Ain Shams Engineering Journal 9 (2018) 1929–1939

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Selfie video based continuous Indian sign language recognition system

G. Ananth Rao [a], P.V.V. Kishore [b,*]

[a] Department of Electronics and Communications Engineering, K.L. University, India
[b] Applied Signal Processing Research Lab, Department of Electronics and Communications Engineering, K.L. University, India

ARTICLE INFO

Article history:
Received 15 July 2016
Accepted 2 October 2016
Available online 24 February 2017

Keywords:
Indian sign language
Sobel adaptive threshold
Morphological differencing
Mahalanobis distance
Multi layered artificial neural networks

ABSTRACT

This paper introduces a novel method to bring sign language closer to real time application on mobile platforms. Selfie captured sign language video is processed by constraining its computing power to that of a smart phone. Pre-filtering, segmentation and feature extraction on video frames creates a sign language feature space. Minimum Distance and Artificial Neural Network classifiers on the sign feature space is trained and tested iteratively. Sobel edge operator's power is enhanced with morphology and adaptive thresholding giving a near perfect segmentation of hand and head portions compensating for the small vibrations of the selfie stick. Word matching score (WMS) gives the performance of the proposed method with an average WMS of around 85.58% for MDC and 90% for ANN with a small variation of 0.3 s in classification times. Neural network classifiers with fast training algorithms will certainly make this novel selfie sign language recognizer application into app stores.

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 a time invariant finger orientations whereas dynamic gestures support a time varying hand orientations and head positions. The proposed four camera model for sign language recognition is a computer vision based approach and does not employ motion or colored gloves for gesture recognition.

An efficient sign language recognition system requires knowledge of feature tracking and hand orientations. Researchers around the world approached gesture classification in two major ways namely glove based and vision based. The former methods uses radio frequency gloves to tackle the problem. The method is less complicated and fast to implement on portable devices with complex hardware problems to attend. Computer vision requires no electronic hardware where advanced image processing algorithms can do hand shape matching and hand tracking on the captured video data. The missing attributes in glove based approach such as facial expressions and sign articulation are handled effectively using computer vision algorithms. Precision hampers the usability of computer vision techniques making it an ingenious research field.

Basically Sign language is used by the hearing impaired people for their communication. Using sign language we can communicate letters words or even sentences of general spoken language by using different hand signs and different hand gestures. This type of communication helps hearing impaired people to talk or express their views. These kind of systems bridge or channel between normal people and hearing impaired people.

Basic sign language system based on the 5 parameters and they are hand and head recognition, hand and head orientation, hand movement, shape of hand and location of hand and head (depends up on back ground). Among the five parameters there are two parameters which are most important and they are hand and head orientation and hand movement in a particular direction. These systems helps in recognizing the sign languages with better accuracy. Hand shapes and head are segmented and obtain feature vectors [1]. these feature vectors which are classified and given to neural networks for training. By advancing some methods in sign

* Corresponding author.
E-mail addresses: ananth.gondu@gmail.com (G.A. Rao), pvvkishore@kluniversity.in (P.V.V. Kishore).
Peer review under responsibility of Ain Shams University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.asej.2016.10.013
2090-4479/

language recognition a large research can be done for human computer interface. The major task in sign language recognition is signer identification, hand shape extraction, hand position extraction, signer facial expressions, body posture of the signer. For an excellent realizable recognition system the above five attributes are to be inputted to the system.

Complex video backgrounds in another gap in sign language recognition where it poses a major challenge to extract signers hand shapes accurately. Most of SLR systems focus on putting a simple constant background with signer's shirt matching the background. In cluttred video backgrounds tracking hands has become simpler but tracking each finger movements remains quite a challenging task. Here researchers believe 3Dimensional body centered space of the signer can be used effectively for extracting finger movements. The 3D locations of fingers are referenced by setting points of knuckles in space. Creating the 3D points as spatial domain information for hand tracking and hand shaping is challenging for computer vision engineers [2].

To make sign language recognition system, we introduce a selfie sign language recognition system capturing signs using a smart phone front camera. The signer holds the selfie stick in one hand and signs with his other hand. The input sign video is processed for corresponding text or voice outputs for normal people to understand a hearing impaired person without the need for an interpreter. Two major problems surfaced during implementation phase. One the signs are preferably single handed and the other video background variations due to the movement of selfie stick in the hand of the signer.

Another most noticeable challenge faced by the researcher in sign language recognition is background of the signer relating to the contrast of the light where signer is present. We have limited our research to simple backgrounds. More research can be taken up on selfie based sign language recognition with real time constraints such as non-uniform background, varied lighting and signer independence, to make the system independent. Object detection (segmentation) is done by applying gradient masking to the image. This is done because the background differs greatly with the object in the image. Obtaining thresholds with different segmentation operators. Tune the system with threshold values and applying the edge to get binary masking of the image. The binary gradient masking is usually dilated using vertical structuring elements like "ball", "disk", "diamond" by horizontal structuring elements.

A sentence in sign language is recorded using a camera and the obtained video is divided into several frames. Each sign in a frame taken out of a set of frames is processed and the features are extracted such that the features apply for nearby preceding and succeeding frames. Our research mainly deals with selfie videos of sign language which are divided into frames for processing. The proposed system concentrates on segmenting hand and head from the given set of frames probably for various signs in the video and features are extracted for various hand and head models.

In Selfie SLRs people can communicate with the help of their phones by recording their sign video and sending it to the other person. The sent video will be decoded according to the system we designed and the signs are inevitably converted to text.

Reyadh [3] worked on an alphabet sign recognition system with a recognition rate with naked hand of 50%, red hand of 75%, black Hand of 65% and white hand of 80%. The sign alphabet images are mapped to histograms to uniquely represent the sign images. KNN algorithm which measures the distance between these sign histograms for classification. The process is simple and effective only on images and produced a very low recognition rate on video data.

In [4], 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 [5] proposed a neuro fuzzy system that deals with images of simple hand signs and succeeded a recognition rate of 90.55%.

Kishore PVV, proposed [6] 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%.

Sign language recognition acts as a machine interpreter (MI) between a mute person and normal person. Active contours energy function is formulated by amalgamating energy function from boundary and shape prior elements. Artificial Neural Network is constructed to classify and recognize gestures from video frames of signers. The proposed VVMI [7] for SLR offers a recognition rate of around 93%.

