

Low cost approach for Real Time Sign Language Recognition

Matheesha Fernando, Janaka Wijayanayaka

Abstract – Sign Language is the language of people who suffer from speech and hearing defects. Still the rest of the world doesn't have a clear idea of sign language. The communication between speech impaired people and other people is very inefficient. To overcome this problem technology can act as an intermediate flexible medium for speech impaired people to communicate amongst themselves and with other individuals as well as to enhance their level of learning / education.

The suggested solutions in the literature for sign language recognition are very expensive for day to day use. Therefore, the main objective of this research is to find out a low cost affordable method of sign language interpretation. This paper discusses the possible ways to deal with the sign language postures to identify the signs and convert them into text and speech using appearance based approach with a low cost web camera. Further this approach will be very useful to the sign language learners to practice sign language.

During the research available human computer interaction approaches in posture recognition were tested and evaluated. A series of image processing techniques with Hu-moment classification was identified as the best approach. The system is able to recognize selected Sign Language signs with the accuracy of 76% without a controlled background with small light adjustments.

Key words – Sign Language Recognition, Contour matching, Hu-moments, YCrCb color space.

I. INTRODUCTION

More than 360 million of the world population suffers from hearing and speech impairments [1]. Sign Language is the language of hearing and speech impaired people. Generally normal people in the society don't know about sign language. In order for an ordinary person to communicate with hearing and speech impaired people, a translator is usually needed to convert the sign language into natural language and vice versa.

The idea of this research is that technology can act as an intermediate flexible medium for hearing and speech impaired people to communicate amongst themselves and with other individuals to enhance their level of learning / education. With the advancement of the technology, a perfect answer for the communication problem of hearing and speech impaired people will be achieved in the near future. When considering the order of activities in sign language interpretation by a human interpreter, he

- See the visual sign
- Identify the sign
- Convert the meaning of the sign into the spoken language
- Speak out the signed message

When automating this process it needs to focus on capturing the visual sign presented by the hearing and speech impaired person. Then have to go through a series of processing and identification process. For that the available technology in the field of image processing is studied.

The available solutions suggested for this problem which will be discussed in the literature are expensive and not affordable for day to day use in social communication. The objective of this research is to identify a low cost, affordable method that can facilitate hearing and speech impaired people to communicate with the world in more comfortable way where they can easily get what they need from the society and also can contribute to the well-being of the society. Another expectation is to use the research outcome as a learning tool of sign language where learners can practice signs.

The research is focused on appearance based approach for recognition with a low cost web camera. Therefore the research problem on recognition increases with low quality, noisy web camera inputs of hand postures.

The rest of this paper organized as follows. Section II briefly introduces the area of study with related works in sign language recognition and human computer interaction. In section III the methodology used is described in detail. Section IV describes the methodology for feature extraction with the experiments results. Section V describes the methodology for classification with the methods followed. Section VI reveals the limitations and drawbacks of the used approach. In section VII possible future improvements are introduced and in section VIII a conclusion about the research findings in presented.

II. RELATED WORK

With the development of information technology new areas of human computer-interaction are emerging. There, human gesture plays a major role in the field of human computer interaction. As sign language is a collection of gestures and postures any effort in sign language recognition is in the field of human computer interaction.

According to reference [2] there are two types of approaches commonly used to interpret gestures for human computer interaction. First category, Data Glove based approach relies on electromechanical devices attached to a glove for digitizing hand and finger motions into multi-parametric data. The major problem with that approach is it requires wearing the devices and will cause less natural behaviors. And also these devices are quite expensive.

The second category, Vision based approach in contrast, require only a camera, thus realizing a natural interaction between humans and computers without the use of any extra

devices. There are two approaches to Vision based hand posture, namely 3D hand model based approach and appearance based approach. As reference [3], 3D hand model based approach tries to estimate the hand parameters by comparison between the input images and the possible 2D appearance projected by the 3D hand model.

