Sign Language Recognition System Simulated for Video Captured with Smart Phone Front Camera

Article · November 2016

DOI: 10.11591/ijece.v6i5.11384

2 authors, including:

P.V.V. Kishore
K L E F (Deemed - to - be -University)
144 PUBLICATIONS **680** CITATIONS

SEE PROFILE

International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 5, October 2016, pp. 2176~2187
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i5.11384

Sign Language Recognition System Simulated for Video Captured with Smart Phone Front Camera

G. Ananth Rao, P.V.V. Kishore
Department of Electronics and Communications Engineering, K.L. University, India

Article Info

Article history:

Received May 20, 2016
Revised Jul 22, 2016
Accepted Aug 10, 2016

Keyword:

Indian sign language
Mahalanobis distance
Mobile platform
Sobel adaptive threshold
Morphological differencing

ABSTRACT

This works objective is to bring sign language closer to real time implementation on mobile platforms. A video database of Indian sign language is created with a mobile front camera in selfie mode. This video is processed on a personal computer by constraining the computing power to that of a smart phone with 2GB ram. Pre-filtering, segmentation and feature extraction on video frames creates a sign language feature space. Minimum distance classification of the sign feature space converts signs to text or speech. ASUS smart phone with 5M pixel front camera captures continuous sign videos containing around 240 frames at a frame rate of 30fps. 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) estimates performance of the proposed method with an average WMS of around 90.58%.

Corresponding Author:

P.V.V. Kishore,
Department of Electronics and Communications Engineering,
K.L. University,
Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, India.
Email: pvvkishore@kluniversity.in

1. INTRODUCTION

Computer vision is a field that features strategies to process, analyze and understand images. It aims to facsimile the potential of human vision by apprehending an image. One of the major hindrance to the applications of computer vision is motion estimation, shape estimation and analysis. Sign language is a computer vision based intact intricate language that engages signs shaped by hand moments in amalgamation with facial expressions and postures. It maps natural way of communication to human signs and gestures enabling hearing impaired people to communicate among them. Dynamic hand movements are involved in gestures and they form signs such as numbers, alphabets and sentences. 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. Approach for identification of sign language can be categorized as gloved based approach and computer vision based approach. The former involves gloves to be employed by signers during the performance of gestures. It minimizes the complexity of segmentation in the stages of processing but complicates the hardware requirements. Vision based approach advances image processing methodologies for hand sign detection and tracking. This approach also adds on signer facial proclamation and is feasible to signer as it does not involve the utilization of gloves. Accuracy hinders the usability of image processing approach making it an enterprising 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 language recognition a large research can be done for human computer interface. The major and foremost difficulty or task in sign language recognition is to identify the signer, facial expressions, posture of the person and not only his hand moments but also his head moments. All these attributes combine together can make a good recognition system.

The next major problem observed in sign language recognition system is to track the signer form the video by avoiding the other variations in the video. This the major task where all the researchers concentrate on. Tracking hand is quite simpler compared to track all the finger variations, which is the most difficult task to perform, a sign language space can be obtained with different entities such as humans or objects stored in it around a 3Dimensional body centered space of the signer. These entities are located with certain locations and later referenced it by pointing to space. To define a model for spatial information, containing entities is another challenge faced by researchers [2].

In selfie sign language recognition system the hand position plays a crucial role because we are implementing for different positions with single hand can cause overlap of hand on face. In selfie the background is not constant which varies constantly which is one of the challenging task in this recognition system.

Another most challenge faced by the researcher in sign language recognition system is background of the signer and also the contrast of the light where signer present. We have worked on simple background. Our research is based up on selfie based sign language recognition with real time constraints such as non-uniform background, varied lighting and to make the system independent to the signer. 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.

P.V.V.Kishore in [3] projected a skeleton of isolated video based ISL identification system in which computational intelligence and image processing methodologies are integrated. Wavelet based video segmentation procedures are adopted for detection of hand shapes and head positions. Elliptical Fourier descriptors achieves unique feature vectors for each individual gesture and linear output membership function based Sugeno fuzzy inference system recognizes these gestures at a rate of 96% for a 80 word and 10 sentence trained system.

