Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 54 (2015) 523 – 531

Procedia Computer Science

Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Continuous Indian Sign Language Gesture Recognition and Sentence Formation

Kumud Tripathi*, Neha Baranwal and G. C. Nandi

Robotics and Artificial Intelligence Lab, Indian Institute of Information Technology, Allahabad

Abstract

Hand gestures are a strong medium of communication for hearing impaired society. It is helpful for establishing interaction between human and computer. In this paper we proposed a continuous Indian Sign Language (ISL) gesture recognition system where both the hands are used for performing any gesture. Recognizing a sign language gestures from continuous gestures is a very challenging research issue. We solved this problem using gradient based key frame extraction method. These key frames are helpful for splitting continuous sign language gestures into sequence of signs as well as for removing uninformative frames. After splitting of gestures each sign has been treated as an isolated gesture. Then features of pre-processed gestures are extracted using Orientation Histogram (OH) with Principal Component Analysis (PCA) is applied for reducing dimension of features obtained after OH. Experiments are performed on our own continuous ISL dataset which is created using canon EOS camera in Robotics and Artificial Intelligence laboratory (IIIT-A). Probes are tested using various types of classifiers like Euclidean distance, Correlation, Manhattan distance, city block distance etc. Comparative analysis of our proposed scheme is performed with various types of distance classifiers. From this analysis we found that the results obtained from Correlation and Euclidean distance gives better accuracy then other classifiers.

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Keywords: Correlation; Indian sign language (ISL); Orientation histogram; Principal component analysis; Gesture Recognition.

1. Introduction

Motion[1] of any body part like face, hand is a form of gesture. Here for gesture recognition I'am using image processing and computer vision. Gesture recognition[2] enables computer to understand human actions and also acts as an interpreter between computer and human. This could provide potential to human to interact naturally with the computers without any physical contact of the mechanical devices. Gestures are performed by deaf and dumb community to perform sign language. This community used sign language[3] for their communication when broadcasting audio is impossible, or typing and writing is difficult, but there is the vision possibility. At that time sign language is the only way for exchanging information between people. Normally sign language is used by everyone when they do not want to speak, but this is the only way of communication for deaf and dumb community. Sign language is also serving the same meaning as spoken language does. This is used by deaf and dumb community all over the world but in their regional form like ISL, ASL. Sign language can be performed by using Hand gesture[4] either by one hand or two hands. It is of two type Isolated sign language and continuous sign language. Isolated

*Corresponding author.
E-mail address: kumudtripathi.cs@gmail.com

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)
doi:10.1016/j.procs.2015.06.060

sign language consists of single gesture having single word while continuous ISL or Continuous Sign language is a sequence of gestures that generate a meaningful sentence. In this paper we performed continuous ISL gesture recognition technique for Indian people. Continuous ISL gestures are mostly made up of two handed and also it is a combination of dynamic[5] as well as static gestures. Therefore it isvery difficult to recognize in real word environment.

In continuous SL gesture recognition[6] system preprocessing, key frame extraction[7], and feature extraction are the major issues which have been solved in this paper. Here we proposed a framework which is flexible towards orientation, size and scaling. In continuous ISL gestures, key frames are extracted using gradient method in which we have measured change in x direction as well as in y direction. With the help of key frames each sentence is divided into sequence of words (Isolated gestures). Features of each isolated gestures are extracted using orientation histogram method. After wards, it has been recognized using various distance based classifier. Finally sign language are combined and translated into audio or text format, so that communication will improve between normal people and hearing impaired community. Almost all gesture has already assigned meaning and grammar is used to create a meaningful sentence from set of recognized gestures.

Organization of paper as follows: section 1 tell us about the introduction of gestures, continuous gestures. In second section we explains about analysis of previous research where we explains what are the works already done and what are the drawbacks. Proposed methodology is explained in 3rd section. In 4th section we give experimental results and what are the findings from those experimental results. Section 5 incorporates conclusion and future work of the paper. End of the paper includes acknowledgement and references.

2. Analysis of Previous Research

There are many sign language recognition[8] technique have already been developed and get prominent results but still it's a challenging research field for the researcher. Most of the work is done for isolated sign language recognition. Very few literature is available in the area of Continuous sign language recognition because of its complex nature.

