International Journal of Pure and Applied Mathematics
Volume 117 No. 20 2017, 9-15
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Automatic Sign Language Finger Spelling Using Convolution Neural Network: Analysis

Beena M.V.
Asst. professor, CSE Dept.
Vidya Academy of Science and
Technology, Thrissur - 680501, India

Dr. M.N. Agnisarman Namboodiri
Dean P G studies,
Vidya Academy of Science and
Technology, Thrissur - 680501, India

Abstract—Very few people understand sign language. More- over, contrary to popular belief, it is not an international language. Obviously, this further complicates communication between the deaf community and the hearing majority. The alternative of written communication is cumbersome, because the deaf community is generally less skilled in writing a spoken language. For example, when an accident occurs, it is often necessary to communicate quickly with the emergency physician where written communication is not always possible. The purpose of this work is to contribute recognizing American sign languages to the field of automatic sign language recognition with maximum efficiency. This paper focuses on the recognition of static gestures of ASL which are collected from Kinect sensor. The most challenging part in the design of an automatic sign language translator is the design of a good classifier that can classify the input static gestures with high accuracy. In the proposed system, design of classifier for sign languages recognition uses CNN architecture from kinect Depth images. The system trained CNNs for the classification of 24 alphabets and 0-9 numbers using 33000 images. The system has trained the classifier with different parameter configurations and tabulated the results. Compared to previous literature the proposed work attained an efficiency of 94.6774% for our classifier. Also created a simple java GUI application to test our classifier. We have designed our network to be light weight so that it can be incorporated easily with embedded devices having limited resources. The result shows that accuracy improves as we include more data from different subjects during training.

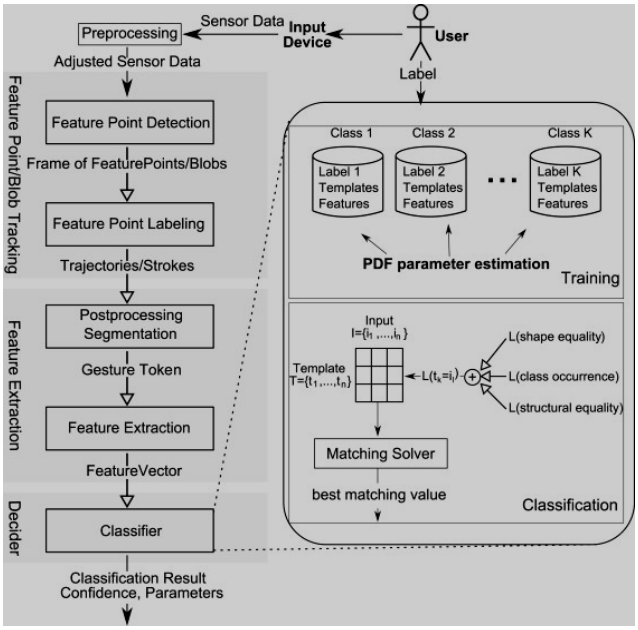

Fig. 1. General System Architecture

Keywords: Artificial Neural Network, ASL, Convolutional Neural Network, Deep Learning, GPU, PDNN, Pyttsx, Theano

I. Introduction

In daily life, the communication between different communities highly depends on human based translation services. The involvement of human expertize is very difficult and expensive for translation. The automatic sign language recognition leads to understand the meaning of different signs without the help from expert persons.

In common, sign language recognition system contains different modules: object tracking, skin segmentation, feature extraction, and recognition. The first two modules are used to extract and locate hands in the video frames. The purpose of next modules stands for feature extraction, classification and recognition. Fig.1 demonstrates a general system architecture overview for a SLR system. Based on segmented hands, we can extract the hand shape and orientation based features. Finally, classifiers are trained to recognize the signs. The sign language to speech converter reduces the bridge between normal people and dump people.

Sign language recognition is still a challenging problem despite of many research efforts during the last few decades. It requires the understanding of combination of multi-modal information such as hand pose and movement, facial expression, and human body posture. Moreover, even same signs have significantly different appearances for different signers and different viewpoints.

In this paper, we focus on American Sign Language (ASL) recognition from static depth images. In all over the world more than hundred sign languages in the world. American Sign Language (ASL) is used throughout U.S. and Canada, as well as other regions of the world, including Western regions of Africa and Southeastern regions of Asia. Approximately 500,000 people use ASL as a primary language in U.S. The fig.2 shows ASL alphabets and numbers. Visual similarity of different signs make it difficult for recognition. So it become a challenging area in computer vision tasks. Depth sensors enable us to capture additional information to improve accuracy and/or processing time.

