Vision Based Real-time Recognition of Hand Gestures for Indian Sign Language using Histogram of Oriented Gradients Features

PRADIP PATEL
Research Scholar,Gujarat Technological University,Ahmedabad, Gujarat, India.
and
NARENDRA PATEL
Department of Computer Engineering,BVM Engineering College,Vallabh Vidyanagar, Gujarat, India.

A sign language is the method of communication used by the deaf people where gestures are used to express meaning. Due to illiteracy of sign language by normal people, there exists communication gap between normal people and deaf people. Very little work has been done for recognition of Indian Sign Language due to lack of standardization and complexity of hand gestures. This resulted in need of automatic system that can recognize Indian Sign Language. Such system, developed via use of image processing and computer vision techniques, will help deaf peoples to communicate with normal people thus filling the communication gap. This paper presents vision based system for real time recognition of hand gesture for Indian Sign Language. The developed system is first trained using training data set. For this, from all the images of dataset, hand region is cropped by performing segmentation using thresholding in YCbCr color space. Histogram of Oriented Gradients (HOG) features of these cropped images are then computed and used to train the classifier. During testing, features of hand region of frames from real time video are presented to classifier for classification. Recognition rates of different classifiers like Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA) are discussed.

Keywords: Sign Language Recognition, Hand Gestures, Segmentation, Feature Extraction, Classification.

1. INTRODUCTION

Sign language is the only medium of communication for deaf people where gestures are used to convey meaning. These gestures are combination of hand shapes, movement, position, palm orientation, arms or body, and facial expressions. In sign language, different gestures are assigned to various alphabets, numbers and words of our language. Badhe and Kulkarni [2015] mentioned gestures of two types: static gestures and dynamic gestures. Static gestures consist of only poses while dynamic gestures often consist of movement of body parts. There are many sign languages in existence all over the world - American Sign Language (ASL), British Sign Language (BSL), Indian Sign Language (ISL) etc. Figure 1 shows representation of ISL alphabets and numbers. Using this sign language deaf people communicates with each other but it is difficult for them to communicate with the normal people as the normal people do not understand sign language. This problem of communication between normal people and deaf people can be solved by using human interpreter as intermediately. But these type interpreters are costly as well as may not available all the time. An alternative solution is to develop a computer based system that can recognize sign-language symbols. Such system can be used as a means of communication with deaf people. Development of an automatic hand gesture recognition system for ISL is more difficult than other sign languages because it includes

—Both single handed and double handed gestures with complex hand shapes.

—Both static and dynamic hand gestures.

—Facial expressions, Head/Body postures.

Figure 1. ISL Signs for Numbers and Alphabets

Sign Language Recognition has been a well researched topic for the ASL, but not so for ISL. There exists some technical challanges for implementing Sign Language Recognition for ISL over other languages like ASL. Due to the complexity of gestures in ISL, hand tracking and segmentation becomes difficult. Furthermore, to do research work in the area of ISL recognition, no standard database is available. So a very smaller amount of research work has been done in this area. As Dixit and Jalal [2013] mentioned, there are two approaches of gesture recognition, hardware based and vision based. Hardware based approach requires signer to wear special hardware like data glove. But this removes naturalness of the system. On the other hand, vision based approach requires the use of image processing and computer vision techniques. These techniques are discussed in details by Gonzalez and Woods [2018] and Forsyth and Ponce [2015]. Due to variable lighting condition as well as dynamic background, vision based approach is very difficult to implement. But still it is found more suitable and practical as compared to hardware based approach. Recently, many researchers are motivated to do research work in this area and, with the improvement of science and technology, have developed few methods to solve the communication problem of deaf people in India. Singha and Das [2013] developed a system that used Eigen vector as feature and Eigen value weighted euclidean distance classifier for 24 different alphabets and achieved 96.25% recognition rate. Kumar et al. [2018] proposed a multimodal framework for Sign Language Recognition system that incorporates facial expression along with sign gesture using two sensors: Leap motion and Kinect. For classification, Hidden Markov Model (HMM) is used. Independent Bayesian Classification Combination (IBCC) approach is used for improve of performance of recognition and achieved recognition rates of 96.05% for single hand gestures and 94.27% for double hand gestures. Rokade and Jadav [2017] applied Euclidean distance transformation on the preprocessed binary image. Row and column projection is applied on the distance transformed image. For feature extraction central moments along with HUs moments are used. The recognition rates achieved are 94.37% for neural network classifier and 92.12% for SVM classifier. Kaur et al. [2017] developed a system that used invariant Krawtchouk moment-based local features and achived 97.9% accuracy. Raheja et al. [2016] performed preprocessing and segmentation in HSV color space. Then features like HuMoments and motion trajectory were extracted and used to train Support Vector Machine. The accuracy achieved was 97.5%. The application proposed by Ansari and Harit [2016] used Microsoft Kinect for capturing image and used Scale invariant feature transform (SIFT) features. It achieved average accuracy rate of 90.68%. A system based on 2D FFT Fourier Descriptors for feature extraction was proposed by Badhe and Kulkarni [2015] where vector codebook is created using LBG and template Matching is done using a simple Euclidean Distance method. The accuracy of the system was 92.91%. Dixit and Jalal [2013] used combination of Hu invariant moment and structural shape descriptors as features. For classification, a multi-class Support Vector Machine (MSVM) is used which

