International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 3 ǁ 2022 ǁ PP. 46-52

SAMBHASHAN – A Sign Language Recognizer and Convertor

Neha A. Dudhane
Bachelor of Engineering – Final Year, Department of Computer Technology
K. D. K. College of Engineering, Nagpur, India

Tejal M. Nirne
Bachelor of Engineering – Final Year, Department of Computer Technology
K. D. K. College of Engineering, Nagpur, India

Khushal U. Demeti
Bachelor of Engineering – Final Year,Department of Computer Technology
K. D. K. College of Engineering, Nagpur, India

Sakshi S. Patkotwar
Bachelor of Engineering – Final Year, Department of Computer Technology
K. D. K. College of Engineering, Nagpur, India

Abhishek P. Nachankar
Professor at Department of Computer Technology, K. D. K. College of Engineering, Nagpur, India.

Abstract— *Sign language is one of the natural forms of language for communication between those people who have speaking and hearing disabilities (Deaf & Dumb). And to communicate with them one should know the sign language as it is the only form of communication they have. Although being one of the oldest languages, still most of the people don't know Sign Language and it's really hard to find affordable interpreters. That's how we came up with an idea to develop a system that can not only recognize the Sign Language but also convert it into human-readable text using Convolutional Neural Network (CNN). In our system once the hand gesture is captures it is filtered and then passed through layers of models to predict its text equivalent. The system will be capable of converting 26 alphabets of American Sign Language (ASL).*

Keywords— *Sign Language, Deaf & Dumb, Convolutional Neural Network (CNN), American Sign Language (ASL).*

Date of Submission: 28-02-2022 Date of acceptance: 09-03-2022

I. INTRODUCTION

According to a survey conducted by World Health Organization (WHO), over 1 million people are dumb and 300 million people are deaf in the world. Deaf & dumb are people with hearing and/or speech disabilities. Hence, these people use a form of visual language for communication and this language is what we call Sign Language. But there are comparatively a lot of people who don't understand sign language and don't take efforts to learn too. This has become the reason for isolation of deaf & dumb people from society and has put a restriction on the communication of deaf & dumb people with the others.

The digital world outside is expanding rapidly to reach the Moon and capture the Sun. But this deaf & dumb community is lacking behind. They don't have much options for the basics of survival i.e. communication. That is why we came up with an idea of developing a user friendly system that will fill the gap of communication between the society and Deaf & Dumb community as it will be a Human Computer Interface (HCI) that will convert Sign Language gestures into its equivalent text. Sambhashan will allow deaf & dumb people to communicate even with those who don't understand sign language.

Just like spoken language, different region uses different sign languages. However, sign languages does not follow the spoken language of corresponding region. The most basic of any sign language is how alphabets are spelled, that is how they are differentiated from one another. Several recognized sign languages are

American Sign Language (ASL), Arabic Sign Language, British, Australian and New Zealand Sign Language (BANZSL), Chinese Sign Language (CSL), French Sign Language (LSF), Japanese Sign Language (JSL) Syllabary, Mexican Sign Language (LSM), Spanish Sign Language (LSE), etc. However most of the alphabets from American Sign Language (ASL) vocabulary are used mostly worldwide. Thus we decided to use American Sign Language (ASL) for Sambhashan.

American Sign Language (ASL) has following main components namely fingerspelling/ hand gestures, palm orientation, movements, location and non-manual signals (NMS). Each of these components plays an important role in defining uniqueness of every sign in sign language. Fingerspelling or hand gestures are the sign make with the use of fingers. Fingerspelling spells words letter by letter. Just like spoken language sign language has its vocabulary. This vocabulary has alphabets from A to Z and numbers from 0 to 9 and several other symbols. Fingerspelling is how one can spell this vocabulary.

Palm orientation is the exact direction your hand is facing during fingerspelling. The orientations can be Palm facing out, Palm facing in, Palm is horizontal, and Palm faces left/right, Palm toward palm, Palm up/down. Several movements such as in a circle, Up and down, Forward, Backward, Tapping, Back and forth and Wiggle can be displayed during signing a particular sign. The location refers to the location where sign is made corresponding to the body. The initial location of a sing can be different from where the sign ends. These locations can be Chin, Shoulder, Front of body, Front left/right of body and Forehead. Lastly, the non-manual signals are the facial expressions or body movements made additional to the signs in order to convey several signs more accurately. Not all the signs require non-manual expressions but are commonly used during communication by the non-deaf & non-dumb society.

