International Journal of Research in Engineering, Science and Management
Volume 4, Issue 8, August 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Conversion of Lip Gestures into Text using Machine Learning Model

Raviteja Avutapalli*
Department of Computer Science and Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India

***Abstract*: Generally, 15% of the total population are having disabilities of speech and are paralyzed. It is difficult for them to convey their feelings or any other information. So, this paper can create an interface to communicate with others. Also, dumb people communicate using lip movements and hand gestures but it is hard to understand. So, the vocal or lip movements are converted into text by the sensors called as Silent Speech Interface and it has given rise to the possibility of speech processing even in the absence of voice. This paper presents an algorithm that records the lip movement by webcam and the recorded signals are analyzed by extracting features using the Local Binary Pattern (LBP) operator and detected using a Haar Cascaded classifier. This development is done by OpenCV software which can be further implemented by hardware design.**

***Keywords*: Haar cascaded classifier, silent speech interface.**

1. Introduction

Around 2.78% of people are dumb in India. Communication is the only medium by which people share their ideas, thoughts and convey the message. But for a dumb person, it is very difficult. They generally use sign language to communicate with others, where it uses gestures by hand than sound to convey. An artificial speech system will be very useful to convey their thoughts to others. The objective is to remove the communication barrier and to make them communicate. So gestures will be a perfect solution for the problem.

Gestures are generally actions and pose that is produced by any part of the body. The actions or movements made by the internal mouthparts for the production of sound are termed mouth gestures. But these cannot be shown by the dumb people. So their communication started with hand gestures and lip gestures. And monitoring these gestures conveying their information can be made possible. But it is difficult to estimate and predict the information very precisely based only on lip gestures because previous researches have proven that lip detection has various limitations like if user lip dimension cannot be drawn exactly or accurately.

Inferior-superior displacements of the upper lip, lower lip, and jaw were converted with a strain-gauge system in 4 normal speaking adults. The upper and lower lip movements were differentiated and compared across conditions during which the jaw was easy to move and when bite blocks were used to fix the jaw at four different vertical positions. However, when the jaw open position was enlarged with the bite blocks, it was found that: Positions of both lips changed for bilabial closure, but the closing movements did not normally maintain consistent proportions between lips across different bite-block sizes; although the lips maintained fairly consistent maximum interlabial opening across many conditions, this opening was reduced in the small bite-block conditions; and in a few cases there was an increase in the duration of lip-closing movements, but these were small and inconsistent. The findings are considered and discussed relative to possible organizational systems that would produce the observed interactions among speech articulators.

2. Methodology

A. Machine Learning model for sign language interpretation using webcam images

A sign is a type of non-verbal communication done with body parts, positions, hand shapes, and movements of the hand or lips arms, facial expressions and used instead of oral communication. Many people use both signs and words during communication. A sign language is a language that uses signs or actions to communicate instead of sounds. According to the above definition, sign language has three major components [1].

The first important component is finger-spelling which means for each letter of the alphabet there is a corresponding sign. This form of communication is used commonly for spelling names sometimes for spelling the location names. At times i.e. not very often this can be used for expressing words for which no signs exist or for emphasizing a particular word. The second vital component of any sign language is word-level sign vocabulary which means for each word of the vocabulary there is a corresponding associated sign in the sign language. The most commonly used type of communication between people with hearing disabilities in combination with facial expression is this type. A third essential component in sign communication is non-manual-features. This mode of communication involves body position, mouth, tongue, eyebrows, and facial expressions. Among all these components most used form of sign language by the deaf community, in

*Corresponding author: tejaavutapalli@gmail.com

reality, is word-level sign language. Hence this paper illustrates and highlights more on frequently used daily words or sentences and their interpretation by Sign Language Interpreter System.

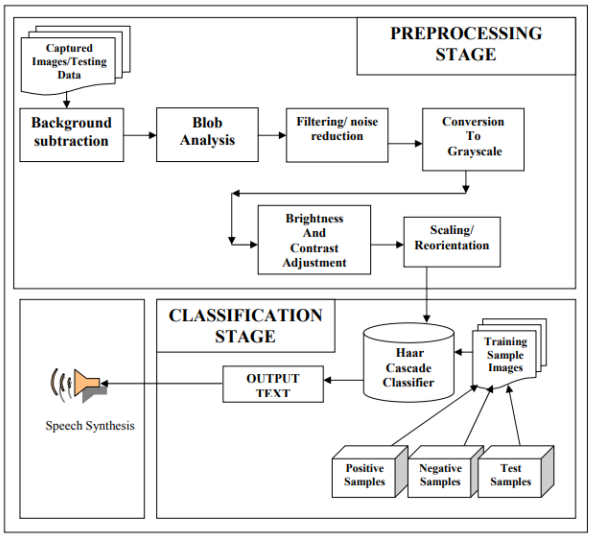

Fig. 1. Sign language interpreter architecture

Advantages:

- Background Subtraction
- Blob Analysis
- Noise Reduction
- Grayscale Conversion
- Brightness and Contrast Normalization
- Image Scaling.