The dynamic time wrapping based level building (LB-DTW) algorithm was proposed in [8] to solve sign sequence segmentation and sign recognition. This LB-DTW introduces two problems in recognition. One is under the bad relationship the recognition rate is very low and HMM was incorporated to improve recognition rate by calculating the similarity between sign model and testing sequence. On the other hand, the grammar constraint and sign length constraint are employed to improve recognition rate. In experiments with a KINECT data set of chines sign language containing 100 sentences composed of 5 signs each, the proposed method shows superior recognition performance and lower computation compared to other existing techniques.

The method proposed in [9] involves extracting the hand gestures form original color images. The segmented hand positions shape modulated using Chan-Vese (CV) active contour model and obtained 92.1% recognition rate.

The extensive literature provides a SLR system design that was never tested for real time application on mobile platforms. In this work we try to simulate this new approach for sign language recognizer. Selfie camera captures sign language continuous videos under simple backgrounds. The signs are single handed simple signs, to facilitate selfie capture by holding selfie stick in the other hand of the signer.

Pre-filtering for removing video capture noise during image acquisition is done using multiple sets of Gaussian filters of zero mean and 0.1–0.5 range variances. Hand and head contour shape extraction uses sobel edge operator enhanced with morphological operation. Shape energy is modelled with discrete cosine transform (DCT) and this feature is optimized with principle component analysis (PCA). Finally each frame is represented with a 1×50 feature vector. An average of 220 frames were detected per sign making the feature matrix 220×50 per video. 20 signs having English alphabets, sentences in regular use are captured. To cut down on classification time minimum distance classifier (MDC) with Euclidian distance is employed at the cost of accuracy. More accurate classifications are obtained with neural networks.

Word matching score (WMS) is the performance estimator when testing the proposed SLR system. It is the ratio of true classifications to total signs inputted to classifier. Multiple testing with 10 different signers has resulted in an average WMS of 93.23%. Further this is the first time in literature this kind of method is proposed for capturing sign videos. But the processing is done using already proposed methods in literature to make the system process sign videos in short period of time.

The paper is prepared as follows. Section 2 gives research methodology, mathematical modelling and derived algorithms for pre-filtering, segmentation, feature extraction and classification. Results and analysis is in Section 3. Section 4 concludes this novel idea of SLR capture with a smart phone selfie front camera.

2. Pre-processing, segmenation, feature extraction and classification

The flow chart 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 in mp4 format having full HD 1920 × 1080 video recording on a 5 M pixel CMOS front camera. Let this 2D video be represented as a 2D frame $\Im(\boldsymbol{x},\boldsymbol{y}) \rightarrow \mathbb{R}^2 \forall(\boldsymbol{x},\boldsymbol{y}) \rightarrow \mathbb{Z}^+$. For video the frame $\Im(\boldsymbol{x},\boldsymbol{y})$ changes with time, which is fixed universally at 30 frames per second. These videos form the database of this work. A 3-fold 2D Gaussian filter $\boldsymbol{K}_{2D}(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2}\boldsymbol{e}^{\frac{-(\boldsymbol{x}-\boldsymbol{m})^2}{2\sigma^2}}$ [10] with zero mean (m = 0) and three variances of σ = 0.01, 0.1, 0.15 smoothens each frame by removing sharp variations during capture. For 2D gradient calculation, two 1D gradients in $\boldsymbol{x}$ and $\boldsymbol{y}$ directions of the frame matrix are computed as follows.

$$\boldsymbol{g}^{\boldsymbol{x}} = \sum_{\boldsymbol{k}=1}^{\boldsymbol{N}} \Im(\boldsymbol{x}-\boldsymbol{k},\boldsymbol{y})\boldsymbol{g}(\boldsymbol{k}) \tag{1}$$

$$\boldsymbol{g}^{\boldsymbol{y}} = \sum_{\boldsymbol{k}=1}^{\boldsymbol{N}} \Im(\boldsymbol{x},\boldsymbol{y}-\boldsymbol{k})\boldsymbol{g}^{\boldsymbol{T}}(\boldsymbol{k}) \tag{2}$$

where $\boldsymbol{g} \rightarrow [+1,-1]$ is the discrete gradient operator. The gradient magnitude $\boldsymbol{G}^{\boldsymbol{xy}}$ gives magnitude of edge strength in sobel edge detector computed as $\boldsymbol{G}^{\boldsymbol{xy}} = \sqrt{(\boldsymbol{g}^{\boldsymbol{x}})^2 + (\boldsymbol{g}^{\boldsymbol{y}})^2}$.

For convenience sobel represented the function using a 2D convolution mask $\boldsymbol{S}^{Mx} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$ and $\boldsymbol{S}^{My} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}^T$. These masks 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 of each 3 × 3 sobel mask is used as threshold. The final binary image is

$$\begin{aligned} \boldsymbol{B}^{x} &= \sum_{\boldsymbol{x}=1}^{\boldsymbol{N}} \sqrt{\left(\boldsymbol{S}^{\boldsymbol{Mx}} \otimes \Im^{\boldsymbol{x}}\right)^2 + \left(\boldsymbol{S}^{\boldsymbol{My}} \otimes \Im^{\boldsymbol{y}}\right)^2} \\ &\geqslant \sum_{\boldsymbol{i}=1}^{\boldsymbol{b}} \sum_{\boldsymbol{x}=1}^{\boldsymbol{N}} \sqrt{\left(\boldsymbol{S}^{\boldsymbol{Mx}} \otimes \Im^{\boldsymbol{x}}\right)^2 + \left(\boldsymbol{S}^{\boldsymbol{My}} \otimes \Im^{\boldsymbol{y}}\right)^2} \end{aligned} \tag{3}$$

where $\boldsymbol{b}$ is the block size. This procedure is fast and reduces background variations automatically without human intervention of selecting a suitable threshold as in case of sobel edge operator. Fig. 2 shows the difference in block thresholding and global thresholding (used 0.2) which failed to handle motion blur.

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

Contour extraction is represented as

$$\boldsymbol{H}^{\boldsymbol{C}}(\boldsymbol{x}) = \left\{\boldsymbol{z}|(\hat{\boldsymbol{M}}_{3\boldsymbol{H}})_{\boldsymbol{z}} \cap \boldsymbol{B}^{\boldsymbol{x}} \neq \varnothing\right\} - \left\{\boldsymbol{z}|(\hat{\boldsymbol{M}}_{3\boldsymbol{H}})_{\boldsymbol{z}} \subseteq \boldsymbol{B}^{\boldsymbol{x}}\right\} \tag{4}$$

$$\boldsymbol{H}^{\boldsymbol{C}}(\boldsymbol{y}) = \left\{\boldsymbol{z}|(\hat{\boldsymbol{M}}_{3\boldsymbol{V}})_{\boldsymbol{z}} \cap \boldsymbol{B}^{\boldsymbol{x}} \neq \varnothing\right\} - \left\{\boldsymbol{z}|(\hat{\boldsymbol{M}}_{3\boldsymbol{V}})_{\boldsymbol{z}} \subseteq \boldsymbol{B}^{\boldsymbol{x}}\right\} \tag{5}$$

$$\boldsymbol{H}^{\boldsymbol{C}}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{H}^{\boldsymbol{C}}(\boldsymbol{x}) \oplus \boldsymbol{H}^{\boldsymbol{C}}(\boldsymbol{y}) \tag{6}$$

Hand and head contours are separated by finding the connected components with maximum number of pixels with a 4 neighbourhood operation on the contour image $\boldsymbol{H}^{\boldsymbol{C}}(\boldsymbol{x},\boldsymbol{y})$. Fig. 3 provides visual conformation of the discussion.

