Conversion of Sign Language into Text and Audio

Janu K.S[1], Ajeesh.T[2], Rohit.R[2], Viju.E[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Rohini College of Engineering and Technology, Kanyakumari- 629 401, Tamil Nadu, India

[2]Department of Computer Science and Engineering, Rohini College of Engineering and Technology, Kanyakumari- 629 401, Tamil Nadu, India

Article Info

Volume 8, Issue 7

Page Number: 69-75

Publication Issue :

May-June-2022

Article History

Accepted: 01June2022

Published: 20June2022

ABSTRACT

Human beings communicate with each other in order to convey their ideas, thoughts, and experiences to the people around them. But among the deaf-dumb people this is not the case. According to WHO survey, around 10-15% of world population are hearing impaired and mute. In this vast growing deaf and dumb people, it is sufficient to bridge the communication gap between mute persons and other people is the crux of this paper. However, deaf and dumb people communicate among themselves using sign languages. This paper proposes a way to convert hand gestures into appropriate text message as well as audio using webcam. In the captured frames, the features will be extracted from the collected segmented data using image pre-processing and feature extraction. In the final step, these features will be fed to supervised models for classification and depicted as message as text and voice.

Keywords: Sign language recognition, Convolutional neural network, Translation, Extraction

I. INTRODUCTION

In the fast-growing world, providing an equitable life for the hearing impaired and dumb people is still a challenge. Deaf-dumb people need to interact with normal people or among themselves for their daily routine. For that they have to use sign language to communicate with other people. Sign language is one of the natural form of communication, but since most people do not understand the particular language, they have to rely on special training or interpreters to understand the language. But the usage of sign language interpreter can be costly. Cheap solution is required so that the deaf-mute and normal people can communicate normally. Sign language uses hand gestures and other means of non-verbal behaviours to convey their intended meaning. It involves combining hand shapes, orientation and hand movements, arms or body movement, and facial expressions simultaneously, to express speaker's thoughts. The idea is to create a sign language to text and speech conversion system, using which the information gestured by a deaf-mute person can be effectively conveyed to a normal person. The main aim of this work is to design and implement a system to translate finger spelling (sign) to speech, and audio using recognition and synthesis techniques. Meanwhile, in addition to the text message and the information, we introduce a concept of voice conversion into our system, reducing the conflicts and this system can be used by both deaf and mute people. The paper is organized as follows: The section II includes study and analysis of related work. The section III briefs about our proposed methodology and its implementation. The section IV shows the experimental results analysis. The section V concludes by comparing our methodology with the present existing approaches and the future work.

II. RELATED WORK

The different types of signs were recognized with different methods by different researchers in which were implemented in different fields. The recognition of various signs were done by vision based approaches, data glove based approaches, soft computing approaches like Artificial Neural Network, Fuzzy logic, Generic Algorithm and others like PCA, Canonical Analysis, etc. The recognition techniques are further divided into three broad categories such as Hand segmentation approaches, Feature extraction approaches and Gesture recognition approaches. Bhumika Gupta [4] proposed a model that perceives static images of the marked letters in order in the Indian Sign Language. Dissimilar to the datasets in other sign languages via hand gestures like the American Sign Language (ASL) and the Chinese Sign language (CSL), the Indian Sign Language (ISL) letter uses both the double hands and as well single hand. Hence, it makes ease the recognition of the gestures by categorizing them into single hand and double handed gesture. For the displaying both the two classes, they used two different features such as HOG and SIFT respectively, which are separated for a set of images which are used for training and are combined into a single matrix. After which, the HOG and SIFT features assigned for the particular input test images are joined with the feature matrices of HOG and SIFT. The resultant classification of the test image is obtained by feeding k-Nearest neighbour classifier with computed correlation for the matrices. Paper [1] demonstrates, a hand free demonstration of Taiwanese data language which uses the wireless system to process the data. To differentiate hand motion, they have used inner sensors put into gloves to show the parameters as given by, posture, orientation, motion, defined of the hand in Taiwanese Sign Language could be recognize in no error. The hand gesture is considered by flex sensor and the palm size considered using the g sensor and the movement is considered using the gyroscope. Input signals would have to be consider for testing for the sign to be legal or not periodically. As the signal which was sampled can stay longer than the pre-set time, the legal gesture sent using phone via connectivity like bluetooth for differentiating gestures and translates it. With the proposed architecture and algorithm, the accuracy for gesture recognition is quite satisfactory. Paper [2] proposed lower the communication gap between the mute community and additionally the standard world. The projected methodology interprets language into speech. The system overcomes the necessary time difficulties of dumb people and improves their manner. Compared with existing system the projected arrangement is simple as well as compact and is possible to carry to any places. This system converts the language in associate text into voice that's well explicable by blind and ancient people. The language interprets into some text kind displayed on the digital display screen, to facilitate the deaf people likewise. In world applications, this system is helpful for deaf and dumb of us those cannot communicate with ancient person. Conversion of RGB to grey scale and grey scale to binary conversion introduced in the intelligent sign language recognition using image processing. Basically any colour image is a combination of red, green, blue colour. A computer vision system is implemented to select whether to differentiate objects using colour or black and white and, if colour, to decide what colour space to use (red, green, blue or hue, saturation, luminosity).

