Journal of Engineering Science and Technology
Vol. 13, No. 8 (2018) 2352 - 2368
© School of Engineering, Taylor's University

SELFIE SIGN LANGUAGE RECOGNITION WITH MULTIPLE FEATURES ON ADABOOST MULTILABEL MULTICLASS CLASSIFIER

G. ANANTHA RAO, P. V. V. KISHORE*

Department of Electronics and Communications Engineering, K.L.E.F
(Deemed - to - be -University), Green Fields, Vaddeswaram, Guntur DT, 522502, India
*Corresponding Author: pvvkishore@kluniversity.in

Abstract

The objective is to take sign language recognition towards real time mobile implementation as a communication link between hearing impaired and normal people. Selfie mode continuous sign language video is the capture method used in this work, where a hearing-impaired person can operate the SLR mobile application independently. To decrease the computations per frame, the algorithms are developed for mobile platforms. The operating algorithm consists of key frame extraction, face detection, hand search space identification, head-hand portion extraction, fuzzy hand – head shape segmentation for multiple features. Hand – head configuration, phraseology, shape signature and orientation features are fused to form a dataset. Training with multiple features on the Adaboost Multilabel Multiclass learning algorithm makes the sign classifier faster. Results show excellent recognition rates and faster recall times when compared to state of the art backpropagation learning algorithm with single and multiple deep layers. The system is tested multiple times with 10 different signers with constant video backgrounds for 10 continuous Indian sign language sentences formed from 282 words.

Keywords: Adaboost classifier, Key frame extraction, Multi feature fusion, Selfie mobile sensor, Sign language recognition.

1. Introduction

Sign language recognition (SLR) is an evolving research area in computer vision. The challenges in SLR are video trimming, sign extraction, sign video background modelling, sign feature representation and sign classification. All the problems [1] are attempted in the past have met considerable amount of success and are instrumental in development of the state of the algorithms for SLR. Gesture recognition uses powerful imaging and artificial intelligence based algorithms for classification [2]. Current trends show an urge to bring gesture recognition into mobile environments [3].

Sign language is a visual mode of communication between two hearing impaired or hard hearing people. The communication foundations are based on finger shapes, hand shapes, hand movements in space with respect to body, hand orientations and facial expressions. The humans are trained exclusively to handle such huge amounts of information for years. For machine translation, the problem transforms into a 2D natural language processing problem. Many 1D/2D/3D models are proposed in literature with little success to bring the model close to real time implementation [4-8].

In this work, the focus is to recognize signs of Indian sign language using 2D selfie video captured using a mobile front camera. Even though the development of a mobile app is far from reality, the objective is to simulate algorithms that can optimally execute on a mobile platform. The primary module is to extract information frames to reduce input video data per frame. A visual attention based framework proposed in this paper is considered for accuracy and computation time. The model works well for constant video backgrounds and we will limit our work to this typical video sets.

Processing entire frame area for signer segmentation increases execution time of the segmentation algorithm. For this, we remove a non-contributing section on the frame by combining face detection and face vicinity searches to decide on which portions of the frame to be discarded. However, the face recognition algorithm used is currently operational on mobile devices. Viola Jones face detection algorithm used for identifying the signer face.

Connected component extraction and sparse representation on the truncated video frame results in hand and face extraction. Fuzzy segmentation extract shapes of hands and face in frames. The next step in the process is to convert the segments into meaningful features that are defines the attributes of sign language. Here we use, Hand – head configuration, phraseology, shape signature and orientation features to characterize a sign. A dataset is created from these attributes and Adaboost algorithm is trained to label each sign separately. Adaboost is fast and takes less time for execution.

Unavailability of benchmark datasets for Indian sign language (ISL) in selfie mode, prompted us to create our own dataset. The dataset is having 10 ISL commonly used conversational sentences consisting of 282 words performed by 3 native ISL users and 7 non ISL users. Training is initiated with 3 native ISL users and tested with 7 non-native ISL users. The performance of the algorithms is measured based on their ability to recall with precision the input query signs in a sentence.

2. Related Work

Sign language recognition (SLR) has transformed with technology upgradation from 1D, 2D to 3D models in the last 2 decades. In 1D, SLR is based on 1D signals acquired from a hand gloves [8] and classified using signal processing methods [9]. In the recent times researchers started using leap motion sensor [10] to extract 1D signals of finger movements and estimate the related gestures of sign language using Hidden Markov Models.

The faster 1D models produce good recognition rates when the emphasis is only on hands. But sign language involves head, torso and face expressions along with hand movements and shapes [11]. 2D video data of signs produces relatively more information compared to 1D data gloves. From 2D capture, one can explore all the elements of a visual language with a constraint on speed and classification accuracy. Again, for 2D SLR HMM is most widely researched classifier with continuous and discrete versions of sign language [12]. More research related material on 2D models and the corresponding research challenges can be found in [13-15]. The other challenge for researchers lies in converting the detected signs into meaningful sentences [11]. The challenging problems in 2D SLR are hand tracking, occlusions on hands and face, background lighting, changing signer backgrounds and camera sensor dynamics.

To reduce the computation time, sign video skimming is applied on the video dataset to reduce the number of frames for processing. This results in KEY frames or action frames containing the hand movements. Visual attention based framework proposed in [16] is used over similar algorithms in the category such as presented in [17, 18]. The algorithm in [16] uses both spatial and temporal attention values from each frame for extraction.

The hand and head segmentation algorithms are directly applied on the Key frames of a set resolution. To increase algorithm speed, the resolution is trimmed down to minimum and the loss in quality is compensated with powerful and slow computer vision models [19, 20]. To increase and improve segmentation speed, we propose to search for hand near the face or head region. Selfie SLR is a one hand sign language model reducing the segmentation complexity. The hand search space is identified with respect to face or head.

The frame is divided into two halves horizontally keeping the face as the centre. A simple frame subtraction [21] on two halves independently between consecutive frames is performed. The movement of hand in any one half produces a positive pixel count enabling to discard the other half. Due to constant background, color and morphological segmentation models work at a faster rate in segmenting hand in the truncated video frame.