P.V.V.Kishore in [4] recommended an efficient theory of segmentation for SLR in which background in videos is assumed to be non static. Profuse attributes of images such as color, texture, and boundary are fused for the level sets energy function detraction .colour frames are extracted and utilized as per the background requirement of the video frames. Spatial information is provided by edge mapping and boundary features are obtained through edge map of image. Gesture shapes are enumerated dynamically and are assumed to be adaptive in the entire process of segmentation of video frames.

Tzuu-Hseng in [5] endorsed a system for the realization of gestures using ABC based Markov Model. Filtered data from AHRS sensors undergo principle component analysis for redundant data elimination and feature vector acquisition. ABC-HMM serves the purpose of classifier with the number of

hidden states obtained from k means algorithm. This work also provides a comparision of achieved recognition rate with various classifiers related to markov models and is found to present encouraging recognition rates compared to other classifiers.

Li Chun wang in [6] embarked a model which could measure the closeness of video in Chinese sign language. This model considers sign semantics which defines the hand shape, location , orientation and vision component which is distance based on VLBP. The experimental results of this model proved effective assessment of measuring similarity.

Sutarman in [7] proposed a Dynamic Malaysian sign language recognition system in which 3D Image data is captured from kinect sensor using skeletal data tracking. Feature extraction is done using the x, y, z coordinates relative to head and spine positions. Spherical coordinate conversion process is obtained using segmentation for dimension matching. Back propagation neural networks are used for classification using node variations in hidden layer. The system identifies 15 Malaysian sign languages with 80.54% accuracy .This system offers an advantage in image acquisition using depth images and by employs infrared light for independency of lighting conditions.

Lubo-Geng in [8] came up with Chinese sign language recognition system in which the hand shape features extracted from depth images and spherical coordinate features extracted from 3D hand motion trajectories form the feature vector. Extreme Learning Machine is employed as classifier.ELM has an accuracy of 82.79% and provides 6 % improvement in identification compared to SVM. This system incorporates spatial and temporal representation which depicts spatial connectivity for recognition and is immune for illumination changes and clustered backgrounds.

Setiawardhana in [9] developed an open CV android application that can elucidate sign language to speech. Viola Jones algorithm and KNN classifier are employed for the system design .Skin colour detection, Thresholding, Noise removal is carried on input images in which signs are based finger alphabets. The system limits its utilization for a distance range of 50 cm at an angle of 0 degrees between the hand and camera, Six to twenty FPS is required for single sign identification.

Neha baranwal in [10] advanced an Indian sign language identification system that helps to understand the inherit meaning of a communication system where gesture play a key role. Novel Gesture Identification mechanism is developed where the compressed data from DWPT extracts its features using PCA and ANN serves as classifier .DWPT offers efficient processing and is advantageous compared to Haar and Wavelet Transforms.

Pratibha Pandey in [11] put forth a review of various approaches for translation of sign language to the world communication language. Various Hand gesture identification methods based on Instrumented Gloves, vision, and color marker are focused in this work. DCT is employed for feature extraction. KOM, CAMSHIFT, TMM models are some of the mechanisms employed to accomplish the task of translation.

Neelesh Sarawate in [12] developed American Sign Language (ASL) word recognition system supported neural networks and a probabilistic model is conferred. A Cyber Glove and Flock of Birds area unit wont to track bending fingers, hand position, and hand orientation. A probabilistic model supported the Markov chains and HMM is employed to method the outputs of the feature vectors. The system has accuracy of 95.4% over a vocabulary of 40 ASL words.

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 [13],[14].

Pre-filtering for removing video capture noise during image acquisition is done using multiple sets of Gaussian filters of zero mean and 0.1 to 0.5 range variances. Hand and head contour shape extraction uses soble 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.