Ankita Saxena[9] *et al.* proposed a fast and efficient technique that is principal component analysis for sign gesture. Here they take 3 frames per second from video and analyses them for static gesture. Overall accuracy of this system is 90%. But this system is highly dependent on background and lighting condition. The author Jung-Bae Kim[10] *et al.* proposed a system that uses Fuzzy Logic and Hidden Markov Model for Korean Sign Language recognition. Using these methods, they Obtain 94% accuracy for 15 KSL sentences. They have rejected meaning less gesture motions such as preparatory motion and useless movement between sign words, using fuzzy partitioning and state automata. They concentrate on two features like speed and velocity for motion of hand. They do sentence based recognition so no need to pause between sign words. This system have high computational burden. Rung-Huei Liang[11] proposed a Data Glove[12] based continuous gesture recognition on real time. For this they are created a large vocabulary Taiwanese sign language interpreter. They solved the problem of key frame extraction using time-varying parameter detection. They detected the discontinuities in frames. They do statistical analysis by four features position, posture, orientation, and motion. For gesture recognition they have used Hidden Markov Model. Average accuracy rate of this system is 80.4%. But limitation of this system is this is person dependent and using gloves which is very expensive and need physical connection between user and computer. To remove few such drawbacks like person dependency, scale, orientation, position dependency we proposed an orientation histogram based framework for continuous ISL gesture recognition.

In this work we present a vision based recognition system with only one camera. Our focus is to solve major problems like key frame extraction, orientation dependency, real time processing. For solving major problem of key frame extraction we used gradient method, in which we see the change of orientation of hand in frames. Because sentence is a collection of meaningful and useless gesture, so need to extract useful gesture by this we reduce computational time and also create database for each gesture separately. This makes my system more flexible by recognizing any new sentence whose gestures are already in database.

3. Proposed Methodology

This paper focuses on the proposed continuous ISL gesture recognition system. Dataset consisting of a collection of signs where single hand or both the hands has been used for performing continuous ISL gestures. Ten sentences

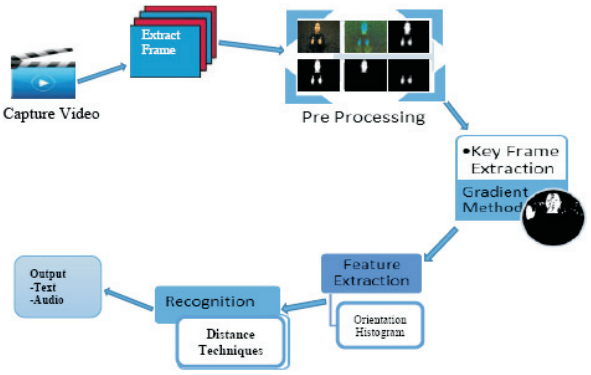

Fig. 1. General diagram of proposed framework.

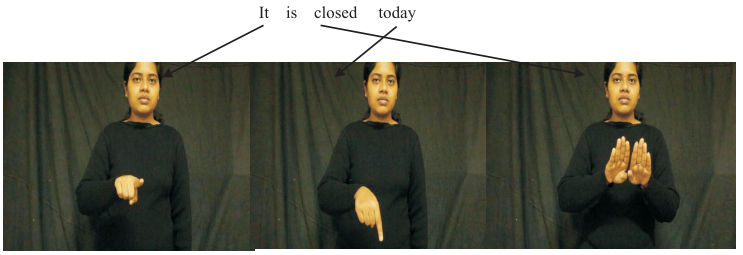

Fig. 2. Gestures of sentence it is closed today.

database has been created. Each sentence consists of two, three and four types of gestures which is shown in Fig. 2. Every sentence is a combination of static and dynamic gestures. Extracting start frame and end frame of each gesture is the main problem in continuous sign language gesture recognition system because it consists of a collection of meaningful gestures and also a vague gestures having no meaning. We deal this problem using gradient based key frame extraction method. Here major change in the gradient shows end of the one gesture and start of another gesture. Key frame helps to break each sentence into sequence of words (isolated gestures) and also obliging for extracting frames of meaningful gestures. Orientation histogram, DWT and PCA is used for extracting features of those frames which comprises of meaningful gestures.General diagram of proposed framework is shown in Fig. 1.

