International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Sign Language Text to Speech Converter using Image Processing and CNN

Mangesh B.[1], Mayur K.[2], Rujali P.[3]

[1]*Student, Dept. of Information Technology, Vidyalankar Institute of Technology, Mumbai, Maharashtra, India*
[2]*Student, Dept. of Information Technology, Vidyalankar Institute of Technology, Mumbai, Maharashtra, India*
[3]*Student, Dept. of Information Technology, Vidyalankar Institute of Technology, Mumbai, Maharashtra, India*

Abstract - *The main form of communication for people who are deaf or hard of hearing is considered to be Sign language. Different innovations in automatic sign recognition try to break down this communication barrier. Our contribution considers a recognition system using the concept of Image processing and Convolution Neural network. As we have used image processing and CNN in our application compared to existing applications we have improved the Integrity and Flexibility of application. Such system will enable common people to understand the sign language with the hearing impaired.*

Key Words: Image Processing, Gesture recognition, Convolution Neural Network, Train and Testing of Gestures, Thresholding

1. INTRODUCTION

In the 21st century field of science and technology, one has reached such a level that people are expecting more comfortable and useful things, which can make their lives easier. Nowadays, homes with voice recognition built in with the sense of gestures have already been conceived. There are video games in the market which can be played with real time gestures and all this has been possible with the advent of the new technology. Even our mobiles have been loaded with all similar technologies. Nevertheless, there are people who are less fortunate than us and are physically challenged, may it be deafness or being aphonic.

Figure 1.1 - Hand Signs.

These people have some expectations from the researchers and mostly from a computer scientist that computer scientists can provide some machine or a model which can help them to communicate and express their feelings with others. The deaf and the mute can only perceive visual, therefore communication is done by visual and sound. This, sign language is a medium for communication between the deaf and the mute. Sign Language. Recognition (SLR) is a tool that executes the conversion of sign language into text and further into speech. Research in SLR started two decades before, all over the world especially in American Sign Languages. Based on statistical world analysis, over 5% of the world's population – 360 million people have hearing disability. The biggest drawback for the deaf is their employment issues. Communication has always played a very vital role in getting the task solved so therefore, the issue. In order to help the deaf people communicate with the ordinary people, we build a system to translate sign language into text and further into speech. This concept proposed a system that can automatically detect hand signs of alphabets in American Sign Language (ASL) that is all the English alphabets. To create spaces between the letters we create extra two signs, i.e., SPACE and OK. The SPACE sign is used to create the space between different recognized words, and the OK sign is used to stop capturing and start to execute a current required function. This system is based on American Sign Language (ASL), which is considered to be a complete language. The main focus of this project is for helping the deaf and the mute by converting hand gestures to speech. Finger sign is a subset of sign language, and uses finger signs to spell words of the spoken or written language. The finger sign recognition task involves the segmentation of finger sign hand gestures from image sequences. ASL (American Sign Language) is the fourth most commonly used language in the USA and is extensively used by deaf people and this language is officially acquired by the deaf society of United States. It is a unique language that highlights signs made by moving the hands. ASL is not defined as the world language but it has its roots in English speaking parts in Canada, few regions of Mexico, and all over United States of America. Human beings are gifted with a voice that allows them to communicate with each other. Therefore, spoken language becomes one of the main key points of humans. Unfortunately, not everybody has this capability because of one sense, i.e., hearing. In India, there are around 5 to 15 million deaf people approx. Sign language is considered to be the basic alternative communication method between the deaf people and several dictionaries of words or single letters have been defined to make this communication strong and effective. Without an interpreter it gets difficult for such a communication to take place. Therefore, a system that converts symbols in sign languages into plain text and further into speech can help with real-time communication.

2. RELATED WORK

Yi Li(2012), This system consist of three components-Hand detection, Finger Identification and Gesture recognition. This system is built on Candescent NUI project, which is freely available online. Open NI framework was used to extract the depth data from the 3D sensor[1]. Zhou Ren, Jingjing Meng, JunsongYaun(2011),The depth sensors like the Xtion Pro Live sensor, have given rise to new opportunities for human -computer interaction (HCI). There is a great progress that has been made by using Xtion Pro Live sensor in human body tracking and body gestures recognition, robust hand gesture recognition which still remains a problem. Compared to the human body, the hand is smaller object, and has more complex articulations. Thus a hand is easily affected by segmentation errors as compared to entire human body[2]. Nobuhiko Tanibata, Nobutaka Shimada, Yoshiaki Shirai(2002), Obtain hand features from sequence of images. This is done by segmenting and tracking the face and hands skin colour. The tracking of elbow is done by matching the template of an elbow shape. The hand features like the area of hand, direction of hand motion etc. therefore are extracted and are then input to Hidden Markov Model (HMM) [3]. Spencer D Kelly, Sarah M Manning, Sabrina Rodak(2008), Recognise hand postures used in various sign languages using novel hand posture feature, Eigen-space Size function and Support Vector Machine(SVM) based gesture recognition framework. They used a combination of Hu moments and Eigen-space size function to classify different hand postures [4].