Reference [2] states that in appearance based method image features are used to model the visual appearance of the hand and then extracted image features from the video input are compared with defined parameters. Appearance based approaches have the advantage of real time performance due to the easier 2D image features that are employed in it. There have been a number of research efforts on appearance based methods in recent years. In reference [4] Freeman and Roth presented a method to recognize hand gestures based on pattern recognition technique employing histograms of local orientation. Some different gestures have very similar orientation histograms which caused to major drawbacks of the method. Orientation histogram method is most appropriate for close-ups of the hand.

In reference [5] Roth and Tanaka have presented a way of classifying images based on their moments with more concern on recognizing shapes of the static postures. In reference [6] Serge and Malik have discussed an approach to measure similarity between shapes and exploit it for object recognition where they attach a descriptor – shape context to each point. The corresponding points in similar shapes will have similar shape contexts. By that they compute the sum of matching errors between corresponding points. This requires more computational power.

Flusser presented a survey on Moment Invariants in Image Analysis on object recognition /classification methods, based on image moments in reference [7]. The paper presents a general theory to construct invariant classification methods. Nianjun, Brian and Lovell have presented an approach to Hand Gesture Extraction by Active Shape Models where a set of feature vectors will be normalized and aligned and then trained by Principle Component Analysis (PCA) in reference [8]. Mean shape, eigenvalues and eigenvectors are computed out and composed of active shape model. Reference [9] presents Chang's approach for Static Gesture Recognition by recognizing static gestures based on Zernike moments (ZMs) and pseudo-Zernike moments (PZMs). This approach takes a step toward extracting reliable features for static gesture recognition. Zhang et al, in reference [10] suggested a Fast Convex Hull Algorithm for Binary Image for pattern recognition. The recognition is achieved by computing the extreme points, dividing the binary image into several regions, scanning the regions existing vertices dynamically, calculating the monotone segments, and merging these calculated segments. Deng and Jason, in reference [11] came up with an idea of shape context based matching with cost matrix for real time Hand Gesture Recognition. There, they translate the edge elements of image shape to a group of feature points with N value.

In reference [12] the ASL is recognized using several feature extraction and machine learning methods such as Histogram technique, Hough transform, OTSU's segmentation algorithm

and a neural network. It uses HSV color spaces and also supports rotation and scale invariant recognition. But its recognition totally depends on the controlled background. In reference [13] a new hybrid approach is introduced to use SURF features with hu moments. Both K-nearest neighbor and support vector machine are used for hybrid classification. And also some other features and classification methods have used to support the system.

Image processing concerned with computer processing of images which includes methods for acquiring, processing, analyzing, and understanding images with high dimensional data from the real world in order to produce numerical or symbolic information. Computer vision requires a combination of low level image processing to enhance the image quality (e.g. remove noise, increase contrast) and higher level pattern recognition and image understanding to recognize features present in the image.

III. SIGN LANGUAGE RECOGNITION

Appearance based method was selected for the research as main objective of the research is to identify a low cost method for sign language recognition. In appearance based method feature extraction and classification are the major components. Feature extraction methods are used to reduce the number of dimensions of an image. A descriptor can be used for that. A descriptor describes an image and if properly used, image can be represented less with dimensions than the image itself. Also it can introduce some useful properties like scale and rotation invariance.

Classification includes a broad range of decision theoretic approaches to the identification of images. It analyses the numerical properties of various image features and organizes data in to categories. The classification represents the task of assigning a feature vector or a set of features to some predefined classes in order to recognize the hand gesture.

As shown in Fig. 1, research is focused on an application that can convert a video signal (processed as sequence of images) into a sequence of written words (text) and speech in real time.

In the real-world, visual information could be very rich, noisy, and incomplete, due to changing illumination and dynamic backgrounds and obstacles, etc. Vision-based systems should be user independent and robust against all these factors. The suggested solution for the communication problem requires real-time facility, with effective as well as cost efficient techniques/algorithms. Therefore robustness, computational efficiency and user's Tolerance were the important challenges need to be considering in conducting this research project.

Overall methodology followed for the sign language classification is summarized in Fig. 2.



