

# Selfie Continuous Sign Language Recognition with Neural Network Classifier

G. Anantha Rao, P.V.V. Kishore, A.S.C.S. Sastry, D. Anil Kumar and E. Kiran Kumar

**Abstract** This work's objective is to bring sign language closer to real-time implementation on mobile platforms with a video database of Indian sign language created with a mobile front camera in selfie mode. Pre-filtering, segmentation, and feature extraction on video frames create a sign language feature space. Artificial Neural Network classifiers on the sign feature space are trained with feed forward nets and tested. An ASUS smart phone with 5M pixel front camera captures continuous sign videos containing an average of 220 frames for 18 single-handed signs at a frame rate of 30 fps. Sobel edge operator's power is enhanced with morphology and adaptive thresholding giving a near perfect segmentation of hand and head portions. Word matching score (WMS) gives the performance of the proposed method with an average WMS of around 90% for ANN with an execution time of 0.5221 s during classification. Fully novel method of implementing sign language to introduce sign language recognition systems on smart phones for making it a real-time usage application.

**Keywords** Indian sign language • Sobel adaptive threshold • Morphological differencing • Artificial neural networks • Word matching score

---

G. Anantha Rao (✉) · P.V.V. Kishore · A.S.C.S. Sastry · D. Anil Kumar · E. Kiran Kumar  
Department of Electronics and Communication Engineering, K L University, Guntur, India  
e-mail: ananth.gondu@gmail.com

P.V.V. Kishore  
e-mail: pvvkishore@kluniversity.in

A.S.C.S. Sastry  
e-mail: ascssastry@kluniversity.in

D. Anil Kumar  
e-mail: danilmurali@kluniversity.in

E. Kiran Kumar  
e-mail: kiraneepuri@kluniversity.in

## 1 Introduction

Sign language is a computer vision-based intact intricate language that engages signs shaped by hand moments in amalgamation with facial expressions and hand shapes. Sign language is a natural language for communication among people with low or no hearing sense. Human speech capture in digital format generates a 1D signal for processing whereas human sign language generates 2D signals from image or video data. Classification of gestures can be identified as both static and dynamic. Static gestures involve time-invariant finger orientations, whereas dynamic gestures support a time-varying hand orientations and head positions. The proposed Selfie video model for sign language recognition is a computer vision-based approach and does not employ motion or colored gloves for gesture recognition.

We introduce a novel sign language recognition system, called selfie sign language recognition system capturing signs using a smartphone front camera. The signer holds the selfie stick in one hand and signs with his other hand.

In the past, researchers attempted this type of work on sign images and videos but the novelty of our method lies in capturing selfie sign videos.

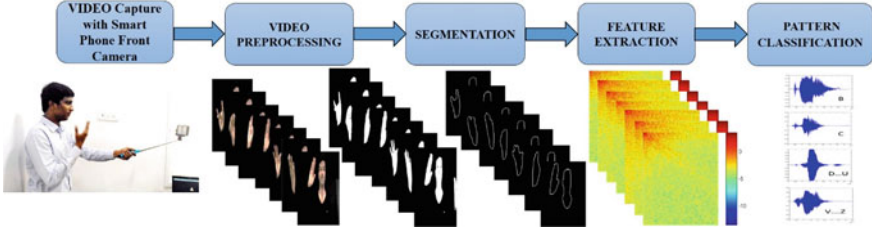
In [1], Mohamed proposed a vision-based recognizer to automatically classify Arabic sign language. A set of statistical moments for feature extraction and support vector machines for classification provided an average recognition rate of 87%. Omar [2] proposed a neuro-fuzzy system that deals with images of simple hand signs and succeeded a recognition rate of 90.55%.

Kishore, proposed [3] 4-Camera model. The segmented hand gestures with extracted shapes created a feature matrix described by elliptical Fourier descriptors which are classified with back propagation algorithm trained artificial neural network. The normal recognition rate in the proposed 4-Camera model for sign language recognition is about 92.23%.