3.1 Dataset acquisition

Dataset has been created using an external camera with the configuration of Canon EOS with 18–55 mm lens, 18 mega pixels, 29 frames per second and resolution is 3920*320 bits/sec. Here we used single camera for creation of gesture dataset. Black background is used for database creation of ISL gestures. Here we concentrate on the upper body part only. Movement of the upper body part is acceptable. Position of camera is very important (camera calibration), for clarity of the dataset and for removing many backgrounds related problems like background noise, body motion etc.

Here we have taken 10 Indian sign language sentences of 5 different people, where each sentence has been recorded 10 times, 6 for training and 4 for testing. Every video is divided into sequence of frames of size 640*480.

Continuous gesture “it is closed today” is made of 3 gestures (“it”, “today” and “closed”) shown in Fig. 2.

3.2 Preprocessing

In this step silhouette images of every hand gestures are created. Here we extracts foreground image from complete image means it removes background of an image and get the skeleton of upper body part. Then hand region are

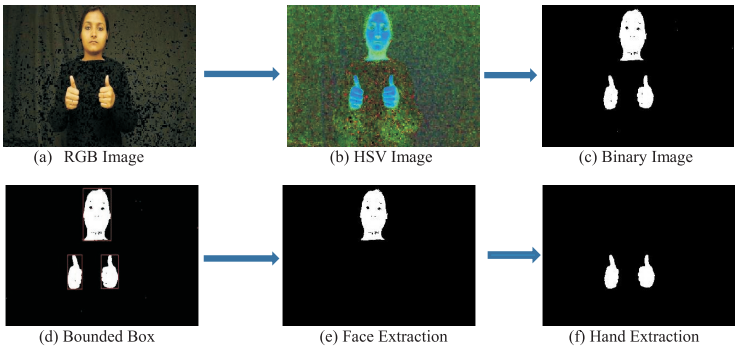

Fig. 3. Pre-processing steps of each RGB frame.

subtracted from these foreground images by eliminating largest connected region which is face. Finally we get the hand portion from upper body.

We first convert each video into sequence of RGB frames. Each frame having dimension 640*480. Skin color segmentation[13] is applied for extraction of skin region which is divided into number of chunks. For finding skin region, each frame is converted into HSV (Hue, saturation, value) plane where only H and S value having threshold ($H > 0.55$ or $S <= 0.20$ or $S > 0.95$) is used for finding non skin region of an image. Then this region mark as zero for extracting skin region. Median filter is applied for preserving outer boundary (edges) of segmented region. It mainly removes salt pepper noise and impulsive noise for edge preservation. Images obtained after median filtering are converted into binary form. At the end of preprocessing subtract largest connected region which is face. Eliminate face region from upper half of the body and will get hand gestures. Each step of pre-processing is shown in Fig. 3 and explained below.

Steps

1. Each video is converted into RGB frames (I). Then RGB frames are converted into HSV plane. Here we consider only H and S part because H and S shows non-skin region. The threshold value of H and S for non-skin region is ($H > 0.55$ or $S <= 0.20$ or $S > 0.95$). Make zeros for all non-skin region in image I and obtained skin colour region in image (I) denoted as I_1.
2. Convert image I into gray form by subtract I_1 image from gray scale of I.
3. Then apply median filter of window size [3, 3] to remove noise from Im image.
4. Calculating the area of each binary region, first we have to label them and then calculate area of each region.
5. Sort areas of all region for extracting largest area. Here we consider face has the largest area. From this extract face region from the whole binary image. Now subtract binary image (bw) from face region to get hands in binary image that is the region of interest.

