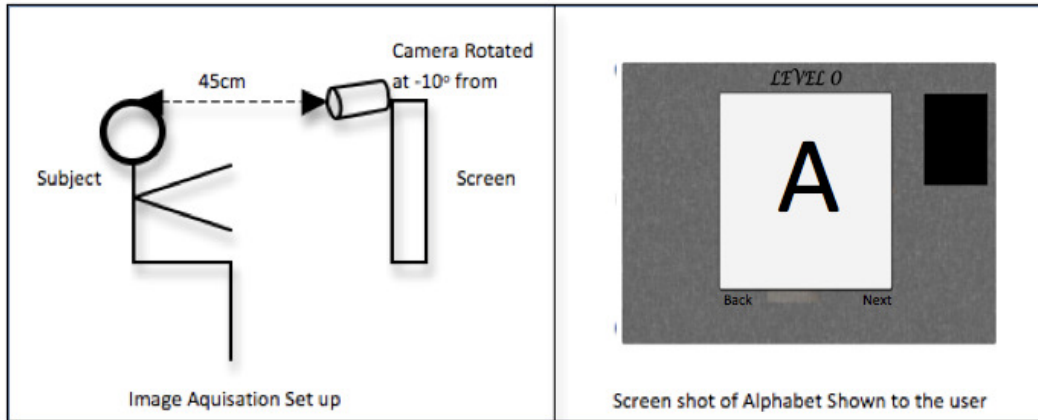


Figure 2: Image Acquisition System

A training database and test database of 120 static images in each was created. Each database comprised of five users and one image per english alphabet for each user. Fig. 3 shows the four samples from the database for images of alphabets A and B.

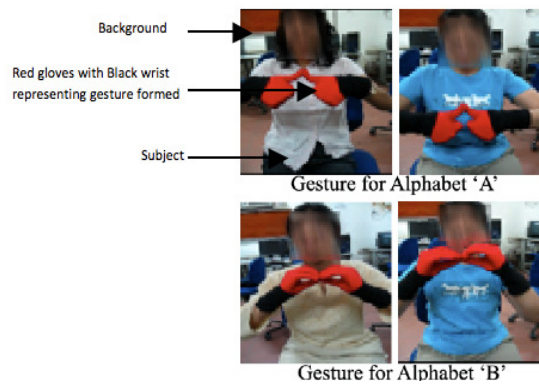


Figure 3. Sample database

4. SEGMENTATION

The image captured (Fig. 3) shows the red gloves representing the gesture and the subject. The background and subject information is not useful for recognition and must be thrown away. Thus pixels representing the hand were be localized in the image and segmented from the background before recognition. In segmentation procedure, usually a variety of restrictions are imposed on - background, user and imaging [6]. Thus, determination of a segmentation cue was critical as to keep the restrictions minimum. Using the skin color is the most natural approach, but it could not

be used as the face regions were also captured along with hands. Eliminating the face by capturing the region below the neck was not considered because some Indian Sign Language comprises of gestures for infront of the face. Another most common approach used is data gloves. Kadous [11] demonstrated a system based on powergloves to recognize a set of isolated Australian sign language. Matsuo et al. [12] used the similar method to recognize signs from Japanese sign language. These setups were connected with high costs. Yoshio [13] used 12 different colors but these colors had to be eliminated from the background. Thus we use single color (bright red) gloves which dealt with above problems. Smooth texture of the gloves was ensured so that texture does not interfere with the edges of the fingers.

Classical approach for color segmentation involves usage of HSI (Hue Saturation Intensity) model. However, it had following problems 1) The HSI model provides greater discrimination between colors [14] and thus in the presence of shadows, when the red color tends to have a black tint the results are poor. 2) Using red color, HSI model gives same value of red close to 0 and 2π radians. So it becomes difficult to learn the distribution due to the hue angular nature that produces samples on both limits [9]. 3) When the saturation values are close to 0, HSI space gives unstable hue and can cause false detections [9]. Therefore, we used the RGB color space model technique for segmentation and determined the thresholds for B and G component using the conversion equations of HSI to RGB. This model allows overlapping shadows to be detected, thus increasing the accuracy of segmentation [14]. Also, the default color space of images is RGB, thus, making it better to use the RGB model for segmentation.

