Selfie Sign Language Recognition With Shape Energy Features And ANN Classifier

G. Anantha Rao, K.Syamala, T.V.S.Divakar

Abstract— This paper introduces an idea of bringing sign language into real time using mobile phone. Smart phone front camera captures Selfie sign videos. The hand shape is extracted from the captured video frames using sobel edge operators and morphological gradients. Two methods of feature extraction are proposed. In first feature extraction method, basic region properties height of hand shape, centroid and area of hand shape and distance of the centroid of hand portion from origin of the frame are used as features. In the second feature extraction method, hand contour DCT treated with PCA is generated as feature vector. Artificial Neural Network (ANN) is used to test the Word matching score (WMS) of the selfie sign language recognition system (SSLR) for the proposed feature vector models. The average WMS with first feature vector is 85.5% and 91% for second feature vector.

Index Terms— Indian Sign Language, Hand shape, Countor, DCT, feature vector, ANN, WMS.

1 INTRODUCTION

Sign language is based on hand moments along with facial expressions. It helps the hearing impaired to communicate with normal people. The gesture are either dynamic or static. The better sign language recognition (SLR) requires tracking and orientation of head and hand.[1]. P.V.V.Kishore proposed video segmentation methods to detect hand shape and head positions using wavelets[2]. Tanibata et.al [3] used orientation of hand portion, the flatness and area of hand portion. Parul et.al [4] used height of hand portion, centroid and area of hand portion and distance of the centroid of hand portion from origin of the frame. In SSLR the videos recorded by the smart phone front camera are converted into text or voice. Pre-filtering is to remove video capture noise during video recording. Hand and head portions are segmented from the frames which are used to generate features. Features are generated in two methods. The performance of SSLR is verified with two feature extraction models by applying to ANN individually[5]. This paper consists of methodology and mathematical models to extract hand shape and contour. Results and discussions, conclusion of proposed SSLR in the following chapters.

2. Feature models and classification

The smart phone front camera is used to capture selfie sign videos with the help of selfie stick. Video database of 18 Indian signs for 10 different signers is created.

2.1. Hand shape and contour generation

Fig.1 shows the flow chart of SSLR. Gaussian filters

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \]

with zero mean and standard deviations \( \sigma = 0.01, 0.1, 0.15 \) are used to remove the capture noise.

\[ g^x = \sum_{m=1}^{N} f(x-m, y) g(k) \]

\[ g^y = \sum_{m=1}^{N} f(x-m, y) g^T(k) \]

where gradient operator g is \([+1,-1]\). Gradient magnitude is given by \( \sqrt{(g^x)^2 + (g^y)^2} \).

\[ S^x = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \]

\[ S^y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \]

are the Sobel 2D convolution masks used to generate binary image with block thresholding.

\[ B^{y} = \sum_{i=1}^{M} (S^x \odot f^x)^2 + (S^y \odot f^y)^2 \geq \sum_{i=1}^{M} (S^x \odot f^x)^2 + (S^y \odot f^y)^2 \]

Where size of block is s. It is a faster method which decreases background variations automatically. Fig.2 shows block thresholding and global thresholding (0.2) binary images.
Morphological gradients on the binary image with connected component analysis generates hand head contours given in the following expressions.

\[ h^C(x) = \left\{ z \mid (\hat{M}_{SH}) \cap B^x \neq \emptyset \right\} - \left\{ z \mid (\hat{M}_{SW}) \subseteq B^x \right\} \]

\[ h^C(y) = \left\{ z \mid (\hat{M}_{SV}) \cap B^y \neq \emptyset \right\} - \left\{ z \mid (\hat{M}_{SV}) \subseteq B^y \right\} \]

\[ h^H(x,y) = -h^C(x) \oplus h^H(y) \]

where \( M_{SV} \), \( M_{SH} \) are line masks in vertical and horizontal directions.

Fig.3. shows the a Frame of sign video, Binary Frame, hand portion Contour and head portion Contour.

Fig.3. Frame no. 78, Binary Frame, hand portion Cont head portion Cont.

Four neighbourhood pixel operation is used to separate Hand and head contours from the contour i.e. \( h^C(x,y)_{\text{hand}} \) and \( h^H(x,y)_{\text{head}} \).