The next step involves feature extraction [22] module. Image feature extraction is widely studied research area after pattern classification. There are generalized feature representation as well as focused representations of objects in an image. However, the feature size generated from these methods are large and are vulnerable to non-rigid video objects. Focused features like shapes, texture, color are used [23] for sign language recognition. These states of the art algorithms work accurately but fail to deliver computational flexibility.

The objective is to select features that represent a sign and is easily distinguishable in closely related sign words and are computationally efficient. The attributes for a selfie sign language recognizer chosen are shape signature [24] for hand and head shapes, hu moments [25] for hand orientations, hand – head distance and hand position vectors for tracking. The chosen attributes perfectly characterize a sign in Indian sign language.

Classifying at faster rate on a huge dataset is a complicated problem. Adaboost [26] classifier is faster and efficient algorithm for large datasets [27]. Inspired from [23] the feature matrix is labelled and inputted to Adaboost classifier for training and testing.

The rest of the paper is organized as follows: section 3 introduces methodology with extensive discussions on math models and their underlying theory. Experiments and outcomes are part of section 4. Section 5 concludes the work.

3. Methodology

Sign language recognition is a mixture of complex hand movements, hand movements with respect to head and torso, finger shapes, facial expressions and body positions. Only vision based sign language recognition systems can capture all the attributes for good classification. Selfie sign language recognizer is introduced for making the system into a real-time application. However, the proposed system introduces an independent signer operation of the application and induces self-checking on the performed signs. In this work, the focus is on recording selfie videos through mobile front camera with simple frame backgrounds and making the algorithms faster for mobile application. MATLAB mobile app is used for testing the proposed system. Figure 1 describes the flow chart of the proposed system.

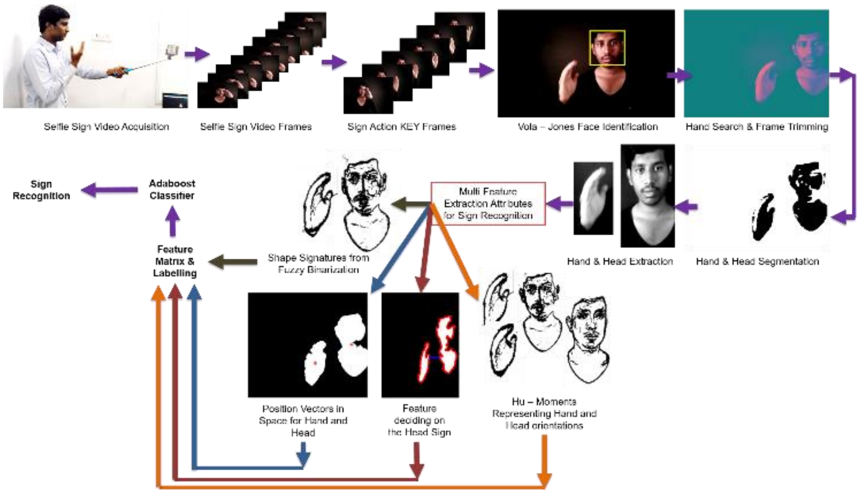

Selfie Sign Video Acquisition
Selfie Sign Video Frames
Sign Action KEY Frames
Vola – Jones Face Identification
Hand Search & Frame Trimming
Sign Recognition
Adaboost Classifier
Multi Feature Extraction Attributes for Sign Recognition
Hand & Head Extraction
Hand & Head Segmentation
Feature Matrix & Labelling
Shape Signatures from Fuzzy Binarization
Position Vectors in Space for Hand and Head
Feature deciding on the Head Sign
Hu – Moments Representing Hand and Head orientations

Fig. 1. Flow chart of the proposed selfie sign language recognizer.

The 4 features being extracted are the attributes a human sign interpreter will use to quickly recall the associated sign label. The sign interpreter uses sparse information from the signer's movements and reconstruct the remaining information in the brains memory to perfectly recall the word. A simpler model is proposed in this work making the attributes representing the sign less complicated. Each frame is represented with only a set of 13 unique values that represent hand shape, head involvement in sign, hand and head tracking, hand and head orientations. Key frame extraction greatly reduces the memory usage as the algorithm runs during query video acquisition stage. Hand and head are viewed as single entity for only first few frames and related head calculations are discarded with reinitialization module running every 100 frames to check the location of the head. If there is change in head location the algorithm initializes the new head position. A theoretical view of the modules is presented in this section.

3.1. Key frame extraction

This procedure is important to discard the action free frames from processing. The method used in this work is based on image signatures [20]. In human visual attention framework, the foreground is more viewed or focused compared to background, is the mainstay of the algorithm. The algorithm is based on discrete cosine transform (DCT), which retains only the transformation of sign.

3.1.1. Video frame spatial saliency maps

For a selfie sign video sequence represented as $V_S\,(x,y,t) \in R^{3\times 3}$ is a color real values in 3 planes, the video frame signature for a color plane p of size $m \times n$ is

$$VFS\left(V_s^p\right) = sign\left(DCT\left(V_S^p\right)\right) \tag{1}$$

where, VFS denotes the video frame signature with $sign(\square)$ is sign determining operator in each of the color channels denoted by p. The inverse DCT of the above sign matrix results in a spatial domain frame denoted as VFS' defined by

$$VFS'\left(V_s^p\right) = IDCT\left(VFS\left(V_s^p(x,y,t)\right)\right) \tag{2}$$

The spatial saliency map $S^s\left(V_S^p\right)$ of video frame V_S^p is calculated as Gaussian distribution on each pixel in all color channels as

$$S^s\left(V_S^p\right) = \frac{1}{2\pi\sigma^2} e \frac{-\left(x^2+y^2\right)}{\sigma^2} \otimes \sum_{p=1}^{3} VFS \circ VFS' \tag{3}$$

where σ is the standard deviation of Gaussian function and (x,y) denote pixel locations. The $'\otimes'$ is the linear convolution operator and $'\circ'$ is the Hadamard product operator. YCbCr color space is preferred because of its closeness to human skin detection. Normalized spatial saliency map is produced in the range $[0,1]$ by dividing $S^s\left(V_S^p\right)$ in each color plane by $\max\left(S^s\left(V_S^p\right)\right)$. Spatial attention value $A^S\,(t)$ is calculated as

$$A^S(t) = mean\left(\frac{S^s\left(V_S^p\right)}{\max\left(S^s\left(V_S^p\right)\right)}\right) \tag{4}$$

$A^S(t) \cong 1$, gives good action frame and $A^S(t) \cong 0$ is a non-action frame. This process takes 0.02 sec on a MATLAB mobile app connected to a desktop server with MATLAB full version. This is important for Selfie mode sign language recognition model. Spatial saliency is insufficient in case of blurring, frame movements due to camera movement and brightness values of the captured frames. These insufficiencies can be handled by computing temporal saliency map and combining with spatial saliency to decide on the Key frames.