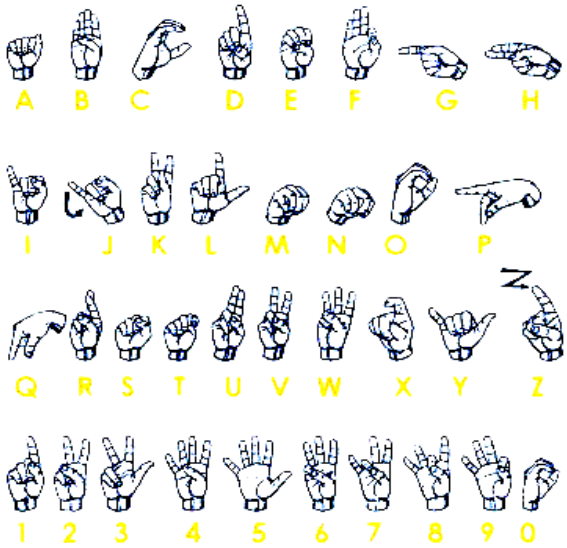

Fig. 2. ASL finger alphabets and numbers

Also, with recent Improvement of GPU, CNNs have been employed to many computer vision problems. The reason for this is the reduced training and testing time when using GPU compared to CPU. So in this work a fast fully parameterizable GPU configuration of CNN is used to train the hand gestures for good feature extraction and classification.

A. Motivation

The various advantages of building such a system includes:

- Sign-to-text/speech translation system or dialog systems which is use in specific public domains such as airports, post offices, or hospitals.
- SLR can help to translate the video to text or speech enables inter communication between normal and deaf people.

B. Problem Statement

Sign language uses lots of gestures so that it looks like a movement language which consists of a series of hands and arms motion. There are different standards for sign languages for different countries. Also to be noted that some unknown words are translated by simply showing gestures for each alphabet in the word.

In addition, sign language also includes specific gestures to each alphabet in English dictionary and for each number between 0 and 9. Based on these sign languages are made up of two groups, namely static gesture and dynamic gesture. Static gesture is used for alphabet and number representation, whereas dynamic gesture is used for specific concepts. Dynamic also include words, sentences etc. Static gesture consists of poses of hand, whereas latter include motion of hands, head or both. Sign language is a visual language and consists of 3 major components, such as finger-spelling, word level sign vocabulary and Non-manual features. Finger-spelling is used to spell words letter by letter whereas latter is keyword based.

But the design of a sign language translator is quite challenging despite of many research efforts during the last few decades. It requires the understanding of combination of multi-modal information such as hand pose and movement, facial expression, and human body posture. Moreover, even same signs have significantly different appearances for different signers and different viewpoints. Also even same signs have significantly different appearances for different signers and different viewpoints. This work focuses on the creation of a static sign language translator by using Convolutional Neural Network. We created a light weight network that can be used with embedded devices having less resources.

C. Objectives

The main objective of this project is to contribute to the field of automatic sign language recognition. We focus on the recognition of the static sign language gestures. This work focused on deep learning approach to recognize 24 alphabets and 0-9 numbers. We created a convolutional neural network classifier that can recognize static sign language gestures with high accuracy. We have trained the network under different configurations, also analyzed and tabulated the obtained results. The result shows that accuracy improves as we include more data from different subjects during training. We have also created a simple java GUI application to test our classifier.

II. LITERATURE REVIEW

Byeongkeun et al. proposed a real-time sign language finger spelling recognition using Convolutional Neural Networks [1] from Depth map. The work focuses on static finger spelling in American Sign Language though small but important part of sign language recognition. Even though it used depth sensors which enable them to capture additional information to improve accuracy and processing time, their caffe architecture is very complicated. [1] A method for implementing a sign language to text/voice conversion system without using handheld gloves and sensors, by capturing the gesture continuously and converting them to voice. In this method only few images were captured for recognition. The design of a communication aid for physically challenged [2] has been created as a prototype.

The system developed under the MATLAB environment. It consists of mainly two phases via training phase and testing phase. In training phase the author used a Feed Forward Neural Network with 200 neurons in the hidden layer and 10 in output which takes 58 epochs. In testing phase a real time footage of sign language gesture is captured, it is then segmented and then compared with the database created. If a match is found by the neural network then text output of the corresponding gesture is produced.

The problem pattern interpretation of hand gestures includes the following issues.