Features are unique representation of objects in this world. Feature is a set of measured quantities in a 1D space represented as $\boldsymbol{F}^{\boldsymbol{V}}(\boldsymbol{x}) = \{\boldsymbol{f}(\boldsymbol{x})|\boldsymbol{x} \subseteq \mathbb{R}\}$, where $\boldsymbol{f}(\boldsymbol{x})$ can be any transformation or optimization model on vector *x*. Top priority in this work is sped of execution of the proposed algorithm. Hence $\boldsymbol{f}(\boldsymbol{x})$ is considered as Discrete Cosine Transform (DCT) along with Principle Component Analysis (PCA). The 2D DCT of hand contour $\boldsymbol{H}^{\boldsymbol{C}}(\boldsymbol{x})$ and head contour$\bar{\boldsymbol{H}}^{\boldsymbol{C}}(\boldsymbol{x})$ is computed as

$$\boldsymbol{F}_{\boldsymbol{uv}}^{\boldsymbol{V}} = \frac{1}{4}\boldsymbol{C}^{\boldsymbol{u}}\boldsymbol{C}^{\boldsymbol{v}} \sum_{\boldsymbol{x}=1}^{\boldsymbol{N}} \sum_{\boldsymbol{y}=1}^{\boldsymbol{N}} \boldsymbol{H}^{\boldsymbol{C}}(\boldsymbol{x}) \cos\left(\boldsymbol{u}\pi\frac{2\boldsymbol{x}+1}{2\boldsymbol{N}}\right) \cos\left(\boldsymbol{v}\pi\frac{2\boldsymbol{y}+1}{2\boldsymbol{N}}\right) \tag{7}$$

where $\boldsymbol{C}^{\boldsymbol{u}} = \boldsymbol{C}^{\boldsymbol{v}} = \frac{1}{\sqrt{2}} \forall(\boldsymbol{uv}) = 0$ and 1 elsewhere. Similar expression with $\bar{\boldsymbol{H}}^{\boldsymbol{C}}(\boldsymbol{x})$ calculated 2D DCT of head contour as $\bar{\boldsymbol{F}}_{\boldsymbol{uv}}^{\boldsymbol{V}}$. Fig. 4 shows a color coded representation of hand DCT features for the frame in a video sequence. The head does not change much in any of the frames captured and hence head contour DCT remains fairly constant throughout the video sequence.

The first 50 × 50 matrix of values possess maximum amount of energy in a frame. But this DCT matrix for every frame consisting of 2500 values representing a sign will cost program execution time. PCA treatment of the matrix $\boldsymbol{F}_{\boldsymbol{uv}}^{\boldsymbol{V}}$, which retains only the unique components of the matrix $\boldsymbol{F}_{\boldsymbol{uv}}^{\boldsymbol{V}}$. The final $\boldsymbol{F}_{\boldsymbol{uv}}^{\boldsymbol{V}}$ is represented as $\boldsymbol{F}_{\boldsymbol{fn}}^{\boldsymbol{V}}$, where $\boldsymbol{fn}$ gives frame number. PCA reduces the feature vector per frame to 50 sample values per frame. Each 50 sample Eigen vector from PCA uniquely represents DCT energy of the hand shape in each frame.

The feature sign matrix $\boldsymbol{F}_{\boldsymbol{fn}}^{\boldsymbol{V}}$ inputs a classifier. Since speed is the prime constraint during mobile implementation, it will be

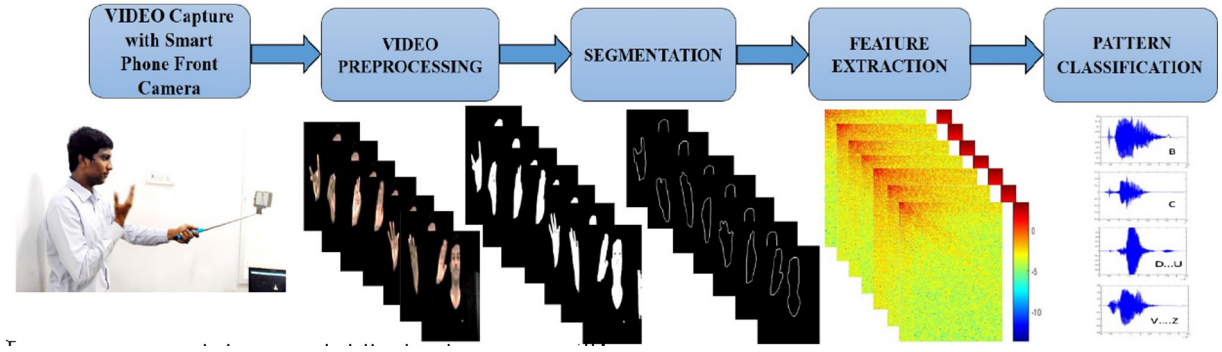


Fig. 1. Flow chart of sign language recognition system with smart phone front camera video capture.

Fig. 2. (a) Block variational mean thresholded frame. (b) Global threshold of 0.2 for sobel mask.

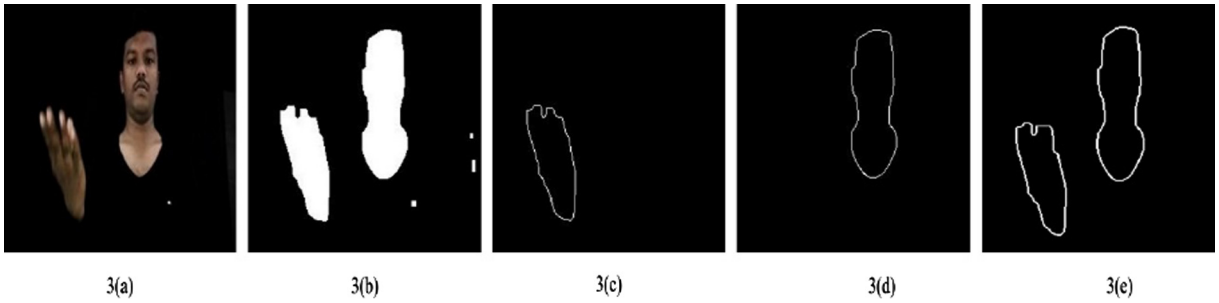


Fig. 3. (a) Capture 98th frame. (b) Segmented frame. (c) Hand contour. (d) Head contour and (e) both contours.