3.1.2. Video frame temporal saliency maps

Optical flow [20] is the state of the art algorithm for determining the temporal changes in consecutive frames. However, optical flow is an iterative algorithm and computation time is large for real time mobile applications. Temporal gradients are the simplest and effective model for calculating the changes in pixels in consecutive frames to approximate the motion saliency and temporal attention value. Temporal attention value $A^T(t)$ is calculated as a mean of normalized temporal saliency

$$A^T(t) = mean\left(T^G\left(V_S^p\right)\right) \tag{5}$$

where $T^G(t)$ is the temporal gradient between consecutive frames in the selfie video. For all frames, $A^T(t) \subset [0,1]$. A high on A^T indicates action frame and low value is a no change frame. This part of the program executes in 0.002sec. Using spatial attention A^S and temporal attention A^T per frame, Key frames are extracted.

3.1.3. Attention value fusion and key sign frames

Spatial and temporal attention curves are formed with A^S and A^T values calculated from each frame. A nonlinear fusion model in [28] is being used to merge the two with weighted values. The weighted fusion function [16] is defined based on attention vector $A = \left[A^S, A^T\right]$ and their corresponding weight vectors $w = [w_s, w_T]$ as $F^{Aw}\left(A^S, A^T\right)$. The key frame index is

$$i_t = \left(F^{Aw}\left(A^S, A^T\right) > Avg\left(F^{Aw}\left(A^S, A^T\right)\right) == 1\right) \tag{6}$$

The entire process is computed in 0.042 sec per frame in selfie sign video. The dataset of selfie sign video consists of 282 words and the continuous recording on mobile with 5MP front camera has resulted in a 4.3 min video sequence. At an average frame rate of 30fps, we have 8109 frames in the first recoding and with the discussed algorithm the resulting sign video has 3666 frames. The redundant and less action frames were reduced by 45.2%. The computation time for the entire program to extract key frame video is 3.4 min, which is less than the entire recording in the video sequence. In real time, the key frame video is extracted for every sec and the resulting Key frame selfie signs are processed further.

3.2. Face tracking and frame pruning

Viola – Jones algorithm in [21] detects and tracks faces in live videos faster and is currently in use on all android mobiles. However, we are interested in cutting the frames spatial regions to reduce computational costs. In a selfie mode video capture, the stick in the signer's hand is aligned to either left or right side of the signer. This is due to large frame space requirements to the signing hand for free movement and discarding the other half. This enables to restrict the sign search space to 60% of the pixels.

Bounding box on the face gives the location of the coordinates $\left(x_{fb}, y_{fb}\right)$in 4 directions. We consider only horizontal coordinates for pruning the video frame. The decision on pruning the left or right side of the frame is based on distance of the bounding box coordinates to the edges of the frame. The hand search space h_{sp} in the video frame is computed as

$$\begin{aligned} & if\ d^{l}\left(x_{fb}, y_{fb}\right) > d^{r}\left(x_{fb}, y_{fb}\right) \text{ then } \mathrm{h}_{sp} = rhp \\ & else\ \mathrm{h}_{sp} = lhp \end{aligned} \tag{7}$$

where $d^{l}\left(x_{fb}, y_{fb}\right)$computes the distance from left bounding box coordinates to starting frame giving left half plane (lhp) and $d^{r}\left(x_{fb}, y_{fb}\right)$is the distance on the right half plane (rhp). This operation took 0.002 sec for extraction.

3.3. Hand-head segmentation

Adaptive histogram based Skin color [29] segmentation is used to separate human in the video with the background. Human skin is better modelled in YCbCr color space compared to RGB and hence RGB colored selfie sign frames are transformed to YCbCr color space. Now, apply the adaptive histogram equalization and set a clip limit to each histogram in a color plane. Then applying the skin color model thresholds in [29], the human portions in the image frame are separated.

3.4. Selfie sign frames

At this stage, the processed frames does not indicate any thing about the sign and is impossible to estimate the sign. From a human sign interpreter view point the fastest model is based on 4 attributes: (1) Hand positions, (2) Head or body involvement in sign, (3) Hand shapes and (4) Hand orientations. These attributes are the necessary features for computer recognition of signs.

3.4.1. Position features – hand and head tracking

Determining hand position in 2D video space and its tracks in corresponding frames help in determining a sign. For example, sign 'morning' and 'evening' have same hand shape whereas the hand moves in opposite direction. To make faster tracking a simple centroid based tracking gradient is proposed in this work. The idea is to identify hand centroids and head centroids between consecutive frames and calculate gradient magnitude in and directions. Here the SLR system must be informed about the direction of hand with outcomes such as up, down, left, right, left diagonal, or right diagonal etc. The feature values are numbers showing the direction of centroid movement along with their magnitude of change. Applying gradients on centroid positions in consecutive frames on hand results in a magnitude vector and direction vector.

3.4.2. Head involvement in sign

We propose to use hand – head boundary distances to measure the involvement of head or face in the sign recognition. Euclidian distance transform is applied to hand and head boundary points to find the minimum distance as feature. The boundary extraction is based on oriented energy approach that can detect and localize boundaries. The distance between hand boundary points and head boundary with a distance threshold gives the involvement of head in the sign. The distance transform is normalized with respect to maximum distance and the range is [0,1]. A value ‘0’ means, head is part of the sign and a value ‘1’ is absence of hand in the frame. This feature also determines the start and end of sign in this framework.