Word matching score (WMS) is the performance estimator when testing the proposed SLR system. It is the ratio of correct classifications to total number of signs. Multiple testing with 10 different signers has resulted in an average WMS of 91.23%. 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 rest of the paper is organized as follows. Section 2 gives research methodoly, mathematical modelling and derived algorithms for pre-filtering, segmentation, feature extraction and classification. Results and analysis is in section 3. Section 4 concludes the proposed novel idea of SLR capture with selfie front camera.

2. PRE-PROCESSING, SEGMENATION, FEATURE EXTRACTION AND CLASSIFICATION

The flow chart of the proposed SLR is shown in Figure 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 5M pixel CMOS front camera. Let this 2D video be represented as a 2D frame $\Im(x,y) \rightarrow \square^2 \ \forall \ (x,y) \rightarrow \square^+$. For video the frame $\Im(x,y)$ 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 $K_{2D}(x) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x-m)^2}{2\sigma^2}}$ [2] with zero mean (m=0) and three variances of $\sigma = 0.01, 0.1, 0.15$ smoothens each frame by removing sharp variations during capture.

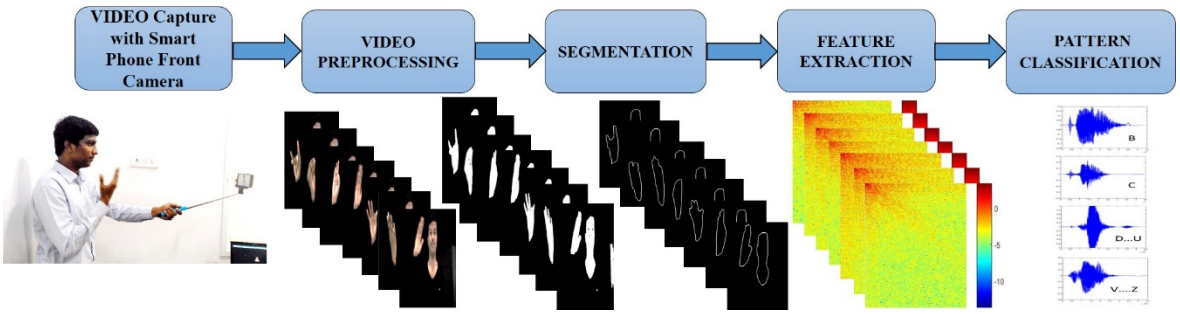

Figure 1. Flow chart of sign language recognition system with smart phone front camera video capture

The smoothed frames in real space $\square$ are treated with a new type of multidimensional sobel mask [15]. From literature the sobel edge operator is a 2D geadient 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.

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

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

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 $G^{xy} = \sqrt{(g^x)^2 + (g^y)^2}$. For convenience soble represented the function using a 2D convolution mask $S^{Mx} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$ and $S^{My} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}^T$. 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

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

Where 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. Figure 2 shows the difference in block thresholding and global thresholding (used 0.2) which failed to handle motion blur.

2(a) 2(b)

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

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 $\boldsymbol{H^C(x) \rightarrow C(B^x)}$ 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 $\boldsymbol{M_{3H}}$ and vertical $\boldsymbol{M_{3V}}$ directions of length 3. Contour extraction is represented as

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

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

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

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 $\boldsymbol{H^C(x,y)}$. Figure 3 provides visual conformation of the discussion.

3(a) 3(b) 3(c) 3(d) 3(e)

Figure 3. (a) Capture 98th frame. (b) Segmented Frame. (c) Hand Contour. (d) Head Contour and (e) both contours

Features are unique representation of objects in this world. Feature is a set of measured quantities in a 1D space represented as $\boldsymbol{F^V(x) = \{ f(x) \,|\, x \subseteq \square \}}$, where $\boldsymbol{f(x)}$ can be any transformation or optimization model on vector $\boldsymbol{x}$. Top priority in this work is sped of execution of the proposed algorithm. Hence $\boldsymbol{f(x)}$ is considered as Discrete Cosine Transform (DCT) [18] along with Principle Component Analysis (PCA) [16]. The 2D DCT of hand contour $\boldsymbol{H^C(x)}$ and head contour $\boldsymbol{\bar{H}^C(x)}$ is computed as

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

Where $\boldsymbol{C^u = C^v = \frac{1}{\sqrt{2}} \forall (u.v) = 0}$ and 1 elsewhere. Similar expression with $\boldsymbol{\bar{H}^C(x)}$ calculated 2D DCT of head contour as $\boldsymbol{\bar{F}^V_{uv}}$. Figure 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.