achieved recognition rate of 96.23%. Chaudhary and Beevi [2017] developed a system in which hand region is segmented using skin segmentation with YCbCr and HSV color models. Histogram of Oriented Gradients (HOG) features are extracted from segmented images and are used to train Support Vector Machine for classification. Adithya et al. [2013] proposed a method for automatically recognizing the fingerspelling in Indian Sign Language. To detect hand region from input images, segmentation based on skin color detection is performed. For feature extraction distance transform based shape feature of the image is used. A feed forward neural network is used for classification and the accuracy achieved was 91.11%. The system proposed by Gupta et al. [2016] first categorizes gestures as single-handed or double-handed. Then feature vector is generated which combines HOG and SIFT features. Classification was performed using K-Nearest Neighbor Classifier that achieved accuracy of 91%. Thus, Indian Sign language recognition is current area of research. In this paper, we propose a computer vision based system for Indian Sign Language. The proposed system provides overall accuracy of 98.70% and

—Performs recognition of hand-gestures from both stored images and live camera.

—Converts all ISL Numbers (0-9) and Alphabets (A-Z) into Text and Voice.

—Recognizes both single handed and double handed gestures with complex hand shapes.

2. PROPOSED SYSTEM

A conceptual block diagram of the system is depicted in Figure 2. As shown in figure, system works in two phases: training and testing. During training phase, the classifier is trained by using

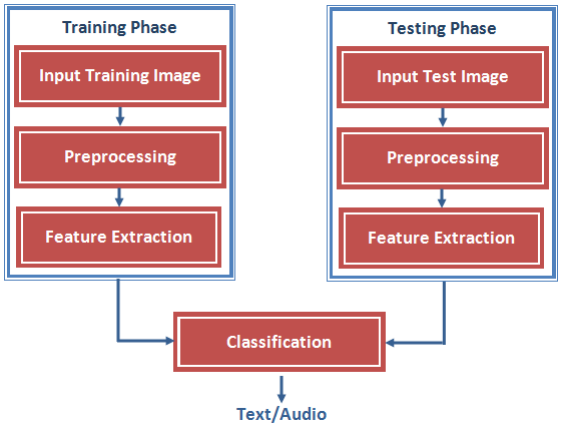


Figure 2. Sign Language Recognition System

the images of training dataset. For this, as no standard database is available for ISL, we have created our own training database of 7920 images. For each alphabet (A-Z) and number (0-9), our database contains 220 images. Few samples of our training database are shown in figure 3. An external webcam is used to capture these images. To train the system, these training images are presented to system along with class labels. Major steps of training phase include dataset creation, preprocessing, feature extraction and classifier training. These steps are discussed in following section.

