Optical Flow Hand Tracking and Active Contour Hand Shape Features for Continuous Sign Language Recognition with Artificial Neural Networks

P.V.V.Kishore and M.V.D.Prasad
Department of Electronic and communication Engineering
K L University
Vaddeswaram , INDIA
pvvkishore@kluniversity.in, mvd_ece@kluniversity.in

D.Anil Kumar and A.S.C.S.Sastry
Department of Electronic and communication Engineering
K L University
Vaddeswaram , INDIA
danilmurali@gmail.com, sastryascs@kluniversity.in

***Abstract* - To extract hand tracks and hand shape features from continuous sign language videos for gesture classification using backpropagation neural network. Horn Schunck optical flow (HSOF) extracts tracking features and Active Contours (AC) extract shape features. A feature matrix characterizes the signs in continuous sign videos. A neural network object with backpropagation training algorithm classifies the signs into various words sequences in digital format. Digital word sequences are translated into text with matching and the suiting text is voice translated using windows application programmable interface (Win-API). Ten signers, each doing sentences having 30 words long tests the performance of the algorithm by computing word matching score (WMS). The WMS is varying between 88 and 91 percent when executed on different cross platforms on various processors such as Windows8 with Inteli3, Windows8.1 with inteli3 and windows10 with inteli3 running MATLAB13(a).**

Index Terms - Sign Language Recognition, Horn Schunck Optical Flow, Hand Tracking, Active Contour Models, Shape Extraction, Artificial Neural Networks.

I. INTRODUCTION

Learning skills of a hearing hindered person are seriously hampered because of the missing hearing sense. From here, a mute person has to depend largely on visual sense and any learning and communication aids will help them learn faster and communicate better. Usually human interpreter trained in sign language understanding acts as a bridge between the normal people (with hearing sense) and mute people (without or with low hearing sense).

The difficulties faced by deaf and elderly community when moving alone and wanted to mingle publicly at government offices, schools, shopping malls and hospitals are indescribable. Having a human interpreter accompanying the deaf person always in a country like India is impractical due to few trained sign language interpreters.

Empowerment of the hearing impaired can be achieved using a mobile based application that can understand sign language and translate it to speech and conversely. The solution proposed is to develop a machine interpreter that can help deaf people to voice themselves at any location in the presence of people with hearing sense.

This machine translation of sign language acts an interpreter between these two groups of humans i.e. with and without hearing sense, intendeds to replace the human interpreter with machine interpreter. Sign language recognition is a major research area that encompasses video image analysis, shape extraction, feature optimization and pattern classification working in tandem to convert video signs into text and voice messages.

Previous research in the area show various methods applied to achieve this objective and to a certain extent achieved by most of the researchers. The entire sign language recognition follows three major methods, tracking sensors [1], glove based sensors [2] and visual sensors [3,4]. The most widely used and most difficult one is sign language recognition using visual sensors and video cameras.

The camera based SLR has two operating modes. The operating techniques are static (image based) and dynamic (video based). In video based the researchers focused on discrete videos with an average running time of 2 to 4 seconds, churning out a frame rate of 30fps. A few researchers implemented the same for continuous signs, where the video lasts for around 3-4 minutes at 30fps putting 7,200 frames for processing.

Early sign language recognition concentrated on articulator units called phonemes [5]. Rationally sign language is understood as a set of dialectal analysis of hands tracking, hands shapes, hands locations, sign articulation, head orientations and facial expressions.

Research on real time American Sign Language recognition shows a wearable computer based video camera [3] using Hidden Makov Model (HMM) recognizes continuous American Signs with good precision. Four HMM states try to capture the signs of American Sign Language producing good recognition rates. But the stability of the method is compromised when the signer changes in the video sequence making it signer dependent.

A signer adaptation model that combines maximum of a posteriori and iterative vector field smoothing [6] which reduces the amount of data to represent a video sign. This method has achieved good recognition rate of around 89% for continuous signs. Review shows a variety of sign language recognition systems with statistical approaches [7], example based approaches [8], finite state transducers [9] showing higher recognition rates close to 90%.

978-1-4673-8286-1/16 $31.00 © 2016 IEEE
DOI 10.1109/IACC.2016.71

Continuous sign recognition from videos calls for solving some major issues related to recognition. Two such issues that are being addressed in this work are hand tracking and hand shape analysis which constitute around 70% to SLR. Remaining 30% is related to movement emptiness, facial tracking and sign corpus identification.