3.4.3. Hand shape features

Here we apply fuzzy -2 partition segmentation model with Gaussian membership functions. The membership functions are designed to partition skin and non – skin pixels in the hand and face image into 2 fuzzy sets. Two sets of rules were designed for segmentation of hand and head regions. Figure 2 shows the results of frames in Selfie SLR.

Every pixel on the shape is kept into a feature vector [30] or a descriptor [31] to represent shape in each frame is used exclusively for gesture classification. Integrating shape feature values over shape 2D spatial domain results in a shape signature value normalized by area of the shape to achieve scale invariance. The range of shape signature is [0,1] based on value of scale.

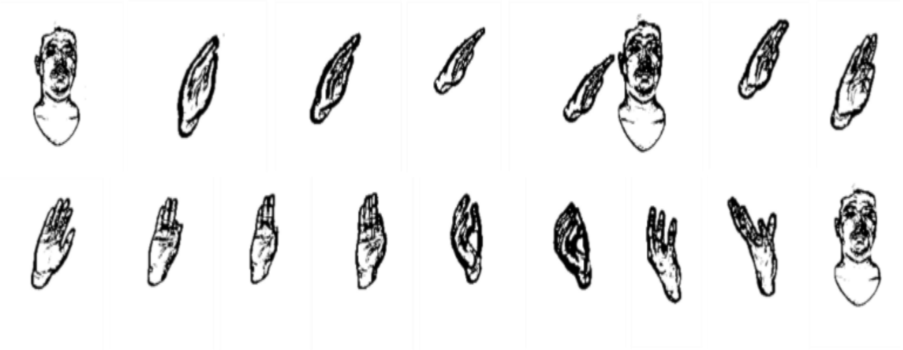


Fig. 2. Shapes of hand fingers from a sign video saying, “Hello, Good Morning”.

3.4.4. Hand orientation features

Hand orientations provide rotation invariant feature of a sign in a signing space. Moments project hand segments on to basis which results in a piecewise continuous linear function in the spatial plane. Sign feature representation with moments is exclusively developed by researchers in [25, 32]. Hu moments in [32] are non – orthogonal centralized moments that are scale, translation and rotation invariant. Hands come in all shapes and sizes and the features describing them may change with signer. Modelling invariance in hand shape features from the hand segments of Fig. 2 is done with Hu moments.

3.4.5. Feature matrix construction

From the 4 features, a complete feature matrix is constructed per frame. From tracking features, we have 4 features measuring gradient magnitude and direction of both hand and head regions. From hand -head distance we have one feature and shape signature gives one feature. Finally, moments will give 7 features towards feature matrix construction. The selfie sign feature matrix consists of 13 feature vectors $f_v \subset \mathbf{R}^{t\times 13}$ per frame. The number of frames per sign is divided with respect to hand – head distance. Each set of frames after division into signs are labelled with the word describing the sign. Recognition of a sign displays a predicted label on the mobile screen or can be converted to voice using a simple text to speech programmable interface application.

3.4.6. Sign classifier: Adaboost multi-label multi-class

Boosting based classifications [23, 26, 27] finds very precise hypothesis from a set of weak hypotheses. Here hypothesis is a classification rule. The set of weak hypotheses are simple rules that generate a predictable classification. Let $\mathrm{T} = \left[\left(f_1, L_1 \right), \left(f_2, L_2 \right), \left(f_3, L_3 \right), \ldots\ldots, \left(f_v, L_v \right) \right]$ be a set of training examples at an instance f_i on i^{th} frame in feature space f with labels L_i on label space L. The algorithm accepts the training samples T along with some class distribution $D = \{1, \ldots., m\} \in \mathbf{R}$ represented as weak learners. On the input, the weak learner computes a weak hypothesis H. Generally, $H : f \to \mathbf{R}$. The interpretation for classification is based on $sign\{H(f)\} = \{+1, -1\} \to \{f_i\}$ for a binary classifier. The $|H(f_i)|$ gives prediction confidence. From this weak hypothesis, through training a strong hypothesis is generated to recognize sign labels $Z_i = H(f_i)$.

4. Experiments and Results

The work in this paper is designed to solve two objectives: (1) To accurately recall selfie sign videos captured on mobile phones and (2) reducing the computation time to operate on mobile phones. For database creation, we used a plain background in the video sequences. The database is a set of most commonly used 282 words in general conversation from Indian sign language. A mobile with 5MP front camera is used. 5 sets of data are captured with black plain background with the signer wearing the black shirt and the remaining 5 with soft backgrounds with no constraint on the signer dress. A 10-set combination of 282 words is tried. We present a set words as an example: “Hello, good morning, I am good. Hope you are doing well, drink tea, eat biscuit, woman is beautiful, men are handsome, Hai Good Morning, I am P R I D H U, Have A Nice Day, Bye Thank You, Morning are happy, afternoons are good and evenings are sad. I am the head of product design group at K.L. University” and so on. The underlined word is shown as an example for the proposed model.

The experimentation is divided into 4 parts: (1) Training and testing with same dataset with constant black backgrounds and (2) Testing with a different signer on black video backgrounds (3) Training and testing with videos with different backgrounds of same signer. (4) Training and testing with two different datasets. Although training and testing is performed on 282 word sentences, only 20-word example is presented to maintain clarity. Figure 3 shows the model of datasets created for testing the proposed method.

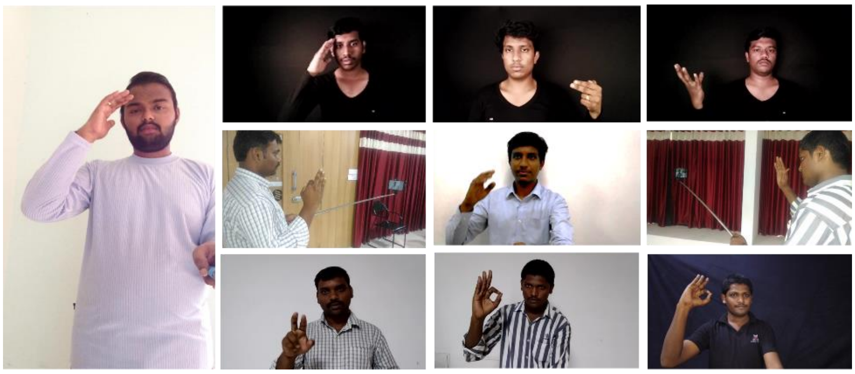


Fig. 3. Database of selfie Indian sign language with simple backgrounds showing frames.

