Volume 10, Issue 3
August 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Speeded Up Robust Features Based Real Time Hand Gesture Recognition for Indian Sign Language

[1]Pradip Patel and [2]Narendra Patel

[1]Research Scholar, Gujarat Technological University, Gujarat, India

[2]Computer Engineering Department, BVM Engineering College, Gujarat, India

E-mail: [1]pradippatel@ldce.ac.in, [2]nmpatel@bvmengineering.ac.in

ABSTRACT

Development of an automatic sign language recognition system to solve the communication problems of speech and hearing impaired people is challenging task. It is modern part of research in the area of human computer interaction. In this paper, we propose such recognition system for hand gestures of Indian Sign Language (ISL). Based on computer vision and image processing algorithms, the system accurately recognizes gestures from stored images and live camera. The recognized gestures are then converted into text and voice. The system is built using the state of art machine learning model - Bag of Visual Words (BOVW) with Speeded Up Robust Features (SURF) features. Four different classifiers such as K-Nearest Neighbors (KNN), Neural Network (NN), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) have been trained. Experiments are carried out on these classifiers with different feature size and results are presented. During experiments it is observed that the system provided invariance with respect to scale, translation and rotation. It achieved recognition rate up to 99.40% using SVM classifier.

Keywords: *Bag of Visual Words, Speeded Up Robust Features, Support Vector Machine, K-Nearest Neighbors*

1. INTRODUCTION

Sign language is used by deaf people as the only way of communication. It includes various gestures consisting of different hand shapes. Further, various alphabets, numbers and words of our language are assigned different gestures in sign language. Basically these gestures [1] fall into one of the two types: static gestures and dynamic gestures. Static gestures consist of only poses while dynamic gestures often consist of movement of body parts. All over the world, for each country, there exists separate sign language. Indian Sign Language (ISL) is used by deaf people in India. Figure 1 shows gesture representation of various alphabets and numbers of ISL. ISL is used by deaf people in India for their internal communication. But these deaf people cannot communicate with others people in the society having speech ability as they are not able to interpret sign language.

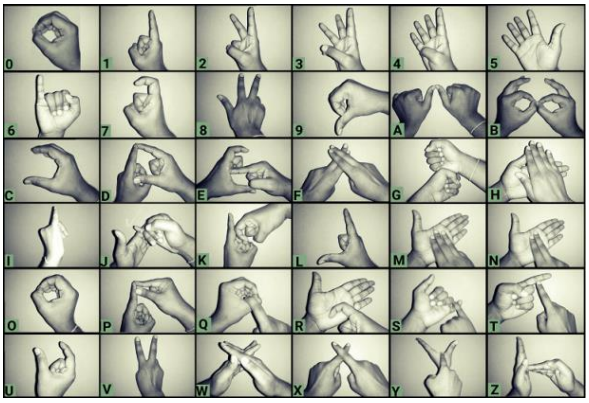

Figure 1. ISL gestures for numbers and alphabets

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. Another option is to build a computer based automatic system that can recognize gestures of sign language and translate them into text or voice. Such system can be used as a means of communication between normal people and deaf people. With comparison to other sign languages, it is very difficult to build such gesture recognition system for ISL because it includes complex hand gestures, body movement and facial expressions.

There exist some technical challenges 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 complicated. Additionally, to carry out research work in the area of ISL recognition, no standard database is available. Therefore, research work of very little amount has been carried out in this field so far. As mentioned in [2], 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 [3] and computer vision [4]. 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. In recent times, numerous research persons are encouraged to carry out their work in this field and, because of the enhancement in technology, have created systems to enable the communication between deaf and normal people in India. A system based on 2D FFT Fourier Descriptors for feature extraction was proposed by Badhe and Kulkarni in [1] 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 [2] used combination of Hu invariant moment