During testing phase an unknown gesture image is presented to system for classification. Testing also involves image acquisition, preprocessing, feature extraction, and sign recognition. Flow of our system during testing phase is shown in Figure 4. During testing, consecutives frames are extracted from live video recorded through camera. Histogram is then computed for each of these frames and if the histogram difference for few consecutive frames is less than some threshold than that frame is considered as input gesture image. Subsequently, preprocessing steps are performed

Figure 3. Training Dataset

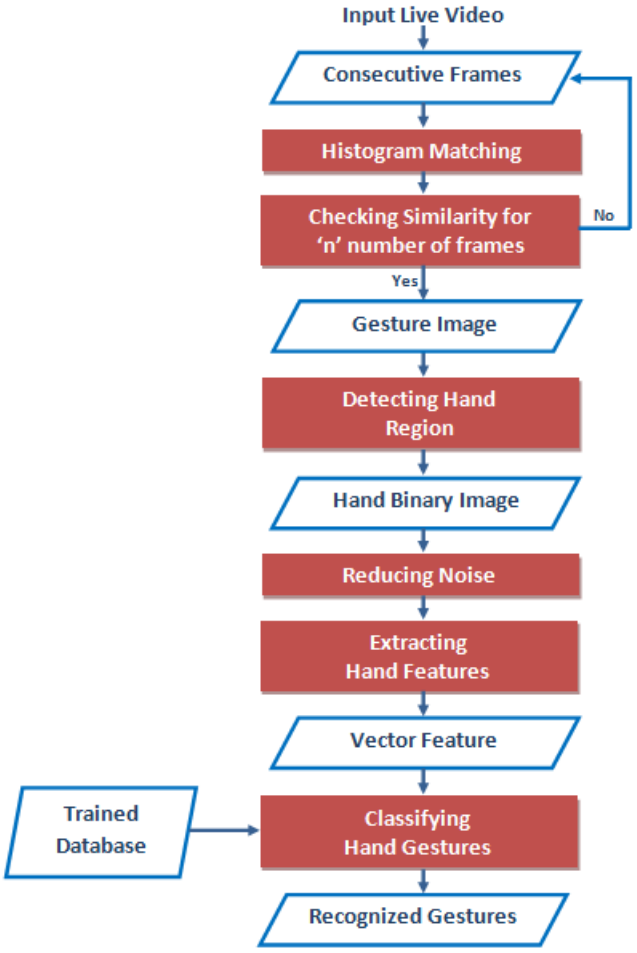


Figure 4. Flow of System during testing

on this input image and Histogram of Oriented Gradients (HOG) features (discussed later) of preprocessed image are given as input to classifier for classification.

2.1 Preprocessing

Preprocessing consist of various operations such as image capturing, segmentation and morphological filtering methods. In the proposed system, after the image is captured from camera, first face region is removed by performing face detection. This steps results in an image with hand region as biggest skin colored object thus simplifying hand detection process. Hand portion is

then detected by applying skin color detection algorithm. As Kolkur et al. [2017] explained, skin color area can be detected by performing thresholding in various colorspace like RGB, HSV and YcbCr. In our experiment, skin color detection by thresholding in YcbCr colorspace resulted in more accuracy than RGB and HSV colorspace as shown in Figure 5. Yusuf et al. [2017] also achieved high accuracy in face detection using skin color detection using YcbCr colorspace. So, in our system, we have used YcbCr colorspace to detect hand area using skin color detection.