The authors of this paper has done a considerable amount of work previously related to static gesture classification for Indian Sign Language with static and dynamic sign videos[10]-[11]. For sign segmentation of video frames the authors used wavelet based image fusion of canny edge operator and morphological differential gradient. Elliptical Fourier descriptors are used to model hand shapes and head portion. They have tested for 80 static signs with neural network classifier [12] and fuzzy inference engine [13] respectively. The percentage of recognition achieved is 87% when neural network classifier is used and 90% when fuzzy inference engine is used as classifier.

Yikai Fang et al [14] proposed a real time adaptive hand gesture segmentation and tracking using motion and color cues with 2596 frames recorded for 6 Gestures. They reported 98% recognition rate for simple backgrounds and around 89% recognition for cluttered video backgrounds. Each frame was processed for around 90 to 110 milliseconds.

G. Fang et al [15] has developed signer independent continuous sign language recognition HMM, Self-Organizing Maps and Recurrent Neural Networks. They used 208 signs form Chinese sign language and reported a recognition rate of around 92%.

Xiying Wang [16] incorporated tracking of deformable human hand for recognizing gestures in sign language. They showed that by incorporating tracking into a sign language recognition system will improve the performance of the classifier [17].

This research proposes a sign language recognition system with hand tracking and shape analysis that builds the feature vector for the classifier. Here we use artificial neural network as a classifier which is trained with error back propagation algorithm.

II. CONTINUOUS SIGN LANGUAGE RECOGNIZER – PROPOSED MODEL

The focus is on hands tracking and hand shapes as features which constitute around 70% of characterization for any sign language. For hand tracking horn schunck optical flow algorithm is used and shape features are extracted in each frame with active contour level set model. The tracking features are set of velocity vectors extracted from each moving hand in the frames. Hand shapes from each frame are represented with hand outliners extracted with shape numbers. Two features from each frame are concatenated to represent signs in each frame. The entire process of the proposed continuous sign language recognizer is presented as a flow chart; the video frames for this work represent a sign with a set of frames. The set of frames can be identified using velocity vectors computed using HSOF algorithm. A start of sign (SOS) set frames is recognized when velocity vectors between frames is maximum. A no sign frame f_{NoSign}^{n-1} will have velocity vectors $u_{NoSign}^{x(n-1)}$ and $v_{NoSign}^{y(n-1)}$ from HSOF algorithm. Similarly the consecutive sign frame f_{Sign}^{n} is having velocity vectors $u_{Sign}^{x(n)}$ and $v_{Sign}^{y(n)}$.

The following formulation decides the SOS

$$S^{SOS} = \begin{cases} u_{NoSign}^{x(n-1)} - u_{Sign}^{x(n)} < T \\ u_{NoSign}^{x(n-1)} - u_{Sign}^{x(n)} \geq T \end{cases} \quad (1)$$

Where T is the velocity threshold. For less hand movements between frames T is very small, whereas larger values of T are produced between sign frames and no sign frames. Similar model is used to decide on end of sign (EOS). In the sequence of frames the first SOS is extracted and the next low threshold difference is marked as EOS. Once EOS is marked, the remaining frames will have almost zero thresholds as there will not be any hand or head movements detected by HSOF. The next SOF will be the max threshold value from producing maximum velocity differences. Figure 1 shows a video sequence indicating the SOS by green border and EOS by red border frames. Close observations of the frames reveal the idea of selecting SOS and EOS. For display in figure only important frames are used.

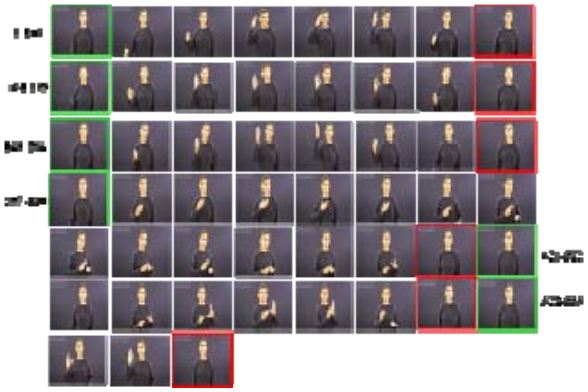


Fig 1. Frames from a sign video sequence showing Start of Sign (SOS) Frame in Green and End of Sign (EOS) Frame in Red. Frame numbers are published extreme left and right of the image. Few important frames are used for representation.