Pure gloves were used with insensity of image almost RGB(240,0,0) under ideal lighting conditions. This intensity could be modified during the real time process due to lighting or shadows. However, intensity of GB channel would not increase drastically unless there was a white spot caused due to reflection of light from the gloves. Keeping uniform lighting without point light sources eliminated such reflections. Since the system was to be used realtime a fast segmentation algorithm based on thresholding of intensity in R,G,B channels was devised. Section 4.1 defines the step by step procedure where $I_G(x,y)$, $I_B(x,y)$ and $I_R(x,y)$ define intensity in the green , blue and red channel respectively on a pixel (x,y). The intensity range was from 0 to 255.

4.1 Segmentation Algorithm

As a preliminary step all pixels that were light colored were removed. Then pixels with $I_G(x,y) > 100$ or $I_B(x,y) > 100$ or high G and B values were set to RGB (0, 0, 0) because a region with such high GB component could not be purely red before image capture. Fig. 4 shows histogram of BG and R channel of the pixels that represented the gloves. It also compares the variation of intensity of the individual channels against each other under both illuminated and shadow region. It was observed that the GB channel were always lower than the B channel. Thus all pixels with $I_B(x,y) > I_R(x,y)$ and $I_G(x,y) > I_R(x,y)$ were set to RGB (0, 0, 0). Fig. 4 also shows that under good illumination the region with R component is higher than or equal to 200. However for the regions of the gloves that lied in shadows of other parts of the hand, the R component decreased but the G and B values also fell sharply, thus this ratio was used to give such pixels a R-value of 200. Then all pixels with $I_R(x,y) > 200$ comprised the final segmented image.

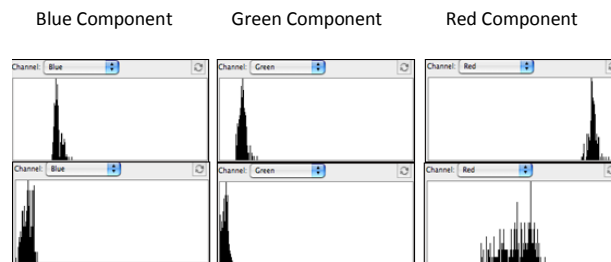








Figure 4. Top 3 histograms showing well illuminated region
Bottom 3 histograms showing shadowed region

4.2 Noise Removal

Anything apart from the hand(s) forming the gesture was considered as noise. The segmentation algorithm generated images but with erroneous and discontinuous regions often around the hand. Sometimes when only one hand was being used to form the gesture the second hand would come in the picture creating noise. Thus following steps were used to remove the noise from gray segmented image:

- Pixels were labelled using 8 connectivity [15]. All pixels that got the same label were considered as one blob.
- The blobs touching the boundary were rejected.
- Total number of pixels (P_{Total}) were found in the segmented image with $I(x,y) > 0$. Now number of pixels in each blob were calculated (P_i). For each i th blob where $P_i \ll P_{Total}$ was rejected from segmented image.
- Regions close to red, such as lips and cheeks of the user were removed by subtracting the background from the image and removing the pixels close to white.
- Only the topmost blob was considered as the final segmented image.

Also few pixels in sharply projected region of the hand may be lost due to reflectance. A 5x5 median filter was used to recover them and smoothen the boundaries and texture, while preserving the edges. The results of segmentation and noise removal step are as shown in Fig. 5. In image of alphabet C we observed the lower hand being removed by noise removal. In image of alphabet F we observe a small patch that was removed by segmentation was filled in by median filter.

Original image	Segmented hand image RGB Image	Segmented hand image Gray Image
 <p style="text-align: center;">D</p>		
 <p style="text-align: center;">K</p>		