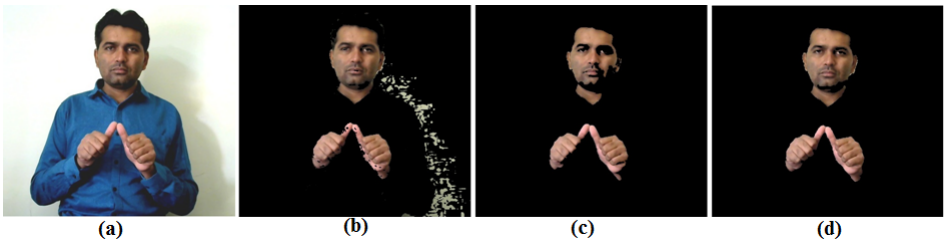


Figure 5. Skin Color Detection (a) Input Image (b) RGB Color Space (c) HSV Color Space (d) YCbCr Color Space

For hand detection, input image is first converted from RGB to YCbCr color space using following equations.

$$Y = 0.299R + 0.587G + 0.114B \tag{1}$$

$$Cr = 128 + 0.5R - 0.418G - 0.081B \tag{2}$$

$$Cb = 128 - 0.168R - 0.331G + 0.5B \tag{3}$$

Then, in order to detect skin color pixels, thresholding is applied. A pixel is set to white if the Y, Cb and Cr values are within a predefined skin color range. Various thresholding values that we have used are as shown in equation 4.

$$75 < Cb < 135 \; and \; 130 < Cr < 180 \; and \; Y > 80 \tag{4}$$

Result of this thresholding operation is binary image. Subsequently, morphological filtering operations are performed on this binary image to remove noise and segmentation errors. Hand region is then detected by finding biggest binary linked object from the image. Eventually, image cropping is performed by finding bounding box of hand region and keeping only that part of image. Finally, the resultant image is scaled to 110 by 110 pixels. Figure 6 shows various preprocessing steps of our system.

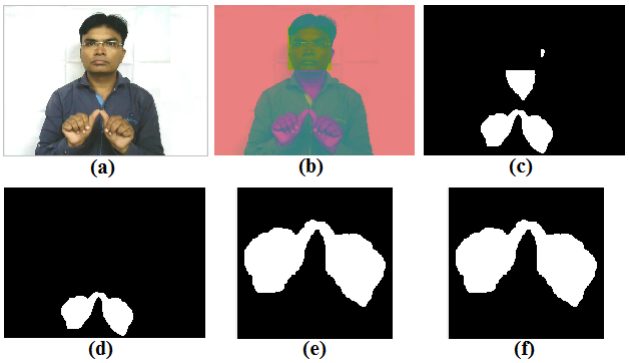


Figure 6. Preprocessing steps (a) Input Image (b) Image in YCbCr color space (c) Image with removed face region (d) Hand Detection using BLOB (e) Scaled Image (f) Filtered Image

2.2 Feature Extraction

Feature extraction is process of extracting image components that are useful for representation and description of shape. It reduces data dimension by encoding related information in a compressed representation and removing less discriminative data. Outputs of this process are image features which are represented as vectors and are given as an input to classifier.

Histogram of Oriented Gradients (HOG), originally proposed by Dalal and Triggs [2005], is frequently used descriptor for recognition of an object in an image. In the system developed by Chaudhary and Beevi [2017] and Gupta et al. [2016], HOG is used as main feature to describe hand shape. This feature is calculated by developing histogram of edge orientation in local regions of image. First, it divides the image into units of small size called cells and for pixels in each cell, it constructs one dimensional histograms of the edge orientations. Then illumination invariance is achieved by normalizing these local histograms for a group of cells which is called block. This normalized histogram forms the descriptors for an image. Steps to calculate HOG features are as below (See Figure 7).

Step 1. Calculate the Gradient Images by using first order sobel operator.
Step 2. Calculate Histogram of Gradients in 3232 cells.
Step 3. Perform 6464 Block Normalization.
Step 4. Calculate the HOG feature vector.