Volume 10, Issue 3
August 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

and structural shape descriptors as features. For classification, a Multi-class Support Vector Machine (MSVM) is used which achieved recognition rate of 96.23%. Singha and Das [5] developed a system that used Eigen vector as feature and Eigen value weighted Euclidean distance classier for 24 different alphabets and achieved 96.25% recognition rate. Joshi et al. [6] presented an ISL recognition using HOG parameters. A combined Taguchi and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) based decision-making technique is applied to determine the values of these parameters. For the acquired ISL complex background dataset, the selected values of parameters are further used to obtain multi-level HOG resulting in the overall accuracy of 92% for 280 features. Rokade and Jadav [7] developed system in which images are first converted into binary form using preprocessing. Euclidean distance based transformation is applied to these binary images. On resultant images, row and column projection is applied. Central moments and HU's moments [8] are then computed as features. The recognition rates achieved are 94.37% for neural network classier and 92.12% for SVM classier. B. Kaur in [9] developed a system that used invariant Krawtchouk moment-based local features and achieved 97.9% accuracy. Raheja et al. [10] performed preprocessing and segmentation in HSV color space. Then features like Hu Moments 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 in [11] used Microsoft Kinect for capturing image and used Scale invariant feature transform (SIFT) features. It achieved average accuracy rate of 90.68%. Chaudhary and Beevi [12] 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. The system proposed by Gupta et al. in [13] 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 Neighbour Classier that achieved accuracy of 91%. In [14], Adithya et al. proposed a method for automatically recognizing the finger spelling 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%.

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 99.40% and

- Recognizes hand-gestures corresponding to all ISL Numbers (0-9) and Alphabets (A-Z) and converts them into Text and Voice.
- Recognizes gestures from stored images as well as live camera.
- Provides invariant with respect to Scale, Translation and Rotation.

2. PROPOSED SYSTEM

Proposed system works in two phases: training and testing. While in first phase of training, classification model is learned using image features of training dataset. Due to unavailability of standard database 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 dataset are shown in Figure 2.

Figure 2. Training dataset

An external webcam is used to capture these images. During the training period, these images are offered to system along with their class labels. Key steps performed during training are preprocessing, feature extraction and classifier learning. These steps are discussed in following section. For the period of testing phase, an unusual gesture image is submitted to system to classify it. Testing stage also includes acquirement of image, preprocessing, extraction of features, and finally classification. Flow of our system during testing phase is shown in Figure 3.

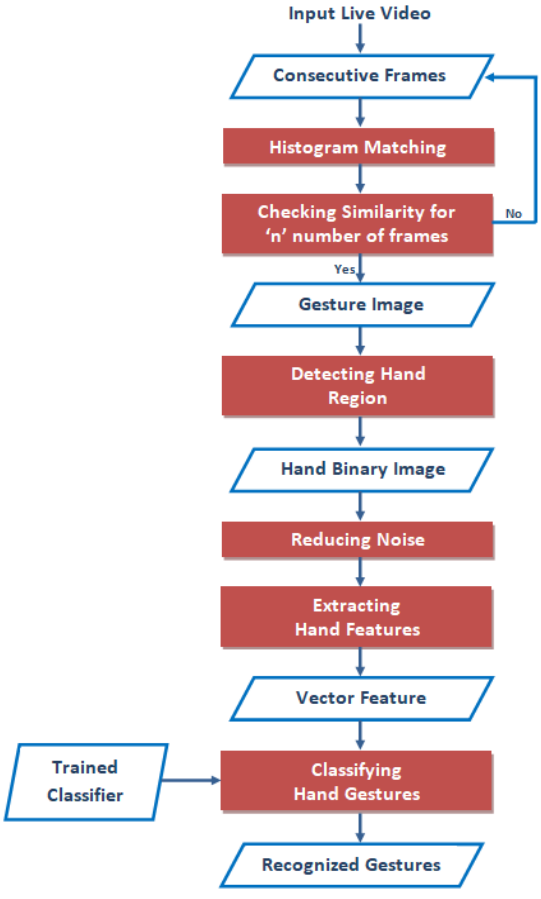

Figure 3. Flow of system during testing

Volume 10, Issue 3
August 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

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 on this input image and Speeded Up Robust Features (SURF) features 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, during experiment we have used RGB camera as well as Microsoft Kinect depth sensor.