A feature matrix is extracted between one SOS and EOS. This feature matrix becomes input to the classifier. Training to the Artificial Neural Network classifier is provided with this feature matrix. Error back propagation algorithm is used as a training algorithm. The error is calculated with gradient descent algorithm. The following ANN model is used for combined feature vector classification as shown in figure 4. The details of weights, biases and learning rate values are randomly initialized and re iteratively updated with a convex cost function such as gradient descent algorithm.

III. RESULTS and Discussion

For testing the above process on the videos of continuous sign language, a camera setup is constructed. To ensure uniform lighting, two 23 luminance bulbs are erected at an

angle of 45 degrees from the signer. A HD Sony camcorder at a distance of 22 meters ensures excellent HD videos of continuous sign language. For experimentation for this model a subset of Indian Sign Language is used. The setup is shown in figure5.

A total of 10 test subjects were used. Each performing the same set of sentences for Indian Sign Language. A continuous sentence comprising 58 words is chosen for our experimentation. The sentence used is ***"Hai, Good Morning, My Name is Kishore, I am Student of K.L.University, Studying final year Undergraduate Engineering, From Department of Electronics and Communications engineering, the college is located at a lush green surroundings, with estimated area of around 50 acres, we are doing this sign language as a part of our final year project, thank you"***. Each video recording of the above sequence of sentences is averaged around 100 sec. This is because, the signers are not regular users of sign language.

Sony camcorder outputs videos with a 30fps HD sequences with a frame size of 640×480 . For a 100 sec sequence we are looking at 3000 frames for a sentence of 58 words. For 10 different signers the figure is 30000 frames. Different signers are selected to make the system signer independent. Of these 5 samples are used for training and 5 will be testing samples. From any 5 samples feature vector is built combining tracking position vectors from HSOF and shape outliner's vectors from AC models, which will train artificial neural network.

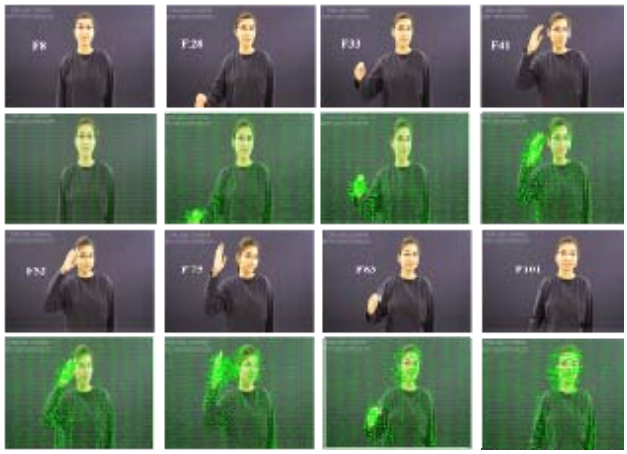


Fig 2. Video frames showing results of HSOF tracking signers hand and head. The green arrows resembling velocity vectors along x and y directions. Here the signer is saying "HAI".

Horn Schunck optical flow is the tracker for hands in the video frames. The algorithm implements on two frames at a time and computes the velocity vectors in x and y directions. From these velocity vectors, we compute the position of the hand in the frame. Multiple positions are extracted due to the some ambiguity in the hand positions. This is due to light intensity variations in the frames during capture. Hence the average position vector is obtained for each hand in the frames. The tracking using Horn Schunck Optical Flow (HSOF) on frames of figure 3 is shown in figure 2.

The same HSOF algorithm can do the job of segmentation of moving objects in a video. The results of HSOF segmentation for a few frames are shown in figure 3. From figure 3(b) and (d) the HSOF segmented hands fails to produce exact contours to represent shape features. Hence in this HSOF algorithm will only track hands. The final track on the continuous video sequence is given in figure 4. Intermediate frames are displayed in figure 4 out of a total frame count of 2998 frames.

(a) (b) (c) (d)

Fig 3. HSOF based sign segmentation. (a) Frame 52 (b) Frame 52 Segmentation output from HSOF (c) Frame 75 (d) Segmentation result of (c) with HSOF.