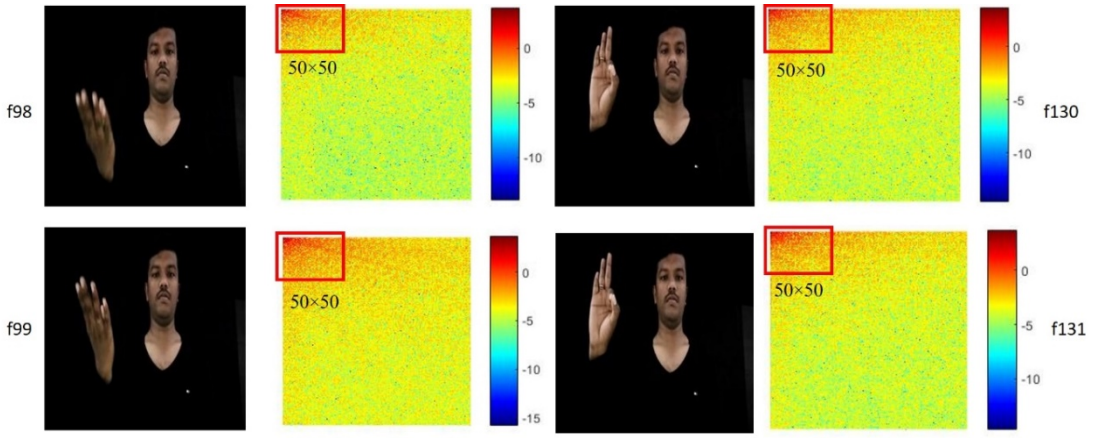

Figure 4. 2D DCT representation of hand contour energy representations

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 F_{uv}^{V}, which retains only the unique components of the matrix F_{uv}^{V}. The final F_{uv}^{V} is represented as F_{fn}^{V}, where 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 F_{fn}^{V} inputs a classifier. Since speed is the prime constraint during mobile implementation, it will be 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 [16] 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

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

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.

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 5M pixel front camera. Sign video capturing is constrained in a controlling environment with room lighting and simple background. The first photo in Figure 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 Figures 5 to 8.

Figure 5. Few sequence of frames for sign ‘HAI or HELLO’

Figure 6. Sign frames for ‘GOOD’ (Top) and ‘Morning’ (Bottom)

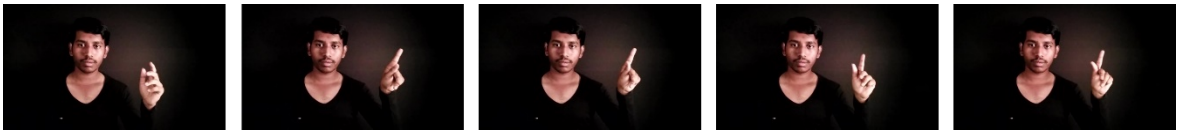

Figure 7. Frames of Sign ‘I’ ‘AM’

The advantage of using front camera is pronounced from Figure 7. Here the signer corrects himself during the signing process which helps to give correct hand sign to the system.

Figure 8. Single frames per sign. (From top left): ‘P’, ‘R’, ‘I’, ‘D’, ‘H’, ‘I AM’, ‘THANK’, ‘YOU’, ‘BYE’, ‘NO SIGN’

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 eq’n 4 to 6 refines the edges of hands and read portions. Figure 9 shows the results of the segmentation process on a few frames. Row (a) has original RGB captured video frames. Row (b) has Gaussian filtered, soble gradiented and region filled outputs of the frames in row (a). The last row contains morphological subtracted outputs of the frames in row (b).