3.3 Key frame extraction

A finite sequence of frames form each video. In which frames could be any gesture frame or non-gesture frame. So we need to extract those frames that belong to any meaningful gesture and remove meaningless frames because these frame creates an extra affliction of processing. In this paper we used gradient method for extracting key frames of each video sequence. We take frames to do segmentation and calculate gradient of each frame. Key frame extraction graph of "How are you?" gesture is shown in Fig. 3. Graph shown in Fig. 4 shows that gesture1 starts in between frame 0^{th} to frame 4^{th} and end at in between frame 254^{th} to frame 258^{th}. From 0^{th} to frame 4^{th}, there is a constant gradient value and similarly from 254^{th} to frame 258^{th} we obtained constant gradient value which shows end of one gesture or start of another gesture. Suppose if gesture 1 ends at frame 256^{th} than this is the start of next gesture. From this we can calculate the total number of frames in each gesture. 15 frames of each gesture are considered from the middle of the each gesture dataset because in between of frames gestures are clearly visible and identifiable. This method has been used for creating database for dividing each continuous gesture into an isolated gestures.

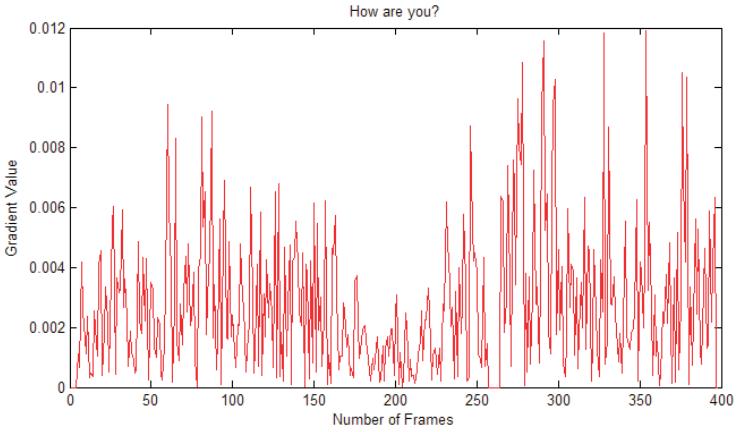

Fig. 4. Key frame extraction of "How are you" gesture.

3.4 Feature extraction

We applied orientation histogram as a feature extraction technique for extracting most appropriate features of each gesture. It provide convenience to even light condition changes and scene illumination changes. The edges of the images of the scene will be still the same. All the continuous ISL gestures have been obtained in normal lighting mode where pixel intensities can be suggested to change the scene lighting. Another advantage which is employed on orientation histogram is translation invariant property. It determines that the same feature vectors will be produced by the same frames at different position of gestures. It is achieved to measure the local orientation histogram for all the frames of the dynamic gestures. Local orientation histogram does not change by the translation of the frame in the gesture. Then dimension of these features are reduced using Principal Component Analysis (PCA). The steps of orientation histogram algorithm are:

i) Subsample the 640*480 image into 60*40 size which reduces space complexity and makes processing time fast.
ii) Finding the edges of an image using 3-tab derivative filter $a = [0\ -1\ 1]$ $b = [01\ -1]$. It helps us for finding the image gradient in a-direction and b-direction.
iii) The gradient in a-direction as well as in b-direction.

$$da = \frac{\delta p(a,b)}{\delta a} = \frac{p(a+1,b) - p(a-1,b)}{2} \tag{1}$$

$$db = \frac{\delta p(a,b)}{\delta b} = \frac{p(a,b+1) - p(a,b-1)}{2} \tag{2}$$

where $p(x, y)$ represents intensity function at (x, y) pixel position.
iv) Find out the gradient direction using arctan2 function which is expressed as:

$$X(a,b) = \text{atan2}\left(\frac{\delta p}{\delta a}\delta p \delta b\right) = \text{atan2}\left(\frac{\delta b}{\delta a}\right) \tag{3}$$

The value of X lies between $[-\frac{\pi}{2}\ \frac{\pi}{2}]$.
v) Magnitude

$$mg(a,b) = \sqrt{da^2 + db^2} \tag{4}$$

vi) Convert these values into a column vector so that the radian values will be converted into the degrees. Here 180 degree is divided into 18 and 36 bins, in 18 bins each bin is of 10° and in 36 bins each bin is of 5°. The polar plot for 18 bins and 36 bins of "How" gesture are shown in Fig. 5. This polar plot shows angle of variation in the hand at the time of performing any gesture.