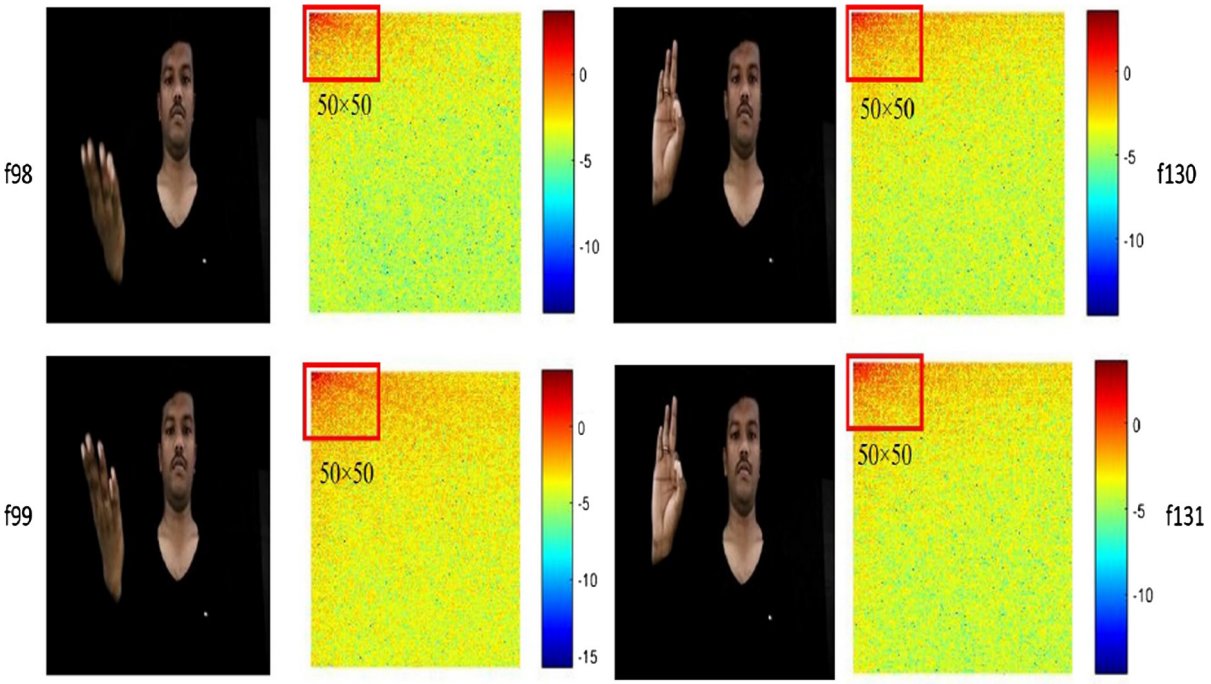


Fig. 4. 2D DCT representation of hand contour energy representations.

reasonable to use minimum distance classifier (MDC). This is one simple classifier that does not require prior training. Mahalanobis distance is the metric that will assign class variables to different sign classes. Mahalanobis distance [11] is chosen for SLR classification on smart phones over Euclidian distance, the former includes inter sample covariance's in different directions during distance calculation. The Mahalanobis distance equals Euclidian distance for uncorrelated data with inter class variance of unity. The squared Mahalanobis distance D_M^2 is given as

$$D_M^2 = (F_{fn}^V - S_C)^T \sum_C^{-1} (F_{fn}^V - S_C) \tag{8}$$

where S_C is mean vector of each sign class defined by F_{fn}^{V}. $\sum_C$ Is the inter class covariance matrix and $\sum_C^{-1}$ is its inverse matrix. Where T is transposition. But faulty distances are measured if the inter class variance is very large. In sign language videos hand shape variations within a particular sign class are very small making Mahalanobis distance ideal for sign language classification. Further study uses artificial neural network based classifier design to test the proposed method.

The feature sign matrix inputs a classifier. Since speed is the prime constraint during mobile implementation due to 4 GB rams on smart phones, we use artificial neural networks for classification with only 3 layers. In the previous work the focus was on a simple classifier that does not require prior training. Mahalanobis distance [11] is chosen in our previous work for SLR classification on smart phones over Euclidian distance, because the former includes inter sample covariance's in different directions during distance calculation.

Faulty distances were measured where the inter class variance is large. In sign language videos hand shape variations within a particular sign class are very small making Mahalanobis distance ideal for sign language classification for the same signer. Signer independent SLR with Mahalanobis distance could not be achieved due huge variations in signer hand and head shapes. Hence the simulations use artificial neural network based classifier design to test the proposed method. The details of ANN with back propagation algorithm are listed in our previous work at [12] and the models used for coding are considered from [13,14]. Further we also study the usefulness of multilayer feed forward ANN's with more than 3 layers along with their performance analysis on sign classifications.

The model of 3 layered artificial neural network is presented in Fig. 5(a). A 3 layered feed forward network of neurons as in Fig. 3 is simulated with features as input to 1st 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 methods, even though there are algorithms for estimation. For this sign classification the estimated neurons are twice the neurons in the input layer for correct classification at reasonably less simulation times. The learning rate is fixed at 0.2, which gives the learning speed of the network. An optimal value is always encouraging. Initial weights and biases are randomly assigned. Error handling rate is 0.01. Error back propagation algorithm trains the network with sigmoid activation function as the deciding function in each neuron. Following similar procedure and increasing the number of hidden layers training is initiated for 2 hidden, 3 hidden and 5 hidden layers. A simulation of this novel method with analysis of the outcomes is considered next.

3. Results and analysis

The front camera video recording of sign language gestures using with smart phones Asus Zen phone II and Samsung galaxy S4 at the end of selfie stick. Both the mobiles are equipped with 5 M pixel front camera. Sign video capturing is constrained in a controlling environment with room lighting and simple background. The first photo in Fig. 1 demonstrates the procedure followed for signers for video capture. The results discussion 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 handled.