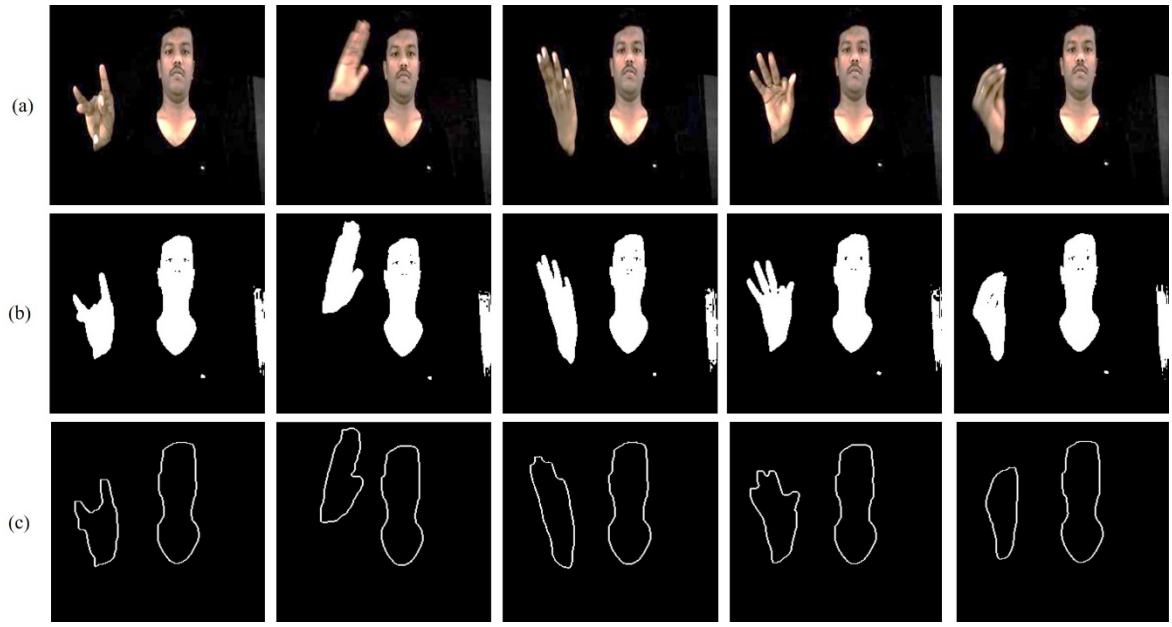

Figure 9. (a) Few frames in RGB format. (b) Their region segments with Gaussian filtering and soble operation. (c) Contours of hands and head produced with morphological subtraction with line structuring elements

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 samples that can effectively use to represent the entire hand and head curvatures. As shown in Figure 4, first 50×50 samples of the DCT matrix were extracted. These 2500 samples out of 65536 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 50×50×220 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 represents 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 $\boldsymbol{M}^{S\%} = \frac{\boldsymbol{Correct\ Classifications}}{\boldsymbol{Total\ Signs\ in\ a\ Video}} \times 100$. Feature matrix has a size of 50×220, 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 and plotted during testing in Figure 10.