3. PROPOSED ARCHITECTURE AND METHODOLOGY

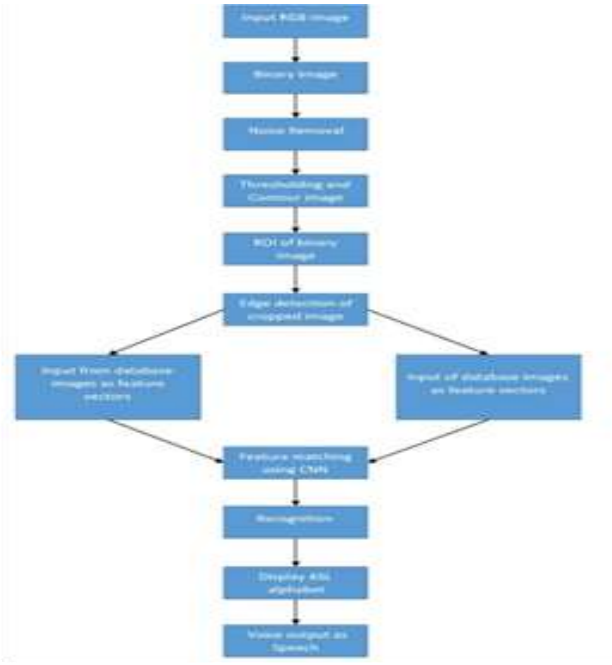


Figure 3.1 – System Architecture

- First we create and store the hand gestures of signs with the help of image processing techniques, i.e., converting RBG(colour) image to grayscale.
- And then again converting that grayscale image to binary image using threshold.
- Then we smooth the image using Gaussian and Median blur technique and to recognize the edges of hand gestures we use contours.
- Then we store these gestures in database and after that we use CNN algorithm on these stored images of hand gestures using Tensorflow and Keras.
- Tensorflow and Keras is used to test and train the system, which then recognizes the gestures which are stored in the database and gives appropriate results when the user runs the application.

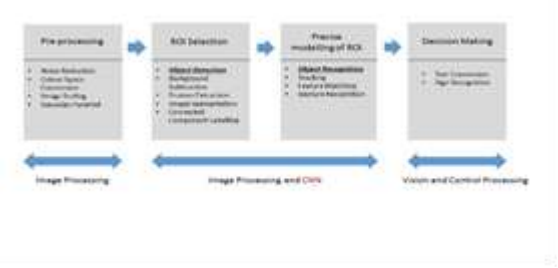


Figure 3.2 – Application Flow [6]

3.1. Image processing:

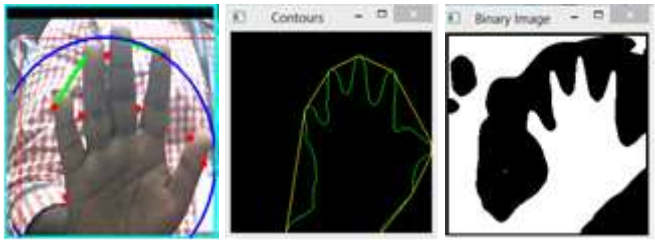


Figure 3.1.1 – Image Processing of hand Gestures

In this application for image processing of hand gestures we have used OpenCV libraries. In image processing of hand gestures, first we have converted the captured hand gesture from RGB to HSV, i.e., in Hue Saturation and Value. After converting the image into HSV we have used Gaussian blur and median blur to smooth the image. Then we do thresholding of an image to split an image into smaller segments or junks using gray scale value to define their boundary. It also reduces the complexity of the data and simplifies the process of recognition and classification. After thresholding to determine the shape of a hand gesture, contour is used as contour are useful tools for shape analysis, object detection and recognition.

