

A Machine Learning Based Approach for the Detection and Recognition of Bangla Sign Language

Muttaki Hasan, Tanvir Hossain Sajib and Mrinmoy Dey*

Department of Electrical and Electronic Engineering
Chittagong University of Engineering and Technology
Chittagong-4349, Bangladesh

tusar4577@gmail.com, tanvirsajibm277@gmail.com, *mrinmoycuet@gmail.com

Abstract—Speech impaired people are detached from the mainstream society due to the lacking of proper communication aid. Sign language is the primary means of communication for them which normal people do not understand. In order to facilitate the conversation conversion of sign language to audio is very necessary. This paper aims at conversion of sign language to speech so that disabled people have their own voice to communicate with the general people. In this paper, Hand Gesture recognition is performed using HOG (Histogram of Oriented Gradients) for extraction of features from the gesture image and SVM (Support Vector Machine) as classifier. Finally, predict the gesture image with output text. This output text is converted into audible sound using TTS (Text to Speech) converter.

Keywords— *SVM; Classification; BdSL; TTS Engine; Feature; HOG; Contouring; Prediction; Recognition rate.*

I. INTRODUCTION

Speech impaired people are usually deprived of normal communication with other people in the society. According to the 2001 Disability Welfare Act Persons with speech impairment are classified as: Loss of one's capacity to utter/pronounce meaningful vocabulary sounds, or damaged, partly or wholly or dysfunctional [1]. Bangladesh has population of 150 million and among them 1.5 million people are hearing impaired [2]. At this age of technology, it is quite essential to make these people feel part of the society by helping them communicate smoothly. Sign Language is the primary means of communication in the deaf and dumb community. Hands are the basic means of communicating using sign languages. Hand shapes, hand movement, palm orientation and hand position are some of the most important components to convey the meaning of a sign.

In Bangladesh a formal sign language has been established only recently. In the year 2000, Centre for Disability in Development (CDD) took the initiative to standardize communication with sign languages in this country. Before this step, there were different local variants and no national dialect existed in their training centre [3]. People in Bangladesh still are ignorant of this mode of communication and thus the deaf children still cannot lead an uncomplicated life here. In this paper we present a vision-based method that detects hand

gestures in different illuminations. For this purpose, we have employed machine learning approach which includes training the classifier using HOG (Histogram of Oriented Gradients) features. Classification has been done using k-NN (k-Nearest Neighbour). Finally, we have measured recognition accuracy by giving test gestures as input to the classifier.

II. RELATED WORKS

Many Researchers have worked on bangla sign language recognition but no researches have been carried out using HOG as feature vector. Rahat Yasir and Riasat Azim Khan implemented two-handed hand gesture recognition using LDA (Linear Discriminant Analysis) and ANN (Artificial Neural Networking) [4]. Clustering and neural networking approach have also been used by Foez M. Rahim, Nasrin Sultana and Tamnun E. Mursalin [5]. SIFT (Scale Invariant Feature Transform) based bangla sign language recognition was done by Farhad Yasir, P. W. C. Prasad, Abir Alsadoon [6]. Histograms of oriented gradients, called HOG, were firstly used by Dalal and displayed an excellent performance in human detection [7]. Zondag et al. used HOG features combination with two variations of Ada Boost algorithm for the construction of a real-time hand detector. It was seen in the experiment that, HOG showed a good performance [8]. Yu et al. [9] used HOG features for characterizing hand shape and constructed a classifier by Support Vector Machine algorithm (SVM) for static hand gesture recognition. In their experimental testing, 9 defined postures were recognized. Sha et al. [10] extended the HOG to refine the best posture region and recognition. Experimental results showed promising performance under various capture conditions. Buehler, M. Everingham, A. Zisserman [11] also developed mechanism where HOG was also chosen to represent the hand shapes for automatically learn a large number of British Sign Language (BSL) signs from TV broadcasts.