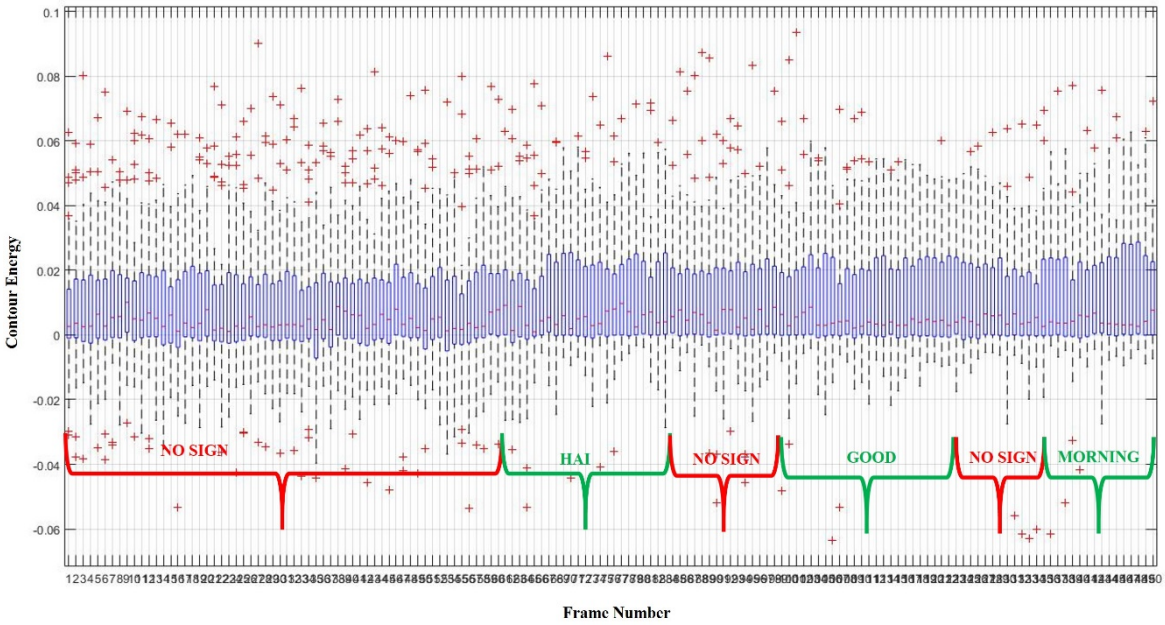

Figure 10. Sample Energy distribution in frames for identification of signs

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'n 8. Test repetition frequency is 10 per sign.

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

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.

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

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, a chart is drawn for 4 signer data in Figure 11. From the chat the following observations can me 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% to 6% if the number of frames in test train data does not match. For perfect matching the WMS is 3% above average.

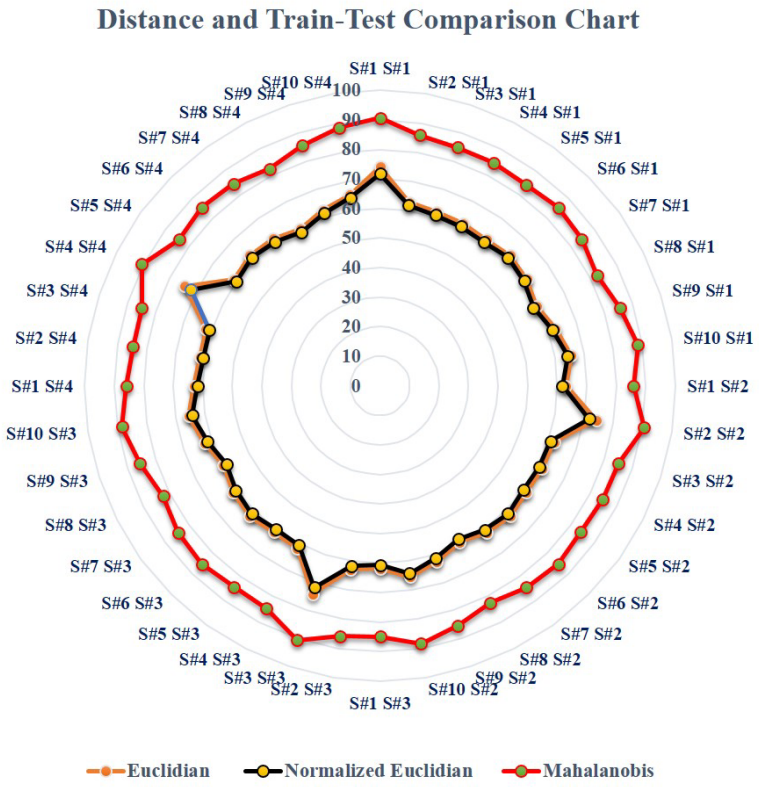

Figure 11. Distance and Train Test Data Comparison chart

Figure 11. Showing average WMS for test train data with respect to 3 distance metrics.