## 2 Pre-processing, Segmentation, Feature Extraction, and Classification

The flowchart of the proposed SLR is shown in Fig. 1. The picture under the first block shows the capture mechanism followed in this work for video capture. Acquired video is in mp4 format having full HD ( $1920 \times 1080$ ) video recording on a 5M pixel CMOS front camera. Let this 2D video be represented as a 2D frame. For video the frame changes with time, which is fixed universally at 30 frames per second. These videos form the database of this work. A threefold 2D Gaussian filter with zero mean ( $m = 0$ ) and three variances in the range 0.1–0.5 smoothens each frame by removing sharp variations during capture.

The smoothed frames in real space  $\mathbb{R}$  are treated with a new type of multidimensional Sobel mask. From literature the Sobel edge operator is a 2D gradient



**Fig. 1** Flow chart of SLR system with smart phone front camera video capture

operator. Gradients provide information related to changes in the data along with the direction of maximum change. For 2D gradient calculation, two 1D gradients in  $x$  and  $y$  directions of the frame matrix are computed as follows:

$$g^x = \sum_{k=1}^N \mathfrak{F}(x-k, y)g(k) \quad (1)$$

$$g^y = \sum_{k=1}^N \mathfrak{F}(x, y-k)g^T(k) \quad (2)$$

where  $g \rightarrow [+1, -1]$  is the discrete gradient operator. The gradient magnitude  $G^{xy}$  gives magnitude of edge strength in Sobel edge detector computed as  $G^{xy} = \sqrt{(g^x)^2 + (g^y)^2}$ . The Sobel masks  $S^{Mx}$  and  $S^{My}$  are sensitive to lighting variations, motion blur, and camera vibrations which are commonly a cause of concern for sign video acquisition under selfie mode. A suitable threshold at the end will extract the final binary hand and head portions. Edge adaptive thresholding is considered with block variational mean on each  $3 \times 3$  Sobel mask is used as threshold. The final binary image is

$$B^x = \sum_{x=1}^N \sqrt{(S^{Mx} \otimes \mathfrak{F}^x)^2 + (S^{My} \otimes \mathfrak{F}^y)^2} \geq \sum_{i=1}^b \sum_{x=1}^N \sqrt{(S^{Mx} \otimes \mathfrak{F}^x)^2 + (S^{My} \otimes \mathfrak{F}^y)^2} \quad (3)$$

where  $b$  is the block size and  $B^x$  is binary video frame or segmented video frame.  $S^{Mx}$  and  $S^{My}$  are Sobel masks in  $x$ - and  $y$ -direction. Figure 2 shows the difference in block thresholding and global thresholding (used 0.2) which failed to handle motion blur.

Sign language is defined by hand shapes. Hand shapes are defined by precise contours that form around the edges of the hand in the video frame. A hand contour  $H^C(x) \rightarrow C(B^x)$  in spatial domain is a simple differential morphological gradient on the binary image  $B^x$ . The connected component analysis separates head and hand contours. Morphological gradient is defined by line masks in horizontal  $M_{3H}$  and vertical  $M_{3V}$  directions having length 3. Contour extraction is represented as

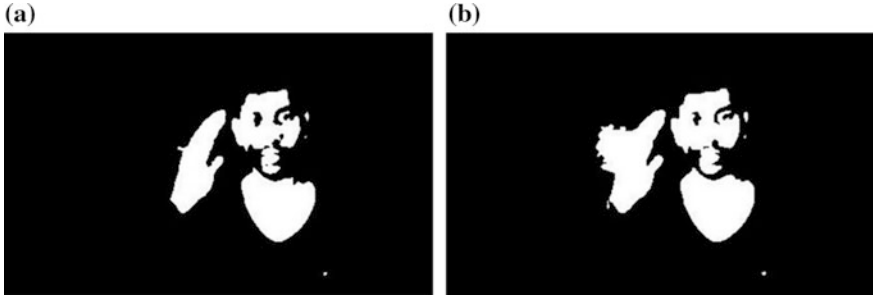
$$H^C(x) = \left\{ z \mid (\hat{M}_{3H})_z \cap B^x \neq \emptyset \right\} - \left\{ z \mid (\hat{M}_{3H})_z \subseteq B^x \right\} \quad (4)$$