2.1.1 RGB Camera

While working with RGB camera, 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. explained [15], skin color area can be detected by performing thresholding in various colorspace like RGB, HSV and YcbCr. In our system, we have used YcbCr [16] colorspace to detect hand area using skin color detection. For hand detection, using equation 1, input image is first converted to YCbCr space from RGB space. Then, to find pixels with having skin color, thresholding is performed. Various thresholding values that we have used are as shown in equation 2.

$$\begin{aligned} Y &= 0.299 * R + 0.587 * G + 0.114 * B \\ Cr &= 128 + 0.5 * R - 0.418*G - 0.081 * B \\ Cb &= 128 - 0.168 * R - 0.331 * G + 0.5 * B \end{aligned} \quad (1)$$

$$75 < Cb < 135 \text{ and } 130 < Cr < 180 \text{ and } Y > 80 \quad (2)$$

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 4 shows these steps of our system.

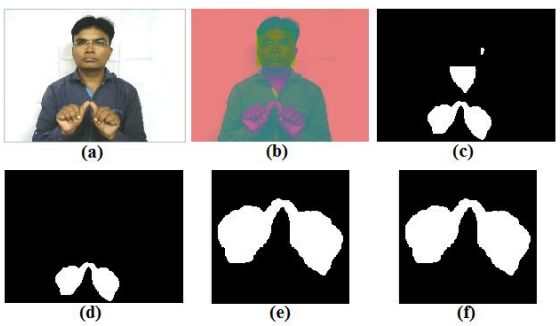

Figure 4. Preprocessing steps (a) Input image (b) Image in YCbCr color space (c) Image with removed face region (d) Hand detection using BLOB (e) Scaled image (f) Filtered image

2.1.2 Kinect Sensor

Microsoft Kinect [11] provides an inexpensive and easy way for real-time user interaction. The hand region is determined by applying thresholds to the hand point in the depth image. First of all, the nearest hand point is found by finding the point with smallest depth value. We then set all the point that has the depth distance to hand point less than a threshold. Figure 5 shows these steps of our system.

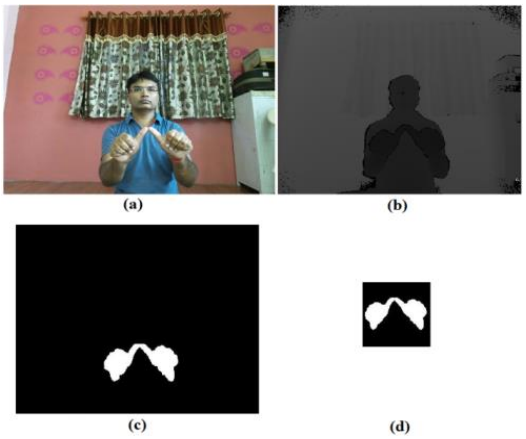

Figure 5. Preprocessing steps (a) Input image (b) Depth image (c) Hand detection by thresholding in depth (d) Cropped image

2.2 Bag of Visual Words (BOVW) Model

This is state of art machine learning model used for image classification having very impressive accuracy. Using this technique for image classification has two steps: Feature Extraction and Codebook Construction.

2.2.1 Feature Extraction

In order to accurately represent the rough shape of gestures, we have used SURF features proposed by Herbert Bay in [17]. These features are based on the orientation of the gradient in localized region.

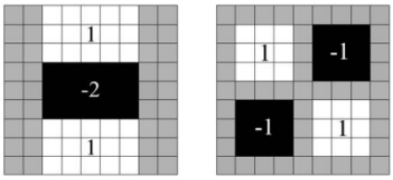

Figure 6. Approximated Gaussian second derivative

To extract SURF features [18], Scale spaces [17] are first computed by using box filters of different sizes. To make the computation fast, integral images are used as filters of different sizes can be applied quickly on them. The output of the 9×9 filters shown in figure 6 is used as the first layer of scale space. The subsequent layers are then obtained by filtering the image with filters of size 9×9, 15×15, 21×21, 27×27,.... Then, to detect interest point [17], non-maximum suppression is applied in a 3×3×3 neighborhood over different scales. Filter size is used as the third dimension. Further, to handle rotation, the dominant direction of the feature is computed and then sampling window is rotated to align with that direction. Square neighborhood interest point is then divided into 16 squares which further divided into 4 squares as shown in Figure 7. Derivatives in x and y direction are

Volume 10, Issue 3
August 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

computed for these squares. The final descriptor v = (∑dx, ∑dy, ∑|dx| ,∑|dy|) with total 64 dimensions is used as final feature vector. Figure 8 shows visual representation of SURF features of image selected from our dataset.