Right hand bounding box is in red color and for left hand green is used. 'X' and 'Y' are position vectors with respect to head of the signer. Head position is marked manually in the first frame. Figure 5 displays tracks in three dimensional space. A feature vector is generated by careful labelling to both hands of the signer. The vector varied from signer to signer as their hand speeds changed during sign acquisition. Hence the tracking feature space is normalized by generating intermediate lost values for fast signers and by removing repeated position vectors for slow signers. Finally for a particular signer with 58 words we have a $1996 \times 58 \times 2$ matrix. Further normalization by temporal averaging, the tacking feature matrix is represented with 1996×58 , giving 1996 tacks for 58 words sequences. By using only tracking information it is impossible to determine the classification problem.

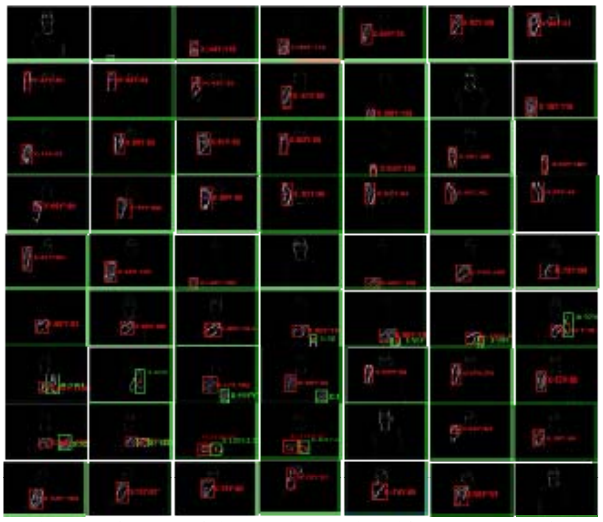
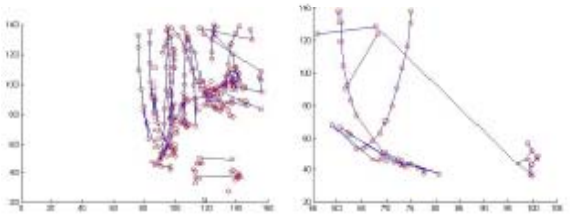


Fig 4. Tracks of Hands in a continuous sign video sequence used for experimentation.

Hence shape information of hands should accompany the tracks to the input of the classifier. Hand shape extraction is accomplished with active contour model. The contour is placed in an area close to the head of the signer. This enables the AC segmentation algorithm to trigger only when there is significant

movement near the torso of the signer. Figure 6 shows the AC segmentation process on a video frame of the signer. The contour is placed close to the torso of the signer to keep the number of iterations for segmentation to a minimum.



a) Right Hand Tracks b) Left Hand Tracks

Fig 5. Hand Tracks of (a) Right Hand of signer (b) Left hand of signer

The segmentation is near perfect and no further processing is required. The head and hand shape boundary numbers are considered as feature vector per frame. A set of frames are presented in figure 11 for the sentence described above. Total numbers for frames in this particular video are around 2998. Only 699 frames have useful information that can be considered as feature vector design. These 699 frames are selected based on frame differencing model used in equation (20). Here threshold is set based on the velocity difference value. Large velocity value changes in frames are retained and those with lesser velocity gradient are discarded.

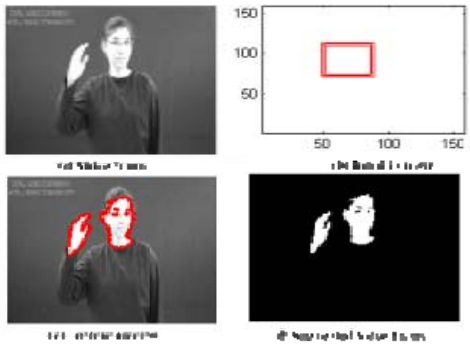


Fig 6. Active contour segmentation on a Video frame

The contours are extracted from the region boundaries of the segments. Hand and head segment contours are manually labelled as Head and hand contours. The extracted contours for a few frames are shown in figure 8. The boundaries of the contours are given unique combination of numbers to uniquely identify a particular shape in the video frame. These labelled shape numbers are fused with tracking features for the same frame. A final feature matrix is an amalgamation of two important characteristics for machine understanding of sign language from video sequences.