1. Identifying and tracking the characteristics of hand gestures.
2. Training the captured gestures using Feed Forward Neural Network
3. Segmentation of the hand gestures which is a continuous Stream.
4. Interpretations of the attribute patterns constituting the Gestural segment.
5. Integrating the concurrent attributes as a whole.

Sruthi Upendran [3] and *et.al* introduced "American Sign Language Interpreter System for Deaf and Dumb Individuals". The discussed procedures could recognize 20 out of 24 static ASL alphabets. The alphabets A, M, N and S couldn't be recognized due to occlusion problem. They have used only a limited number of images.

The same can be implemented using an optimized approach by implementing the famous viola jones algorithm with LBP feature for hand gestures recognition in a real time environment. By using this algorithm created Indian sign language interpreter with android implementation [4]. The advantage of this approach is that it takes less computational power to detect the gestures.

Another work related to this field was creating sign language recognition system by using pattern matching [5]. The main aim of this proposed work is to create a system which will work on sign language recognition. Many researchers have already introduced about many various sign language recognition systems and have implemented using different techniques and methods. This proposed system is focusing on an approach which is to put the SLR system which will work on Signs as well as Text (which will understandable by deaf and dumb persons and also by normal persons). The main task will be performed in two ways by the system. It will take input by the user in the form of text which will be then perform matching with the sign and vice-versa.

The first way is when user will give the input as a text, it will perform matching with the already created database entries with its corresponding signs and then system will output that sign to the requesting user. The same technique is used to process letters, numbers as well as words and eventually phrases. The second way includes the concept of image processing [4], the input given by another user as a sign (which will be in image format) will be processed by the system on the basis of the outer portion of the fingers and hands portion of the image .If the sign is valid then it will generate its text format which will be output on screen to user.

Chenyang Zhang, Yingli Tian [5] *et.el* "multi-modality American Sign Language recognition" .The main enlightened features of the system is twofold: 1) it consider about multiple signal modalities including the sequence of depth images, RGB image-based hand shapes, facial expression attributes and key point detections for both body joints and facial landmarks. 2) By learning from signing sequences performed by fluent ASL signers and annotations provided by professional linguisticians, our system can recognize different components such as English words and special ASL grammar components, such as facial expressions or head movements that have grammatical meaning during sentences.

Real time tracking [16] of gestures from hand movement is difficult than the face recognition [6]. A comparison has been done in [7] for still and moving image recognition. Sign language recognition has been extensively tried using mathematical models. [8] Explains the deployment of Support Vector Machine for sign language recognition.

III.IMPLEMENTATION

A. Dataset

For the system implementation, 33000 images have collected from the available dataset, using Creative Senz3D depth camera of the resolution of 320x240. Compared to RGB images, more information can collect from depth images. So in the proposed system, we have taken the advantage of Kinect images for attaining maximum efficiency.

The dataset consists of 1,000 images for each of the 33 different hand signs from five subjects. 33 hand signs include all the finger spellings of both alphabets and numbers except J and Z which require temporal information for classification.

Since (2/V) and (6/W) are differentiated based on context, only one class is used to represent both one alphabet and one number.

The collected dataset images need to be modified according to our needs. Our aim was to create a light weight CNN classifier that can be used with resource constrained embedded devices. Se we downscaled the dataset image to 28x28 gray scale images. This will help to reduce the number of input nodes in the first layer. Here we are using only 784 features from each image for both training and testing. For simplicity we have pickled all the images using pythons pickle function.

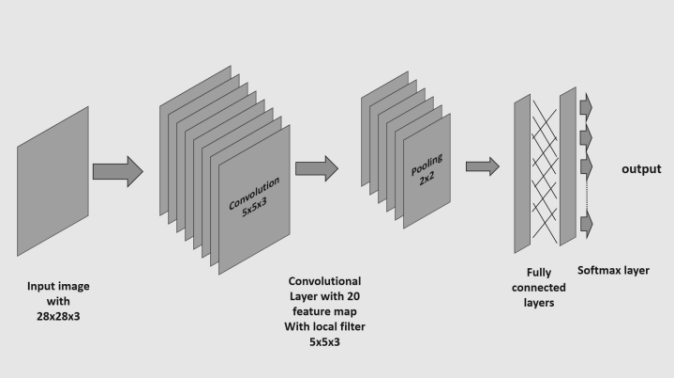

Fig. 3. Convolution with single layer

B. Classification

Architecture: In our work we used PDNN implementation of the CNN. PDNN is a Python deep learning toolkit developed under the Theano environment [14]. The architecture consists of a single convolutional layer with 20 feature maps, local filter having a size of 5x5, and a pooling size of 2x2. The input to the architecture has one feature map with a dimension of 28x28. The output of the network are flattened. The number of targets or output classes is 33. We have trained our network with FC hidden layers ranging from 1 to 4. We have used a learning rate of 0.1 for our model.