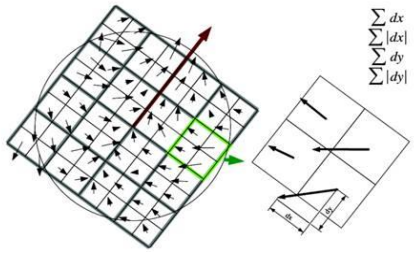

Figure 7. A graphical representation of the SURF descriptor.

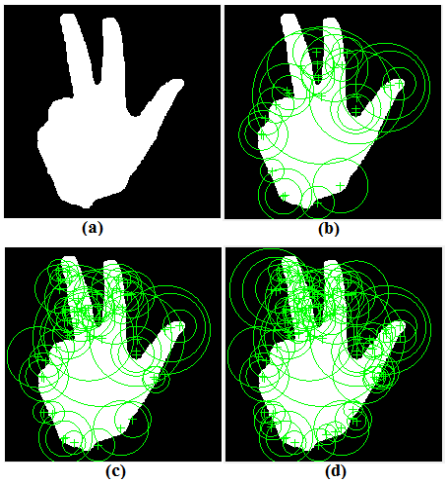

Figure 8. SURF visualization (a) Input image (b) 25 Key points (c) 50 Key points and (d) 75 Key points

2.2.2 Codebook Construction

The vectors generated in the feature extraction step are now converted into the codewords. Codewords are nothing but vector representation of similar patches. Then codebook is created via use of the k-means clustering algorithm (See Figure 9). Each cluster center produced by k-means becomes a codeword. And the number of clusters becomes the codebook size. So summarizing this step, each patch in an image is mapped to a certain codeword through the clustering process and the image can be represented by the histogram of the codewords.

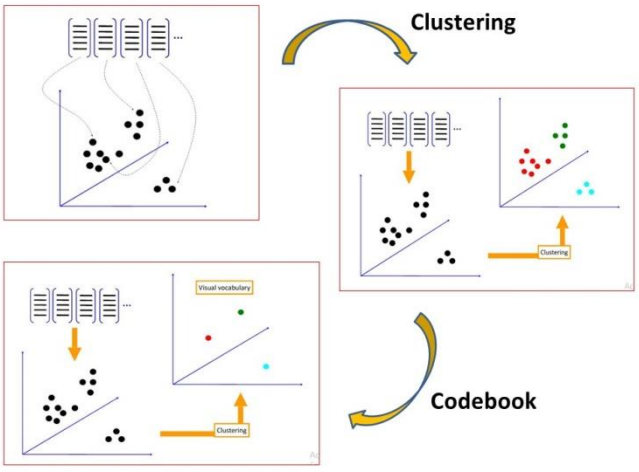

Figure 9. Codebook Construction

All the images of training set are represented into a histogram of codewords. It is done by first applying the keypoint detector to every training image, and then matching every keypoint with those in the codebook. The result of this is a histogram where the bins correspond to the codewords, and the count of every bin corresponds to the number of times the corresponding codeword matches a keypoint in the given image.

2.3 Classification

During training, the histograms of the training images are used to learn a classification model. During testing, test image is first represented by a histogram of codewords which is then given to classifier for classification. Here we have carried out experiments with four different classification models: K-Nearest Neighbors (KNN), Neural Network (NN) Support, Vector Machine (SVM), and Linear Discriminant Analysis (LDA).

2.3.1 Support Vector Machine

The SVM [10] is defined by hyper plane which separates two classes. Once trained, SVM classifies an input as belonging to one of the two classes. For multi-class problem, Multiclass Support Vector Machine (MSVM)[2] is used where the problem is divided into several two-class problems that can be solved directly by using multiple SVMs.

2.3.2 K-Nearest Neighbors

The KNN [13] classifier classifies an input element as a member of the class whose elements are maximum in number around it. It is defined by parameter k. Total k elements nearest to the input element are considered for the classification. Euclidean [3] is used as the default distance.