The feature vector for both training and testing is built based on velocity vector gradients. Lesser gradients marks Star of sign (SOS) and End of Sign (EOS) frames. All the middle frames will have a feature vector. In this work we have 58 words and hence we have 58 feature vectors. Each feature vector is represented by variable number of samples with both tracking and shape numbers. For a complete 58 word sentence we have 699 frame video.

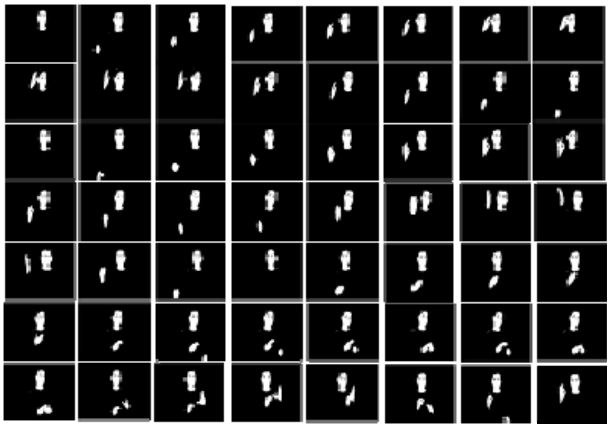


Fig 7. Segmentation outputs of a few important frames form the video sequence. The frames are layered in horizontal format.

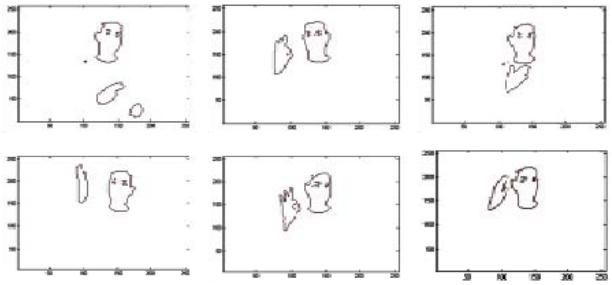


Fig 8. Extracted shape contours for few frames

A complete tracking feature matrix is created from 6 position vectors obtained from tracking i.e. right hand position (x, y), left hand position and Fixed head position. Similarly hands and head are represented with 52 shape numbers per frame constitute shape feature matrix. Concatenating the two produces a 58×699 feature matrix that trains the back propagation neural network. The entire process of feature vector design is mapped diagrammatically.

The target matrix consists of 58 words in the sentence in the order of sequence described previously. To improve the efficiency of the training program, four more samples are added to the one derived feature matrix earlier. The training input vector is 290×699 whereas the target is 58×699 . Both the matrices are supplied as input to the neural network with 290 inputs and 58 targets. Gradient descent error is transmitted to update the weights after every iteration as described in section 4. Log sigmoid activation function is used in all the layers. The model of neural network object is created in MATLAB.

A good number of hidden neurons will speed up the training process and 350 hidden neurons are chosen for the input vector of 290 elements. There is no cap on the hidden neurons expect for the fact that there are algorithms and formulae to decide on the number of hidden neurons. In this simulation hidden neurons are chosen a little above input neurons.

In back propagation training the networks weights and biases are continuously updated with the mean square error to minimize network performance. The mean square error

tolerance is fixed at 0.01 for training the samples. The learning rate and momentum factor were chosen as 0.241 and 0.5. The training graph between mean square error and epochs for the neural network object is shown in figure 9.

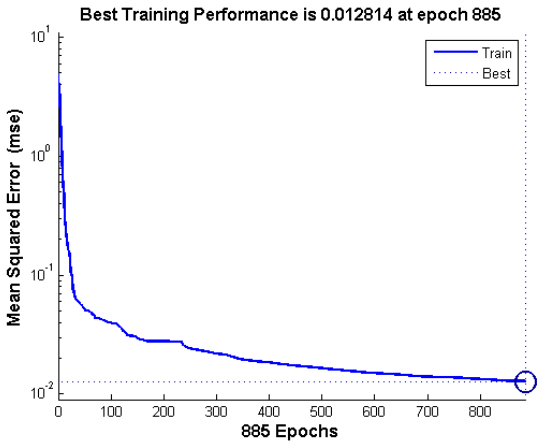


Fig 9. Training plot for a mean square error tolerance fixed at 0.01

The word matching score given by

$$\omega_{ms} = \frac{Word\ Matching}{Total\ Words} \times 100 \qquad (2)$$

For individual words in the sentence, word matching score is computed with 5 samples for training and remaining 5 samples and the 5 already trained ones are used for testing the trained network. Table I gives values of WMS for the proposed method against the three other methods. The experimentation was done 5 times. The WMS values in the table-I are averaged values over 5 times. Each time training is accomplished with same training set for all models of sign language systems in table.