Fig 1: Outline of the proposed method

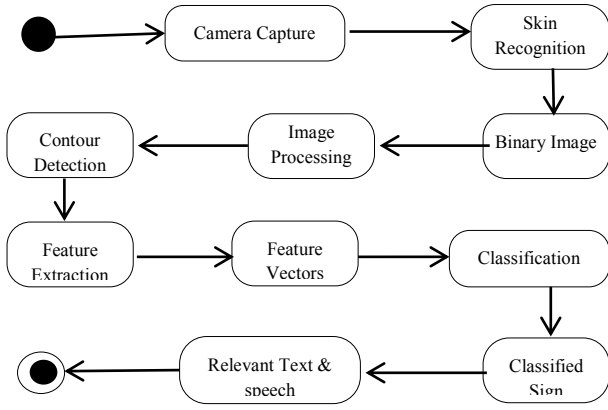


Fig 2: Classification process

In this approach the movement of the hand is recorded by a camera and the input video is decomposed into a set of features by taking individual frames into account. The video frames contain background pixels other than the hand, as the hand will never fill a perfect square. These pixels have to be removed as they contain arbitrary values that introduce noise into the process.

The basic idea is to get the real time image and then extract the predefined features and then compare the feature vectors against the features extracted from stored template sign images.

A. Image Capturing Phase

Only a web camera was used to make this system a low cost, affordable application and that made the capturing process simple. As the normal web camera images are very poor in quality it requires more image processing techniques to deal with the issue.

B. Hand Segmentation

The hand must be localized in the image and segmented from the background before recognition. Color is the selected feature because of its computational simplicity, its invariant properties regarding to the hand shape configurations and due to the human skin-color characteristic values. The method is based on a more rule based model of the skin-color pixels distribution. As reference [14] states in an RGB format, for each pixel in the image Red, Green and Blue color components has a value stating the intensity of that color in the pixel. Since different people have different hand colors it is incorrect to rely on the intensity. Therefore, it is required to convert the RGB representation to an intensity free color space model. Hue-Saturation-Value (HSV) representation and YCrCb color space are the best options available.

The advantage of YCrCb over RGB is that it can reduce resolution of the color channels without altering the apparent resolution of the image, since the main Y (luminance) signal is untouched. YCrCb is derived from the corresponding RGB space as follows,

$$\begin{aligned}
 Y &= 0.299R + 0.587G + 0.114B \\
 Cb &= 0.564(B - Y) + 1/2 \text{ full scale} \\
 Cr &= 0.713(R - Y) + 1/2 \text{ full scale}
 \end{aligned}$$

After number of iterations of experiment the suitable color ranges for human skin in both HSV and YCrCb were figured out as follows,

$$hsv_min = (0, 30, 60), hsv_max = (30, 150, 255)$$

$$YCrCb_min = (0, 131, 80), YCrCb_max = (255, 185, 135)$$

Since human skin color has a wide range, these figures are for more general purpose to consider as default range. Binarizing (converting the image to black and white) is more desirable since it reduces the amount of information contained in the source image. Thresholding can be used in order to binarize the image.

C. Image Pre-processing Phase

Image processing techniques were used to improve the quality of the image. Morphological transformations are used to get a more clear hand image. The basic morphological transformations are called dilation and erosion, and they arise in a wide variety of contexts such as removing noise, isolating individual elements, and joining disparate elements in an image. The skin recognized image was processed with both dilation and erosion to get image clearer as shown in Fig. 3.

IV. FEATURE EXTRACTION

Under feature extraction Hand outline, Hand contour, Multi scale color features, Scale Invariant Feature Transform (SIFT), SURF (Speeded Up Robust Feature), Haar-like features, Histograms of Oriented Gradients (HOG), Local Orientation Histogram, Hough Transform (HT) are some of available research findings in efficient feature extraction.

In the feature extraction phase what is most important is to get possible precise features as output. A very simple method is to compare each pixel location with each other and sum all the differences. This method is not realistic as it is not going to work on images that are not the same size or orientation. And also for the same hand posture there will be different images with small variations. Also it is computationally expensive for larger images. For an image of 100 by 100 pixels there are already 10,000 dimensions. Therefore features selected for classification are hand contour, orientation histogram, convex hull, convexity defects and hu moments.