2.3.3 Neural Network

Neural Network [19] consists of connections among layers of artificial neurons with each layer is totally connected to the subsequent layer. The neural network is trained using algorithm of back propagation where input-output training pairs are given as input propagating backwards the difference of actual and projected results. It makes an effort to shrink the difference until it sufficiently well learns the training data.

2.3.4 Linear Discriminant Analysis

The LDA [20] is another classifier used to solve multi-class classification problem. When an input is presented to LDA, it calculates probability with which an input belong to each class. The class having highest probability becomes the output class. The LDA model uses Bayes Theorem to estimate the probabilities.

3. EXPERIMENT RESULTS

We have implemented our system in MATLAB 2019b 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 ISL gestures of 26 alphabets (A to Z) and 10 numbers (0 to 9). 160 images out of 220 images were used in training while remaining 60 images were used for testing.

Volume 10, Issue 3
August 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Figure 10 represents gesture wise recognition rates of different classifiers for feature vector size of 200. Performance measured during experiments in terms of different parameters is shown in Table 1.

Parameter	SVM	KNN	NN	LDA
Accuracy	0.9940	0.9884	0.9907	0.9731
Error	0.0060	0.0116	0.0093	0.0269
Sensitivity	0.9940	0.9884	0.9907	0.9731
Specificity	0.9998	0.9997	0.9997	0.9992
Precision	0.9941	0.9888	0.9909	0.9747
False Positive Rate	0.0002	0.0003	0.0003	0.0008
False Negative Rate	0.0060	0.0116	0.0093	0.0269
F1_score	0.9940	0.9885	0.9907	0.9734

Table 1. Performance of proposed system

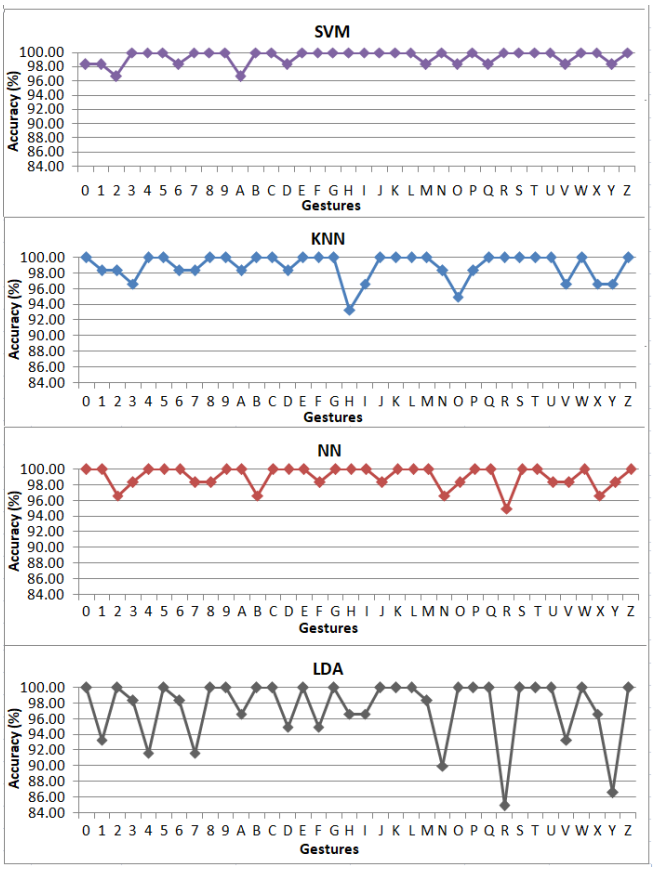

Figure 10. Gesture wise recognition rates (Feature size = 200)

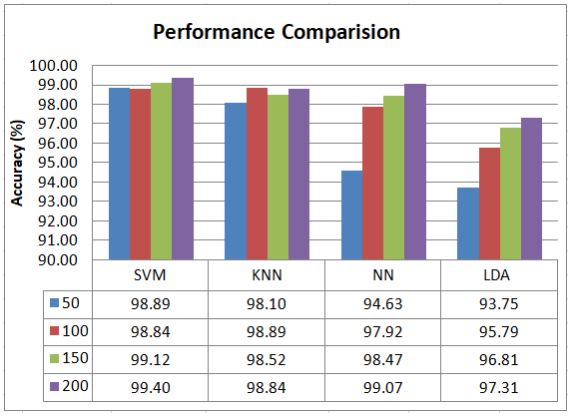

Figure 11. Recognition rates for feature vectors with different size

We also measured recognition rates of different classifiers for feature vectors with size 50, 100, 150 and 200 (see Figure 11). With the help of developed GUI, the system can recognize gesture from stored image as shown in Figure 12 and from real time camera as shown in Figure 13. Table 2 depicts computed training and classification time for different classifiers.