All the above components of Sign language together define uniqueness of a sign. Sambhashan has used combination of the fingerspelling/ hand gestures and palm orientation of American Sign Language (ASL) which is sign language developed in the United States (US). Sambhashan will be capturing these hand gestures with the help of video camera of the device. These images then will go through several filters to ensure sharp borders of the hand gestures. Sambhashan is trained with this dataset so that it can determine to which model it belongs to. Sambhashan has created two layers of models such as one is having all the alphabets and the other is having groups of the alphabets having confusing signs and then only predict the output thus ensuring accuracy of the system.

II. EXISTING SYSTEM

A tremendous amount of research and work is done throughout the years on Sign Language Recognition and converting it into human readable form i.e. text. There are several other existing systems that can successfully recognize signs and convert them into its equivalent text. After doing our literature survey, we found out that the recognizing hand gesture is the most important task in sign language recognition as it is one of its most useful applications.

The sign language recognition system can be categorized as follows based on its implementation approach: Glove based system and Vision based system. This categorization of the sign language recognition system is done on the basis of how the input data is collected for the system. As the name itself suggests the glove based systems collect the input data with sensors located on gloves whereas the vision based systems collect their input data using computer vision devices such as camera sensors.

A. Glove based system

Over the past few years, scientists and researchers have come up with several amazing techniques and inventions for sign language and/or hand gesture recognition. Glove based system is one of those inventions where a hand glove with five sensors detecting movement of every finger can recognize the hand gesture or sign made by that hand. During the survey it was found that the researchers at a university named Comenius University Bratislava situated in Slovakia in Central Europe discovered a system named WaveGlove, a glove based system that can recognize hand gestures.

This totally new system was introduced by a paper pre-published on platform named arXiv, which stated that this system is a glove based hand gesture recognition system with an excellent capability of handling several sensors, five to be precise at the same time. These sensors are situated on every finger of the hand to capture and monitor movement of every finger. Use of multiple sensors rather than a single sensor allowed the system to create a richer dataset for the training purpose of the system ensuring higher accuracy rates. Matej Kralik, one of the researchers of the system stated to a platform TechXplore that they created two datasets containing over 11000 hand gesture samples, First dataset containing the 8 gestures of the whole hand movements and the second dataset containing 10 complex gestures of movements of every finger of the hand moving differently that the other, thus creating two whole different vocabularies of fingerspelling/ hand gestures.

But one of the highlighted drawbacks of the system is that the person needs to wear these gloves during the conversation every time. The cost efficiency of the system is also considered to be low, which cannot be ignored.

Fig.1. WaveGlove Prototype

B. Vision based system

The computer vision based systems are the most reliable and comparatively less expensive ones. These systems are basically the ones that work on the computer vision. Computer vision is the devices like camera sensors. The hand gestures are captured with such 2D or 3D camera sensors as images or videos. This data is then processed by computer and thus the recognition and conversion of hand gestures into text takes place. However, this approach of Sign language recognition system has several challenges. These challenges can be addressed as lighting variation, background issues, the effect of occlusions (which is the process of challenging the network to learn not to rely on canonical features by blocking portion of an input image during training time), complex background, processing time affected due to resolution and frame rate and foreground or objects presenting the same skin color tone or otherwise appearing as hands. Based on these factors the vision based system can further distinguished as Color- based approach, Appearance-based approach, Motion-based approach, Skeleton-based approach and 3D-Model-based approach.

a) Color-based approach: One of the color-based approaches for sign language recognition system is by using a colored glove. The glove is colored with several colors and is captured with a camera. The colors on the gloves help the camera sensors to determine the shape, position and location of the fingers and palm of the hand thus determining and extracting the geometric shape of the hand gesture. One of the applications of this color-based recognition with glove system other than sign language recognition is interaction with 3D models, allowing some advanced processing such as zooming, moving, writing and drawing using virtual keyboard with excellent flexibility. The main advantage of this system is its simplicity and cost effectiveness compared to that of the multi-sensor based gloves. However, even in this approach it is essential to wear these gloves during fingerspelling which is a really inconvenient.

Fig.2. Color-based Glove

Another color-based approach of sign language recognition system is by using skin color detection. Skin color detection is one of the most popular methods in color-based approach and is used in a wide range of applications, such as object classification, degraded photograph recovery, person movement tracking, video observation, Human Computer Interface (HCI) applications, facial recognition, hand segmentation and gesture identification. Skin color detection system uses two methods. In the first method the hand gesture is determined by pixel based skin detection, in which each pixel of an image is determined into skin or not, individually from its neighboring pixel. Whereas, the second method is region skin detection, in which the skin pixels are processed based on the information like pixel intensity and pixel texture values.