III. METHODOLOGY

In this paper, 26 ISL alphabets are used along with customized symbols which is to be recognized in real-time using a webcam. For this purpose, here the webcam inbuilt system is used. The algorithm is developed on top of a Python-based OpenCV wrapper. The entire system was developed using images that are converted in RGB format. It recognizes ISL gestures taken from static images. The system comprises of following steps that are, Data acquisition, Data pre-processing, Feature extraction, Classification as shown in the figure 1.

A Sign Language Recognition System

Sign language recognition for the gestures has different approaches to acquire data:

- Glove based approach
- Vision based approach

Glove based approach:

It uses electromechanical devices such as sensor glove to extract hand configuration, position and its features. The task will be simplified during segmentation process by wearing glove. The drawback of using this approach is that the individual who is using has to wear the sensor hardware along with the glove during performing the operation.

Vision based approach:

In vision-based approach, computer camera is the input device for observing the information from hands or fingers in order to detect and by using subsequent algorithm the problem is solved. This technique is easier to the signer since there is no need to wear any extra hardware or gloves. The main challenge of vision-based hand detection is to cope with the large variability of human hand's appearance due to a huge number of hand movements, to different skin- colour possibilities as well as to the variations in viewpoints, scales and speed of the camera capturing the scene. However, there are also accuracy problems related to image processing algorithms and these problems are yet to be modified.

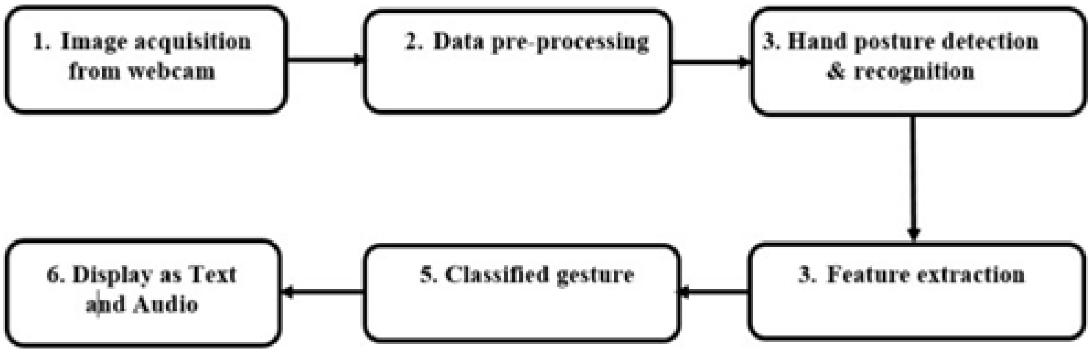
1. Image acquisition from webcam
2. Data pre-processing
3. Hand posture detection & recognition
3. Feature extraction
5. Classified gesture
6. Display as Text and Audio

Fig 1: Overview of the system

1. Data acquisition

As there is lack of research in this field, proper raw image dataset for ISL is not available. So, based on paper [3] we created our own data set of ISL which contains classes in which each class has 1600 images. These images are included in training as well as in testing database. The captured image is adjusted such a way to get the required clarity of image. All images for separate classes are captured via system's webcam. All gestures with the classes can be seen in the figure 3.

2. Data pre-processing

In this phase, all the images that are captured and stored in database are pre-processed so that they can be used for feature extraction. This involves three steps, Segmentation, skin detection and filtering based on paper [5]. Segmentation is the process in which image is converted into small segments so that the more accurate image attribute can be extracted. We can use Adaboost face detector to differentiate between faces and hands as both involve similar skin-color. We can also extract necessary image which is to be trained by applying a filter. The filter is applied using OpenCV. Finally, the whole image, which is RGB is converted into grey scale image, thus extracting various features of our image.

3. Feature extraction:

The pre-processed images are fed to the keras CNN model like paper [8]. The model that has already been trained generates the predicted label, which in our case is called class. All the gesture labels are assigned with a specific probability. The label with the highest probability is treated to be the predicted label.

4. Display as Text & Audio

The model accumulates the recognized gesture to words. The predicted label is already been initialized to particular message by the database. According to the label, the specific message is displayed as text. The recognized words are converted into the corresponding speech using pyttsx library. The text to audio result is a simple work around but is a valuable feature as it gives a feel of an actual verbal conversation which inhibits the use for mute people.