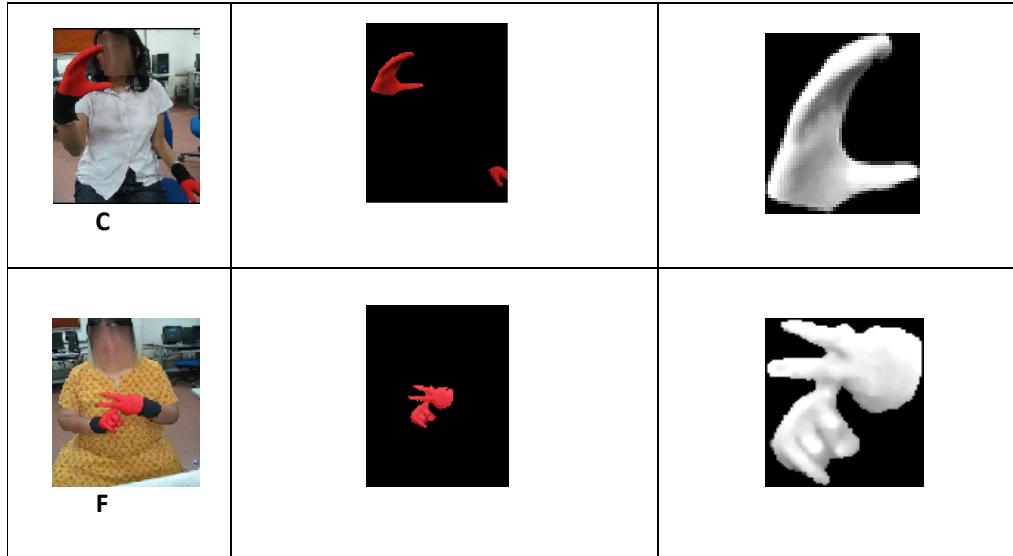


Figure 5. (a) shows the captured images of alphabets D, K, C and F.
 (b) Segmented image (c) After removing noise.

Only the boundary information was useful in recognition. A sobel filter was applied to extract the edges of the image. These edges served as input for feature extraction phase-I.

5 STAGE-I FEATURE EXTRACTION AND RECOGNITION

5.1 Feature Extraction

A feature vector describing the overall shape of the segmented edges was extracted. It contained following features:

- *Convex hull*: A slight variation of descriptors defined in section 2.3 of [16] was used, which included eccentricity, normalized lengths of principal axis and solidity of the convex hull. All features were calculated using 8 connectivity. . Experimental results show that using convex hull instead of original input stroke improves the precision of primitive shape recognition.
- *Aspect Ratio*: Ratio of height to width of the bounding box. It was strong in demarcating tall and wide gestures.
- *Number of points where the boundary curve changed*: k-curvature algorithm [10] is used taking k as 20 and $\theta = 140$. Usually many peaks and valleys are obtained in a single region . These are required to be suppressed. The total no of suppressed peaks and valleys were used as a feature. Fig. 6 shows the peaks and valleys for alphabet A.
- *Holes*: Holes are black region enclosed by a gray region. Those larger than threshold were considered.
- *Centre of Mass*: The 2D coordinates of centre of mass were particularly useful in rejecting features formed in opposite directions or with wrong hands.

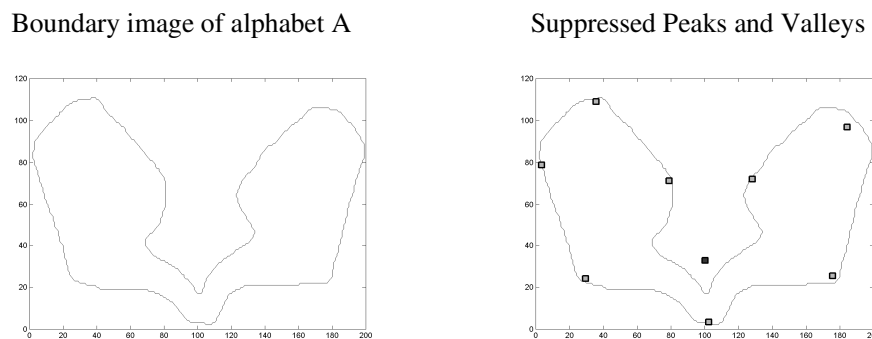


Figure 6. Results for peak and valley detection for alphabet A

The time complexity to calculate all the features is $O(n)$ time suitable for realtime feature extraction.

5.2 Recognition

Recognition involved comparing feature vectors of an input image and a base image. If the distance between two feature vectors was within a threshold range, then gestures were considered matched.

Threshold values were set by using values obtained by 5 images for each alphabet from the training database. Lower and upper thresholds were calculated as follows:

If ' f ' is the feature vector for one alphabet then for each coefficient of the vector

$$\text{variance}_f = \text{variance of feature vector}(f)$$

$$\text{ThreshLower} = \text{lowest value} - \text{variance}_f$$

$$\text{ThreshUpper} = \text{highest value} + \text{variance}_f$$

Testing involved the calculation of feature vector for each image and each coefficient was checked to lie between the corresponding *ThreshUpper* and *ThreshLower*. Even if one coefficient failed to lie with in the range, the gesture was classified as incorrect. If the gesture passes stage-I it was sent to stage-II.