3.1. Visual analysis

Each video sequence is having a meaningful sentence. The following sentence is "Hai Good Morning, I am P R I D H U, Have A Nice Day, Bye Thank You". There are 18 words in the sentence. The words in the training video are sequenced in the above order but testing video contains same words in different order. Classification of the words is tested with Euclidian, Normalized Euclidian and Mahalanobis distance functions. Few frames of the video sequence are in Figs. 6–9.

The advantage of using front camera is pronounced from Fig. 8. Here the signer corrects himself during the signing process which helps to give correct hand sign to the system.

Filtering and adaptive thresholding with sobel gradient produces regions of signer's hands and head segments. Morphological differential gradienting with respect to line structuring element as in Eqs. (4)–(6) refines the edges of hands and read portions.

Fig. 10 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 is uses orthogonal basis functions that represent the signal energy with minimum number of frequency domain

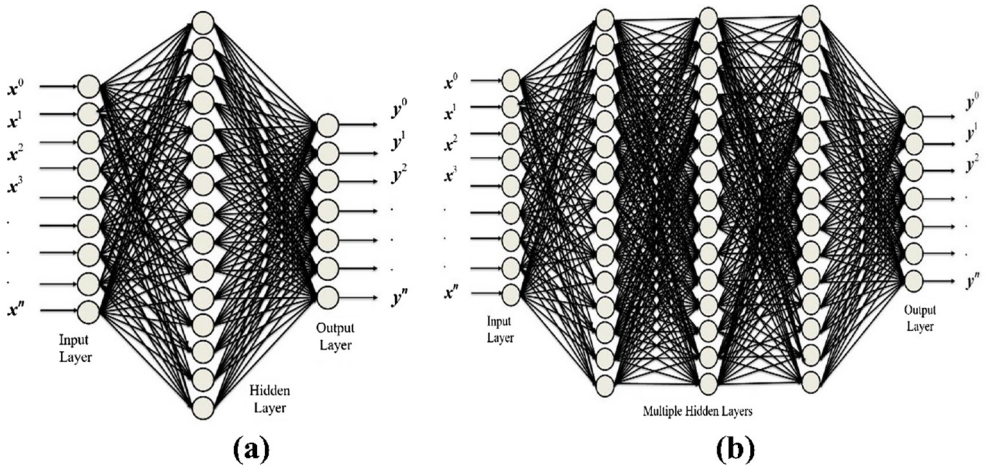


Fig. 5. (a) Conventional 3 layer neural network model used for sign classification. (b) Multi layered ANN with 3 hidden layers and 2 input and output layers.

Fig. 6. Few sequence of frames for sign 'HAI or HELLO'.

Fig. 7. Sign frames for 'GOOD' (Top) and 'Morning' (Bottom).

Fig. 8. Frames of Sign 'I' 'AM'.

Fig. 9. Single frames per sign. (From top left): 'P', 'R', 'I', 'D', 'H', 'I AM', 'THANK', 'YOU', 'BYE', 'NO SIGN'.

samples that can effectively use to represent the entire hand and head curvatures. As shown in Fig. 4, first 50×50 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 50×50 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 a multi dimension feature matrix of this size takes longer execution periods. Hence PCA treats each frames 50×50 energy features by computing Eigen vectors and retaining the principle components to from a 50×1 vector per frame. A combination of these features represent a sign in a video sequence. When no hand is detected in the frame it is considered as 'No Sign'. These particular frames are detected as their feature matrix is having only head contour energy samples.

The training vector contains a few head only sample values for such 'No Sign' detection. Three classifier are compared to test the execution speeds matching that of smart phone execution. 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).

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 $M^{S\%} = \frac{Correct\ \ Classifications}{Total\ \ Signs\ \ in\ \ a\ \ Video} \times 100$. Feature matrix has a size of 50×220, each row representing a frame in the video sequence.

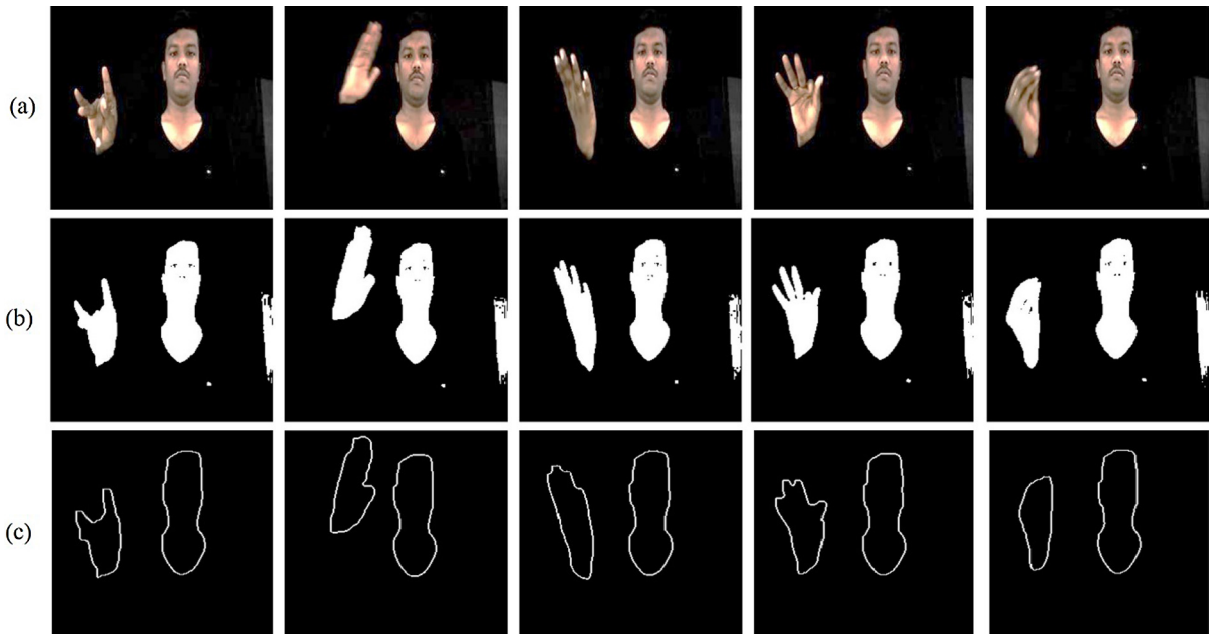


Fig. 10. (a) Few frames in RGB format. (b) Their region segments with Gaussian filtering and sobel operation. (c) Contours of hands and head produced with morphological subtraction with line structuring elements.