A. Contour Detection

Detecting contours of the hand image means finding the edges that have high contrast pixels than its neighbors. Output of contour finding is a sequence of boundary pixels. Edge tracking algorithms can be used as a support for contour finding as shown in Fig. 4. In the real world a hand image may have dark places and shadows not only at the ends but also at the middle of the hand. It will result in a set of contours in various lengths.

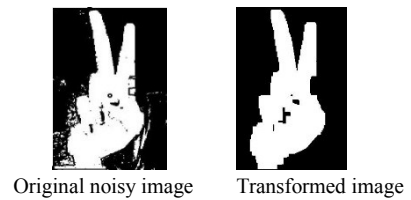


Fig 3: Morphological transformation

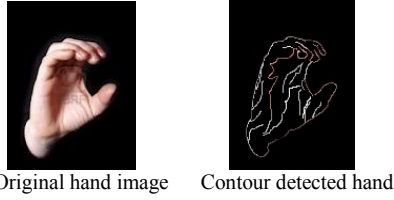


Fig 4: Process of getting a contour

Further in feature analysis the local features are then represented by the local extreme of the outline, whereas there are two different kinds of extreme, the peaks and the valleys. The peaks are usually found at the finger tips, whereas the valleys are rather found in the regions where two fingers join the palm of the hand. Advantage of such features is the quick exclusion of inappropriate gestures, using the number of peaks and valleys as indicators.

The convex hull for the hand shape is computed and the poly line minimum area covering box and the rectangle is identified as shown in Fig. 5.

Then center of the box is computed. The Region of Interest (ROI) of the image is set to the minimum rectangle as shown in Fig. 6. Here, in the algorithm, Region of Interest (ROI) plays a major role as it is the area subjected to matching.

V. CLASSIFICATION

At the first stage of research practicing and testing on available image processing techniques were conducted and the investigations were carried out to identify the best methods and algorithms for sign language recognition. Different methods and algorithms in the literature of human computer interaction such as template matching, contour matching, defects computation and shape context matching were implemented and tested with modifications to apply them into sign language recognition domain as follows. There Normalized template matching was tested with equation (1) as shown in Fig. 7.

$$(x, y) = \sqrt{\sum_{x', y'} T(x', y') \cdot \sum_{x', y'} I(x + x', y + y')^2} \quad (1)$$

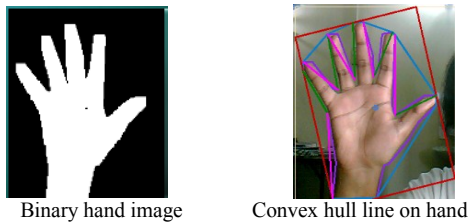


Fig 5: convex hull line

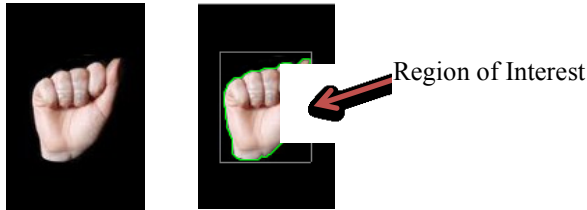


Fig 6: Region of interest identification

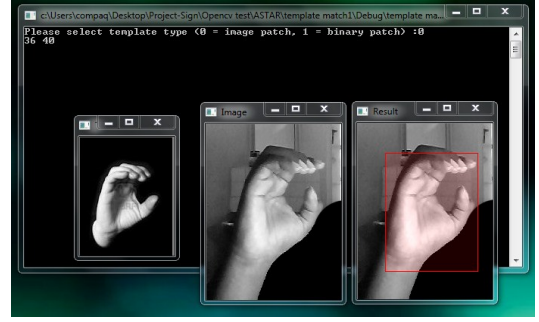


Fig 7: Template matching

As the results were not satisfactory, and the conclusion derived is that template matching is not suitable for real time systems with lot of background noise.