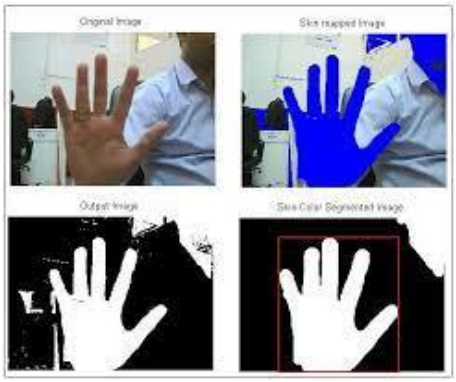

Fig.3. Skin Detection Hand Gesture Recognition

Several color formats can be used in color-based approach such as:

- Red, Green and Blue (RGB)
- Hue and Saturation (HSV, HSI and HSL) and
- Luminance (YIQ, Y–Cb–Cr and YUV)

b) Appearance-based approach: This approach can be implemented in both of the 2D static model and 3D motion models. This approach extracts the features of hand to model the visual appearance and then compare these values with the values extracted from the input images. This approach is executed in real time because of use of 2D images rather than 3D images which tend to take more time in processing due to excessive data. In addition to this, this method can also additional features such as determining skin tone which can help in accuracy and can be used in several other applications other than sign language recognition.

c) Motion-based approach: The motion-based recognition system recognizes the gestures through a series of image frames. The system uses several sensors to detect the motion. But several issues encounters if there is more than one gesture are active during recognition process. The dynamic background too has a negative effect on the processing of the system. In addition, there may be loss of gestures due to occlusions between the tracked hand gestures. Or there can be error in detection of region from tracked hand gesture or the distance between the gesture and the region can also affect the results of the system. The loss of data cannot be afforded as data here is the most crucial and important factor of the system.

d) Skeleton-based approach: Skeleton-based recognition systems reduce complexities in the detection process by specifying some model features of the data. With the help of parallel Convolutional Neural Network (CNN) the images are captured focusing on the skeleton of the hand. The most common features that are focused in this system are the joint orientation, the space between joints, the skeletal joint location and degree of angle between joints. Thus, reducing unwanted data such as skin color and hand palm boundaries. But in order to capture images with such high definition, depth camera sensors are used which comes with high costing. The overall system too is quite complex.

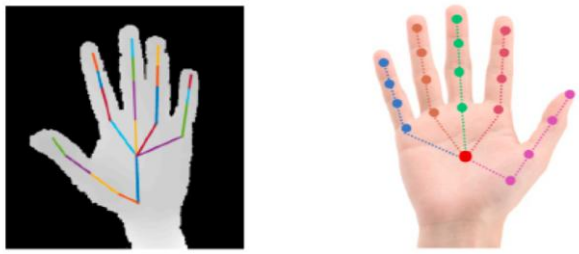

Fig.4. Skeleton-based Approach

e) 3D-model-based approach: The 3D model based systems essentially depends on 3D Kinematic hand model which gives a high degree of freedom, where hand parameter estimation is obtained by comparing the 2D image input data projected by 3D hand model. In this model the depth parameter is added to increase the accuracy of the system. The system provides a 3D hand appearance. With the use of high definitive kinetic camera sensors and the Graph CNN (Graph Convolutional Neural Network) this system detects the hand gestures in 3D format. In addition, the 3D model introduces human hand feature as pose estimation by forming

volumetric or skeletal or 3D model that identical to the user's hand. In order for the system to give such accuracy, there is need of wide range of dataset which reduces the reliability of the system.

III. PROPOSED SYSTEM

There is no single sign language. Different countries use different sign languages. For example British Sign Language (BSL), American Sign Language (ASL) and French Sign Language (LSF). American Sign Language is used all around the world including different countries of Asia and Africa. American Sign Language is the most popular sign language after French Sign language. ASL is derived from French Sign Language and local sign language. ASL grow early in 19th century in American School for the Deaf (ASD). ASL uses different hand sign and movement to communicate with deaf & dumb people. As other languages ASL use its own grammar. It is not easy to understand and learn sign language for normal person. Sambhashan is useful for those who are not able to communicate with deaf people using sign language.

Sambhashan use American Sign Language (ASL) to convert sign language to text for user. There are different components of sign language such as fingerspelling, facial expression and body gesture. Using fingerspelling based American Sign Language (ASL) Sambhashan translate sign language to text. Sambhashan uses the concept of Artificial Neural Network (ANN). Artificial Neural Network is the subfield of machine learning and has different application like speech recognition, face recognition and character recognition. There are different types of Artificial Neural Network such as Modular Neural Network (MNN), Recurrent Neural Network (RNN) and Convolution Neural Network (CNN). Different types of neural networks are used for different purpose. Sambhashan make use of Convolution Neural Network (CNN). CNN takes the input image recognizes them and then shows the appropriate output. There are different layers in Convolution Neural Network (CNN).