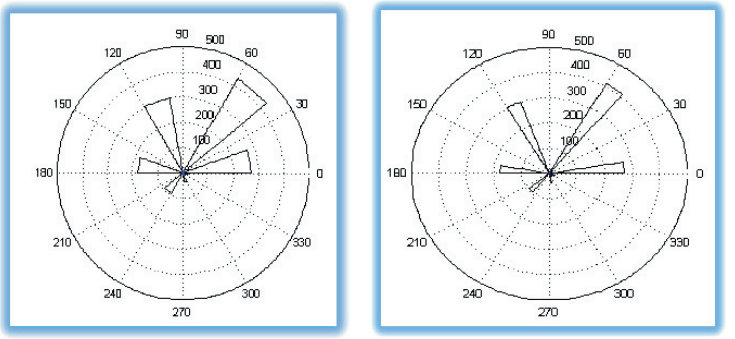

Fig. 5. Polar plot of orientation histogram of "How" gesture.

Principal component analysis

It is basically used for finding patterns in input data for highlighting similarities and differences between them and for reducing dimension of data set. After extracting these patterns form the data then we have to compress the data using PCA[7]. Finding patterns from the large data set is very difficult. So for analysing data set pca is a strong tool. PCA is a non-parametric, simple method for extracting significant information from confused data. Eigen value generated from pca gives projection direction of confused data set.

3.5 Classification

Here gesture recognition is done using different distance Metrics like Euclidean Distance, City Block Distance, Chess Board Distance, Mahalanobis Distance, Correlation Distance, and Cosine Distance. After pre-processing each probe sentences are divided into sequence of isolated gestures. Then extract features of each gesture and matched the number of frames have been successfully classified. After classification process a text formation has been performed.

Euclidean distance:

Here Euclidean distance is used to find the distance between projected trained image (r) and projected test image (s). Keep minimum distance value from the all trained image and discard the rest. Distance between two points p and q in Euclidean space is called Euclidean distance. Euclidean space is of n-dimension if $r = (r_1, r_2, \ldots, r_n)$ and $s = (s_1, s_2, \ldots, s_n)$, here r and s are called as Euclidean Vectors.

$$E(r,s) = \sum_{1=1}^{n} (|r_i - s_i|^2)^{1/2} \tag{5}$$

Mahalanobis distance:

When we have two vectors projected trained image (r) and projected test image (s) of the same distribution with the covariance matrix S then the Mahalanobis distance will be:

$$M(r,s) = \sqrt{(\bar{r} - \bar{s})' Co^{-1} (\bar{r} - \bar{s})} \tag{6}$$

Covariance is a measure of the random variables from two ordered pair of sets either data move in the same direction.

$$Co(r,s) = 1/r \left(\sum_{i=1} (r_i - \bar{s})(r_i - \bar{s}) \right) \tag{7}$$

City block distance:

The shortest distance between two points is along the hypotenuse, which is calculated by using Euclidean distance. Instead of this city block distance is measured as the length in x direction added with length in y direction. The distance measured by city block is always greater than or equal to zero. If points are same than distance is zero otherwise greater than zero. Most of the times city block distance gives similar result with Euclidean distance.

$$CT(r,s) = \sum |r_i - s_i| \tag{8}$$

Chess board distance:

This distance comes into focus from game of chess where kings take minimum number of moves to go from one square to another in chessboard. The chessboard distance between two vectors projected trained image (r) and projected test image (s) is represented by:

$$CB(r,s) = \mathrm{Max}_i(|r_i| - |s_i|) \tag{9}$$

It gives some times better result than Euclidean distance.