Contour matching was tested where it compared the number of points match at the sequences. To make the contour features more powerful, filtration of hand out line contour was used. The biggest contour by the area that each contour covers was computed for filtering process. As the experiment results shows in Fig. 8 the biggest contour is very much close to hand out line.

Contours can be represented by sequences in which every entry in the sequence encodes information about the location of the next point on the curve. As shown in Fig. 9, the contours were identified for the given image and the template image and then matching was done by comparing number of points match.

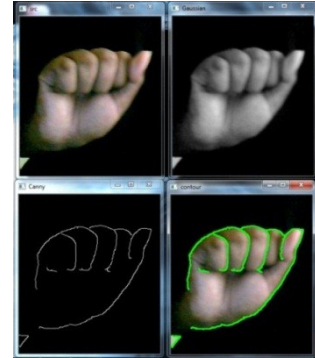


Fig 8: Biggest contour calculation

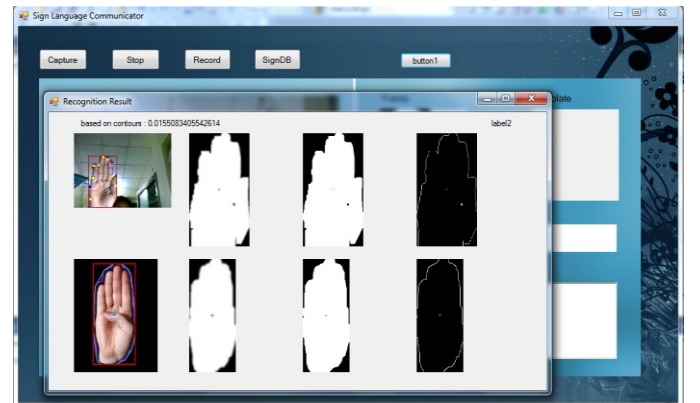


Fig 9: Contour matching

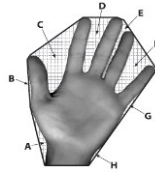


Fig 10: defects of a hand

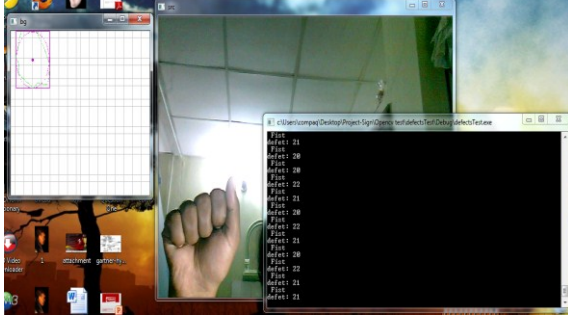


Fig 11: Classification by number of defects

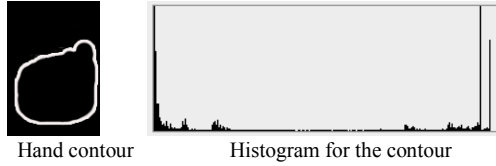


Fig 12: Histograms and contour

Shape match with contour was introduced considering the outer contour as a shape and did the shape matching. These approaches provided some successful results.

Then defects computation for the given real time image was done, where it calculates the total area which is differed from the template, and that was used for classification. By the amount of defects we can get an idea what is the state of the hand as shown in Fig. 10.

As shown in Fig. 11 'src' window shows the real time captured hand sign and 'bg' window shows the hand out line and minimum rectangle and ellipse. This method is limited to few completely different signs.

According to reference [4] Histogram based classification was implemented with binary image, as histograms can be used to represent the color distribution of an object, an edge gradient template of an object etc. Here, the histogram was generated for the given binary images and the histograms were normalized before matching. As the whole image is used for histogram construction the computational efficiency is very less and the real time concept didn't actually work with this model. As a solution, histograms were computed for contours considering the drawbacks of the previous version with the idea to reduce the image data for histogram construction. Histogram for a contour is shown in Fig. 12 bellow.