$$H^C(y) = \left\{ z \mid (\hat{M}_{3V})_z \cap B^x \neq \emptyset \right\} - \left\{ z \mid (\hat{M}_{3V})_z \subseteq B^x \right\} \quad (5)$$

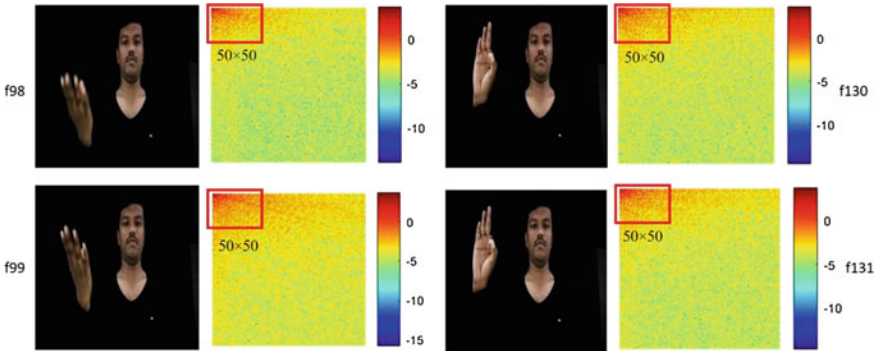
$$H^C(x, y) = H^C(x) \oplus H^C(y) \quad (6)$$

where  $H^C(x)$  is hand contour in  $x$ -direction and  $H^C(y)$  in  $y$ -direction. Hand and head contours are separated by finding the connected components with maximum number of pixels with a 4 neighborhood operation to form a contour image  $H^C(x, y)$ .

Features are unique representation of objects in this world. Feature is a set of measured quantities in a 1D space represented as  $F^V(x) = \{f(x) \mid x \subseteq \mathbb{R}\}$ , where  $f(x)$  can be any transformation or optimization model on vector  $x$ . Here  $f(x)$  is considered as Discrete Cosine Transform (DCT) along with Principle Component Analysis (PCA). Figure 3 shows a color-coded representation of hand DCT features  $F_{uv}^V$  for the frame in a video sequence. The head does not change much in any of the

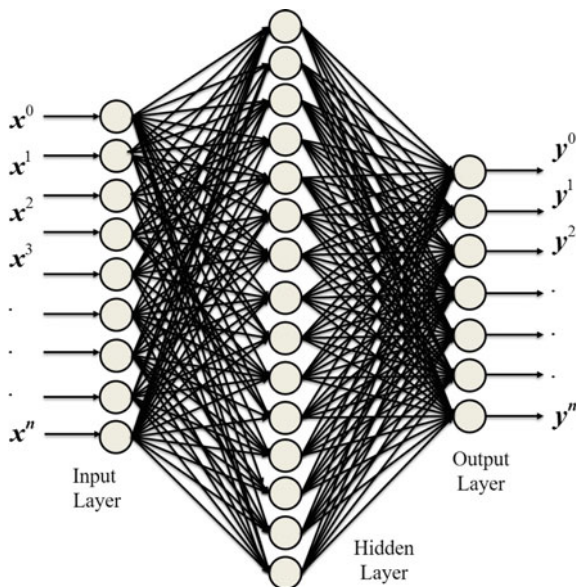


**Fig. 2** **a** Block variational mean thresholded frame. **b** Global threshold of 0.2 for Sobel



**Fig. 3** 2D DCT representation of hand classification

**Fig. 4** Neural network model used for sign contour energy representations



frames captured and hence head contour DCT remains fairly constant throughout the video sequence.