III. METHODOLOGY

We have chosen for the implementation is supervised machine learning for sign language recognition. We have used SVM as classifier for prediction of the desired gesture. The whole process has been divided into six steps. We will perform the following steps for our applications:

- (a)Pre-process an image
- (b)Segment an image
- (c)Extract the features
- (d)Training classifier with feature vector
- (e) Recognition and Prediction
- (f) Audio Output

We have used OpenCV as image processing library and python as programming language. An USB webcam has been used for capturing the gesture images. In training phase a person has to provide sample image of his hand gesture so that the reference template model or database can be created. In training phase images we have collected hand gesture of 16 Bangla Sign Language expressions from Bangla Sign Language Dictionary. Each of the images has been resized by 200×200 pixels and saved to the database. Each gesture contains 20 sample images which have been taken from different angle for better accuracy. So, in total we have collected total 320 sample images. These 320 images and saved into a database after pre-processing.

IV. PRE-PROCESSING

We collected gesture image using a program which captures 16 images of a gesture upon key press. LEDs illumination has been used for proper thresholding. These LEDs brighten the surface of desired target. So it becomes easier to detect the contour. We have used binary thresholding technique for this purpose. The flowchart of capturing the gesture image datasets are given below:

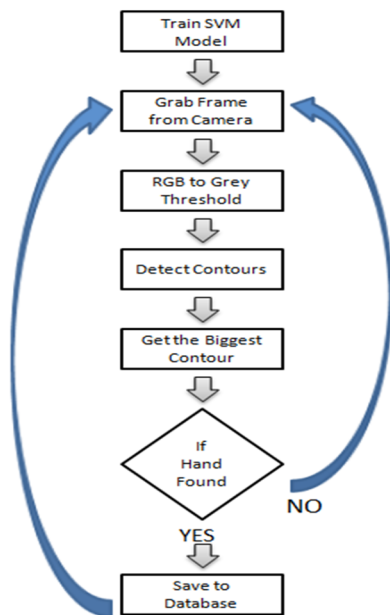


Fig. 1. Flow Chart of Building Gesture Database.

If the key 'c' is pressed on the keyboard the program starts to capture image. All images are captured from different angles for proper training and better recognition. The angle of the camera plays a crucial point. It has to be set in such a way that it directs towards the hand of the user. The database has been

created using the BdSL sign language dictionary and hand gestures have been given accordingly.

If the key 'n' is pressed the program tells the user to change and start capturing a new gesture and it continues capturing image until 20 images have been captured. After capturing the samples the images are then stored in the database.

The user can control number of gestures and training datasets.

V. SEGMENTATION

In the segmentation step, we need to extract the regions of interest of an image and isolate each one as a unique object of interest. This has been done by a python program which uses OpenCV library to detect hand, crop the portion of hand region and save it to the database. Hand is detected using contouring process. We first find out all contours of the image and then validate the largest contour to detect the hand. A bounding rectangle around the hand is drawn using the contours. The area of the rectangle is our area of interest. Next the area is cropped by 200×200 pixels. The cropped image is saved to the training folder. The cropping has been done using python program which uses OpenCV function to crop a picture by the defined size.



Fig. 2. Training Image Datasets for classifier.

VI. TRAINING THE CLASSIFIER

After getting the feature inside the image, we continue with the next step. We need to extract all the features of each one detected object; a feature is a vector of characteristics of objects. We have used HOG as our feature vector. HOG feature, however, divides an image into many units which called cells. Firstly, gradient orients or edge orients and angles of each pixel are calculated all over the image using varieties of masks Sobel masks. The histograms of gradient directions

over each pixel of the cell are then accumulates. Some cells compose a region called block. Then an image can be regarded as a connection of many blocks. The concatenated histograms of the whole block forms the vectors of HOG. Now we need to find the HOG descriptor of each cell. For that, we find Sobel derivatives of each cell in X and Y direction. Then find their magnitude and direction of gradient at each pixel. This gradient is quantized to 16 integer values. Divide this image to four sub-squares. For each sub-square, calculate the histogram of direction (16 bins) weighted with their magnitude. So each sub-square gives a vector containing 16 values. Four such vectors (of four sub-squares) together give us a feature vector containing 64 values.