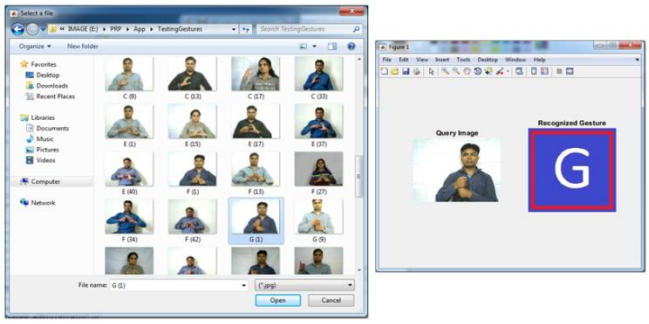

Figure 12. Sign language recognition from stored image

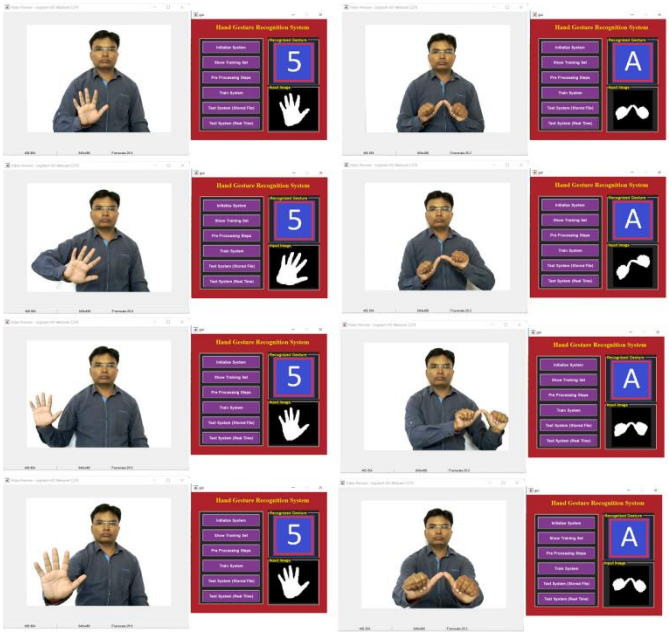

Figure 13. Real time sign language recognition from camera

Classifier	Training Time (Seconds)	Classification Time (Seconds)
SVM	9.859	0.320
KNN	0.326	0.007
NN	1.256	0.004
LDA	7.544	0.007

Table 2. Timing calculation (Feature size = 200)

4. CONCLUSION

In this paper we proposed vision based real-time hand gesture recognition system for Indian Sign Language. System is built using state of art machine learning model: bag of visual words. This bag is created using SURF features. Using these features, four different classifiers namely SVM, KNN, NN and LDA are trained. Experiments are carried out on these classifiers with different feature size. From the results

it is observed that the proposed system achieved accuracy up to 99.40% using SVM classifier and 200 feature size. The proposed system also provided invariance with respect to scale, translation and rotation.

REFERENCES

[1] P. C. Badhe and V. Kulkarni, "Indian Sign Language Translator using Gesture Recognition Algorithm", IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), Bhubaneswar, India, pp. 195-200, 2015.

[2] K. Dixit and A. S. Jalal, "Automatic Indian Sign Language Recognition System", 3rd IEEE International Advance Computing Conference, Ghaziabad, India, 2013.

[3] Wilhelm B. and M.J. Burge, "Principles of Digital Image Processing: Advanced Methods", Springer, 2013.

[4] R. Szeliski, "Computer Vision: Algorithms and Applications", 2nd Edition, Springer, 2021.

[5] J. Singha and K. Das, "Recognition of Indian Sign Language in Live Video", International Journal of Computer Applications, Volume 70, Number 19, pp. 17-22, 2013.