A. Convolution Neural Network:

Unlike regular Neural Networks, in the layers of CNN, the neurons are arranged in 3 dimensions: width, height, depth like paper [6]. The neurons in a layer will only be connected to a small region of the layer (window size) before it, instead of all of the neurons in a fully-connected manner. Moreover, the final output layer would have dimensions (number of classes), because by the end of the CNN architecture we will reduce the full image into a single vector of class scores.

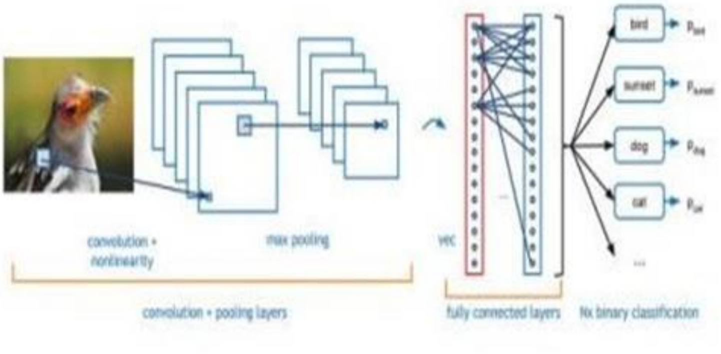

Fig 2: Convolutional Neural Network

1. **Convolution Layer:** In convolution layer we take a small window size [typically of length 5*5] that extends to the depth of the input matrix. The layer consist of learnable filters of window size. During every iteration we slid the window by stride size [typically 1], and compute the dot product of filter entries and input values at a given position. As we continue this process well create a 2-Dimensional activation matrix that gives the response of that matrix at every spatial position. That is, the network will learn filters that activate when they see some type of visual feature such as an edge of some orientation or a blotch of some colour
2. **Pooling Layer:** We use pooling layer to decrease the size of activation matrix and ultimately reduce the learnable parameters. There are two type of pooling:
   a) **Max Pooling:** In max pooling we take a window size [for example window of size 2*2], and only take the maximum of 4 values. Well lid this window and continue this process, so well finally get a activation matrix half of its original Size.
   b) **Average Pooling:** In average pooling we take average of all values in a window. 12

3. **Fully Connected Layer:** In convolution layer neurons are connected only to a local region, while in a fully connected region, well connect the all the inputs to neurons.
4. **Final Output Layer:** After getting values from fully connected layer, well connect them to final layer of neurons [having count equal to total number of classes], that will predict the probability of each image to be in different classes.

B. Tensorflow:

Tensorflow is an open source software library for numerical computation. First we define the nodes of the computation graph, then inside a session, the 13 actual computation takes place. Tensorflow is widely used in Machine Learning.

C. Keras :

Keras is a high-level neural networks library written in python that works as a wrapper to TensorFlow. It is used in cases where we want to quickly build and test the neural network with minimal lines of code. It contains implementations of commonly used neural network elements like layers, objective, activation functions, optimizers, and tools to make working with images and text data easier.

D. OpenCV :

OpenCV (Open Source Computer Vision) is an open source library of programming functions used for real-time computer-vision. It is mainly used for image processing, video capture and analysis for features like face and object recognition. It is written in C++ which is its primary interface, however bindings are available for Python, Java and MATLAB/OCTAVE.

E. MATLAB:

MATLAB is used for various works regarding pattern recognition , gesture detection etc. [17] .In this work ,for the implementation of training and predicting the flex sensor values, MATLAB neural network tool is used which have the provision to implement various algorithms and methods such as back-propagation neural network , particle swarm optimization etc. [18]. The input values, output values and training samples can be easily inserted and the neural network tool can be used for training the data and simulate the results. The predicted values can also be plotted so that the difference between the predicted values and actual values is understood.

F. Algorithm

Real time sign language conversion to text and voice Start

S1: Show the hand gesture via webcam to adjust with the skin complexion and the lighting conditions.

S2: Apply the data augmentation to the dataset to expand it and therefore reduce the over fitting.

S3: Split the dataset into train, test datasets. S4: Train the CNN model to fit the dataset.

S5: Generate the model such in a way which includes accuracy, error and confusion matrix.

S6: Execute the prediction file - this file predicts individual gestures, converts them into words, displays the words as text, plays the voice output.