6 STAGE-II FEATURE EXTRACTION AND RECOGNITION

In stage-II more specific features were calculated and recognition was done on the basis of absence or presence of feature. Either the edge image or the gray segmented image was used to calculate the features.

6.1 Feature Extraction

Feature extracted described that part of the gesture that actually formed the alphabet. Following features were calculated.

- *Peaks and valleys*: Using the algorithm described in [10] taking k as 25 and $\theta = 75$ degrees sharp peaks or position of the fingertips were calculated. This was used for set of alphabets like A, C, E, F, L, T, W, X, Y, Z.

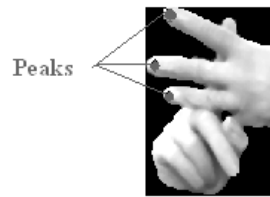


Figure 7. Showing Peaks in alphabet E

- *Hough Transform*: As described in [17]. It is used to detect the orientation of the longest line with the vertical axis in the segmented hand region. This was used in alphabets like P and Y. Fig. 8.



Figure 8. Longest Line in Alphabet P using Hough Transform

- *Edge Map*: To detect whether the back of the wrist or front was presented in the gesture, edge map was used. Leaving out the boundary, summation of length of all the other edges was normalized and edge map was characterized as sparse or dense. To obtain all edges except the boundary the result of sobel operator was subtracted from result of canny operator. This was used for alphabets like A, S, G, Q. Fig. 9 shows a sparse edge map for alphabet A depicting it was a back handed alphabet.



Figure 9. Edge Map for Alphabet A

- *Chain Codes*: They were calculated as described in [18]. These were used to detect if the front or back of the hand is projected in gestures that presented a spread hand. This was used for set of alphabets like D, E, F, U, V, M, N. Where as in alphabets like L, K, T, Z they are used for determining the shape.
- *Region of Edges*: In alphabets like M, N, R, U, V the edges in region where fingers were placed above the palm were extracted. As shown in Fig. 10 for alphabet R.

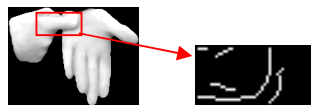


Figure 10. Edge extracted for Alphabet R.

- *Shape of holes*: Three kinds of shapes were determined Circular (B,O,P), Rectangular (D,U) and triangular (V) as shown in Fig. 11 . For alphabet B which has two holes, it was also determined if they were approximately in the same vertical line.



Figure 11. Shape of holes

- *Smoothness of the curve*: To find out the area of fingers in gestures like G, S, Q where closed fists were presented we used opening morphological operation to find the smoothness of the curve.

6.2 Recognition

For recognition, features were calculated one by one and if any feature was not present image was discarded preventing further computation. Based on this image was classified as incorrect and a feedback was provided. MATLAB 7.0 was used for the implementation.

7 RESULTS

The system was tested for 115 gesture images. Table 1 shows the results obtained and Fig. 12 the sample results:

Table 1. Results Obtained

Alphabet	% rejection of correct samples	% acceptance of incorrect samples
A-Z (except H, I, J, S, G, W)	0%	0%
S, G, W	0%	1.6%

7.1 Efficiency of the System

The system gave accurate results for the images in database. The probability that a genuine gesture is rejected or not recognized was very less. Algorithms developed were simple and gave good experimental results. The processing was done on a small selected area. This reduced the processing time to a great extent with the entire process taking time in order of milliseconds to complete the recognition on a 2 Hz intel core 2 Duo processor , 1 GB 667 MHz DDR2 SDRAM.

8 CONCLUSION AND FUTURE WORK

In this study, the tutor developed is designed as a single user interface module, which is simple and easy to learn as well as easy to use. It gave highly reliable results in an environment well illuminated by artificial lights, when tested on database of five users. The images of the gestures that serve as the input to the module are manually obtained from avi videos of continuous gestures. The time taken for processing was suitable for real time processing as the computationally intensive features were calculated only when required. The advantage of using a two-step recognition approach is that the module can be used for both recognizing and verifying

the gestures. Moreover, latex gloves were used, which increased cost-effectiveness of the system. The GUI built presented challenging environment to the user by keeping their scores, providing entertainment along with education and communication means.