We use three performance evaluators for validating the results. They are precision – recall curves, percentage recognition rate and the computation time per sign. For a strong hypothesis H resulted from Adaboost training and testing for an input feature f_i on a trained distribution D with $Z_i = H\left(f_i\right)$ predicted labels. The following are the metrics

$$Precision\left(H,D\right)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|L_i \cap Z_i|}{|Z_i|} \quad (8)$$

$$Recall\left(H,D\right)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|L_i \cap Z_i|}{|L_i|} \quad (9)$$

$$\%\ \ Recognition=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|L_i \cap Z_i|}{|L_i \cup Z_i|} \quad (10)$$

4.1. Exp-1. Identical signer dataset for training and testing

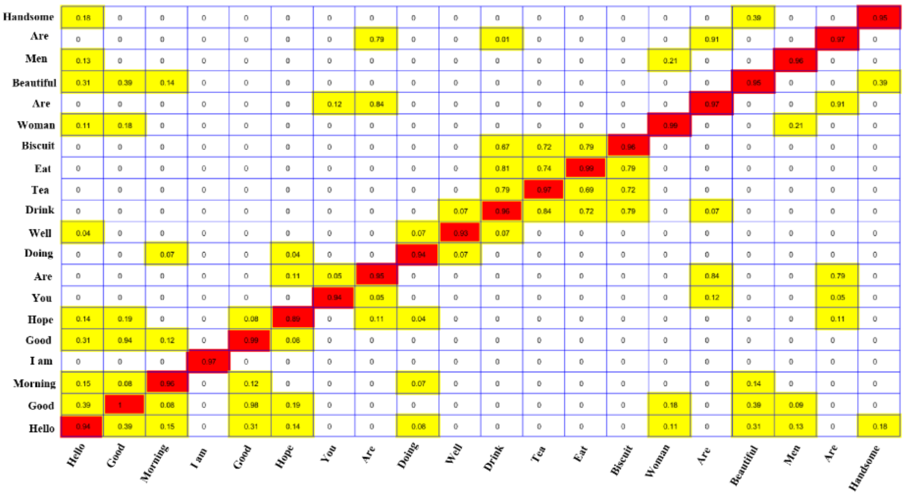
In exp-1, training and testing is initiated on a 282-words on a selfie video captured with black background and a dark shirt on the signer. We designed 10-282 word sentences from the same signer twice in a row. The 10 sentences which contain the same 282 word combinations forming different meaning. At this point we are interested in faster processing with reasonable precision. The average length of video sequence is 8109 frames per 282-word sentence from 5 different signers. With Key frame extraction, the average frames are reduced to 3666 frames per 282 words. We have 3 fast moving native signers and 2 non-native slow signers and hence the key frame algorithm used has reduced the processing time by 45%. The computation time for the entire video sequence on Key frame extractor is around 3.4min or 0.0622sec/frame. For real time operation, the Key frame extraction algorithm is made to run after 30sec video captures in the background to separate action frames from redundant frames.

For a video sequence in exp-1 with 3666 frames and 282 words, we have a feature training set of size and sign labels of size strings with 282 words. During training, each feature in a frame is concatenated with the corresponding label as an example set. Training is initiated with Adaboost classifier with an initial LUT weak hypothesis and Gaussian distribution with parameters. Training is given for feature vector of size and the average time is 0.0401 sec per frame for 100 iterations on the training set.

Testing on the same dataset as used in training has resulted in a 100% recognition rate calculated from Eq. (10). The recall time during testing on the same dataset is 0.0044sec per frame and for the entire video it has been measured as 16.1304 seconds.

Exp-1 uses one dataset for training and the remaining 9 datasets for testing. Training and testing with same dataset has always resulted in a 100% recognition rate with zero false predictions. With the remaining 9 sentences, the testing resulted in a variation in recognition rate between 98% to 91%. The variation is due to the number of frames the sign occupied at different locations in the test sentence. For example, training sentence is “hello, good morning, I am good” and if the testing sequence is “good morning, I eat biscuit”, then good and morning are displaced resulting in an overall small misclassification between frames at 4%.

Figure 4 shows the average confusion matrix generated from calculations using Eq. (10) . The recognition rate is averaged across 10 sentences with 20 signs on the same signer videos for training and testing.