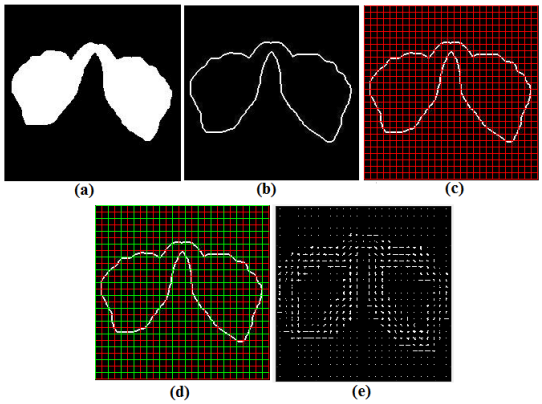


Figure 7. (a) Input Image (b) Image Gradient (c) Image division into cell (d) Formation of block (e) Visualization of HOG Feature

Dimension of HOG feature vector depends on the size of cell. For a given image, visual representation along with corresponding length of HOG features for different cellsize is shown in Figure 8. In the proposed system, we have used cellsize of 32x32 resulting in HOG feature vector of length 144.

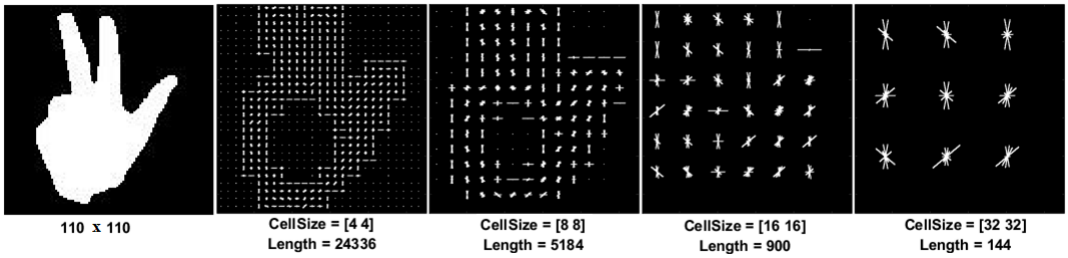


Figure 8. Visualization of HOG features for different cell size

2.3 Classification

After feature extraction, to classify the input signs into one of possible classes, a classifier is used. Classifier, during the training phase, is trained by using the feature vectors obtained from the training database. When an input, images or real time video, is presented for testing to the classifier, it identifies the class corresponding to the sign. The performance of the classifier is measured in terms of accuracy of recognition. We have tested our system with three classifiers: Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA).

2.3.1 *Support Vector Machine.* The SVM, used by Vinh and Tri [2015], is a popular classifier with supervised learning and is defined by hyper plane which separates two classes. It is originally developed by Vapnik and colleagues at bell laboratories as a binary classifier. When an input is presented to SVM, it predicts which of two possible classes forms the output. Multiclass Support Vector Machine (MSVM) is used to solve multi-class problem. It divides problem into several two-class problems that can be solved directly by using multiple SVMs.

2.3.2 *K-Nearest Neighbors algorithm.* The KNN is a object classification method based on nearest training samples based on features. An element is classified as belonging to a class for which maximum elements are near around it. For classification, k elements which are nearest to the test element are considered. When value of k is passed as 1 to the classifier, only one nearest element is considered. The test element is classified as a member of the class of this nearest element. Euclidean is used as the default distance.

2.3.3 *Linear Discriminant Analysis.* The LDA, described by Kumar [2018], is the preferred linear classification technique for more than two classes. It makes predictions based on estimation of the probability that a new set of inputs belongs to each class. The class that has the highest probability becomes the output class. The LDA model uses Bayes Theorem to estimate the probabilities.

3. EXPERIMENTAL RESULTS

We have implemented the proposed system in MATLAB 2018a on a system with Windows 10 operating system, Intel core i5 processor and 8GB RAM. To capture images Logitech webcam with resolution 640x480 is used. Our system is able to recognize gestures of 26 alphabets (A to Z) and 10 numbers (0 to 9) of Indian Sign Language. To measure the accuracy of our system, out of 220 images for each gesture, 160 images were used for training and 60 images were used for testing. Out of these 60 images, number of correctly recognized images using different classifiers are shown in Table 1. Figure 9, 10 and 11 represents gesture wise recognition rates of LDA, SVM and KNN respectively.