3.2. Convolution Neural Network Algorithm:

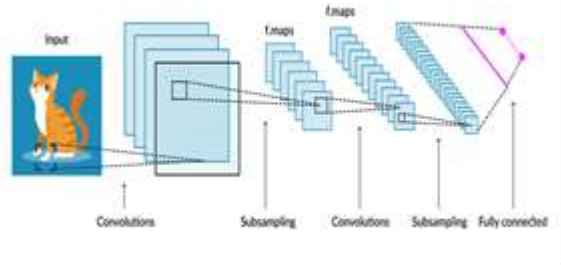


Figure 3.2.1 – CNN block Diagram [5]

For hand gesture recognition and classification, we have used CNN architecture. Convolutional Neural Networks are implemented successfully for human gesture recognition in recent times and also in image recognition and classification. Research says there has been work done in the field of sign language recognition with the help of deep language of CNN's, with input recognition that is sensitive to more than just pixels of the images. The camera makes the process much easier to develop characteristic depth and motion profiles for each sign language gesture so therefore, it senses depth and contour. The advantage of CNN is its abilities to learn features as well as the weight corresponding to each feature. CNNs seek to optimize some objective function, especially the loss function. We utilized the softmax-based loss function:

$$Loss = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{f_{i,y_i}}}{\sum_{j=1}^{C} e^{f_{i,j}}}\right) \quad (1)$$

$$f_j(z) = \frac{e^{z_j}}{\sum_{k=1}^{C} e^{z_k}} \quad (2)$$

N = total number of training examples
C = total number of classes

Equation (2) is the softmax function. It takes the feature vector z for the given training example, and squashes its value to a vector of [0, 1] valued real numbers summing to 1 [7]. Equation (1) takes the mean loss for each training [7].

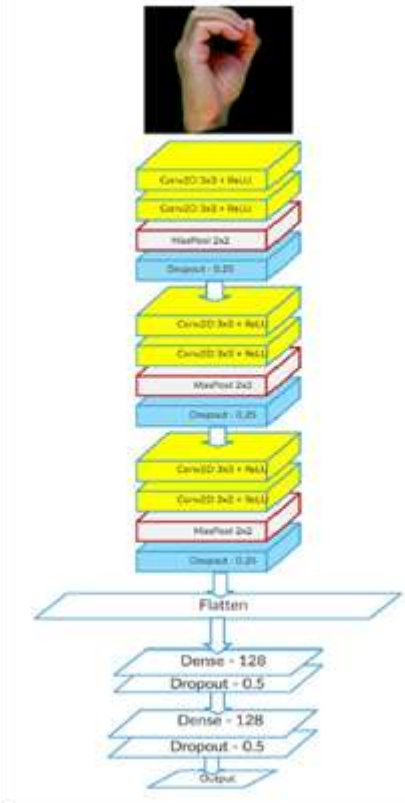
3.3. Dataset:

We basically supervised learning, and learning by training the network with our own set of sign dataset. We have classified letters and digits, i.e. A-Z and 0-9. We initially trained and tested on self-generated dataset of images we took ourselves. The images were trained and tested with the help of Google Collab. This dataset consists of 1200 images of each alphabet and the digits 0-9. Additionally pipeline was created, so that people are able to generate and continue to add the gesture in dataset.

4. IMPLEMENTATION DETAILS

In these section, the implementation of Sign language text to speech converter using image processing and OpenCv has been described. In section A of this topic, the actual methodology used has been described with respect to all the modules. Section B are the snapshots of the application with their description showing the implemented application in detail.

A) Methodology Used



B) Figure 4.1 – CNN architecture [9]

CNN contains four types of layers: convolution layers, pooling/subsampling layers, nonlinear layers, and fully connected layers. It captures various image features and complex non - linear features and interactions. The softmax layer is used to recognize the hand gesture/signs. We have used common CNN architecture, consisting of multiple convolutional and dense layers. The architecture consists of 3 groups of 2 convolutional layers followed by max pool layer and dropout layer, and 2 groups of fully connected layer followed by dropout layer and one final output layer.