This feature vector has been used to train our classifier. We have used SVM (Support Vector Machine) as our classifier. SVM is a discriminative classifier which classifies test data by optimal hyper plane. The operation of the SVM algorithm is based on finding the hyper plane that gives the largest minimum distance to the training examples. The problem of finding the optimal hyper plane is an optimization problem and can be solved by optimization techniques.

We can define the hyper planes H such that:

$$w * x_i + b \geq +1 \text{ when } y_i = +1 \quad (1)$$

$$w * x_i + b \leq -1 \text{ when } y_i = -1 \quad (2)$$

H_1 and H_2 are the planes:

$$H_1 : w * x_1 + b = +1$$

$$H_2 : w * x_1 + b = -1$$

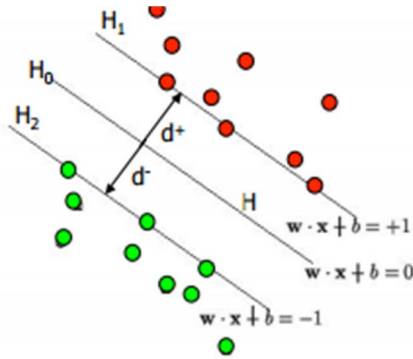


Fig. 3. Optimal hyper plane of Support Vector Machine.

The points on the planes H_1 and H_2 are the tips of support vectors. The plane H_0 is the median in between where,

$$W * x_i + b = 0 \quad (3)$$

Here:

W is a weight vector,

x_i is input vector,

b is bias.

d^+ = the shortest distance to the closest positive point.

d^- = the shortest distance to the closest negative point.

We have used k-NN which is one of the simplest of classification algorithms available for supervised learning. In k-NN a 2d feature space is considered where all training datasets are projected. These datasets are categorized into class. To classify a new data into the feature space the best method is to check nearest neighbour. For those who are nearest neighbour get higher weights while those are far away get lower weights. Then we add total weights of each family separately. Whoever gets highest total weights, new-comer goes to that family.

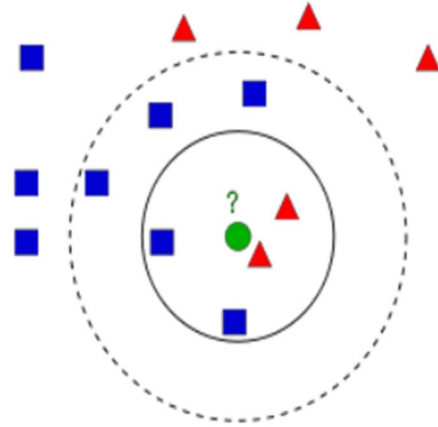


Fig. 4. Feature space of k-NN.

VII. TESTING THE CLASSIFIER

About 320 have been used to train the SVM classifier and among them 64 datasets have been used for testing purpose. The whole datasets are divided into 9 classes. Each class contains 20 samples. Cross validation is used to evaluate our model and number of cross validation folds are 5. The cross validation score was found 0.94 which has been done using a python program. It also supported multiclass. This proves that our model fitted properly.

VIII. RECOGNITION AND PREDICTION

In the recognition process testing images are filtered using image processing method and HOG features are extracted from those images which gives us feature vector. These feature vectors are used as an input on each SVM classifier. k-NN method is applied on the testing feature vector which measures the weight of the feature vector with each of the two classes of SVM. If the response of the prediction goes with the positive class the testing image is considered to be a member of that positive class. On the other side negative response discard the testing image from the classifier. Finally the testing image which receives a positive response from the classifier is shown through the output window.