Table I. Word Matching Scores for individual words in the sentence used in this work

Words	Word Matching Score-Proposed, HSOF+AC+ANN	Sobel+ DCT+ ANN [18]	Sobel+Hough Transform + ANN[19]	Active Contour Model for seg & tracking+ANN[11]
Hai	90	50	60	90
good	100	70	80	100
morning	100	70	70	100
my	80	20	30	50
name	80	60	60	70
is	100	100	100	100
Kishore	70	30	40	80
I	70	10	20	60
am	60	10	20	70
a	100	60	80	100
student	90	60	60	80
of	100	100	100	100
K	100	50	80	100
L	100	80	100	100
University	80	40	50	70
Studying	90	60	80	80
final	80	20	50	80
year	90	80	70	90
Undergraduate	70	20	10	60
Engineering	80	20	30	50
From	100	40	50	90
Department	70	20	40	50
of	100	100	100	100
Electronics	90	50	40	70
and	100	100	100	100
Communications	100	80	90	90
engineering	80	20	30	50
the	100	100	100	100
college	80	20	30	70
is	100	100	100	100
located	80	30	50	80
at	100	100	100	100
a	100	100	100	100
lush	80	30	40	70
green	80	10	20	60
surroundings	100	100	100	100
with	100	100	100	100
estimated	70	20	30	70
area	80	30	40	60
of	100	100	100	100
around	100	100	100	100
Fifty	100	100	100	100
acres	80	10	20	70
we	90	40	50	70
are	100	100	100	100
doing	80	40	40	70
this	100	100	100	100
sign	100	50	60	80
language	90	70	80	80
as	100	100	100	100

a	100	100	100	100
part	100	50	60	80
of	100	100	100	100
our	80	10	20	60
final	80	20	50	80
year	90	80	70	90
project	100	40	50	70
thank you	100	20	20	50
Total	**90.172**	**58.448**	**65**	**82.586**

The average word matching score is 89.482% for the entire classification process which is on par with other researchers for American Sign Language [20] and Chinese Sign Language in [11]. The results indicate that by employing multiple features in continuous sign language recognition process, the word classification rates can be improved drastically when compared to single feature SLR models. Finally the recognized text words input the voice application programmable interface available in windows OS to produce voice from text.

IV. Conclusion

This work gives a multi feature model for recognizing continuous gestures of Indian sign language. Videos of continuous signs are captured for 58 words forming meaningful sentences. Horn Schunck optical flow algorithm extracts tracking features of both hands providing position vectors of hands in each frame. Active Contour model on each frame extracts hand shapes features along with head portion. The combined feature matrix having tracking and shape features train the back propagation neural network. The classified signs are mapped to text from the target matrix of ANN and converting those text inputs to voice commands with windows text-to-speech application programmable interface. Validating the proposed model by computing the word matching score for each word recognized by the neural network. The word matching score over multiple instances of training and testing of the neural network resulted in around 90%. This work can be extended to include other characteristics of continuous sign language.

REFERENCES

[1] Zahid Halim, Ghulam Abbas, "A Kinect-Based Sign Language Hand Gesture Recognition System for Hearing- and Speech-Impaired: A Pilot Study of Pakistani Sign Language", Assistive Technology, Vol. 27, No. 1, pp.34-43, 2015.

[2] Gaolin Fang and Wen Gao, "Large Vocabulary Contineous Sign language Recognition Based on Trnsition-Movement Models", IEEE Transaction on Systems, MAN, and Cybernetics, Vol.37, No.1, pp 1-9, 2007.

[3] T.Starner and A.Pentland "Real-Time American Sign Language Recognition from video using Hidden Markov Models", Technical Report, MIT Media laboratory Perceptual computing section, Technical Report number.375,1995.

[4] Ming-Hsuan Yang and Narendra Ahuja, "Extraction of 2D Motion Trajectories and its Application to Hand Gesture Recognition", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.24, No.8, pp.1061-1074, 2002.