Convolution Layer:

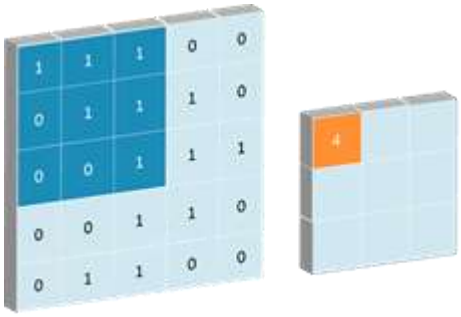


Figure 4.2 – Conv2D layer [5]

Briefly, some background a convolution layer scans a source image with a filter of, for instance, 5×5 pixels, to extract features which can be important for classification. This filter is additionally called the convolution kernel. The kernel also contains weights, which are tuned within the training of the model to realize the foremost accurate predictions. During a 5×5 kernel, for every 5×5 pixel region, the model computes the dot products between the image pixel values and therefore the weights defined within the filter. A 2D convolution layer means the input of the convolution operation is three dimensional, as an example, a colour image which features a worth for each pixel across three layers: red, blue and green. However, it's called a 2D convolution because the movement of the filter across the image happens in two dimensions. The filter is meet the image 3 times, once for every of the three layers. After the convolution ends, the features are down sampled, then an equivalent convolutional structure repeats again. At first, the convolution identifies features within the original image, then it identifies sub-features within smaller parts of the image. Eventually, this process is supposed to spot the essential features which will help classify the image.

Pooling/Subsampling:

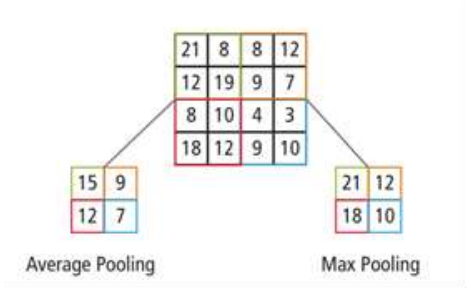


Figure 4.3 – Pooling layer [6]

Pooling layer is another building block of CNN. It reduces the spatial size of representation to decrease the quantity of parameters and computation within the network. Pooling layer operates on each feature map independently.

Fully Connected Layer:

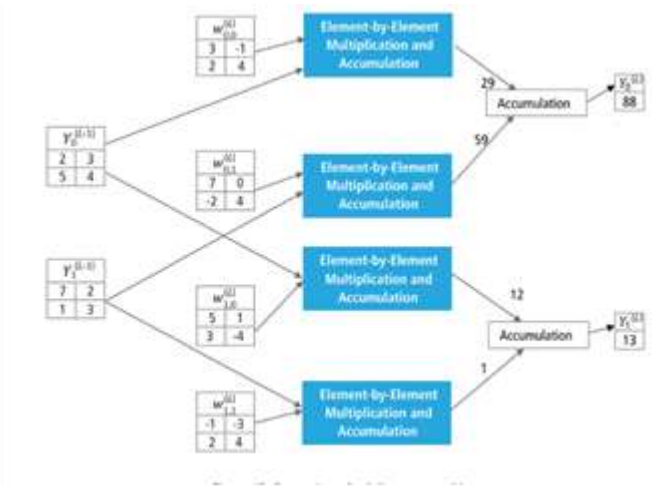


Figure 4.4 – Fully connected layer [6]

The objective of fully connected layer is to need the results of pooling process and use them to classify the image into label. The output of pooling is flattened into one vector of values where each value is representing a probability that a specific feature is belonging to a label. For instance, if the image is of a dog, features representing things like whiskers or fur should have high probabilities for the label "dog". The fully connected a part of the CNN network goes through its own backpropagation process to work out the foremost accurate weights. Each neuron receives weights that prioritize the foremost appropriate label. Finally, the neurons "vote" on each of the labels, and thus the winner of that vote is that the classification decision.

5. RESULTS AND DISCUSSIONS

Figure 5.1 – Setting Hand Threshold

In gesture creation, we have to set the hand coordinates as shown in fig. While setting the coordinates of hand gesture, it is converted to binary image for minimizing the background disturbance.

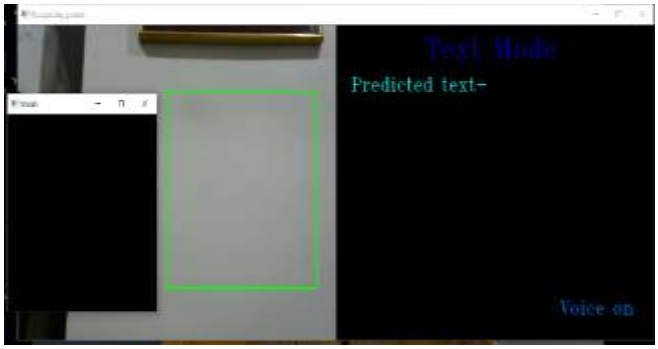


Figure 5.2 – Final layout of Application

In the above fig. the final layout of application is shown. The layout of application consist of Threshold window and Gesture recognition window.