However, it used red colored gloves as the segmentation cue, imposing a constraint that user must not be wearing red color clothes. Also due to manual extraction of static frames from the gestures the system could not be fully automated and only individual alphabets could be learnt. Overcoming these constraints presents the basis for the future work.

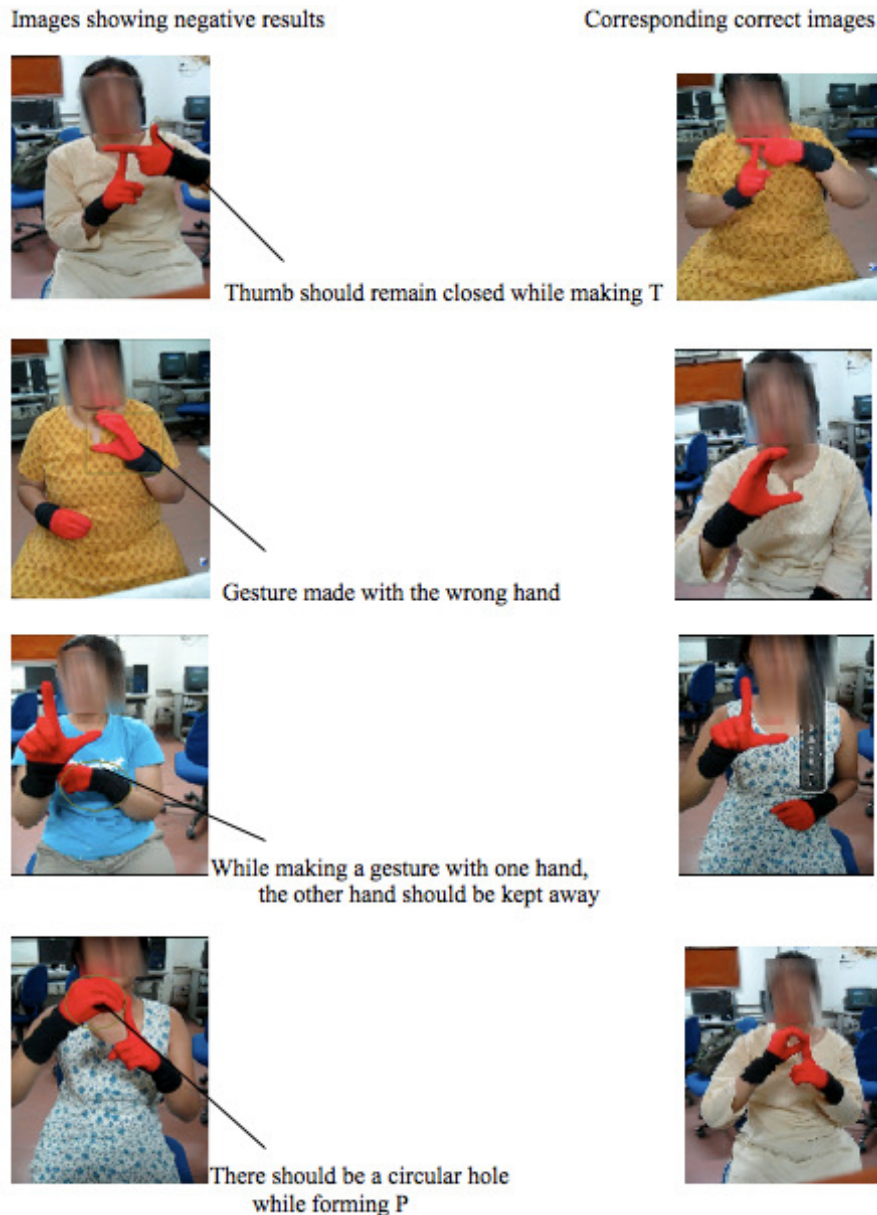


Figure 12. Screen shot of sample output

Usage of stripes of red green blue on the gloves can relax the constraint from red color clothes to clothes with alternate red green blue stripes as the later is less common. Secondly, integrating a capturing device with the module, it could be possible to perform word recognition (i.e sequence of gestures), using motion segmentation and hand tracking. In this future process, the image frames will be captured at 10 frames per second. The constant background can be deleted from the frames reducing the image size. Now, only those image frames will be of interest in which the hand region is fairly stationary, with respect to hand region in other frames. This can be determined by comparing the segmented images pixel by pixel. The frames with change less than a threshold for 3 continuous frames will be considered as a static image. Now, our recognition algorithm can be applied to the frame/static image thus obtained. The GUI can now have three levels for the user: first level for the alphabets, second for dictionary words and last for random spelling words. The system can be extended to incorporate non-standardized gestures as well. Further, the system could be modified to recognize other gestures like numerals in addition to alphabets and so on to be a self-sufficient tutor.