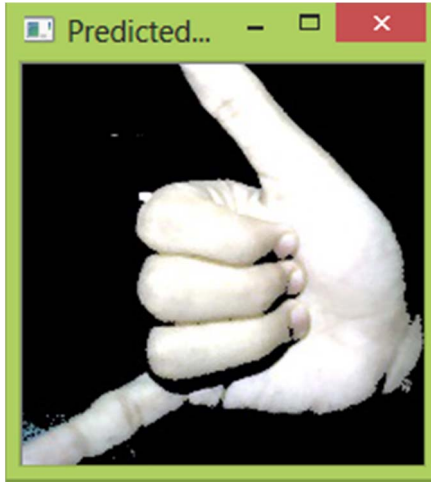


Fig. 5. Predicted Output.

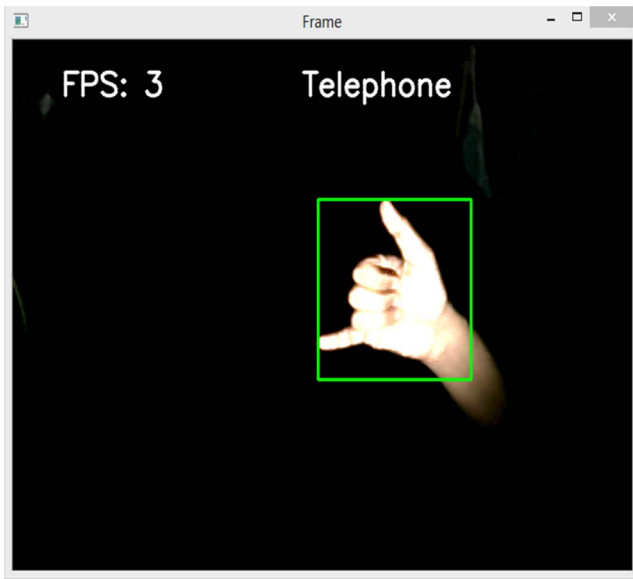


Fig. 6. Detection of Gesture with label.

IX. AUDIO OUTPUT

Each of the testing images has their corresponding level. This level contains the meaning of the predicted image as text. The text is taken as a string by the python compiler. The string is given as input to the TTS Engine using pyttsx python library. The string is hence converted to audible sound using TTS. As a result when the predicted image is shown the meaning of the sign can also be heard as audible sound.

X. RESULTS










This research approach used 16 signs of expressions. The signs have been chosen from “বাংলা ইশারা ভাষার অভিধান” published by BdSL[12]. Signs have been chosen in such way that it includes single handed gesture along with two handed gesture. The training dataset was prepared using different background. Each of these signs takes 20 samples for data training in the SVM classifier. For sign of expression we have








used $16 \times 20 = 320$ images for training and applied k-Nearest Neighbour algorithm.

If we are able to insert more sample image for the sign we could acquire more robust SVM classifier which enhances the recognition rate. Positional variation of the sample image gives accurate prediction. It was seen that the classifier gives better performance where background reflection didn't occur. Moreover it worked well in black background and white foreground. Proper illumination to the region of interest gives perfect prediction.

We used the test vector to obtain an approximation error of our model. To get the error estimation, we need to predict all the test vector features to obtain the SVM prediction results and then compare these results to the original labels. We used the predict function to predict results. An accuracy test was performed on prediction and the following results were found:

TABLE I. PREDICTION TEST

Sign of Expression	Table Column Head		
	Meaning	Recognition Rate	Result
	South (□□□□□)	90.13%	Good
	North (উত্তর)	90.25%	Good
	Three (তিন)	85.20%	Moderate
	Bless (আশির্বাদ)	82.33%	Moderate
	Telephone (দূরভাষা)	91.23%	Good
	Victory (বিজয়)	87.33%	Moderate
	One (এক)	83.20%	Moderate
	Nice (সুন্দর)	85.50%	Moderate
	Time (সময়)	87.50%	Moderate