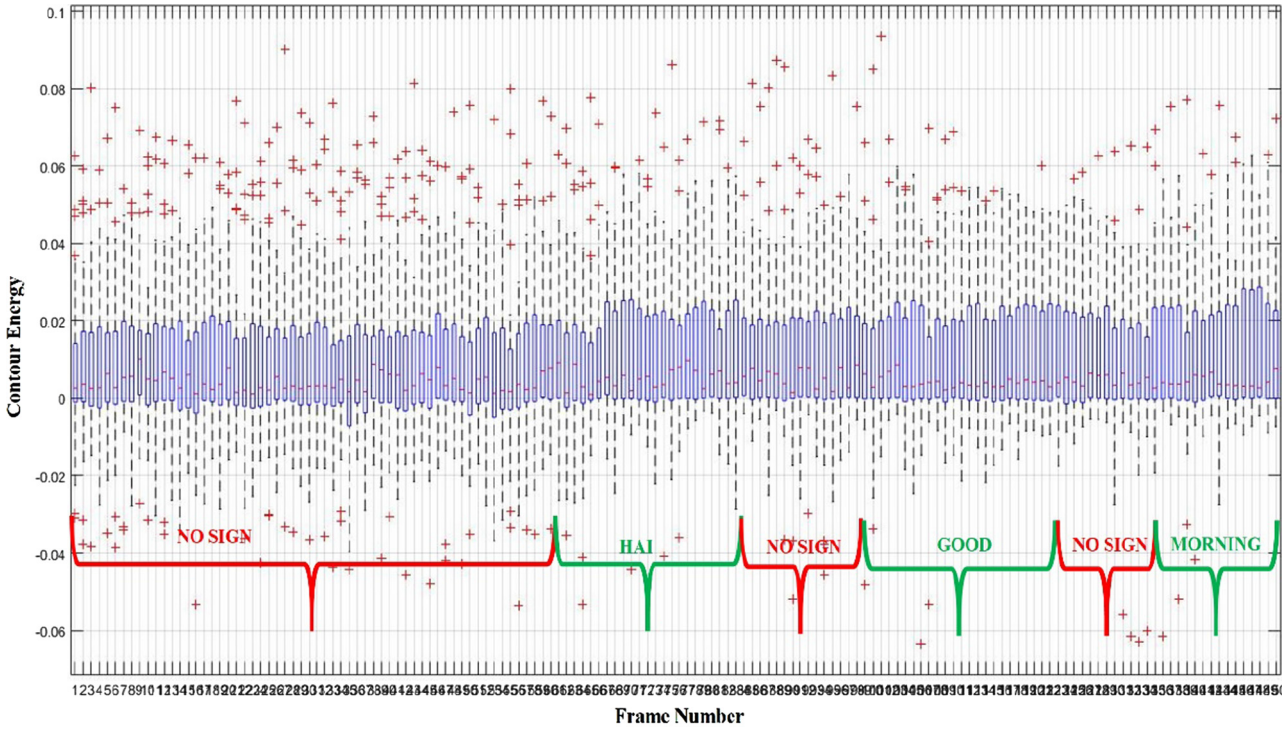


Fig. 11. Sample Energy distribution in frames for identification of signs.

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 and plotted during testing in Fig. 11.

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 the inter class variance considerations as in Eq. (8). Test repetition frequency is 10 per sign.

For more exhaustive testing different signer's videos were used against the previous training sample and the results were tabulated in Table 2 for all three distance measures.

To find the average WMS for all the different signer video samples and their performance against same sample train-test data and different sample train test data with respect to three distance metrics, From the analysis the following observations can me

Table 1
The performance of three minimum distance classifiers.

Signs	Euclidian distance classifer	Normalized euclidian distance	Mahalanobis distance classifier
HAI	80	70	90
GOOD	70	70	90
MORNING	80	80	80
I AM	60	60	80
P	90	90	100
R	90	90	100
I	90	90	100
D	90	90	100
H	90	90	100
U	90	90	100
HAVE	50	50	80
A	90	80	100
NICE	60	50	80
DAY	70	70	80
BYE	50	50	90
THANK	50	50	90
YOU	60	50	80
Average WMS	74.11	71.76	90.58

Table 2
The performance of three minimum distance classifiers with different testing video.

Signs	Euclidian distance classifer	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

made: (i) for same test train samples all distance metrics show good WMS. (ii) For different train test data, Mahalanobis distance performance is better for all test data. Further WMS decreases by 4–6% if the number of frames in test train data does not match. For perfect matching the WMS is 3% above average.

To further improve performance of the classification process an artificial neural network to complete the task of identifying and classifying 18 gesture signs. The input layer 18 neurons and 10 output neurons in the hidden layer in neural network. This section presents the results of the trials to categorize gestures of selfie sign language with neural network classifier. The training set of data consists of 18 video sequences of size 256×256. We have a total of 18 different continuous signs by 10 different signers taking the sign count to 180.

The network is trained and tested for different samples of the database [9]. Table 3 gives the details of the training and testing process. The first two columns give number of frames for training and testing. The third column shows the neural network architecture created for training and testing. Column four shows the output confusion matrix of testing. The last column shows the recognition rates calculated from testing the corresponding networks.

Network training used error backpropagation algorithm with sigmoid activation function given by $\boldsymbol{s}(\sum \boldsymbol{net}) = \frac{1}{1+\boldsymbol{e}^{-\sum \boldsymbol{net}}}$, where $\sum \boldsymbol{net}$ is the sum of neural networks inputs multiplied with weights. Initial weight matrix is randomly chosen for each test. Therefor multiple trains of the network were needed to get the correct estimate. The target matrix is binary having the same size as the input matrix. Learning rate and momentum factor were 0.2 and 0.9 respectively. Error tolerance for all training and testing phase is set at 0.01.

For training 2 sets of videos with a total of 524 frames 262 frames each continuous sign sequence from two different signers were chosen. 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 789 frames trained the ANN and from that 3 sets, 2 sets are tested i.e. 524 frames. The number of hidden neurons were 125 and found an increase in WMS at 85.5%. Similar trend observed in results shown at the row 3 of Table 3. The WMS significantly improved for higher sample training with a compromise in speed. Hidden neurons were 200 in the last testing phase.

Multiple layers were introduced for training and testing. 2 and 4 hidden layer feed forward back propagation networks with same parameters showed improvement in terms of WMS at 86.4% and 96.9% respectively. The WMS is better than the network with single hidden layer. But training was relatively slow compared to single hidden layer. Training for single hidden layer network took 9.213 s for a total of 36 epochs, whereas for 2 hidden network it was 1800.23 s at 18 epochs. The number was huge for 4 hidden network at 3800.52 s at 12 epochs. The number of epochs decreased sharply but the lost training time.