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 was recorded with 10 different signers. Pre-filtering, segmentation and contour detection are performed with Gaussian filtering, soblel 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. 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.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] K. Li, *et al.*, "Sign Transition Modeling and a Scalable Solution to Continuous Sign Language Recognition for Real-World Applications," *ACM Transactions on Accessible Computing (TACCESS)*, vol/issue: 8(2), pp. 7-23.
[2] W. W. Kong and S. Ranganath, "Towards subject independent continuous sign language recognition: A segment and merge approach," *Pattern Recog.* Vol/issue: 47(3), pp. 1294–1308, 2014.
[3] P. V. V. Kishore, *et al.*, "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*, vol/issue: 5(5), pp. 3742-3756, 2013.
[4] P. V. V. Kishore and M. V. D. Prasad, "Optical Flow Hand Tracking and Active Contour Hand Shape Features for Continuous Sign Language Recognition with Artificial Neural Networks," *International Journal of Software Engineering and Its Applications*, vol/issue: 9(12), pp. 231-250, 2015.
[5] T. Hseng, *et al.*, "Recognition System for Home-Service-Related Sign Language Using Entropy-Based K-Means Algorithm and ABC-Based HMM," in *IEEE transactions on systems, man, and Cybernetics: systems*, vol/issue: 46(1), pp. 150-162.
[6] L. C. Wang, *et al.*, "Similarity Assessment Model for Chinese Sign Language Videos," *IEEE Transactions On Multimedia*, vol/issue: 16(3), pp. 751-761, 2014.
[7] Sutarman, *et al.*, "Recognition of Malaysian Sign Language Using Skeleton Data with Neural Network," in *2015 International Conference on Science in Information Technology (ICSITech)*, pp. 231–236, 2015.
[8] L. Geng, *et al.*, "Chinese Sign Language Recognition with 3D Hand Motion Trajectories and Depth Images," in *11th World Congress on Intelligent Control and Automation Shenyang*, China, June 29 – July 4, pp. 1457–1461, 2014.
[9] Setiawardhana, *et al.*, "Sign Language Learning based on Android For Deaf and Speech Impaired People," in *International Electronics Symposium (IES), IEEE*, pp. 114–117, 2015.
[10] N. Baranwal, *et al.*, "Indian Sign Language Gesture Recognition Using Discrete Wavelet Packet Transform," in *2014 International Conference on Signal Propagation and Computer Technology (ICSPCT)*, pp. 573-577, 2014.
[11] P. Pandey and V. Jain, "Hand Gesture Recognition for Sign Language Recognition: A Review," in *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol/issue: 4(3), pp. 464-470, 2015.
[12] N. Sarawate, *et al.*, "A real-time American Sign Language word recognition system based on neural networks and a probabilistic model," in *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 23, pp. 2107-2123, 2015.
[13] T. Enikuomehin and J. Sadiku, "Text Wrapping Approach to natural Language Information retrieval using significant Indicator," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol/issue: 2(3), pp. 136-142, 2013.
[14] N. Pwint Oo and N. L. Thein, "Design and Evaluation of Android Slide Kyeboard for Myanmar Language," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol/issue: 1(2), pp. 119-125, 2012.
[15] W. Gao, *et al.*, "An improved Sobel edge detection," *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, Chengdu, pp. 67-71, 2010.
[16] J. Mei, *et al.*, "Learning a Mahalanobis Distance-Based Dynamic Time Warping Measure for Multivariate Time Series Classification," in *IEEE Transactions on Cybernetics*, vol/issue: 46(6), pp. 1363-1374, 2016.

BIOGRAPHIES OF AUTHORS

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 degree 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 Research Head at K.L.University, ECE Department. He is currently engaged in a project sponsered by Department of Science and Technology,SEED, Government of India, under the Technology for Disabled and Elderly scheme, related to sign language 3D invarient models for Indian sign language. 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 70 research papers in Various National and International journals and conferences including IEEE, Springer and Elsevier.