EasyChair Preprint

№ 2541

# Sign Language Translator using Deep Learning

Uday Patil, Saraswati Nagtilak, Sunil Rai, Swaroop Patwari, Prajay Agarwal and Rahul Sharma



February 4, 2020

# Sign Language Translator using Deep Learning

Uday Patil
*Dept of Information Technology,Smt Kashibai Navale College of Engineering . Savitribai Phule Pune University*
Pune, India
upatil395@gmail.com

Saraswati Nagtilak
*Prof at Information Technology,Smt Kashibai Navale College of Engineering . Savitribai Phule Pune University*
Pune, India
sanagtilak@sinhagad.edu

Dr. Sunil Rai
*Prof. at MITSOE, MITADT University*
Pune, India
drsunilrai2017@gmail.com

Swaroop Patwari
*Dept of Information Technology,Smt Kashibai Navale College of Engineering . Savitribai Phule Pune University*
Pune, India
swaroop07patwari@gmail.com

Prajay Agarwal
*Dept of Information Technology,Smt Kashibai Navale College of Engineering . Savitribai Phule Pune University*
Pune, India
prajayagarwal24@gmail.com

Rahul Sharma
*Dept of Information Technology,Smt Kashibai Navale College of Engineering . Savitribai Phule Pune University*
Pune, India
rahul.n.sharma511@gmail.com

***Abstract*** **- People affected by speech impairment can't communicate using hearing and speech, they rely on sign language for communication. Sign language is used among everybody who is speech impaired, but they find a hard time communicating with people which are non-signers (people aren't proficient in sign language). So requirement of a sign language interpreter is a must for speech impaired people. This makes their informal and formal communication difficult. There has been favorable progress in the field of gesture recognition and motion recognition with current advancements in deep learning. There is also been quite a significant advancement in computer vision which would enable us to easily track the hand gestures. The proposed system tries to do a real time translation of hand gestures into equivalent English text. This system takes hand gestures as input through video and translates it text which could be understood by a non-signer. There will be use of CNN for classification of hand gestures. By deploying this system, the communication gap between signers and non-signers. This will make communication speech impaired people less cumbersome.**

**Keywords - CNN (Convolutional Neural Network), gesture recognition, motion recognition, deep learning.**

## I. INTRODUCTION

The speech impaired people around the world communicate with others using the sign language. It is essential that the people that they are communicating with know sign language. But more often then not the other person doesn't understand sign language. Then the speech impaired people are dependent on sign language interpreter who would then translate the sign language to the people who don't understand sing language. The speech impaired people face huge problem conversing with the non-signers (people who don't understand sign language). This is pandemic problem not just around the world but in India also.

There is a standardized sign language namely Indo-Pakistani Sign Language which is practiced in India. These speech or hearing impaired people face a lot of difficult in their day to day life. Speech impaired people heavily rely on speech interpreters for medical, legal, educational and training sessions.

When the non-signers has to visit to a clinic for treatment of diseases he can't communicate what has happened if the doctor diagnosing him isn't proficient in sign language. Then the disease of the speech impaired may escalate if the symptoms aren't been conveyed properly to the doctor and he can't prescribe proper medicine. The presence of interpreter is imperative for diagnosing the disease for conveying the symptoms. But there are not much interpreter available.

The similar problem is being faced by signers in during educational purposes if the speech impaired hasn't understood a topic and has a doubt then he can't ask the doubts if the faculty doesn't understand sign language. So speech impaired people lots of problem in their day to day life if their interpreters isn't there.