Cosine distance:

It is a complement of the cosine similarity in positive space. It gives distance in terms of angular cosine in between two vectors r and s. Cosine similarity is a measure of similarity between two vectors projected trained image (r) and projected test image (s) by calculating the cosine of the angle between them. It can be represented as:

$$CS(r,s) = \cos(Q) = \frac{r \cdot s}{|r|\,|s|} \tag{10}$$

Cosine distance can be represented as:

$$CD(r,s) = 1 - \frac{r \cdot s}{|r|\,|s|} \tag{11}$$

Correlation distance:

Statistical dependency between random variables like projected trained image (r) and projected test image (s) is measured by using correlation distance.

$$CO(r,s) = 1 - \frac{(r - \mathrm{mean}(r)) \cdot (s - \mathrm{mean}(s))}{|(r - \mathrm{mean}(r))| \cdot |(s - \mathrm{mean}(s))|} \tag{12}$$

4. Experimental Results and Analysis

Experiments are performed on 10 type of sentences. Each sentence having 2, 3 or 4 gestures. Here each continuous gesture is made of static as well dynamic gesture for performing experiments. Each sentence will be recorded 10 times, 6 times for training and 4 sentence is used for testing. In each sentence we consider 20 frames of each gesture out of n number of frames (n is vary from sentence to sentence) for training and 10 frames of each gesture for testing. Here only those frames are considered which are present at the middle because it consists of most informative frames. Here experiments are performed on 18 bins as well as on 36 bins which means 180° is divided into 8 parts each part is of 10°. Similarly for 36 bins.

From experimental results we found that the results obtained from Euclidean distance and correlation have higher recognition rate than other distance based classifiers like city-block distance, chessboard distance etc. Table 1 and 2 shows that the orientation histogram with 36 bins give higher accuracy then 18 bins because it measures angle of change of hand much appropriate. In 36 bins histogram there is a 5 degree resolution and in 18 bins there is a 10 degree resolution. From above table we also seen that the recognition rate of some gestures is much higher than some other gestures. This will happen because some gestures are of similar type like “you” and “I” etc. Those gestures are of same type it gets misclassified and give wrong results.

Table 1. Classification results of OH (18 bins) with PCA at various distance based classifier.

Sentences	Euclidean distance	Mahalanobis distance	City block distance	Chessboard distance	Cosine distance	Correlation distance
How are you?	91%	82%	82%	84%	92%	91%
I am agree.	90%	82%	79%	82%	88%	90%
Are you coming?	93%	78%	81%	81%	90%	88%
I am studying.	90%	79%	85%	82%	89%	90%
It is closed today.	89%	84%	83%	83%	86%	91%
Are you hearing?	90%	78%	81%	80%	89%	90%
You do not have a car?	91%	79%	78%	79%	90%	92%
Your home is big or small?	82%	67%	71%	80%	80%	80%
Why are you sad?	93%	82%	79%	78%	87%	90%
Is this your room?	91%	80%	76%	78%	87%	88%

Table 2. Classification results of OH (36 bins) with PCA at various distance based classifier.

Sentences	Euclidean distance	Mahalanobis distance	City block distance	Chessboard distance	Cosine distance	Correlation distance
How are you?	93%	82%	82%	84%	92%	93%
I am agree.	90%	82%	79%	82%	88%	90%
Are you coming?	93%	90%	87%	89%	90%	91%
I am studying.	92%	82%	85%	90%	89%	90%
It is closed today.	93%	84%	83%	83%	86%	91%
Are you hearing?	90%	78%	81%	80%	89%	90%
You do not have a car?	91%	79%	78%	79%	90%	92%
Your home is big or small?	85%	71%	75%	80%	80%	82%
Why are you sad?	93%	82%	79%	81%	87%	91%
Is this your room?	93%	80%	76%	78%	87%	88%

5. Conclusion

The Proposed framework for continuous ISL gesture gives satisfactory performance including features obtained using orientation histogram with PCA where both the hands have been used for performing any gesture. In continuous ISL recognition system key frame extraction is the foremost step. It helps us for extracting the continuous gesture into isolated gestures also it show how many number of frames an isolated gesture will have. After hand segmentation we applied OH for extracting features of hands for training dataset as well as for testing. Minimum distance shows maximum classification rate from probe dataset to training dataset. Here classification accuracy is measured with the maximum number of matched frames. Experimental results shows that the designed method gives satisfactory results with Euclidean distance and correlation. Results are also tested using normal webcam and get appropriate results.