Feature Extraction: We extracted 784 feature vector from each preprocessed depth image. The images are then grouped to training, validation and testing and they were pickled respectively.

Training: We train and test neural networks in twenty four different operating modes. We trained the model by varying the hidden layers from 1 to 4 and by varying the epochs from 500 to 1000 in each case. The number of nodes in each hidden layer also varies. Out of the 1000 samples of each image we used 900 images for creating the training set, 60 for validation and 40 for testing. The data sets were pickled and were used for training. Figure 4 shows the flowchart of the training process. After the model was trained the net parameters were saved so that it can be used in the testing phase for testing the accuracy of the model and also for classification of the input symbols.

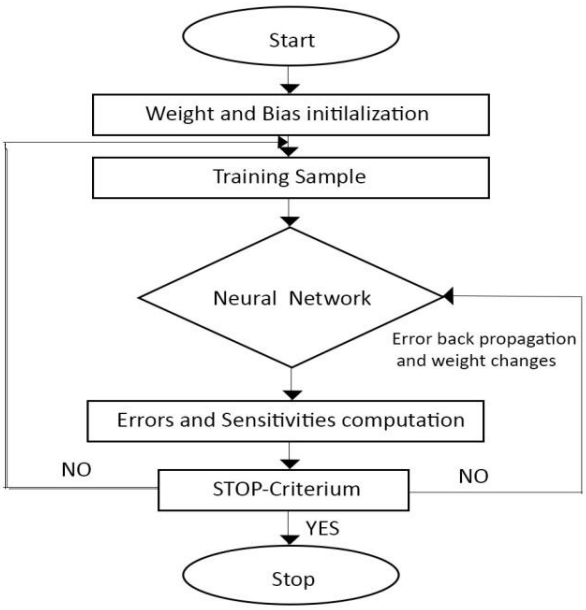

Fig. 4. Training phase

Testing: In the testing phase the accuracy of our trained model is tested. The saved network parameter is loaded to test the dataset and determined the accuracy. The method used for all test cases and tabulated the obtained accuracies. A simple java GUI application is created to test out classifier for the purpose of static sign language translation. The application allows the user to select the images of the sign language static gestures that need to be classified. The application used our trained CNN network to classify these symbols and produce their corresponding labels. The labels are then turned to their corresponding alphabets/numbers and they are grouped to form words or sentences. The words/sentences are then spoken out using the python pyttsx text to speech module. The Screen shot of the GUI is shown in fig.5.

IV. EXPERIMENTAL RESULTS

As mentioned in Sec.3, we train and test for twenty four different experimental settings. The results are shown in figure 6 .Our system achieves 94.6774% accuracy when training and validation data have samples corresponding to the test subject with training set having 33000 images and validation set having 1980 and the test set having 1320 images. In this experiment, we have used 98% of images in the dataset. Out of 1000 images belonging a class 900 images are used for training, 60 for validation and 40 for testing. We have trained the model for 33 symbols. We considered all possible combinations of subjects for training, validation, and test and the final reported accuracy is the average of all. We have trained the model with different configurations. We have trained the model up to 4 hidden layers. For each hidden layer we have used epochs ranging from 500 to 1000. For all the cases we have used sigmoidal function as the activation function. We have used a learning rate of 0.1 for all of the above cases.