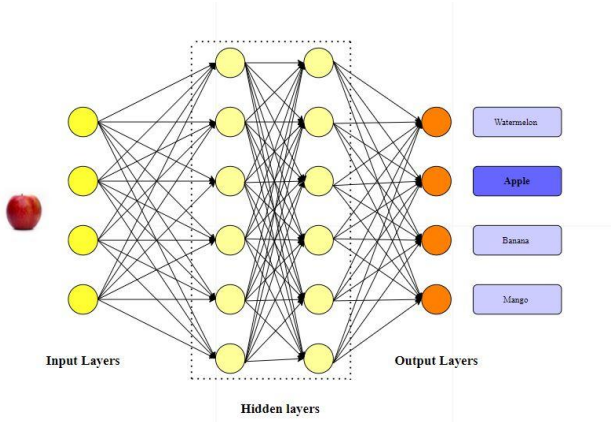

Fig.1. Convolution Neural Network (CNN)

Sambhashan is designed and implemented to recognize alphabet from A to Z using fingerspelling based American Sign Language (ASL).System uses web cam to take hand gestures as input. These sign processed for some feature extraction using different models. This feature are compared with the testing dataset and then finally the recognize hand gestures are converted to sign. This sign recognition system will help the deaf & dump people to communicate with normal people. According to the processing, Sambhashan have following different models:

A. Data Collection

Dataset is the main component of machine learning. A dataset is the collection of small separate data treated as a single unit. This dataset are used to train an algorithm to find a particular goal inside the dataset. There are different datasets are available but all the datasets are in the form of RGB value. Sambhashan has its own dataset and to create a dataset we used open computer vision (OpenCV). OpenCV is an open source library used for real time computer vision. OpenCV are used for different purpose such as object recognition, image processing and video capturing. Using OpenCV we capture images to create dataset. Approximate 400 images of every alphabet are captured in American Sign Language (ASL). This images are captured and then stored in different folders of the dataset i.e. folder A saves the image of alphabet A.

B. Image Processing

When we capture the images we define an area called as Region Of interest (ROI). This Region of Interest is represented by blue bordered square. This Region of Interest (ROI) covers the hand gesture in the image. From the whole image we extract the Region Of interest (ROI) which is in RGB (Red, Green, and Blue) form. This colored image is not a good idea to feed neural network as input because colored images have too

much parameters which are not required for projects such as sign language detection. Another reason is if we give colored images as input it requires more computation time .It will also take a lot of time for training models.

First we convert this colored image to black and white images. After this process three color channel i.e. red, green and blue are converted into a single channel. To convert sign language to text we don't require too much features of hand we only require some main features. We use thresh holding technique for image processing. Thresh holding is a type of image segmentation that coverts grayscale image to binary images. Thresh holding is an easy and effective way to separate background and foreground image. For extracting the main features of hands we use the Gaussian filter on black and white image. Gaussian filter is linear filter and basically used to remove contrast and boundaries detection. Finally we get the final images with boundaries. This image will easy to understand by the neural network.

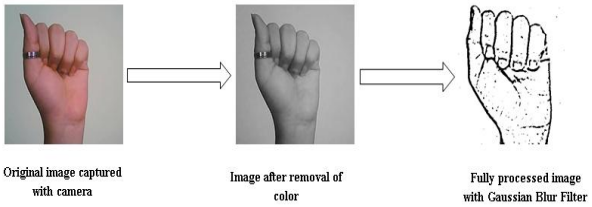

Fig.2. Image Processing

C. Training Model

A model is created using TensorFlow which will recognize the real-time input of the hand signs of the user and will display the corresponding letter of the sign language. TensorFlow is an open source library for machine learning. TensorFlow is developed by Google. It also provides different library and extension for training models. This is an easy and simple library to train and deploy the models. Tensorflow has wide range of application in medical, education and research. Keras is an open source library used as an interface for TensorFlow. Keras is written in python and used to create neural network. Keras is integrated with TensorFlow .It is used by many companies like Netflix and Uber for their product recommendation using neural network. Using Keras and TensorFlow we train a model. This model follows two layer approaches to predict the final symbol. First layer of this approach is Convolution Neural Network (CNN) and second layer are the classifiers.