ACKNOWLEDGEMENTS

The authors are deeply indebted to Mr. Akash Tayal (Lecturer, Electronics department, IGIT Delhi) for his stimulating suggestions in all the phases of the project. The authors express their gratitude to the teacher at Delhi School of Deaf for teaching us the correct sign language. The authors profoundly thank all the students who participated and helped to gather data for the study.

REFERENCES

- [1] D. McNeill and E. Levy. "Conceptual Representations in Language Activity and Gesture", pages 271-295. John Wiley and Sons Ltd, 1982.
- [2] "Sign Language Gestures", Informational article resource for Sign Language, http://www.prairiedogmarketing.com/Sign_language_gestures.htm#
- [3] Vasishta, M., J. C. Woodward, and K. L. Wilson (1978). "Sign Language in India: Regional Variation within the Deaf Population". *Indian Journal of Applied Linguistics* 4 (2): 66-74.]
- [4] Ethnologue gives the signing population in India as 2,680,000 in 2003. Gordon, Raymond G., Jr. (ed.) (2005). *Ethnologue: Languages of the World, Fifteenth edition.*. Dallas, Tex.: SIL International.
- [5] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. "A linguistic feature vector for the visual interpretation of sign language". In *Proc. ECCV*, pages Vol I: 390-401, 2004.
- [6] Vladimir I. Pavlovic, Rajeev Sharma, Thomas S.Huang, "Visual Interpretation of Hand Gestures for Human – Computer interaction: A review", Department of Electrical and Computer engineering, University of Illinois.
- [7] William T. Freeman, Michal Roth, "Orientation Histograms for Hand Gesture Recognition", TR94-03 December 1994.
- [8] R. K. McConnell. "Method of and apparatus for pattern recognition". U. S. Patent No. 4,567,610, Jan. 1986.
- [9] Cristina Manresa, Javier Varona, Ramon Mas and Francisco J. Perales, "Real-Time Hand Tracking and Gesture Recognition for Human-Computer Interaction, Electronic Letters on Computer Vision and Image Analysis" 0(0):1-7, 2000.
- [10] Shahzad Malik, "Real- time Hand Tracking and Finger Tracking for Interaction" CSC2503F Project Report. 18. December. 2003.

- [11] M.W. Kadous, "Machine recognition of Auslan signs using PowerGloves: towards large-lexicon recognition of sign language", Proc. Workshop on the Integration of Gesture in Language and Speech, pp. 165-174, 1996.
- [12] H. Matsuo, S. Igi, S. Lu, Y. Nagashima, Y. Takata, and T. Teshima, "The recognition algorithm with non-contact for Japanese sign language using morphological analysis", Proc. Int'l Gesture Workshop, pp. 273-284, 1997.
- [13] Yoshio Iwai, Ken Watanabe, Yasushi Yagi, and Masahiko Yachida, "Gesture recognition Using Colored Gloves", Department of Systems Engineering, Osaka University, Toyonaka, Osaka 560,
- [14] Tze Ki Koh, Nicholas Miles, Steve Morgan, Barrie Hayes-Gill, "Image segmentation using multi-coloured active illumination", Journal of Multimedia, Vol. 2, No. 6, November 2007.
- [15] Rafael C. Gonzalez, Richard E. Woods. "Digital Image Processing", pages 644-645. Prentice Hall Inc, 2001.
- [16] Jukka Iivarinen, Markus Peura, Jaakko Särelä, and Ari Visa Helsinki, "Comparison of Combined Shape Descriptors for Irregular Objects", University of Technology Finland Lappeenranta, University of Technology, Finland
- [17] D. P. Argialas and O. D.Mavrantza, "Comparison of Edge Detection and Hough Transform Techniques for the extraction of Geologic Features", Laboratory of Remote Sensing, School of Rural and Surveying Engineering, National Technical University of Athens,Athens.
- [18] Abdoloh Chalechale, Golshah Naghdy, Prashan Premaratne, and Alfred Mertins, "Cursive Signature Extraction and Verification", School of Electrical, Computer, Telecommunication, Engineering University of Wollongon, Australia.