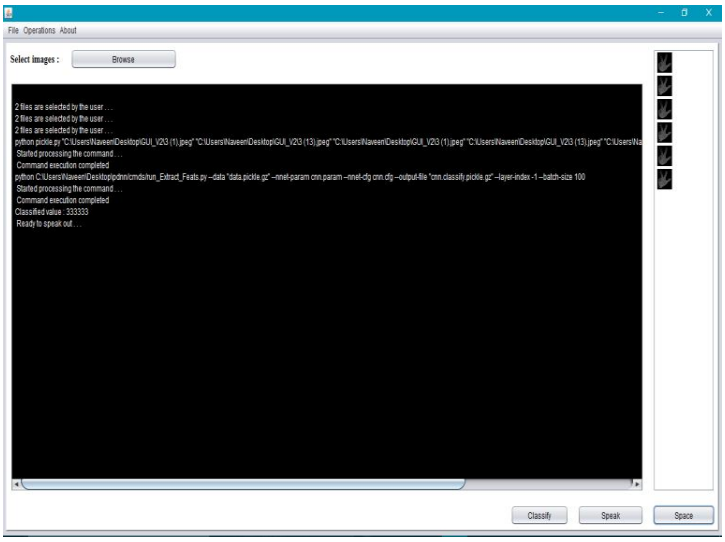

Fig. 5. GUI Interface of application

Hidden Layer\Epoch	500	600	700	800	900	1000
1	94.5161	94.3548	94.6774	94.5161	94.0322	94.0323
2	93.0645	93.3870	92.4193	92.4193	93.0645	93.0645
3	91.6129	92.9032	93.5483	93.7096	94.0322	92.2580
4	79.1935	92.7419	92.4193	94.3548	94.5161	94.5161

Fig. 6. Accuracy table

It is clear that the accuracy of the model can be improved by increasing the number of samples in the training set. In our work, GPU enabled system with Theano is used for training. Theano offers more features in complicated optimization algorithms like conjugate gradient, CNN etc. For fast calculation and implementations of mathematical operations theano includes CUDA code generators and n-dimensional (dense) arrays located in GPU memory with Python bindings.. The processing time is about 5 sec for each epoch using Nvidia GeForce GTX 970 M. Without GPU it was found that each epoch has taken about 1.30 minutes on Intel 6th generation 17 processor. So the training time was drastically reduced by using Theano with GPU. The accuracy vs epoch plot for our experiment is shown in fig. 7. From this we can see that for each hidden layer configuration the accuracy of the model increases with epoch. Then it reaches a peak value then it start decreasing as we increase the number of epochs. We have achieved a high accuracy of 94.6774% for the model with 1 hidden layer and 700 epoch. We have used up to 4 hidden layers. The hidden layer configuration used for our experiment are, for one hidden layer we have used 5x5 convolution mask with 20 local filters and a max-pooling layer of 2x2.The accuracy vs hidden layer plot is shown in fig.8. The calculated values of precision and recall of alphabets and numerals for the model with the highest accuracy is shown in fig. 9.

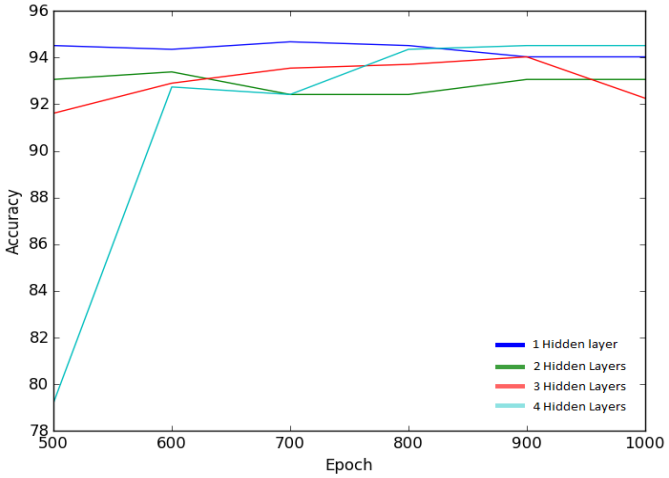

Fig. 7. Accuracy vs Epoch

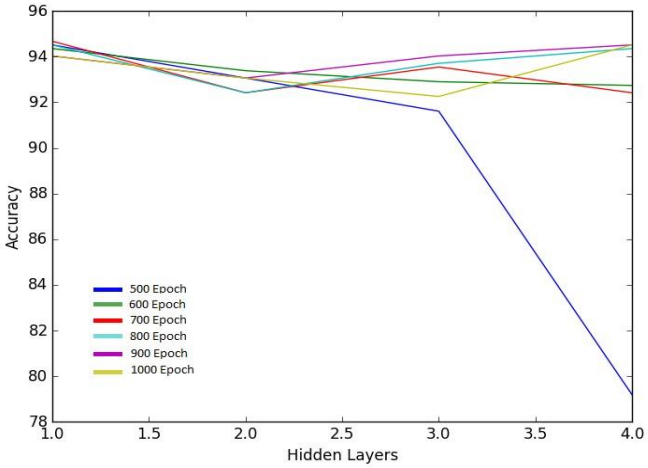

Fig. 8. Accuracy vs Epoch

V. Conclusion

Communications between deaf-mute and a normal person have always been a challenging task. The goal of this project is to reduce the barrier of communication by contributing to the field of automatic sign language recognition.