The first  $50 \times 50$  matrix of values possess maximum amount of energy in a frame. The boxed region in Fig. 4 is the  $50 \times 50$  feature matrix. This matrix of  $50 \times 50$  is reshaped to from a  $1 \times 2500$  feature vector. But for every frame, a 2500 value feature vector for frame will cost program execution time. PCA treatment of the matrix  $F_{uv}^V$ , retains only the unique components of the matrix  $F_{uv}^V$ . The final feature vector  $F_{uv}^V$  is represented as  $F_{fn}^V$ , where  $fn$  gives frame number. PCA reduces the feature vector per frame to from  $50 \times 50$  sample values per frame. Each 50 sample Eigen vector from PCA uniquely represents DCT energy of the hand shape in each frame.

The details of ANN with backpropagation algorithm are listed in our previous work at [4] and the models used for coding are considered form [5].

The model of artificial neural network is presented in Fig. 4. A 3-layered feedforward network of neurons is simulated with  $F_{fn}^V$  features as input to first layer. Then number of neurons in input layer is estimated from the samples obtained from PCA treated DCT energy matrix. The numbers of output neurons are equal to number of signs to be recognized by the network. Hidden layer neurons are estimated through trial and error method, and an optimum no. of neurons are selected to achieve accuracy at higher execution speech. For this sign classification, the estimated neurons are twice the neurons in the input layer for correct classification at reasonably less simulation times.

### 3 Results and Analysis

The front camera video recording of sign language gestures with smart phones Asus Zen phone II and Samsung Galaxy S4 at the end of selfie stick are used in experiment. Both the mobiles are equipped with 5M pixel front camera. Sign video capturing is constrained in a controlled environment with room lighting and simple background. The first photo in Fig. 1 demonstrates the procedure followed by signers for video capture. The discussion on results is presented in two sections: quantitative and qualitative. Quantitative analysis provides visual outcomes of the work and qualitative analysis relates to various constraints on the algorithm and how are these constraints are handled.

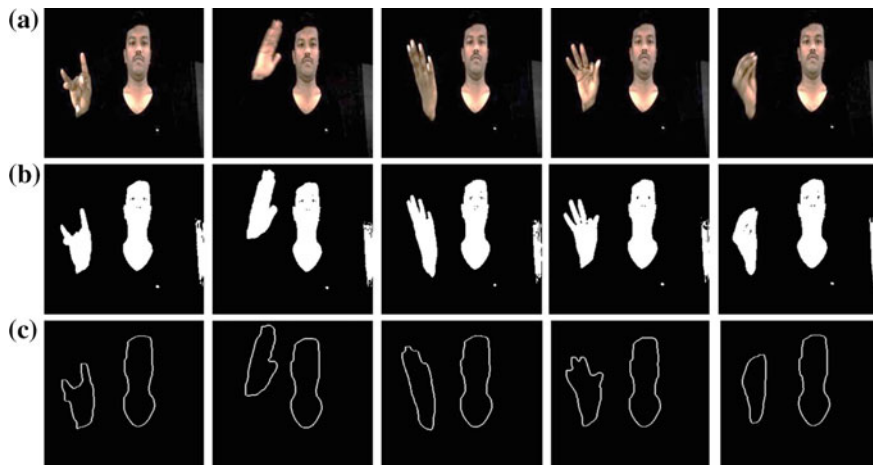
#### 3.1 Visual Analysis

Each video sequence is having a meaningful sentence. The following sentence “Hai Good Morning, I am P R I D H U, Have A Nice Day, Bye Thank You” is used for training and testing. There are 18 words in the sentence. The words in the training video are sequenced in the above order but the testing video contains same words in different order.