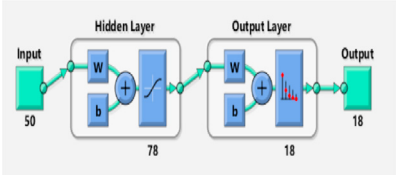
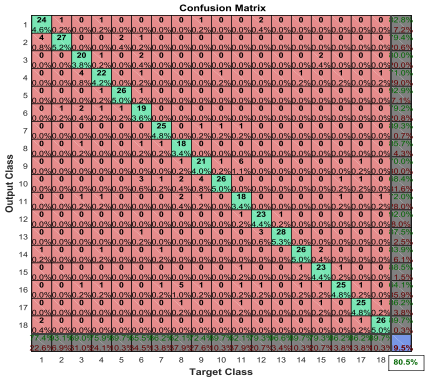
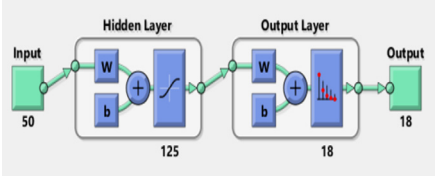
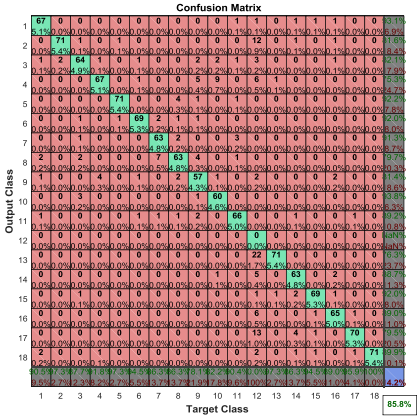
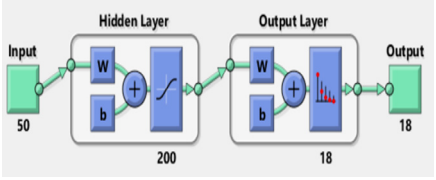
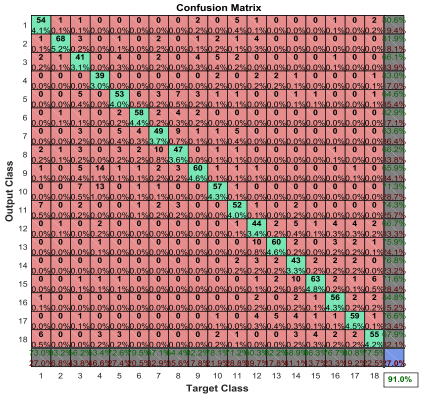
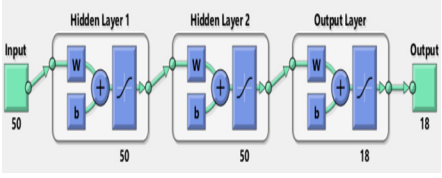
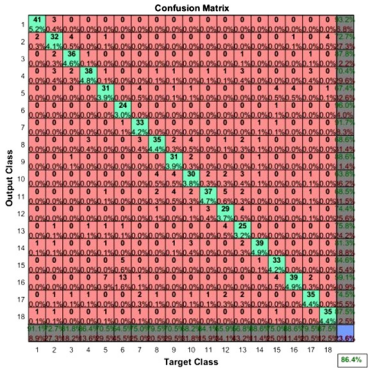
Testing output simulations were really fast and accurate for a previously unseen frames as well. In all the above cases testing times were 0.00432 s. This is due to the use of same number of neurons in both input and hidden layers. The average recognition rate (WMS) is 90% for the total classification method which is on par with other researchers for American Sign Language [15] and Chinese Sign Language in [16]. 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. The results of the testing are reported in Table 4.

Table 4 summarizes the entire testing of the proposed gesture classification process. The last row giving the average values for the entire gesture recognition process using ANN.

Table 4 values are standardized with respect to 100 hidden neurons, 50 input neurons and 18 output neurons. Fig. 12 shows true positive rate for all classes is around 0.8–0.9 and in some cases is equal to unity. This plot reflects samples for penultimate row in Table 4.

Two classifiers based on WMS with same training and testing inputs and different training and testing inputs are compared to understand the classifiers and the type of data they use for better performance. Minimum Distance Classifier (MDC) with Mahalanobis distance produces a 85.5% WMS with an execution time on 4 GB RAM computer with MATLAB 2015 of 0.4823 s for a data set of 1313 frames. The ANN based classifier with same parameters produced a 90% WMS at 0.5452 s. When tested with a different data set of continuous signs MDC registered a WMS of 59.66% and problems arise if the test data set frame number does not match the train data set. Whereas ANN handled these problems nicely because of training. Training does not depend on frame number, it depends on the sample size extracted from each frame. ANN's in the recent years have become faster and can be considered for smart phone based sign language recognition system design.

Table 3
Details of Training and testing of sign videos under simple backgrounds with Different Samples and their recognition rates.

Training samples	Testing samples	Network architecture	Output confusion matrix	WMS (%)
18 Signs 524 (2 Sets) Frames	18 Signs (2 Sets) 524 Frames			80.5
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
18 signs (5 Sets) 1313 Frames	18 Signs (3 Sets) 524 Frames			86.4

(*continued on next page*)

Table 3 (*continued*)

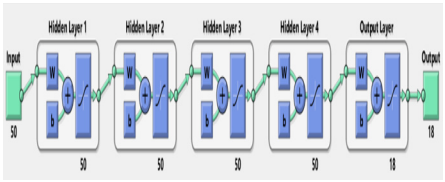
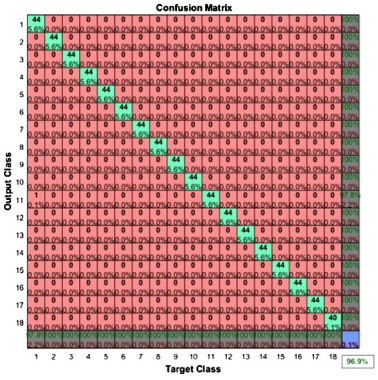
Training samples	Testing samples	Network architecture	Output confusion matrix	WMS (%)
18 signs (5 Sets) 1313 Frames	18 Signs (3 Sets) 524 Frames			96.9

Table 4
Gesture classification results summary.