Through this work, a CNN classifier is constructed which is capable of recognizing static sign language gestures. A basic GUI application is created to test our classifier in this system.. The application allows users to select the static sign gestures as input and it will speak out the words or sentences corresponding to the gesture. We have trained our model for 33 symbols which include alphabets and numbers. We were able to achieve an accuracy of 94.6774% for our CNN classifier.

Class	Precision	Recall
A	100%	80%
B	100%	100%
C	80%	100%
D	95.2380%	100%
E	80.3333%	100%
F	80%	100%
G	100%	100%
H	100%	100%
I	100%	70%
K	100%	100%
L	100%	100%
M	73.0769%	95%
N	100%	95%
O	100%	75%
P	100%	100%
Q	100%	100%

Class	Precision	Recall
R	100%	100%
S	100%	75%
T	80%	100%
U	100%	100%
X	100%	95%
Y	86.9565%	100%
1	100%	100%
2	100%	100%
3	95.2388%	100%
4	90.9090%	100%
5	100%	100%
6	100%	100%
7	100%	100%
8	100%	100%
9	100%	50%

Fig. 9. Precision and Recall

References

[1] Kang, Byeongkeun, Subarna Tripathi, and Truong Q. Nguyen. "Real- time sign language fingerspelling recognition using convolutional neural networks from depth map." arXiv preprint arXiv: 1509.03001 (2015).

[2] Suganya, R., and T. Meeradevi. "Design of a communication aid for phys- ically challenged." In Electronics and Communication Systems (ICECS), 2015 2nd International Conference on, pp. 818-822. IEEE, 2015.

[3] Sruthi Upendran, Thamizharasi. A," American Sign Language Interpreter System for Deaf and Dumb Individuals", 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 978-1-4799-4190-2,2014 IEEE.

[4] Pramada, Sawant, Deshpande Saylee, Nale Pranita, Nerkar Samiksha, and M. S. Vaidya. "Intelligent Sign Language Recognition Using Image Processing." IOSR Journal of Engineering (IOSRJEN) 3, no. 2 (2013):45-51.

[5] Chenyang Zhang, Yingli Tian, Matt Huenerfauth," multi-modality american sign language recognition" ICIP 2016,978-1-4673-9961-6.

[6] Thad Starner & Alex Pentland ,(1995), Real-Time American Sign Language From video using Hidden markov models (IEEE), pp. 265-271.

[7] D. Kumarage, S. Fernando, P. Fernando, D. Madushanka& R. Samarasinghe,(2011) "Real time sign language recognition using still image comparison & motion recognition" sixth international conference on industrial and information systems, srilanka

[8] M.S. Sinith, Soorej G Kamal, Nisha B.,Nayana S, KiranSurendran, & Jith P S, (2012), "sign language recognition using support vector machine" international conference on advances in computing and communications,(IEEE), pp 122-129

[9] Yeo, Hui-Shyong, Byung-Gook Lee, and Hyotaek Lim."Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware." Multimedia Tools and Applications 74, no. 8 (2015):2687-2715.

[10] Garg, Pragati, Naveen Aggarwal, and Sanjeev Sofat. "Vision based hand gesture recognition." World Academy of Science, Engineering and Technology 49, no. 1 (2009): 972-977.

[11] Amruta S. Talrejaa, Darshana Tekadea, Shailesh Bharada and Lovely Mutnejab."Sign Language Recognition System by Pattern Match- ing."International Journal of Innovative and Emerging Research in En- gineering, vol. 2, Issue 2, 2015

[12] Nagarajan, S., and T. S. Subashini. "Static hand gesture recognition for sign language alphabets using edge oriented histogram and multi class SVM." International Journal of Computer Applications 82, no. 4 (2013).

[13] Pigou, Lionel, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. "Sign language recognition using convolutional neural networks." In Computer Vision-ECCV 2014 Workshops, pp. 572-578. Springer International Publishing.

[14]http://deeplearning.net/software/theano/
[15] http://note.sonots.com/SciSoftware/haartraining.html
[16] Anju .M. Nair, S. Joshua Daniel, "Design Of Wireless Sensor Networks For Pilgrims Tracking And Monitoring", International Journal of Innovations in Scientific and Engineering Research (IJISER), Vol.1, No.2, pp.82-87, 2014.