Then the moment based recognition was implemented and tested according to reference [15]. The general, (p, q) moment of a contour is defined in equation 2.

$$m_{p,q} = \sum_{i=1}^n I(x,y) x^p y^q \quad (2)$$

Here p is the x -order and q is the y -order, where order means the power to which the corresponding component is taken in the sum just displayed. The summation is over all of the pixels of the contour boundary (denoted by n in the equation). Here the test image and the template images are divided into channels of BGR and contours are calculated for each channel and the template image's relevant channel contour shape is compared with the test image's relevant channel contour shape. Then the difference to the template is calculated and the best matching is decided as the lowest difference. Moment based recognition showed more accuracy in identifying the signs than all other attempts mentioned before. Counting on the success of moment based recognition then hu moments were tested for better results.

Hu moments use the central moments where it is basically the same as the moments just described except that the values of x and y used in the equation 3 are displaced by the mean values.

$$\mu_{p,q} = \sum_{i=0}^n I(x,y) (x - x_{avg})^p (y - y_{avg})^q \quad (3)$$

where $x_{avg} = m_{10}/m_{00}$ and $y_{avg} = m_{01}/m_{00}$

The idea here is that, by combining the different normalized central moments, it is possible to create invariant functions representing different aspects of the image in a way that is invariant to scale, rotation. Seven hu moment calculations presented by Hu (1962) in reference [15] are shown below.

$$\begin{aligned} h_1 &= \eta_{20} + \eta_{02} \\ h_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\ h_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\ h_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\ h_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \\ &\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})(3\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \\ h_6 &= (\eta_{20} - \eta_{02})(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\ h_7 &= (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})(3\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \\ &\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})(3\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \end{aligned}$$


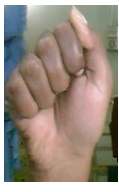




At the matching phase hu moments were calculated for both pre-processed real time image frame and the template signs and best match is identified.

Hu moment matching showed satisfactory level of accurate results than other tested methods. In finding the actual output of hu moment matching, table 1 display some real time generated hu moment matching result values.

As shown in table 1, real time images used have taken in a normal background. As the results are accurate, it derives that hu moment calculations can be used for real time recognition.

For the testing, out of 50 signs, 5 signs from each 10 users (A, B, C, D, V Signs). 8 signs were stored and used as templates including the testing 5 and 3 other signs (A, B, C, D, L, P, V, Y). Tested signs are from the alphabet of ASL(American Sign Language), 12 signs failed to identify clearly. That gives an accuracy level of 76%.

Table 1: Hu moment calculation results

Sign	Template image	Real time image	Result values
A			A :- 0.08956746514951
			V :- 0.27577140714475
			L :- 0.34207054932742
			B :- 0.24181056482163
			D :- 0.37176152384964
			C :- 0.18130743044722
			P :- 0.25942242245184
D			Y :- 0.43210990521291
			A :- 0.49651217137801
			V :- 0.34664529950186
			L :- 0.72065353243200
			B :- 0.55167658949923
			D :- 0.24146058292755
			C :- 0.32495304940430
V			P :- 0.25311353806331
			Y :- 0.42539378543313
			A :- 0.77538997758076
			V :- 0.15136323806656
			L :- 1.0413004951077
			B :- 0.32666217539152
			D :- 0.47280905471127
			C :- 0.52194655877205
			P :- 0.48663355955186
			Y :- 0.69101854910652

VI. LIMITATIONS AND DRAWBACKS

As this is a real time application it is highly depends on the user's behavior. Being an appearance based solution the system is very sensitive to light condition of the environment. In this research only the postures of sign language signs were considered based on the idea that postures are fair enough to identify a certain level of sign language. The spatial information for the hand signs also was not considered in to account for classification and as there are some occasions where same sign posture is made for different meanings with spatial differences.

VII. FUTURE IMPROVEMENT POSSIBLE

Sign language is a very vast area with its grammar and literature component. For the research mainly ASL alphabet is used as the test cases. In future, this project can be advanced to convert words, phrases and simple sentences. Further, this project only looks at the hand postures as it is the major part of sign language and this can be extended to identify hand gestures with spatial information. There, a state machine for gesture recognition can be implemented with a machine learning algorithm where it will result in an intelligent sign language recognition system. Movements of the hand sign can be detected by taking the absolute difference of the center of ROI box.