2.2. First feature extraction model

As most of the sign are based on hand gestures [6], the hand and head portions are separated. The following features are extracted using region props. Feature vector generated for each frame. i. Average Height(AH) of a Sign: AH of a hand segment in the segmented hand frame based on pixel scanning method is given by

\[ \text{AH} = \frac{P}{Q} \]

Where \( P \) is no. of black pixels of hand shape frame and \( Q \) is no. columns which have more than 0 black pixel. Now, the feature vector is given by the above four samples.

ii. Area A: A is white pixels count pixel value ‘1’.

iii. Centroid \( = (X_h, Y_h) = (\text{Round} (\Sigma (\text{white boundary pixels x coordinate values x}_{\text{wp}})/A), \text{Round} (\Sigma (\text{white boundary pixels y coordinate values y}_{\text{wp}})/A)) \), i.e

\[ X_h = \Sigma x_{\text{wp}} / A \quad \text{and} \quad Y_h = \Sigma y_{\text{wp}} / A \]

iv. The Euclidian distance between the top left pixel to centroid is given by

\[ d = (X_{ht}^2 + Y_{ht}^2)^{1/2} \]

2.3. Second feature extraction model

Hand contour DCT \( h^D(x) \) is given by

\[ F_n^y = \frac{1}{4} \sum_{x,y=0}^{H \times W} h^D(x,y) \cos\left(\frac{2\pi x + 1}{2W}\right) \cos\left(\frac{2\pi y + 1}{2W}\right) \]

here \( C^w = C^y = \frac{1}{\sqrt{2}} \forall (u,v) = 0 \) and 1 elsewhere.

The color coded DCT of hand contour is shown in fig.4. The maximum energy is concentrated in first 50x50 matrix. These 2500 costs more execution time. To reduce the execution time 50x50 feature matrix is optimized with PCA to generate feature vector with 50 samples [7].

2.3. Classification

The single hidden layered ANN shown in fig.5. is used for the sign classification. The extracted feature vectors are used to train and test the ANN FFBP algorithm with sigmoid activation function. Sigmoid activation function is given by

\[ s(\sum \text{net}) = \frac{1}{1 + e^{-\sum \text{net}}} \]

where \( \sum \text{net} \) represents addition of inputs multiplied with weights.

3. RESULTS AND DISCUSSION

The smart phone front camera is used to capture selfie sign videos by holding the mobile phone with selfie stick. We created the video database of 18 Indian signs for 10 different signers.

3.1. Visual Analysis

The selfie sign video database is created for “Hai, Good, Evening, Nice, To, Meet, YOU, I, Am, D, H, R, U, V, A, Thank, You, Bye”. These eighteen words are kept sequential for training and different order for testing video. The sample frames and segmented hand head portions, contours are shown in the fig.6.
In the next section the performance of SSLR with the two feature models is analyzed using mahalanobis distance classifier.

### 3.2. SSLR Performance: Word Matching Score (WMS)

Word matching score is given by

\[ \text{WMS, } \% M = \frac{\text{Correct Classifications}}{\text{Total tested Signs}} \times 100 \]

The results of classification are presented in this section. Training and testing is done with 18 signs of 10 different signers (180 signs). We observed the recognition rates with different number of sets for training and testing. The average recognition rate achieved with the first method of feature extraction is 85.5% shown in table 1 and with the second method of feature extraction is 91% shown in Table 2. The recognition rate is improved with the contour energy features. But training is relatively slow compared to first feature model. Training with first feature model took 0.272 s for a total of 68 epochs and contour energy features took 9.781 s for a total of 28 epochs.

### Table 1

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
<th>Network Architecture</th>
<th>Output Matrix</th>
<th>Confusion Matrix</th>
<th>WMS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 sets each with 18 signs</td>
<td>6 sets each with 18 signs</td>
<td></td>
<td></td>
<td></td>
<td>85.5%</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
<th>Network Architecture</th>
<th>Output Matrix</th>
<th>Confusion Matrix</th>
<th>WMS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 sets each with 18 signs</td>
<td>6 sets each with 18 signs</td>
<td></td>
<td></td>
<td></td>
<td>91.0%</td>
</tr>
</tbody>
</table>

### 4. CONCLUSION

The SSLR system is simulated and tested. Created video database of 18 Indian signs for 10 different signers. The hand shape and contours are generated. The features vectors are obtained in two methods, one with the height of hand shape, centroid and area of hand shape and distance of the centroid of hand portion from origin of the frame and the other with Hand contour energies optimized with PCA. The performance of SSLR system for the two methods of feature vector generation is compared with the word matching score using FFBP ANN. The word matching score with the contour energy features is improved by 5% compared to that of the feature vector with height of hand shape, centroid and area of hand shape and distance of the centroid of hand portion from origin of the frame. Further work needs the improvement in feature set and the classifier models.

### Acknowledgments

We are very much thankful to GMR Institute of Technology for providing research facilities.

### REFERENCES