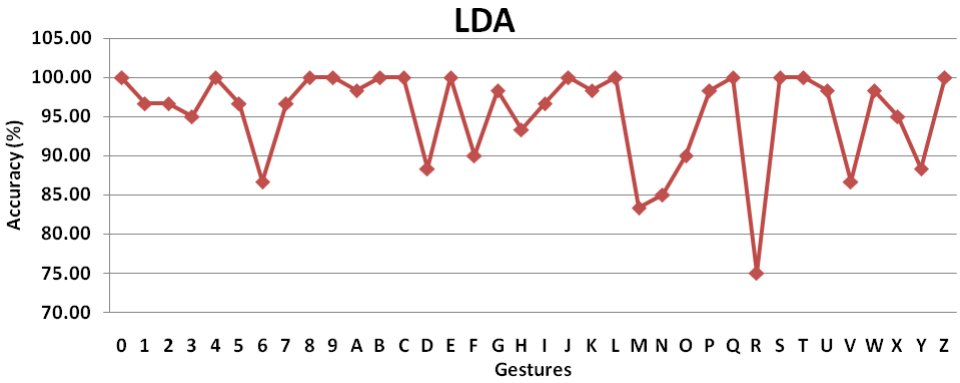


Figure 9. Gesture wise recognition rate of LDA

Gesture	LDA	SVM	KNN
0	60	60	60
1	58	59	60
2	58	59	52
3	57	59	60
4	60	60	60
5	58	60	60
6	52	60	60
7	58	59	59
8	60	60	60
9	60	60	60
A	59	59	60
B	60	60	60
C	60	60	60
D	53	59	53
E	60	58	60
F	54	59	57
G	59	60	60
H	56	60	60
I	58	58	59
J	60	60	60
K	59	60	60
L	60	59	60
M	50	58	52
N	51	57	56
O	54	58	59
P	59	60	58
Q	60	59	60
R	45	60	58
S	60	60	60
T	60	60	60
U	59	60	60
V	52	60	59
W	59	60	60
X	57	52	57
Y	53	60	56
Z	60	60	60

Table I: Gesture wise correct Recognition of LDA, SVM and KNN

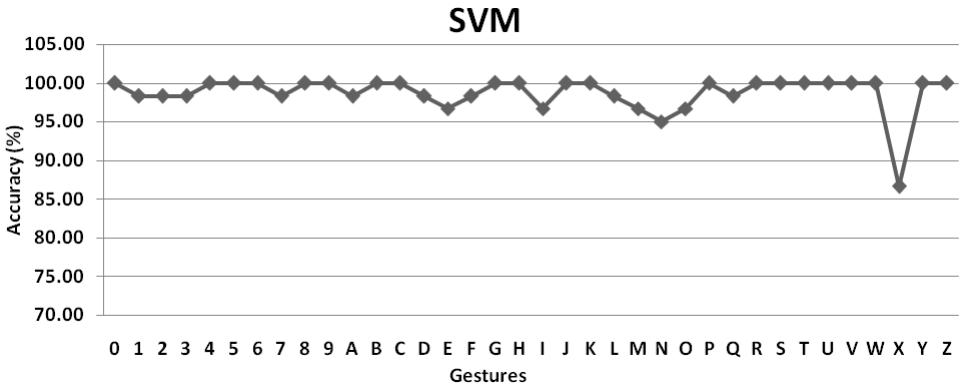


Figure 10. Gesture wise recognition rate of SVM

During experiment, for all three classifier, recognition rate for single handed gestures is better than double handed gestures (see Figure 12). For single handed gestures, recognition rate of LDA