There isn't any infrastructure available for speech impaired people to communicate with non-signers without the interpreter. There is not a pathway created for automation of sign language translation. So that's why there is need of automation of sign language translation so which would result in convenient communication between speech impaired people and a non-signer without the need of an interpreter for translation. We wish to automate the process of translation of sign language into a form that can be understood by the non-signer, for this we will require other subjects such as Image Segmentation, Object Tracking. In the area of Computer Vision and Pattern Recognition, image segmentation plays an important role as a preliminary step for high level image processing. Segmentation subdivides an image into regions or objects ,one needs to isolate the regions and find relation among them. The process of separation of such objects is referred as image segmentation [1]. Object tracking is an important task within the field of Computer Vision, video analysis has generated a great deal of interest in object tracking algorithm. There are three key steps in video analysis: detection of moving object, tracking of such moving object and analysis of object tracks to recognize their behavior. Thus use of object tracking is pertinent in our model [2].

## II. MOTIVATION

According to a survey conducted by Ministry of Statistics and Programme Implementation in India there are approximately 4.2 million people suffer from either speech or hearing impairments. Speech impaired people heavily rely on speech interpreters for medical, legal, educational and training sessions. However they face problems when speech interpreters are unavailable. The focus of this project is to automate interpretation using deep learning.

## III. EXISTING SYSTEM

The existing system for communication between signer and non-signer takes place through an interpreter. The interpreter is well versed with Indo-Pakistani Sign Language. When the signer wants to convey a message to the non-signer he/she makes the appropriate sign language hand gestures to the interpreter and then the interpreter translates the hand gestures made by the signer into equivalent English.

TABLE 1: LITERATURE SURVEY

| Sr. No. | Literature Survey | | |
|---|---|---|---|
| | ***Title of the paper*** | ***Core Idea*** | ***Merits*** |
| 1 | An Adaptive Color Image Segmentation (Electronic letters on Computer Vision and Image Analysis-2005) | Feature Extraction using various Image Segmentation Techniques and a Neural Network | Doesn't require a priori information about the number of objects in the image. |
| 2 | Object Tracking : A Survey (ACM Computing Survey-2006) | Object Tracking Algorithms discussion on the basis of performance | Qualitative comparisons of the tracking algorithms. |
| 3 | Very Deep Convolutional Networks for Large-Scale Image Recognition (ICLR-2015) | Evaluation of Convolutional networks depth | Improvement on the prior-art configurations. |
| 4 | Deep learning (Nature-2015) | Using chain rule to train neural net. | Discover intricate structure in large data sets. |
| 5 | Hand Gesture Recognition with 3D Convolutional Neural Networks (IEEE-2015) | Spatio-temporal data augmentation method | Robustness ,Correct Classification rate of 77.5%. |
| 6 | Introduction to multi-layer feed-forward neural networks (Chemometrics and Intelligent Laboratory Systems-2016) | Multi-Layer feed forward neural network | Learning, Robustness, and Non-Linearity |
| 7 | Hand Gesture Recognition Using Deep Learning (IEEE-2017) | Hand Gesture Recognition using a CNN model | Works for both static and dynamic Hand gestures. |
| 8 | American Sign Language Recognition using computer vision (IEEE-2018) | Deep Learning approach | Architecture for robust and accurate model for Hand gesture recognition |

Similarly, the non-signer communicates to signer through the interpreter. First the non-signer tells the interpreter what he/she wants to communicate to the signer then the interpreter through the hand gestures tell the signer. This process is quite tedious and time consuming and the signer may suffer if there is an absence of an interpreter. The existing system requires automation which would help the signers in communicating with the non-signers in absence of an interpreter. The standard sign language Indo-Pakistani can be automated with the help of computer vision, deep learning models which will classify the hand gestures which would aid in the reducing the communication gap between the signers and non-signers.

## IV. PROPOSED METHODOLOGY

As discussed earlier the core of the project is the construction, training and evaluation of a model that can classify signs of a signer with acceptable accuracy. With the advances in computer hardware and processing power of it is now feasible for us to train a neural network, over a gigantic dataset that can be used to accurately classify signs. Our system involves Gathering data, training Deep Learning model to recognize ISL alphabets and building user interface. Initial processing shall be done using OpenCV [3] a popular library for computer vision developed by Intel. OpenCV [3] enables conversion of video to frames. It also allows to downscale image and make it grayscale so that skin tone should have no impact on final result as given by a neural network [4].