This work has been enhanced by creating dataset with different background and different illumination conditions. Here we applied more appropriate features which incorporate shape of hand in the time of acting gestures, speed of performing each gesture etc. There are various other classifiers like Hidden Markov Model (HMM), Support vector machine (SVM) has been applied for classification.

Acknowledgements

We would like to thank all other researchers of our robita lab of Indian Institute of Information Technology, Allahabad, for their comments and suggestions. We also thank our technical staff of robita lab for their help in data collection.

References

[1] M. Elmezain, A. Al-Hamadi, J. Appenrodt and B. Michaelis, A Hidden Markov Model-based Continuous Gesture Recognition System for Hand Motion Trajectory, *19th International Conference on IEEE, Pattern Recognition, 2008, ICPR 2008*, pp. 1–4, (2008).

[2] V. Athitsos, Quan Yuan and S. Sclaroff, A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation", *IEEE Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 31, pp. 1685–1699, September (2009).
[3] A. Nandy, S. Mondal, J. S. Prasad and P. Chakraborty, Recognizing & Interpreting Indian Sign Language Gesture for Human Robot Interaction, *International Conference on Computer and Communication Technology (ICCCT), (2010)*, pp. 712–717, 17–19 September (2010).
[4] P. Morguet and M. Lang M, Comparison of Approaches to Continuous Hand Gesture Recognition for a Visual Dialog System, *IEEE International Conference on IEEE Acoustics, Speech, and Signal Processing, 1999, Proceedings, 1999*, vol. 6, pp. 3549–3552, 15–19 March (1999).
[5] Yi Li, Weidong Chen and Yang Zheng, Dynamic Hand Gesture Recognition using Hidden Markov Models, *7th International Conference on IEEE Computer Science & Education (ICCSE)*, pp. 360–365, 14–17 July (2012).
[6] M. K. Bhuyan, FSM-based Recognition of Dynamic Hand Gestures via Gesture Summarization Using Key Video Object Planes, *International Journal of Computer and Communication Engineering*, vol. 6, pp. 248–259, (2012).
[7] Yi Li, Weidong Chen and Yang Zheng, Dynamic Hand Gesture Recognition using Hidden Markov Models, *7th International Conference on IEEE Computer Science & Education (ICCSE)*, pp. 360–365, 14–17 July (2012).
[8] M. K. Bhuyan, Mithun Kumar Kar and Debanga Raj Neog, Hand Pose Identification from Monocular Image for Sign Language Recognition, *Proceedings of IEEE International Conference on Signal and Image Processing Applications (ICSIPA 2011), Malaysia*, pp. 378–383, November (2011).
[9] A. Saxena, D. K. Jain and A. Singhal, Sign Language Recognition using Principal Component Analysis, *Fourth International Conference on IEEE Communication Systems and Network Technologies (CSNT) 2014*, pp. 810–813, 7–9 April (2014).
[10] Jung-Bae Kim, Kwang-Hyun Park, Won-Chul Bang and Z. Z. Bien, Continuous Gesture Recognition System for Korean Sign Language based on Fuzzy Logic and Hidden Markov Model, *IEEE International Conference on Fuzzy Systems, 2002, FUZZ-IEEE'02, Proceedings of the 2002*, vol. 2, pp. 1574, 1579 doi: 10.1109/FUZZ.2002.1006741, (2002).
[11] Rung-huei Liang and Ming Ouhyoung, A Real-time Continuous Gesture Recognition System for Sign Language, *IEEE International Conference on Automatic Face and Gesture Recognition*, Japan, pp. 558–567, (1998).
[12] D. Mazumdar, A. K. Talukdar and K. K. Sarma, Gloved and Free Hand Tracking based Hand Gesture Recognition, *1st International Conference on IEEE Emerging Trends and Applications in Computer Science (ICETACS), 2013*, pp. 197–202, 13–14 September (2013).
[13] W. Werapan and N. Chotikakamthorn, Improved Dynamic Gesture Segmentation for Thai Sign Language Translation, *7th International Conference on IEEE Signal Processing, 2004, Proceedings, ICSP'04*, vol. 2, pp. 1463–1466, 31 August – 4 September (2004).