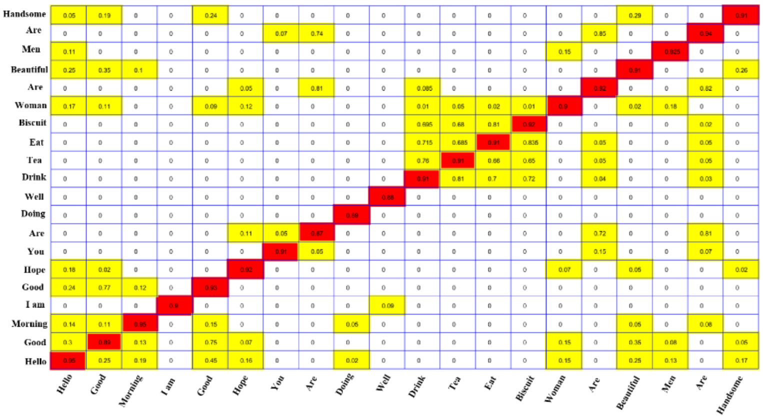
	Hello	Good	Morning	I am	Good	Hope	You	Are	Doing	Well	Drink	Tea	Eat	Biscuit	Woman	Are	Beautiful	Men	Are	Handsome
Handsome	0.18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.39	0	0	0.95
Are	0	0	0	0	0	0	0	0.79	0	0	0.01	0	0	0	0	0.91	0	0	0.97	0
Men	0.13	0	0	0	0	0	0	0	0	0	0	0	0	0	0.21	0	0	0.96	0	0
Beautiful	0.31	0.39	0.14	0	0	0	0	0	0	0	0	0	0	0	0	0	0.95	0	0	0.39
Are	0	0	0	0	0	0	0.12	0.84	0	0	0	0	0	0	0	0.97	0	0	0.91	0
Woman	0.11	0.18	0	0	0	0	0	0	0	0	0	0	0	0	0.99	0	0	0.21	0	0
Biscuit	0	0	0	0	0	0	0	0	0	0	0.67	0.72	0.79	0.98	0	0	0	0	0	0
Eat	0	0	0	0	0	0	0	0	0	0	0.81	0.74	0.99	0.79	0	0	0	0	0	0
Tea	0	0	0	0	0	0	0	0	0	0	0.79	0.97	0.69	0.72	0	0	0	0	0	0
Drink	0	0	0	0	0	0	0	0	0	0.07	0.96	0.84	0.72	0.79	0	0.07	0	0	0	0
Well	0.04	0	0	0	0	0	0	0	0.07	0.93	0.07	0	0	0	0	0	0	0	0	0
Doing	0	0	0.07	0	0	0.04	0	0	0.94	0.07	0	0	0	0	0	0	0	0	0	0
Are	0	0	0	0	0	0.11	0.05	0.95	0	0	0	0	0	0	0	0.84	0	0	0.79	0
You	0	0	0	0	0	0	0.94	0.05	0	0	0	0	0	0	0	0.12	0	0	0.05	0
Hope	0.14	0.19	0	0	0.08	0.89	0	0.11	0.04	0	0	0	0	0	0	0	0	0	0.11	0
Good	0.31	0.94	0.12	0	0.99	0.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I am	0	0	0	0.97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Morning	0.15	0.08	0.96	0	0.12	0	0	0	0.07	0	0	0	0	0	0	0	0.14	0	0	0
Good	0.39	1	0.08	0	0.98	0.19	0	0	0	0	0	0	0	0	0.18	0	0.39	0.09	0	0
Hello	0.94	0.39	0.15	0	0.31	0.14	0	0	0.06	0	0	0	0	0	0.11	0	0.31	0.13	0	0.18

Fig. 4. Confusion matrix for Exp-1 averaged over 10 different sentences formed from 20 words.

4.2. Exp-2. Dissimilar signer dataset for training and testing

In this experiment, we apply same 282 – word sentences for training and testing but with the test sequence is captured by a different signer. Here no constraints are placed on hand speed during signing. The native signer-2 dataset has captured an average of 8211 frames for the 282 – word 10 sentences. After key frame extraction, the number of frames were 3695 as against 3666 for signer -1 in the exp-1.

The recognition time varied slightly per frame compared to exp-1. Figure 5 gives the average confusion matrix from the 2 native and 2 non-native signers. The non-native signers are slow in hand movements and took on an average 258 frames more than native signers.



	Hello	Good	Morning	I am	Good	Hope	You	Are	Doing	Well	Drink	Tea	Eat	Biscuit	Woman	Are	Beautiful	Men	Are	Handsome
Handsome	0.05	0.19	0	0	0.24	0	0	0	0	0	0	0	0	0	0	0	0.29	0	0	0.91
Are	0	0	0	0	0	0	0.07	0.74	0	0	0	0	0	0	0	0.85	0	0	0.94	0
Men	0.11	0	0	0	0	0	0	0	0	0	0	0	0	0	0.15	0	0	0.925	0	0
Beautiful	0.25	0.35	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.91	0	0	0.26
Are	0	0	0	0	0	0.05	0	0.81	0	0	0.085	0	0	0	0	0.92	0	0	0.82	0
Woman	0.17	0.11	0	0	0.09	0.12	0	0	0	0	0.01	0.05	0.02	0.01	0.9	0	0.02	0.18	0	0
Biscuit	0	0	0	0	0	0	0	0	0	0	0.695	0.68	0.81	0.92	0	0	0	0	0.02	0
Eat	0	0	0	0	0	0	0	0	0	0	0.715	0.685	0.91	0.835	0	0.05	0	0	0.06	0
Tea	0	0	0	0	0	0	0	0	0	0	0.76	0.91	0.66	0.65	0	0.05	0	0	0.06	0
Drink	0	0	0	0	0	0	0	0	0	0	0.91	0.81	0.7	0.72	0	0.04	0	0	0.03	0
Well	0	0	0	0	0	0	0	0	0	0.88	0	0	0	0	0	0	0	0	0	0
Doing	0	0	0	0	0	0	0	0	0.89	0	0	0	0	0	0	0	0	0	0	0
Are	0	0	0	0	0	0.11	0.05	0.87	0	0	0	0	0	0	0	0.72	0	0	0.81	0
You	0	0	0	0	0	0	0.91	0.05	0	0	0	0	0	0	0	0.15	0	0	0.07	0
Hope	0.18	0.02	0	0	0	0.92	0	0	0	0	0	0	0	0	0.07	0	0.06	0	0	0.02
Good	0.24	0.77	0.12	0	0.93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I am	0	0	0	0.9	0	0	0	0	0	0.09	0	0	0	0	0	0	0	0	0	0
Morning	0.14	0.11	0.95	0	0.15	0	0	0	0.06	0	0	0	0	0	0	0	0.05	0	0.08	0
Good	0.3	0.89	0.13	0	0.75	0.07	0	0	0	0	0	0	0	0	0.15	0	0.35	0.08	0	0.05
Hello	0.95	0.25	0.19	0	0.45	0.16	0	0	0.02	0	0	0	0	0	0.15	0	0.25	0.13	0	0.17

Fig. 5. Confusion matrix for Exp-2 averaged over 10 different sentences formed from 20 words and 4 different signers.