Number of frames for training	Number of epochs for training	Number of unknown samples for testing	Number of correctly recognized samples	Recognition rate (%)
220	25	189	152	80.47
524	33	220	189	85.90
789	36	220	198	90.18
1313	51	1111	1027	92.49
2846	**145**	**1740**	**1566**	**90**

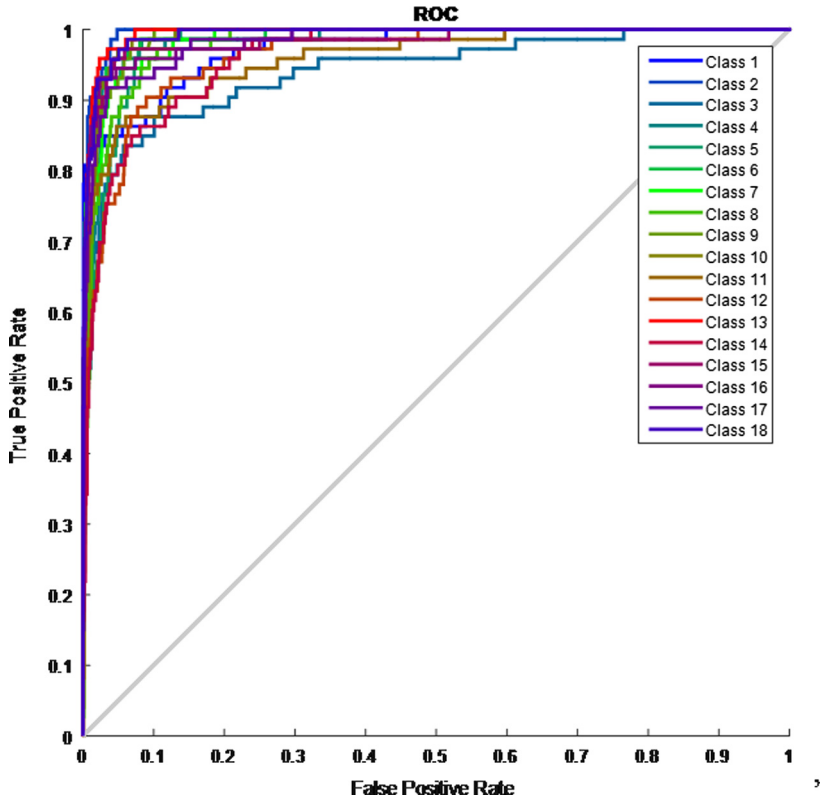


Fig. 12. ROC for penultimate row in Table 2.

4. Conclusion

A novel idea of putting sign language into smart phones is simulated and tested. 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. Pre-filtering, segmentation and contour detection are performed

with Gaussian filtering, sobel with adaptive block thresholding and morphological subtraction respectively. 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. Similarly multilayer ANN with 2 and 4 hidden layer registered larger training times compared to single hidden layer networks. But at the same time the WMS accuracy showed improvement with less number of epochs and testing times. 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.

Acknowledgment

We would like to thank K.L. University, Cams Department for providing facilities and the students of ECE department for volunteering in sign language database creation.

References

[1] Li Kehuang, Zhou Zhengyu, Lee Chin-Hui. Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications. ACM Trans Access Comput (TACCESS) 2016;8(2):7–23.
[2] Kong WW, Ranganath Surendra. Towards subject independent continuous sign language recognition: a segment and merge approach. Pattern Recog 2014;47 (3):1294–308.
[3] Naoum, Reyadh, Hussein H. Owaied, and Shaimaa Joudeh. Development of a new Arabic sign language recognition using k-nearest neighbor algorithm." (2012).
[4] Mohandes, Mohamed. "Arabic sign language recognition." International conference of imaging science, systems, and technology, Las Vegas, Nevada, USA. Vol. 1. 2001.
[5] Al-Jarrah Omar, Halawani Alaa. Recognition of gestures in Arabic sign language using neuro-fuzzy systems. Artificial Intelligence 2001;133(1):117–38.
[6] 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.
[7] Kishore, P. V. V., A. S. C. S. Sastry, and A. Kartheek. "Visual-verbal machine interpreter for sign language recognition under versatile video backgrounds", Networks & Soft Computing (ICNSC), 2014 First International Conference on. IEEE, 2014.
[8] Yang Wenwen, Tao Jinxu, Ye Zhongfu. Continuous sign language recognition using level building based on fast hidden Markov model. Pattern Recognition Letters 2016.
[9] Kishore, P. V. V., and P. Rajesh Kumar. "Segment, Track, Extract, Recognize and Convert Sign Language Videos to Voice/Text". International Journal of Advanced Computer Science and Applications (IJACSA) ISSN (Print)-2156 5570 (2012).
[10] Neha Baranwal, Neha Singh and G.C.Nandi "Indian Sign Language Gesture Recognition Using Discrete Wavelet Packet Transform" in 2014 International Conference on Signal Propagation and Computer Technology (ICSPCT), 2014, pp. 573577.
[11] Mei J, Liu M, Wang YF, Gao H. Learning a mahalanobis distance-based dynamic time warping measure for multivariate time series classification. IEEE Transactions on Cybernetics 2016;46(6):1363–74.
[12] Kim N. Euclidian distance minimization of probability density functions for blind equalization. Journal of Communications and Networks Oct. 2010;12 (5):399–405.
[13] V. N. Kumar and K. V. L. Narayana, "Development of an ANN-Based Pressure Transducer", in IEEE Sensors Journal, vol. 16, no. 1, pp. 53–60, Jan. 1, 2016.
[14] S. K. Gharghan, R. Nordin, M. Ismail and J. A. Ali, "Accurate Wireless Sensor Localization Technique Based on Hybrid PSO-ANN Algorithm for Indoor and Outdoor Track Cycling" in IEEE Sensors Journal, vol. 16, no. 2, pp. 529–541, Jan. 15, 2016.
[15] Zamani, Mahdi, and 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.
[16] Zhang, Jihai, Wengang Zhou, and 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.

G. Anantha Rao received B.Tech Degree from, GMRIT, JNTU, Hyderabad, In 2007. M.Tech. Degree from STIET, JNTUK, Kakinada, India In 2011, Pursuing Ph.D. In The Department of Electronics and Communication Engineering, KL University, Vijayawada, India. Currently He is working as Assistant Professor in The Department of Electronics and Communication Engineering, AIET, Vizianagaram, INDIA. His research interest includes on Signal Processing, Image and Video Processing.

P.V.V. Kishore is having Ph.D degree in electronics and communications Engineering from Andhra University College of engineering in 2013. He received M.Tech from Cochin University of science and technology in the year 2003.He received B.Tech degree in electronics and communications engineering from JNTU, Hyd. in 2000. He is currently full professor and Image,signal and speech processing Head at K.L.University, ECE Department. His research interests are digital signal and image processing, Artificial Intelligence and human object interactions. He is currently a member of IEEE. He has published 60 research papers in Various National and International journals and conferences including IEEE, Springer and Elsevier.