First layer is Convolution Neural Network. After applying the thresh holding and Gaussian blur filter for feature extraction this pre-processed image is passed to first layer which is Convolution Neural Network (CNN) model. This layer predicts the output using the dataset. There are seven layers present in this Convolution Neural Network (CNN) model. Each layer has its own role to predict the final result. First layer is 1^{st} convolution layer which takes an input image of resolution 128 X 128. Using 32 filter weights in first convolutional layer this image is processed and result an output of 126 X 126 pixel image. Second layer of the model is 1^{st} pooling layer which down sampled the image using max pooling of 2 X 2 which result an image of 63 x 63 pixel image.

Output of the second layer is providing as an input for the third layer. Third layer is 2^{nd} convolution layer which converts the 63 X 63 pixel resolution image into 60 X 60 pixel image using 32 filter weight. These images are served as an input to the fourth layer of the model. Fourth layer of the model is 2^{nd} pooling layer which is used to down sample the image using the max pooling of 2 X 2. This layer reduces the size of image to 30 X 30 resolutions. Fifth layer of this model is 1^{st} densely connected layer this layer takes the input of fifth layer to fully connected layer with 128 neurons. Now the output of 2^{nd} convolution layer is reshaped to an array of 28800 values. Sixth layer of the CNN model is 2^{nd} densely connected layer. This layer takes the input of 1^{st} densely connected layer to fully connected layer with 96 neurons. The output of the 2^{nd} densely connected layer is fed as an input to the final layer. This final layer has the number of neurons as the number of classes we describe.

The second layer is used as classifiers. After applying thresh-holding and Gaussian blur filter we use the first approach of algorithm to verify and predict the symbols. During testing we find out some similar symbols which are confusing for the model. This symbols are grouped as {D, R, U}, {T, K, D, I} and {S, M, N}.We train different models for each of this groups. After image processing the image is predict using the first layer of approach. If the output symbol is not belongs to this group then the model will return the predicted text and if the predicted symbol is belong to this group then the second layer model will predict the text.

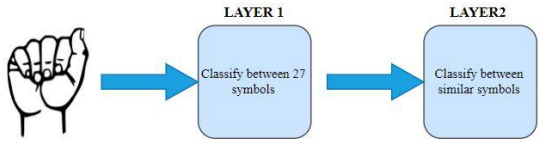

Fig.3. Two-Layered Approach

D. Testing Model

After image processing the image is fed to the model for testing the system. The prediction layer estimates how likely the image will fall under one of the classes. We use softmax function to classify the output. This function produces the output between 0 and 1. Softmax function is mainly used in output layer of Convolution Neural Network for classification. Softmax function is used for binary and multi class classifiers. Softmax function returns the probability distribution for each of the class. This probability distribution is range between 0 and 1 .The sum of each probability distribution is always 1. Even after using the softmax function the actual output value is not accurate. We use cross-entropy for more accuracy. The value of cross-entropy function is exactly 0 when it gets the accurate symbol. TensorFlow has inbuilt function to calculate the cross-entropy.

For user experience we had created a Graphical User interface. This user interface has a frame which takes input from the user. After applying thresh-holding and Gaussian filter the image is fed to the model for prediction. At last the model predicts the word and displays it on user interface. We use the finger spelling for sentence formation. Whenever the system get a specific value and no other similar words are detect then the value is printed and added to the current string. Whenever the frame detects the plain background it considers it as end of the string. If the current buffer is empty then no blank space is detected

.

REFERENCES

[1]. Pujan Ziaie, Thomas M¨uller , Mary Ellen Foster , and Alois Knoll"A Na¨ive Bayes Munich,Dept. of Informatics VI, Robotics and Embedded Systems,Boltzmannstr. 3, DE-85748 Garching, Germany.
[2]. https://docs.opencv.org/2.4/doc/tutorials/imgproc/gau sian_median_blur_bilateral_filter/gausian_median_blur_bilateral_filter.html
[3]. Mohammed Waleed Kalous, Machine recognition of Auslan signs using PowerGloves: Towards large- lexicon recognition of sign language.
[4]. Aeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convol utional-Neural-Networks- Part-2/
[5]. Byeongkeun Kang , Subarna Tripathi , Truong Q. Nguyen "Real-time sign language fingerspelling recognition using convolutional neural networks from depth map" 2015 3rd IAPR Asian Conference.
[6]. Pattern Recognition (ACPR)
[7]. https://opencv.org/
[8]. https://en.wikipedia.org/wiki/TensorFlow
[9]. https://en.wikipedia.org/wiki/ConvolutionalNeuralN etwork