4.3. Exp-3 and 4. identical signer dataset for training and testing without background constraints

In exp-3, we capture the 282 – word sentence with no constraints on background and on the dress of the signer in selfie mode. The basic image processing process for head and hand segmentation was long and color thresholds are modified for distinguishing human skin under a different background. We focused on plain white backgrounds such as walls for this capture. Again 3 native and 2 non-native signers were used. The average recognition rate in Exp – 4 is 85%. The unconstraint's put too much pressure on the computation resources. The computations times are high and compared to previous 3 experiments and recognition rate decreased by 11% from the 1st experiment.

4.4. Comparison with other publications [33-35]

Three classifiers that are simple and accurate are used against Adaboost in the proposed SLR model with selfie video capture. They are minimum distance classifier with Mahalanobis distance (MD) which is simple and no training is required. Artificial neural network trained with backpropagation neural networks (BNN) algorithm and Deep backpropagation neural nets (DBNN) are used as artificial intelligence based algorithms for comparison. In the first comparison on accuracy with the help of recognition rates for all 10 sentences is calculated. Figure 6 shows the average recognitions obtained on 10 sentences for 10 different signers in all 4 experiments per word recognized. For 282 signs, we can see that Adaboost is accurate in recognition and is on par with the training model classifiers. Average recognition for Adaboost multi class classifier in this work is 89.82% on overall datasets. Deep ANN with 5 hidden layers with the same features averaged at 91.22% and whereas ANN with 1 hidden layer produced 88.25% recognition. MD was the least with 78.23%.

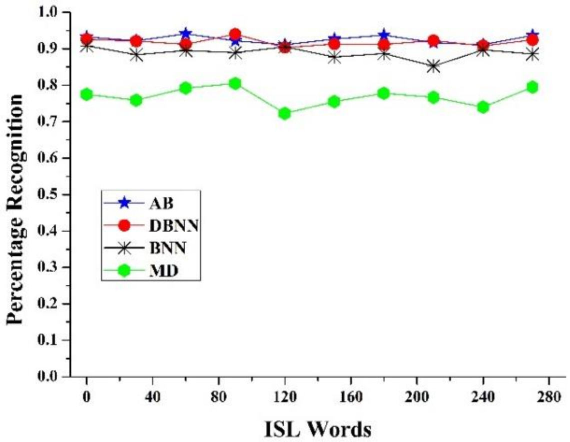


Fig. 6. Recognition rate comparison between AB, DBNN, BNN and MD classifiers

A computation time comparison plot is shown in Fig. 7. For commonality in samples for time plot, we calculated the total time by adding all the times for the entire 282-word sentence during execution and divided the total time per frame. AB is simple algorithm that takes less computation time per frame, which is around 50.21sec. MD takes 49.44sec per frame but the accuracy is very poor as no training is involved. Final performance estimation is through precision – recall curves. Figure 8 shows the plots computed and averaged over all 4 experiments per signer.

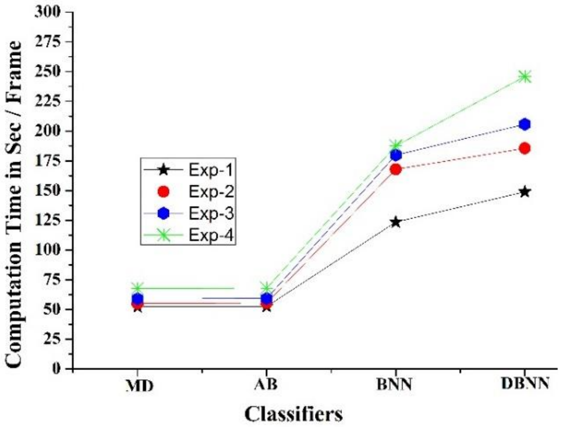


Fig. 7. Computation times per frame between classifiers for selfie sign language recognizer.

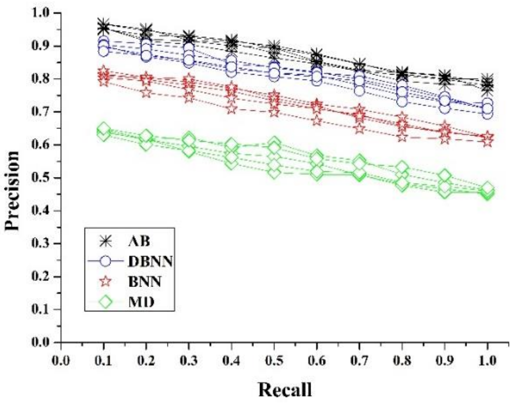


Fig. 8. Precision – recall plot for selfie sign language classifiers.

5. Conclusions

We proposed a framework for sign language recognition with selfie mode video capture. Our approach has advantages in the form of self-checking during signing and implementation on mobile platforms. At the core are the algorithms that can reduce the computation times at high accuracy. We present, key frame extraction, frame pruning, hand – head segmentation, human like feature extraction and recognition. Human like features are those used by human interpreters to recall signs accurately. Adaboost multi label multi class algorithm is used for pattern recognition of signs. 4 experiments were designed based on the dataset with simple and unconstraint video backgrounds and signer. A comparison is initiated based on classifiers performance at recognition rates, computation times and precision – recall plots for Adaboost, Minimum Distance, Backpropagation neural networks and Deep backpropagation neural networks. We found AB is faster and produces accuracies that are around 90% on the 282-word length sentences of Indian sign language. Further the work can be extended for a real time larger selfie sign data by initiating a convolutional neural network based training to bring the sign language translation into more reality.