Stop

IV. RESULTS AND DISCUSSION

Sample images of different ISL signs were collected using the webcam using image acquisition toolbox on MATLAB .About thousand (1000) data samples (with each sign count five and ten (5-10) were collected as training data. The reason for this is to make the algorithm very robust for images of the same database in order to reduce the rate of misclassification. Examples of the images collected is shown in fig.no.3

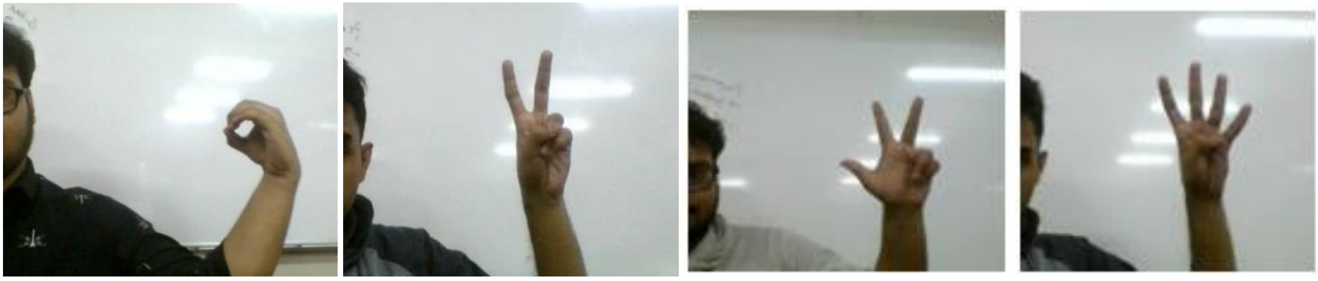

Fig 3: Real time recognition of gestures results

V. CONCLUSION AND FUTURE WORKS

This paper finds the simple and the most optimal approach for creating a vision-based application for sign language to text/speech conversion. The performance of our system is evaluated on a dataset of 1000 images for every gesture.. The experimental results show that our approach performs well and is fit for the real-time applications. The main objective is achieved, that is the need of interpreter is eliminated. The barrier of communication between among deaf and dumb people is also solved. The other issue that people might face is regarding their proficiency in knowing the ISL gestures. Bad gestures will not predict the correct output. The project can be extended to other sign languages by building the corresponding datasets using CNN models. In future, ToF cameras can be used to address the background complexity and improve the robustness of hand detection. It can also be enhanced by building a web or a mobile application for the users to conveniently use it. This project can be improved to achieve higher accuracy in future even in case of complex backgrounds and in low light conditions thus aims at societal contribution.

VI. ACKNOWLEDGEMENT

We thank the institute Rohini College of Engineering for the wonderful opportunity. We also express our gratitude to the department of Computer Science and engineering for facilitating the process

VII. REFERENCES

[1]. L. Ku, W. Su, P. Yu and S. Wei, "A real-time portable sign language translation system," 2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS), Fort Collins, CO, 2015, pp. 1-4, doi: 10.1109/MWSCAS.2015.7282137.

[2]. S. Vigneshwaran, M. Shifa Fathima, V. Vijay Sagar and R. Sree Arshika, "Hand Gesture Recognition and Voice Conversion System for Dump People," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019, pp. 762-765, doi: 10.1109/ICACCS.2019.8728538.

[3]. Z. Wang et al., "Hear Sign Language: A Real-time End-to-End Sign Language Recognition System," in IEEE Transactions on Mobile Computing, doi: 10.1109/TMC.2020.3038303.

[4]. B. Gupta, P. Shukla, and A. Mittal. K-nearest correlated neighbour classification for Indian sign language gesture recognition using feature fusion. In 2016 International Conference on Computer Communication and Informatics (ICCCI), pages 1– 5, 2016.

[5]. S. S Kumar, T. Wangyal, V. Saboo and R. Srinath, "Time Series Neural Networks for Real Time Sign Language Translation," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, 2018, pp. 243-248, doi: 10.1109/ICMLA.2018.00043.

[6]. P. Vijayalakshmi and M. Aarthi, "Sign language to speech conversion," 2016 International Conference on Recent Trends in Information Technology (ICRTIT), 2016, pp. 1-6, doi: 10.1109/ICRTIT.2016.7569545.

[7]. Archana S. Ghotkar, Rucha Khatal, Sanjana Khupase, Surbhi Asati & Mithila Hadap, “Hand Gesture Recognition for Indian Sign Language”, IEEE International Conference on Computer Communication and Informatics (lCCCI), pp: 1- 4,(2012).

[8]. M. Khan, S. Chakraborty, R. Astya and S. Khepra, "Face Detection and Recognition Using OpenCV," 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2019, pp. 116-119

[9]. Singh, Sanjay & Pai, Suraj & Mehta, Nayan & Varambally, Deepthi & Kohli, Pritika & Padmashri, T. (2019). Computer Vision Based Sign Language Recognition System.

[10]. S. S Kumar, T. Wangyal, V. Saboo and R. Srinath, "Time Series Neural Networks for Real Time Sign Language Translation," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL,2018,pp.243248,doi:10.1109/ICMLA.2018.000 43.

[11]. Hu Peng, “Application Research on Face Detection Technology based on OpenCV in Mobile Augmented Reality”, International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 8, No. 2 (2015).