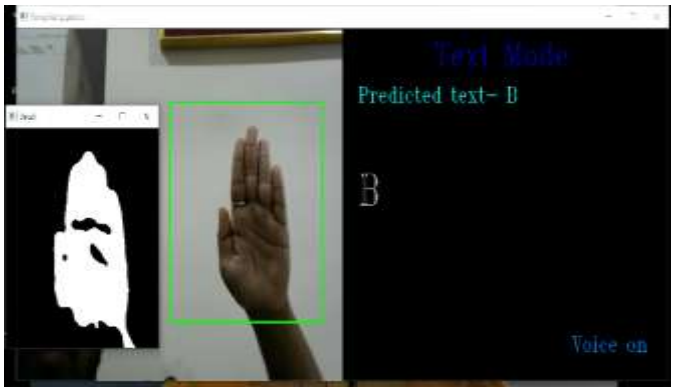


Figure 5.3 – Gesture sign alphabet

In above fig. we can see the binary image and gesture of "B" Alphabet and its result in Text Mode.

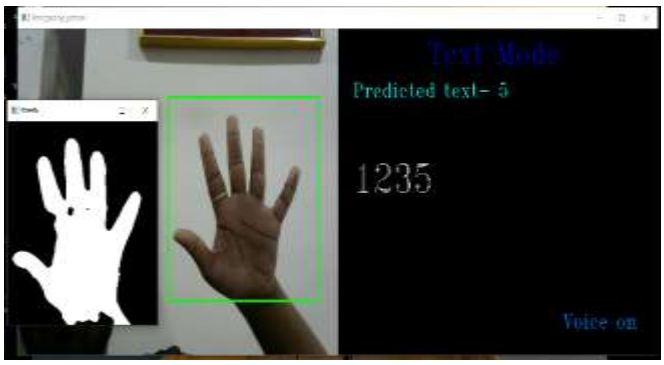


Figure 5.4 – Gesture sign alphabet

In above fig. we can see the threshold and gesture of "5" number. In Text mode we can see the result of 5 and previous hand gestures.

As we have trained and tested our own dataset, the accuracy rate of an application is approximately 85%. And as we can see it's a research based project the purpose of this project and also the accuracy rate will be fulfilled satisfactorily in the coming updates.

6. CONCLUSIONS AND FUTURE SCOPE

The system will provide an interface which will easily communicate with deaf people by signing recognition. The system is not applied only in family environment, but also can apply in public. For the social use, this technique is extremely helpful for deaf and dumb people. We will build a simple gesture recognizer based on OpenCV toolkit and integrated it into Visionary framework. As a yes gesture we'll price and down hand motions regardless of which hand is employed.

The project focuses on distinguishing among various different alphabets of English language. Future work may include recognition of all the English alphabets and numbers. Furthermore, we may move on to recognition of words, from as large as a dictionary as possible.

REFERENCES

[1] Yi Li (2012),"Hand Gesture Recognition Using Kinect".

[2] Zhou Ren, JingjingMeng, JunsongYaun(2011),"Robust Hand Gesture Recognition with Kinect Sensor"

[3]]Nobuhiko Tanibata, Nobutaka Shimada, Yoshiaki Shirai(2002),"Extraction of Hand Features.

[4] Ferdousi, Z., "Design and Development of a Real Time Gesture Recognition System", U.M.I. Publishers, June 2008.

[5] https://missinglink.ai/guides/keras/keras-conv2d-working-cnn-2d-convolutions-keras

[6] Using Convolutional Neural Networks Image Recognition by Samer Hijazi, Rishi Kumar, and Chris Rowen, IP Group, Cadence.

[7] Real-time American Sign Language Recognition with Convolutional Neural Networks Brandon Garcia Stanford University Stanford, CAbgarcia7@stanford.edu Sigberto Alarcon Viesca Stanford University Stanford, CA.

[8] Subha Rajam, P. and Balakrishnan, G.(2011), "Real Time Sign Language Recognition System to aid Deaf-dumb People", IEEE, pg: 737- 742,2011.

[9] Sharmila Gaekwad, Akanksha Shetty, Akshaya Satam, Mihir Rathod, Pooja Shah(2019), "Recognition of American Sign Language using Image processing and Machine Learning", IJCSMC, pg: 352-357,2019.