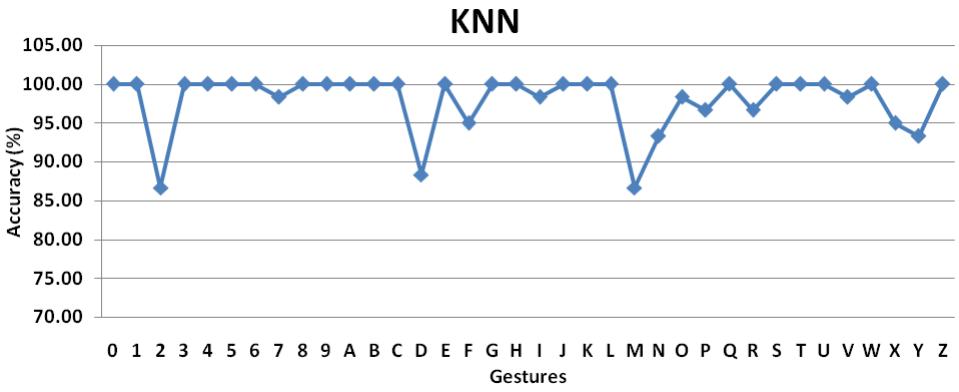
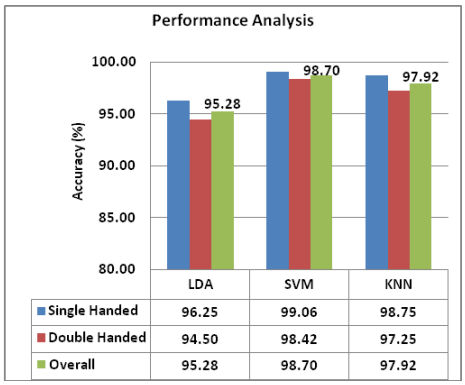


Figure 11. Gesture wise recognition rate of KNN

is 96.25%, SVM is 99.06% and KNN is 98.75%. Similarly, for double handed gestures, recognition rate of LDA is 94.50%, SVM is 98.42% and KNN is 97.25%. Overall recognition rate of LDA is 95.28%, SVM is 98.70% and KNN is 97.92%. In all experiment it is observed that the recognition rate of SVM is better than that of LDA and KNN. So, we have used SVM as classifier in real time recognition system. With the help of developed GUI, the system can recognize gesture from stored image as shown in Figure 13 and from real time camera as shown in Figure 14.



	LDA	SVM	KNN
Single Handed	96.25	99.06	98.75
Double Handed	94.50	98.42	97.25
Overall	95.28	98.70	97.92