[6] G. Joshi, S. Singh and R. Vig, "Taguchi-TOPSIS based HOG parameter selection for complex background sign language recognition", Journal of Visual Communication and Image Representation, Volum 71, No 102834, 2020.

[7] Y. I. Rokade and P. M. Jadav, "Indian Sign Language Recognition System", International Journal of Engineering and Technology, Volume 9, Number 3, 2017 .

[8] M. Hu, "Visual Pattern Recognition by Moment Invariants," IRE Trans. Inf. Theory, Volume 8 (2), pp.179–187, 1962.

[9] B. Kaur, G. Joshi and R. Vig, "Indian Sign Language Recognition Using Krawtchouk Moment-Based Local Features", The Imaging Science Journal, Volume 65, Number 3, pp.171-179, 2017.

[10] J. L. Raheja, A. Mishra and A. Chaudhary, "Indian Sign Language Recognition Using SVM", Pattern Recognition and Image Analysis, Volume 26, Issue 2, pp.434-441, 2016.

[11] Z. A. Ansari and G. Harit, "Nearest Neighbour Classification of Indian Sign Language Gestures Using Kinect Camera", Indian Academy of Sciences, Volume 41, Number 2, pp.161-182, 2016.

[12] D. Chaudhary and S. Beevi, "Spotting and Recognition of Hand Gesture for Indian Sign Language Using Skin Segmentation With YCbCr and HSV Color Models Under Dierent Lighting Conditions", International Journal of Innovations and Advancement in Computer Science(IJIACS), Volume 6, Issue 9, 2017.

[13] B. Gupta, P. Shukla and A. Mittal, "K-Nearest Correlated Neighbor Classification for Indian Sign Language Gesture Recognition Using Feature Fusion", International Conference on Computer Communication and Informatics, Coimbatore, India, pp. 1-5, 2016.

[14] Adithya, Vinod, and U. Gopalakrishnan, "Artificial Neural Network Based Method for Indian Sign Language Recognition", Conference on Information and Communication Technologies, IEEE, pp.1080-1085, 2013.

[15] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat and J. Jatakia, "Human Skin Detection Using RGB, HSV And YCbCr Color Models", Advances in Intelligent Systems Research, Volume 137, pp.324-332, 2017.

[16] T. Sokhib and TK . Whangbo, "A Combined Method of Skin and Depth based Hand Gesture Recognition", International Arab Journal of Information Technology, Volume 17, Issue 1, pp. 137-145, 2020.

[17] H. Bay, T. Tuytelaars and L. Van Gool, "SURF: Speeded Up Robust Features", European Conference on Computer Vision, pp 404-417, 2006.

[18] D. Silva, A. Sousa, A and V. Costa, V. "A Comparative Analysis for 2D Object Recognition: A Case Study with Tactode Puzzle-Like Tiles", Volume 7, Issue 4, No 65, 2021.

[19] PN. Huu , QT. Minh and HL.The, "An ANN-based gesture recognition algorithm for smart-home applications", Ksii Transactions on Internet and Information Systems, Volum 14, Issue 5, pp. 1967-83, 2020.

[20] M. Kumar, "Conversion of Sign Language into Text", International Journal of Applied Engineering Research Volume 13, Number 9, pp. 7154-7161, 2018.

AUTHOR PROFILES

Pradip Patel is currently working as Assistant Professor at L. D. College of Engineering, Ahmedabad. He is Pursuing Ph.D. from Gujarat Technological University, Ahmedabad. He received the degree in M.E. Computer Engineering from BVM Engineerinc College, Vidyanagar, Guajrat. His area of interest are Image Processing, Computer Vision, Machine Learning and Deep Learning. He has many publications in this areas. In addition to this he has guided many M.E. students.

Dr. Narendra Patel is currently working as Professor at BVM Engineering College, Vidyanagar, Gujarat. He received his PhD degree from Sardar Vallabhbhai National Institute of Technology, Surat. He has got vast knowledge of research in the field of computer science and engineering. His key interest area are Image Processing, Deep Learning and Computer Vision. He has many publications in reputed journals in these areas. He has guided many M.E. and PhD students.