VIII. CONCLUSION

The research has dealt with the sign language gesture/posture and identified the sign and converted that sign into text and speech. The main objective of the research was to find a best fitting algorithm for sign language posture recognition. There number of methods were designed and

developed and tested and developed an algorithm by combining different methods. There hu moments, contour computation, histograms, convex computation, defects computation was mainly used.

The research was conducted to find out the best approach for a low cost, affordable sign language recognition system. Digital image processing techniques combined with hu-moment classification was found out to be the best approach. The outcome of the project provides an advancement of technology for the communication between a hearing and speech impaired person and a normal person. And also mainly the project can be used as a learning tool for sign language where hearing and speech impaired people can practice sign language using the application.

Reference

- [1] <http://www.who.int/mediacentre/factsheets/fs300/en/>
- [2] Harshith.C, Karthik.R.Shastry, Manoj Ravindran, M.V.V.N.S Srikanth, Naveen Lakshmikanth, "Survey on various gesture recognition Techniques for interfacing machines based on ambient intelligence", International Journal of Computer Science & Engineering Survey (IJCSES) Vol.1, No.2, (November 2010)
- [3] Goh Chong Yang, Chua Hock Chuan, "Marker-Based 3D Gesture Recognition For Interactive Applications", Proceedings of the URECA@NTU 2009-10.
- [4] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition", IEEE Intl. Wkshp. on Automatic Face and Gesture Recognition, Zurich, (June 1995)
- [5] M. Roth, K. Tanaka, C. ssman, W. Yezauris, "Computer Vision for Interactive Computer Graphics", IEEE Computer Graphics and Applications, May-June, 1998, pp. 42-53, (1998)
- [6] Serge Belongie, Jitendra Malik, "Shape Matching and Object Recognition Using Shape Contexts", IEEE Transactions on pattern analysis and machine intelligence, Vol.24, No.24, (2002)
- [7] Jan Flusser, "Moment Invariants in Image Analysis", World Academy of Science, Engineering and Technology, (2005)
- [8] Nianjun Liu, Brian C. Lovell, "Hand Gesture Extraction by Active Shape Models", Digital Image Computing: Techniques and Applications, 2005. DICTA '05. Proceedings, (2005)
- [9] Chin-chen chang, jiann-jone chen, wen-kai tai and chin-chuan han, "New Approach for Static Gesture Recognition", Journal of information science and engineering, (2006)
- [10] Xianquan Zhang and Zhenjun Tang, Jinhui Yu, Mingming Guo, "A Fast Convex Hull Algorithm for Binary Image", Informatica Oct2010, Vol. 34 Issue 3, p369, (2010)
- [11] Lawrence Y. Deng, Jason C. Hung, Huan-Chao Keh, Kun-Yi Lin, Yi-Jen Liu, and Nan-Ching Huang, "Real-time Hand Gesture Recognition by Shape Context Based Matching and Cost Matrix", Journal of Networks, Vol. 6, No. 5, May 2011, (2011)
- [12] Vaishali.S.Kulkarni et al., "Appearance Based Recognition of American Sign Language Using Gesture Segmentation", International Journal on Computer Science and Engineering (IJCSE), 2010
- [13] J. Rekha, J. Bhattacharya and S. Majumder, "Hand Gesture Recognition for Sign Language : A New Hybrid Approach", Proceedings of 15th International Conference on Image Processing, Computer Vision and Pattern Recognition IPCV'11, WorldComp'11, CSREA Press, July 18-21, Las Vegas, Nevada, USA, 2011.
- [14] X. Zabulis, H. Baltzakisy, A. Argyroszy, "Vision-based Hand Gesture Recognition for Human-Computer Interaction", World Academy of Science, Engineering and Technology, (2006)
- [15] Ming-Kuei Hu, "Visual pattern recognition by moment invariants", Information Theory, IRE Transactions on 8(2):179-187, (1962)