Nomenclatures

$A^S(t)$	Spatial attention value
$A^T(t)$	Temporal attention value

Abbreviations

SLR	Sign Language Recognition
ISL	Indian Sign language
HMM	Hidden Markov Model
DCT	Discrete Cosine Transform
AB	Adaboost
MD	Mahalanobis Distance
BNN	Backpropagation neural network
DBNN	Deep backpropagation neural nets
VFS	Video Frame Spatial

References

1. Parton, B.S. (2005). Sign language recognition and translation: A multidisciplined approach from the field of artificial intelligence. *Journal of deaf studies and deaf education*, 11(1), 94-101.
2. Mitra, S.; and Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3), 311-324.
3. Raffa, G.; Nachman, L.; and Lee, J. (2017). *U.S. Patent Application No.* 15/397, 511.
4. Liu, Z.; Huang, F.; Tang, G. W. L.; Sze, F.Y.B.; Qin, J.; Wang, X.; and Xu, Q. (2016). Real-time Sign Language Recognition with Guided Deep Convolutional Neural Networks. *Proceedings of the 2016 Symposium on Spatial User Interaction*, 187.
5. Chen, F.S.; Fu, C.M.; and Huang, C.L. (2003). Hand gesture recognition using a real-time tracking method and hidden Markov models. *Image and Vision Computing*, 21(8), 745-758.
6. Cavender, A.; Vanam, R.; Barney, D.K.; Ladner, R.E.; and Riskin, E.A. (2008). MobileASL: Intelligibility of sign language video over mobile phones. *Disability and Rehabilitation: Assistive Technology*, 3(1-2), 93-105.
7. Starner, T.; Weaver, J.; and Pentland, A. (1998). Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12), 1371-1375.
8. Kushwah, M. S.; Sharma, M.; Jain, K.; and Chopra, A. (2017). Sign Language Interpretation Using Pseudo Glove. *Proceeding of International Conference on Intelligent Communication, Control and Devices*, 9-18.
9. Kumar, P.; Gauba, H.; Roy, P. P.; and Dogra, D. P. (2017). Coupled hmm-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, 86, 1-8.
10. Mapari, R. B.; and Kharat, G. (2016). American Static Signs Recognition Using Leap Motion Sensor. *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, 67.
11. Auti, A.; Amolic, R.; Bharne, S.; Raina, A.; and Gaikwad, D. P. (2017). Sign-talk: hand gesture recognition system. *International Journal of Computer Applications*, 160(9).
12. Belgacem, S.; Chatelain, C.; and Paquet, T. (2017). Gesture sequence recognition with one shot learned CRF/HMM hybrid model. *Image and Vision Computing*, 61, 12-21.
13. Sun, S.; Luo, C.; and Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36, 10-25.
14. Mohandes, M.; Deriche, M.; and Liu, J. (2014). Image-based and sensor-based approaches to Arabic sign language recognition. *IEEE Transactions on Human-Machine Systems*, 44(4), 551-557.
15. Koller, O.; Forster, J.; and Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141, 108-125.

16. Ejaz, N.; Mehmood, I.; and Baik, S.W. (2013). Efficient visual attention based framework for extracting key frames from videos. *Signal Processing: Image Communication*, 28(1), 34-44.
17. Muhammad, K.; Sajjad, M.; Lee, M.Y.; and Baik, S.W. (2017). Efficient visual attention driven framework for key frames extraction from hysteroscopy videos. *Biomedical Signal Processing and Control*, 33, 161-168.
18. Wang, W.; Shen, J.; Yang, R.; and Porikli, F. (2017). A unified spatiotemporal prior based on geodesic distance for video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1), 20-33.
19. Kishore, P.V.V.; Sastry, A.S.C.S.; and Kartheek, A. (2014). Visual-verbal machine interpreter for sign language recognition under versatile video backgrounds. *Proceedings of International Conference on Networks and Soft Computing* (*ICNSC*), 135-140.
20. Kishore, P.V.V.; Prasad, M.V.D.; Kumar, D.A.; and Sastry, A.S.C.S. (2016). Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks. *Proceedings of the 2016 IEEE 6th International Conference on Advanced Computing* (*IACC*), 346-351.
21. Viola, P.; and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR2001*), 1, 1-9.
22. Hashiyama, T.; Mochizuki, D.; Yano, Y.; and Okuma, S. (2003). Active frame subtraction for pedestrian detection from images of moving camera. *IEEE International Conference on Systems, Man and Cybernetics*, 1, 480-485.
23. Qi, C.; Zhou, Z.; Sun, Y.; Song, H.; Hu, L.; and Wang, Q. (2017). Feature selection and multiple kernel boosting framework based on PSO with mutation mechanism for hyperspectral classification. *Neurocomputing*, 220, 181-190.
24. Nixon, M.S.; and Aguado, A.S. (2012). *Feature extraction and image processing for computer vision*. Academic Press.
25. Priyal, S.P.; and Bora, P.K. (2013). A robust static hand gesture recognition system using geometry based normalizations and Krawtchouk moments. *Pattern Recognition*, 46(8), 2202-2219.
26. Wu, B.; Ai, H.; Huang, C.; and Lao, S. (2004). Fast rotation invariant multi-view face detection based on real adaboost. In *FGR' 04 Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition*, 79-84.
27. Hastie, T.; Rosset, S.; Zhu, J.; and Zou, H. (2009). Multi-class adaboost. *Statistics and Its Interface*, 2(3), 349-360.
28. Borji, A.; and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1), 185-207.
29. Kakumanu, P.; Makrogiannis, S.; and Bourbakis, N. (2007). A survey of skin-color modeling and detection methods. *Pattern recognition*, 40(3), 1106-1122.
30. Cao, X.; Wei, Y.; Wen, F.; and Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2), 177-190.
31. Kishore, P.V.V.; Prasad, M.V.D.; Prasad, C.R.; and Rahul, R. (2015). 4-Camera model for sign language recognition using elliptical Fourier

descriptors and ANN. *2015 International Conference on Signal Processing and Communication Engineering Systems* (*SPACES*), 34-38.

32. Dahmani, D.; and Larabi, S. (2014). User-independent system for sign language finger spelling recognition. *Journal of Visual Communication and Image Representation*, 25(5), 1240-1250.
33. Rao, G. A.; and Kishore, P.V.V. (2017). Selfie video based continuous Indian sign language recognition system. *Ain Shams Engineering Journal*. Available online 24 February 2017 (in press).
34. Rao, G. A.; Kishore, P.V.V.; Kumar, D.A.; and Sastry, A.S.C.S. (2017). Neural network classifier for continuous sign language recognition with selfie video. *Far East Journal of Electronics and Communications*, 17(1), 49-71.
35. Rao, G.A.; and Kishore, P.V.V. (2016). Sign language recognition system simulated for video captured with smart phone front camera. *International Journal of Electrical and Computer Engineering*, 6(5), 2176.