Disadvantages:

- Difficult to learn
- Difficult to understand
- Lots of effort required.

B. Designing a verbal deaf talker system using mouth gestures

When people communicate, the brain has to accomplish a series of tasks. If someone asks a question, for instance, the brain has to understand what has been heard by the sound acquisition system, put things together, tell the muscles how to move, and then perform an action, which may be speech or gestures. Gestures and speech are expressions, which are commonly used in communication among human beings. Learning of their use begins with the first years of life.

Every normal human being sees, listens, and further reacts to situations by talking himself or herself out. But some unfortunate ones are disadvantaged by this valuable ability. This creates a gap between normal human beings and the deprived ones (deaf people) [2]. In the last few years, there has been an increased interest among researchers in the field of gesture recognition to translate it into an understandable form for normal people. Passing through various researches that have been conducted in this field and gaining some knowledge, this paper tends to overcome deaf people's problems of speaking by designing a verbal system that uses mouth gestures.

Gesture sign recognition is a research area for many years.

There are different methods for hand sign recognition employing different techniques. Traditionally, gesture recognition technology was divided into many categories, vision-based, glove-based, and some other approaches. For vision-based methods, the image is captured through a web camera. Input images are applied to image preprocessing and segmentation. In which, object and background are separated. The resultant image has shown some features. Then feature extraction and recognition were done by using PCA (Principle component analysis). Finally, this result is converted into text and voice.

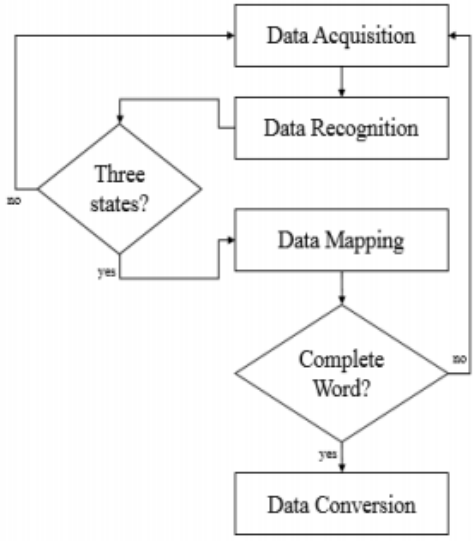

Fig. 2. Verbal deaf talker flow chart

Advantages:

- Ease of use
- Doesn't require any Knowledge
- Provides high rate of communication with deaf people.

Disadvantages:

- Signal problems leads to insufficient voice
- Require more Normalization
- Detecting voice takes lots of time

C. Sign language recognition using facial expression

Vision-based practices have always been forming a foundation and it is the basis to be used as a primary communication method that allows impaired hearing people to communicate with others in their daily life. It is the common and fundamental means of communication bridge among the hearing deficient person [3]. The field of vision-based systems is enormous and the problems faced by recognition systems are immense. Further, the system being related to human-computer interaction (HCI), possesses great importance in making such communication real-time and effective.

Facial expressions are virtually ignored because of their interpretation of the character, various complexities, and for not proper understanding. Any facial based recognition system incorporates with a vision based hand recognition system could

prove to be beneficial for deaf and dumb people. For this purpose, we use the Viola-Jones algorithm followed by Ad boost techniques for better thresholding. In this methodology the bounding box of the lip can be determined by cropping the region from the original image for recognition.

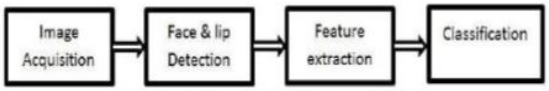

Fig. 3. Block diagram of the facial expression system

Advantages:

- Convenient to use
- High rate of use

Disadvantages:

- Minimum knowledge is required on signal language
- Hard to learn some signals
- Detecting the signals leads to cumbersome

D. A hand gesture recognition based communication system for silent speakers

In recent years, hand gesture recognition is mainly used in human Computer interactions. They play a vital role in gaming and control application like tele-robotics, 3-D mouse, and virtual reality controlling [4]. Beyond this, it can also be used in application aiding the physically challenged community like dumb people. Hand-gesture recognition is the primary requirement for conversion of sign language to speech.

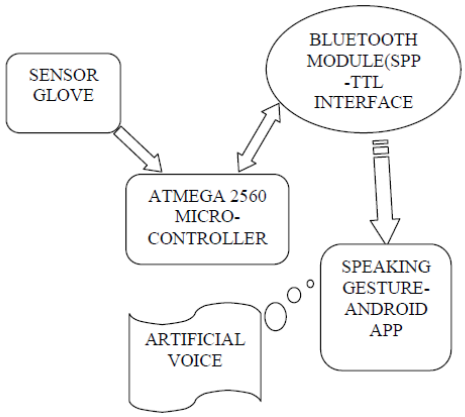

Fig. 4. Hand gesture system design

Advantages:

- Easy communication
- Desired outcomes
- More accurate.