For the hand detection we use YOLO [5] an object detection algorithm developed by Joseph Redmond to locate gesture being performed in the video frame. This step is essential as there may be noise i.e. unwanted spatial features which are irrelevant for the hand gesture classification. YOLO [5] is used as it is extremely fast as it is written in C which is much closed to machine code. In the best condition it can detect objects in up to 45 frames a second. YOLO [5]is used to define area of interest.

For gesture classification i.e. to identify which gesture has been made by the signers, we use CNN. Out of many CNN architectures VGGNet [6] shall be used since it has state of the art 7.3% error rate. Using CNN, ISL alphabets shall be continuously classified until a word is formed. These words when combined forms sentence for communication.

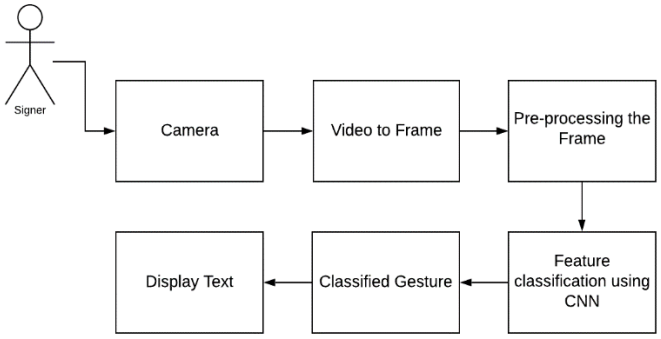


Fig 3.1. Proposed System

Hand Gestures are given as input in to the system in the video format. Then the video would be converted into individual frames. After that pre-processing would be carried out on the frames. In pre-processing the frames are converted from RGB to Greyscale and the noise in the image is reduced and the size is reduced. The size is reduced that the computational power required will be less. Then those frames would be fed into the trained neural network. At last the neural network would classify the hand gesture in the frame into equivalent text.

During training, the input to our ConvNets is a fixed-size of 224 × 224 RGB image. The only preprocessing we do is subtracting the mean RGB value, computed on the training set, from each pixel. The image is passed through a stack of convolutional (conv.) layers, where we use filters with a very small receptive field: 3 × 3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations we also utilize 1 × 1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution[7], i.e. the padding is 1 pixel for 3 × 3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers (not all the conv. layers are followed by max-pooling). Max-pooling is performed over a 2 × 2 pixel window, with stride 2.

Some of the limitations faced by our model are firstly, with facial features and skin tones. While testing with different skin tones, the model dropped accuracy if it hadn't been trained on a certain skin tone and was made to predict on it. Secondly, the model also suffered from loss of accuracy with the inclusion of faces, as faces of signers vary, the model ends up training incorrect features from the videos. So the videos had to be trimmed to only include gestures which were only extended up till the neck.

Thirdly, the model also performed poorly when there was variation in clothing. Maybe using a ROI to isolate hand gestures from the images would help accuracy, but for the context of this paper, a consistent full-sleeved shirt was used in all the gesture recordings.

## V. Dataset Formation

The Dataset available is of American Sign Language, but the model we are proposing will classify Indian Sign Language and the dataset for Indian Sign Language is not available. The process of creating the dataset by capturing the images is a huge task.

So, we will create our own dataset. The images that we are going to capture for the dataset, then object tracking will be applied. The model requires changes in the background of the images so that accuracy is maintained. For this to achieve we need to capture a large number of images having different backgrounds. The model suffers in accuracy when the complexion skin tone changes. So, the dataset will have images with various backgrounds.

Also the dataset will comprise of 26 classes. Each class belonging to each alphabet in the English language. Every class in the dataset have distinct hand gesture so that the model can easily classify hand gestures. Every class has no upper bound for the number of images of hand gestures. We are keeping a minimum of 2000 plus images.