Sign of Expression	Table Column Head		
	Meaning	Recognition Rate	Result
	Illness (অসুস্থতা)	80.25%	Average
	Taxi (গাড়ী)	90.74%	Good
	Clean (পরিষ্কার)	85.3%	Moderate
	Village (গ্রাম)	84.66%	Moderate
	Spy (গুপ্তচর)	83.33%	Moderate
	Founder (প্রতিষ্ঠাতা)	85.33%	Moderate
	Today (আজ)	92.33%	Good

Testing Result on our testing dataset gave average recognition rate of 86.53%. This result shows a great potential of HOG feature based method can meet the requirements of hand detection.

XI. CONCLUSION

In this paper we propose a method for 16 predefined gesture recognition using HOG features. Considering large dataset and computational efficiency SVM is an efficient approach for decision making. Many previous works has been done for detection and recognition of hand gestures but HOG feature along with audio output enables the speech impaired to communicate more efficiently. We firstly construct HOG feature based hand detector for bangla sign language detection. The final testing result shows that our method exhibits a good performance in testing. However this method is suitable for static gesture recognition. For dynamic gesture recognition we have taken first frame as input as dynamic gesture is a combination of consecutive static gesture. The number of training image can be increased for high accuracy of recognition. The recognition rate of gestures can also be increased by proper positioning and orientation of hand and proper lighting. In future we would like to introduce a portable

sign language recognizer that can take hand gesture as input and gives the meaning of the gesture as an audio output.

ACKNOWLEDGMENT

The authors would like to thank the department of Electrical & Electronic Engineering (EEE) of Chittagong University of Engineering and Technology (CUET), Chittagong-4349, Bangladesh.

REFERENCES

- [1] "POPULATION MONOGRAPH BANGLADESH" volume - 5 Bangladesh Bureau of Statistics Bangladesh, BBS, November 2015, chapter-1, pp. 2.
- [2] *Society For Education & Care of Hearing Impaired Children of Bangladesh* [Online], Available at <http://www.hicarebd.org/>
- [3] Najeefa Nikhat Choudhury and Golam Kayas, "Automatic Recognition of Bangla Sign Language," B. Sc. Thesis, Dept .Comput. Eng. Brac University, Dhaka, 2012, pp.7.
- [4] Rahat Yasir, Riasat Azim Khan, "Two-Handed Hand Gesture Recognition for Bangla Sign Language using LDA and ANN," IEEE-2014, November 6-7, 2015, Hiroshima, Japan.
- [5] Foez M. Rahim, Nasrin Sultana and Tamnun E. Mursalin, "Intelligent Sign Language Verification System - Using Image Processing, Clustering and Neural Network Concepts,"
- [6] Farhad Yasir, P.W.C. Prasad, Abir Alsadoon, "SIFT Based Approach on Bangla Sign Language Recognition," *IEEE 8th International Workshop on Computational Intelligence and Applications*, November 6-7, 2015, Hiroshima, Japan
- [7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, 2005.
- [8] J. A. Zondag, T. Gritti and V. Jeanne, "Practical study on real-time hand detection," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–8, 2009.
- [9] R. Yu and G. Chengcheng, "Hand gesture recognition based on HOG characters and SVM," *Bull. Sci. Technol.*, vol. 27, no. 2, pp. 211–214, 2014.
- [10] L. Sha, G. Wang, A. Yao, et al., "Hand posture recognition in video using multiple cues," in *Proc. IEEE International Conference on Multimedia and Expo.*, pp. 886–889, 2009.
- [11] P. Buehler, M. Everingham and A. Zisserman, "Learning sign language by watching TV (using weakly aligned subtitles)," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2009, pp. 2961–2968, 2009.
- [12] *Bangla Sign Language Dictionary*, 2nd ed., Bangladesh Sign Language Committee, 1997, pp. 92-117.