Figure 12. Recognition rates LDA, SVM and KNN classifiers

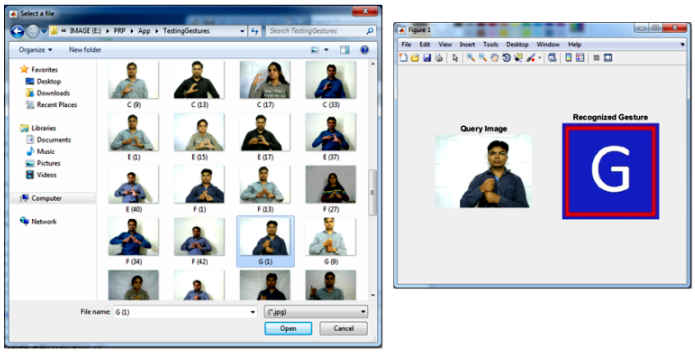


Figure 13. Sign Language Recognition from Stored Image

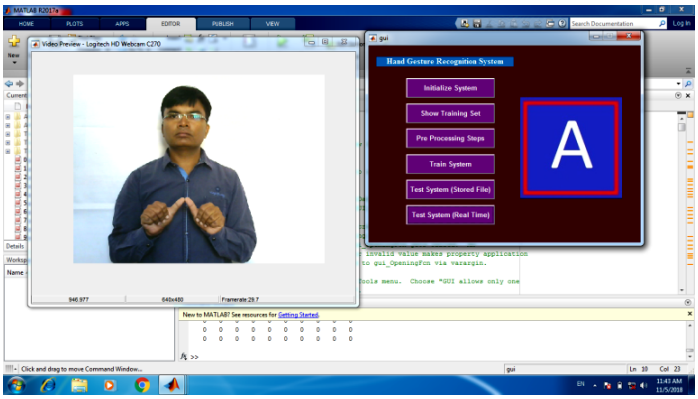


Figure 14. Real Time Sign Language Recognition from Camera

4. CONCLUSION AND DISCUSSION

In this paper we proposed vision based real-time hand gesture recognition system for Indian Sign Language. The system works in two phases, Training and Testing. Each of these two phases includes various modules like Preprocessing, Feature Extraction and Classification. Preprocessing consists of hand detection by segmentation in YCbCr color space and then noise removal by filtering functions to improve the quality of images. During feature extraction, dimension reduction of images is performed by extracting HOG features. These HOG features are used to train classifier. In the developed system, we have evaluated performance of three different classifiers LDA, SVM and KNN. Out of these three classifiers, SVM provided better accuracy. During experiment, it is observed that recognition rate for single handed gestures comes out to be better compared to double handed gestures for all three classifiers.

References

Adithya, Vinod, and Gopalakrishnan, U. 2013. Artificial neural network based method for indian sign language recognition. In *Conference on Information and Communication Technologies (ICT 2013)*. IEEE, pp.1080–1085.

Ansari, Z. A. and Harit, G. 2016. Nearest neighbour classification of indian sign language gestures using kinect camera. *Indian Academy of Sciences Volume 41, Number 2*, pp.161182.

Badhe, P. C. and Kulkarni, V. 2015. Indian sign language translator using gesture recognition algorithm. In *IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*. Bhubaneswar, India, pp. 195–200.

Chaudhary, D. and Beevi, S. 2017. Spotting and recognition of hand gesture for indian sign language using skin segmentation with ycbcr and hsv color models under different lighting conditions. *International Journal of Innovations and Advancement in Computer Science(IJIACS) Volume 6, Issue 9*.

Dalal, N. and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 886–893.

Dixit, K. and Jalal, A. S. 2013. Automatic indian sign language recognition system. In *3rd IEEE International Advance Computing Conference (IACC)*. Ghaziabad, India.

Forsyth, D. A. and Ponce, J. 2015. In *Computer Vision A Modern Approach*. Pearson Education. 2nd Edition.

Gonzalez, R. C. and Woods, R. E. 2018. In *Digital Image Processing*. Pearson Education. 4th Edition.

Gupta, B., Shukla, P., and Mittal, A. 2016. K-nearest correlated neighbor classification for indian sign language gesture recognition using feature fusion. In *International Conference on Computer Communication and Informatics (ICCCI)*. Coimbatore, India, pp. 1–5.

Kaur, B., Joshi, G., and Vig, R. 2017. Indian sign language recognition using krawtchouk moment-based local features. *The Imaging Science Journal Volume 65, Number 3*, pp.171–179.

Kolkur, S., Kalbande, D., Shimpi, P., Bapat, C., and Jatakia, J. 2017. Human skin detection using rgb, hsv and ycbcr color models. *Advances in Intelligent Systems Research Volume 137*, pp.324–332.

Kumar, M. 2018. Conversion of sign language into text. *International Journal of Applied Engineering Research Volume 13, Number 9*, pp. 7154–7161.

Kumar, P., Roy, P. P., and Dogra, D. P. 2018. Independent bayesian classifier combination based sign language recognition using facial expression. *Information Sciences Volume 428*, pp. 30–48.

Raheja, J. L., Mishra, A., and Chaudhary, A. 2016. Indian sign language recognition using svm. *Pattern Recognition and Image Analysis Volume 26, Issue 2*, pp.434441.

Rokade, Y. I. and Jadav, P. M. 2017. Indian sign language recognition system. *International Journal of Engineering and Technology Volume 9, Number 3*.

Singha, J. and Das, K. 2013. Recognition of indian sign language in live video. *International Journal of Computer Applications Volume 70, Number 19*, pp. 17–22.

Vinh, T. Q. and Tri, N. T. 2015. Hand gesture recognition based on depth image using kinect sensor. In *IEEE 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science*. Ho Chi Minh City, Vietnam.

Yusuf, A., Mohamad, F., and Sufyanu, Z. 2017. Human face detection using skin color segmentation and watershed algorithm. *American Journal of Artificial Intelligence Volume 1, Issue 1*, pp.29–35.