In our model if there is slight variation in the hand gestures made then the accuracy differs accordingly. So, to cope up with this unstable accuracy problem, we are inserting the images in the dataset with the movement in the hand gestures.


Fig 5.1. Hand gestures for alphabets A, B & C respectively

Here above images showcase the hand gestures for Indo-Paskistani sign languages for alphabets A, B and C. Large amount of images will maintained in the database for each alphabet to enhance the accuracy. Dataset is the main aspect in our deep learning model's learning phase as it is the base from which it learns, understands and classify different alphabets hand gestures.

## VI. Conclusion

The rudimentary idea of communication of speech impaired through sign language is discussed and the problem they face when there is absence of interpreter. So this project aims to implement a sign language translator which will facilitate them to communicate with non-signers without any complications. They can also communicate in the absence of a sign language interpreter. For sign language translator concepts of YOLO for object detection and convolutional neural network for accurately and precisely classifying the hand gestures. Sign language translator is implemented using python 3.7. OpenCV used for video to frame conversion and down scaling the frames. The proposed system provides real time and accurate translation of hand gestures.

## VII. Potential Improvements

First potential improvement is that we can make words by alphabet gesture classification by adding a gesture for word break. This can work by giving a definite gesture for it then making that gesture when a signer wants to end a word. So then the signer by making gestures for alphabets while using word break hand gesture to make words and then sentences. This would be quite beneficial for the signer to communicate with the non-signers.

Second potential improvement would be to classify hand gestures for words. Which is a daunting task as it requires the use of RNN (Recurrent Neural Network) [8]. The RNN is similar to other neural network but with a major advantage

which is memory. So it feeds the outputs a result back into itself. The classification of words requires movement of hand gestures which could be classified by RNN. RNN requires lot of processing power and large storage as it stores the output for the next epoch.

One of the potential improvements would be to experiment with different RNN architectures for the output of the pool layer. Including GRU [8] and Independent RNN's. In terms of CNN improvements, using Capsule Networks instead of Inception may yield better results than Inception.

## REFERENCES.


[1] Karen Simonyan & Andrew Zisserman " Very Deep Convolutional Networks for Large-Scale Image Recogntion" ICLR, pp 730-734, 2015

[2] LeCun, Y., Bengio, Y., & Hinton, G." Deep learning." vol 521 Nature, 521 pp 436-444 May 2015.

[3] Daniel Svozil, Vladimir KvasniEka, JiE Pospichal "Introduction to multi-layer feed-forward neural networks", Chemometrics and Intelligent Laboratory Systems 39 pp 43-62 1997.K. Elissa, "Title of paper if known," unpublished.

[4] Alper Yilmaz, Omar Javed, Mubarak Shah, "Object Tracking : A Survey", ACM Computing Survey, Vol 38 No.4, Article 13, December 2006.

[5] Deshmukh K.S., Shinde G. N, "An Adaptive Color Image Segmentation", Electronic letters on Computer Vision and Image Analysis 5(4) : pp 12-23, December 2005.

[6] Soeb Hussain, Rupal Saxena, Xie Han, Jameel Ahmed Khan, Prof. Hyunchal Shin, "Hand Gesture Recognition Using Deep Learning", IEEE International Conference for Design, pp 48-49, 2017.

[7] Humphries, T., & Padden, C.." Learning American sign language.",Englewood Cliffs, N.J: Prentice Hall, pp 35-48, 1992.

[8] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz ,"Hand Gesture Recognition with 3D Convolutional Neural Networks", IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 1-7, 2015.

[9] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov. " Improving neural networks by preventing co-adaptation of feature detectors", arXiv:1207.0580v1 [cs.NE], pp 56-73, July 2012.

[10] Hsien-I Lin, Ming-Hsiang Hsu, and Wei-Kai Chen ,"Human Hand Gesture Recognition Using a Convolution Neural Network", IEEE International Conference on Automation Science and Engineering (CASE) Taipei, Taiwan, pp 1038-1043, 2014.