Classification of the words is tested with Euclidian, Normalized Euclidian, and Mahalanobis distance functions from our previous work on this video sequence. Filtering and adaptive thresholding with Sobel gradient produces regions of signer’s hands and head segments. Morphological differential gradient with respect to line structuring element as in Eqs. 4–6 refines the edges of hands and head portions. Figure 5 shows the results of the segmentation process on a few frames. Row (a) has original RGB captured video frames. Row (b) has Gaussian-filtered, Sobel-gradiented and region-filled outputs of the frames in row (a). The last row contains morphological subtracted outputs of the frames in row (b).

The energy of the hand and head contours gives features for sign classification. 2D DCT calculates energy of the hand and head contours. DCT uses orthogonal basis functions that represent the signal energy with minimum number of frequency domain samples that can effectively use to represent the entire hand and head curvatures. As shown in Fig. 3, first  $50 \times 50$  samples of the DCT matrix were extracted. These 2500 samples out of 65,536 samples are enough to reproduce the original contour using inverse DCT. This hypothesis is tested for each frame and a decision was made to consider only 2500 samples for sign representation.

With  $50 \times 50$  feature matrix per frame and an average number of frames per video at 220 frames, the feature matrix for the considered 18 signs is a stack of  $50 \times 50 \times 220$  matrix. Initiating the classifier with a multidimensional feature matrix of this size takes longer execution periods. Hence, PCA treats each frame of size  $50 \times 50$  energy features by computing Eigen vectors and retaining the principle components to from a  $50 \times 1$  vector per frame. The training vector contains



**Fig. 5** **a** Few frames in RGB format. **b** Their region segments with Gaussian filtering and Sobel operation. **c** Contours of hands and head produced with morphological subtraction with line structuring elements

only a few head sample values for such ‘No Sign’ detection. Euclidian distance, Normalized Euclidian distance, and Mahalanobis distance classifies the feature matrix as individual signs. The next section analyses the classifiers performance based on word matching score (WMS). And proposes to use ANN to trade accuracy over speed.

### 3.2 Classifiers Performance: Word Matching Score (WMS)

Word matching score gives the ratio of correct classification to total number of samples used for classification. The expression for WMS is  $M^{S\%} = \frac{\text{Correct Classifications}}{\text{Total Signs in a Video}} \times 100$ . Feature matrix has a size of  $50 \times 220$ , each row representing a frame in the video sequence. To test the uniqueness of the feature matrix for a particular sign or no sign, energy density variations of the 50 samples for first 150 frames is computed.

Exclusive testing with three distance measure on a sign video having 18 signs consisting of 220 frames provides an insight into the best distance measure for sign features. Table 1 gives details of the metric  $M^{S\%}$  for three distance measures. The average classification rate with same training feature for testing individual frames is around 90.58% with Mahalanobis distance. The low scores recorded by Euclidian distance (74.11%) and normalized Euclidian Distance (71.76%) compared to Mahalanobis is due to the inter class variance considerations in Mahalanobis. Test repetition frequency is 10 per sign. To further improve performance of the classifier the distance classifier is replaced by artificial neural network. The input layer has 50

**Table 1** The performance of three minimum distance classifiers with different testing videos

Signs	Euclidian distance classifier	Normalized Euclidian distance	Mahalanobis distance classifier
HAI	70	60	80
GOOD	60	60	80
MORNING	70	70	80
I AM	50	40	80
P	80	80	90
R	80	80	100
I	80	80	100
D	80	80	100
H	80	80	90
U	80	80	90
HAVE	40	40	80
A	60	80	90
NICE	50	40	80
DAY	60	60	80
BYE	40	40	80
THANK	40	40	80
YOU	50	40	80
Average WMS	62.94	61.76	85.88


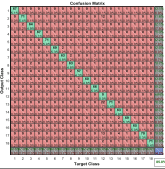

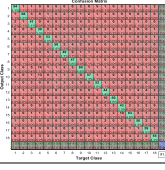
neurons that feed the hidden layer with 100 neurons are assigned in the output layer represent each sign.