[5] W. Stokoe, D. Casterline, and C. Croneberg. A Dictionary of American Sign Language on Linguistic Principles. Gallaudet College Press, Washington D.C., USA, 1965.

[6] Yu Zhou and Xilin Chen, "Adaptive sign language recognition with Exemplar extraction and MAP/IVFS", IEEE signal processing letters, Vol 17, No.3, pp.297-300, 2010.

[7] Och J., Ney. H, Discriminative training and maximum entropy models for statistical machine translation. In: Annual Meeting of the Ass. For Computational Linguistics (ACL), Philadelphia, PA, pp. 295–302,2002.

[8] Sumita, E., Akiba, Y., Doi, T., et al., 2003. A Corpus-Centered Approachto Spoken Language Translation. Conf. of the Europ. Chapter of the Ass. For Computational Linguistics (EACL), Budapest, Hungary, pp. 171–174.

[9] Casacuberta, F., Vidal, E., Machine translation with inferred stochastic finite-state transducers. Computational Linguistics, Vol.30, No.2, pp.205–225, 2004.

[10] Kishore, P.V.V, Rajesh Kumar, P, "Segment, Track, Extract, Recognize and Convert Sign Language Videos to Voice/Text", International Journal of Advanced Computer Science and Applications (IJACSA) ISSN (Print)-2156, Vol.3,No.6, pp.35-47,2012.

[11] Kishore, P.V.V., Sastry, A.S.C.S., Kartheek, A., "Visual-verbal machine interpreter for sign language recognition under versatile video backgrounds", 2014 First International Conference on Networks & Soft Computing (ICNSC), pp.135-140, 19-20 Aug 2014.doi: 10.1109/CNSC.2014.6906696.

[12] Kishore, P. V. V., S. R. C. Kishore, and M. V. D. Prasad. "Conglomeration of Hand Shapes and Texture Information for Recognizing Gestures of Indian Sign Language Using Feed forward Neural Networks." International Journal of engineering and Technology (IJET), ISSN: 0975-4024, Vol.5,No.5, pp.3742-3756,2013.

[13] Kishore, P.V.V.; Prasad, M.V.D.; Prasad, C.R.; Rahul, R., "4-Camera model for sign language recognition using elliptical fourier descriptors and ANN," 2015 International Conference in Signal Processing And Communication Engineering Systems (SPACES), pp.34-38, 2-3 Jan. 2015, doi: 10.1109/SPACES.2015.7058288.

[14] Yikai Fang, Kongqiao Wang, Jian Cheng and Hanqing Lu, "A Real-Time Hand Gesture Recognition Method", In Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2007), Beijing, China, July 2007, pp. 995-998.

[15] G. Fang, W. Gao, X. Chen, C. Wang, and J. Ma, "Signer-Independent Continuous Sign Language Recognition Based on SRN/HMM," Proc. Gesture Workshop, pp. 76-85, 2001.

[16] Xiying Wang, Xiwen Zhang and Guozhong Dai. Tracking of Deformable Human Hand in Real Time as Continuous Input for Gesture based Interaction, 2007 International Conference on Intelligent User Interfaces (IUI 2007), Hawaii, USA.

[17] L. Brkthes, P. Menezes, E Lerasle and J. Hayet "Face tracking and hand gesture recognition for human-robot interaction" Pmcwdingr of the 2004 IEEE InternatIona1 Confennee on RobotIco 6 Automation New Orleans. April 2004.

[18] P.V.V. Kishore, P. Rajesh Kumar, A. Arjuna Rao, "Static Video Based Visual-Verbal Exemplar for Recognizing Gestures of Indian Sign Language", International Journal of Digital Image Processing, ISSN 0974 – 9586, vol. 3, No. 9, pp. 530-537, 2011.

[19] Qutaishat Munib et.at., (2007). American Sign Language (ASL) recognition based on Hough Transform and Neural Networks, Expert Systems with Applications 32, Science Direct, Elsevier ltd., pp24-37.

[20] Paulraj M P,Sazali Yaacob, Hazry Desa Hema, C.R.Wan, Mohd Ridzuan, Wan Ab Majid, "Extraction of Head and Hand Gesture Features for Recognition of Sign Language" 2008 IEEE International Conference on Electronic Design, December 1-3, 2008, Penang, Malaysia.