For training 2 sets of videos at the same time with a total of 524 frames from continuous sign sequence of two different signers were chosen for testing with same set with 78 hidden neurons has resulted in a WMS of 80.5%. Putting more number of hidden neurons will further increase the WMS, but reduces speed of execution and they are optimized for this set at 78. In the next phase of experimentation 789 frames trained the ANN and from that 3 sets, 2 sets are tested, i.e. 524 frames. The number of hidden neurons was 125 and we found an increase in WMS at 85.5%. Similarly results for experimentation 3 are shown at the last row of Table 2. The WMS significantly improved for higher sample training with a compromise in speed. Hidden neurons were 200 in the last testing phase.

The average recognition rate was 90% for the total classification method which is on par with other researchers for American Sign Language [6] and Chinese Sign Language [7]. To standardize the entire algorithm, the number of hidden neurons are taken as 100 and testing is carried out with all other values being constant from previous testing's. Minimum Distance Classifier (MDC) with Mahalanobis distance produces a 85.5% WMS at 0.4823 s for a data set with 1313 frames. The ANN-based classifier with same parameters produced a 90% WMS at 0.5452 s.



**Table 2** Details of training and testing sign videos under simple backgrounds with different samples and their recognition rates

Training samples	Testing samples	Network architecture	Output confusion matrix	WMS (%)
18 Signs 789 (2 Sets) frames	18 Signs (3 Sets) 524 frames			85.5
18 signs (5 Sets) 1313 frames	18 signs (3 Sets) 789 frames			91.0

ANN's in the recent years have become faster and can be considered for smart phone based sign language recognition system design.

### 4 Conclusion

A novel idea of putting sign language into smart phones is simulated and tested in this work. Sign video capture using selfie stick is being introduced for the first time in the history of computerized sign language recognition systems. A formal database of 18 signs in continuous sign language were recorded with 10 different signers. Hand and head contour energies are features for classification computed from discrete cosine transform. Execution speeds are improved by extracting principle components with principle component analysis. Euclidian, normalized Euclidian, and Mahalanobis distance metrics classify sign features. Mahalanobis distance reached an average word matching score of around 90.58% consistently when compared to the other two distance measures for the same train and test sets. Mahalanobis distance uses inter class variance to compute distance which is required in sign language recognition due to the fact that no two signers in this world will not perform same sign similarly. For different train and test samples ANN outperformed MDC by an upward 5% of WMS for ANN. Further studies are required for improving the performance of ANN's to be put to use in smart phone based SLR with front camera video capture.

## References

1. Mohandes, Mohamed.: Arabic sign language recognition. International conference of imaging science, systems, and technology, Las Vegas, Nevada, USA. Vol. 1. (2001).
2. Al-Jarrah, Omar, Alaa Halawani.: Recognition of gestures in Arabic sign language using neuro-fuzzy systems. Artificial Intelligence, pp 117–138, (2001).
3. Kishore, P. V. V., et al.: 4-Camera model for sign language recognition using elliptical fourier descriptors and ANN. Signal Processing And Communication Engineering Systems (SPACES), 2015 International Conference on. IEEE, (2015).
4. Rao, Ananth. G., Kishore, P. V. V.: Sign Language Recognition System Simulated for Video Captured with Smart Phone Front Camera.” International Journal of Electrical and Computer Engineering (IJECE) vol. 6, 2176–2187, (2016).
5. Kumar, V. N., Narayana, K. V. L.: Development of an ANN-Based Pressure Transducer. in *IEEE Sensors Journal*, vol. 16, pp 53–60, (Jan. 1, 2016).
6. Zamani, Mahdi., Hamidreza Rashidy Kanan.: Saliency based alphabet and numbers of American sign language recognition using linear feature extraction” Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on. IEEE, (2014).
7. Zhang, Jihai., Wengang Zhou, Houqiang Li.: A new system for chinese sign language recognition” Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on. IEEE, (2015).