Disadvantages:

- Lot of effort needed
- Analysis is more cumbersome
- Connection failed issues.

3. Proposed System

The proposed model is based on the concepts of OpenCV image recognition like Haar cascade classifiers. The detection of an image using Haar feature-based cascade classifiers is an efficient object detection method proposed by Paul Viola and Michael Jones. The process is a machine learning-based approach where a cascade function is trained from tons of positive and negative images. Now it is used to detect objects in other images.

At this moment face detection is performed. In the first instance, the algorithm requires a lot of positive images and negative images and also needs images without faces to train the classifier. Subsequently, the features are extracted from it. After extracting the features, information related to queries is obtained. And the detected lip sample is compared with samples in the database and the corresponding word is displayed.

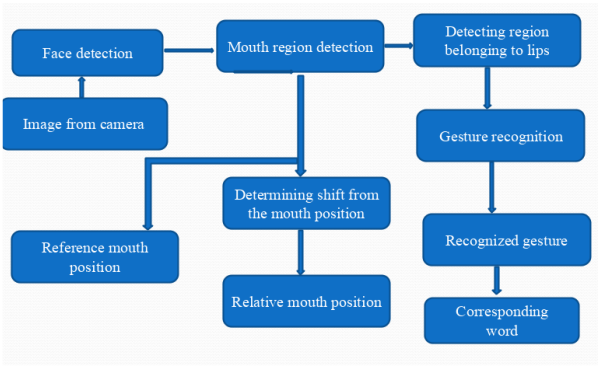

Fig. 5. Proposed methodology

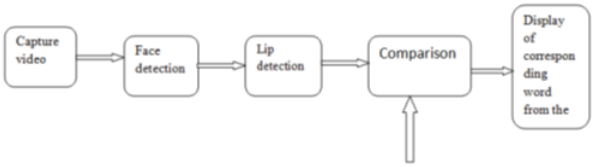

Fig. 6. Block diagram of proposed model

INPUT: Image of user's face as captured by webcam.

OUTPUT: From the captured image, webcam displays the corresponding text according to certain lip gesture identified.

A. Haar Feature Extraction

Haar-like features are digital image features used in object recognition. These were used in the first real-time face detector.

Viola and Jones considered the idea of using Haar wavelets and refined by developing the so called Haar-like features which checks out adjacent rectangular regions at a specific desired location in a detection window, adds up the pixel intensities in each region and then calculates the difference between these sums. The obtained difference is then used to categorize subsections of an image.

In the detection of Viola-Jones object detection framework, a window of the target size is moved over the input image and for each subsection of the image the Haar-like feature is found to be derived and calculated. This obtained difference is now compared to a learned threshold that separates non objects from objects. Such a Haar-like feature is only a weak learner or classifier, a large number of Haar-like features are necessary to describe an object with sufficient accuracy [8]. The Haar-like features are therefore organized in something called a classifier

cascade to form a strong classifier.

The advantage of a Haar-like feature over other features is its calculation speed. Initially, the algorithm needs a lot of positive images images of faces and negative images 28 images without faces to train the classifier. There need to extract features from it. They are just similar to convolution kernel. Each feature is a single value obtained by subtracting sum of pixels under the white rectangle from sum of pixels under the black rectangle.

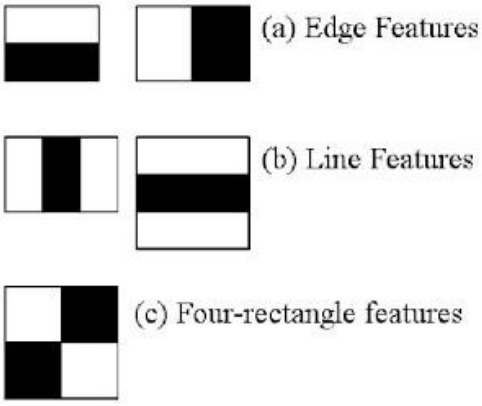

Fig. 7. Haar features

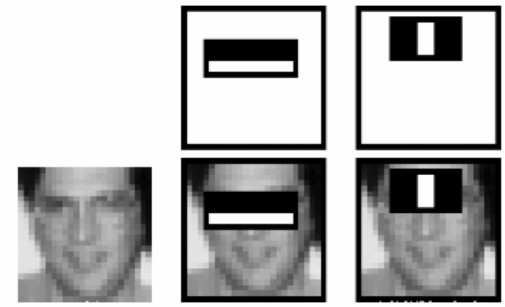

Fig. 8. Feature extraction

Speech processing is one of the largest growing research areas in signal processing. Every year on an average billions of pounds are being spent on supporting research in speech processing. The ultimate aim of this research is to provide an interactive man-machine communication. Speech is a special form of communication medium; it conveys not only the meaning but it also expresses the emotion of the speaker and individual information about the speaker.

The human apparatus is worried and concerned with speech production and perception is complex and uses many important organs - The lungs, mouth, nose, ears controlling muscles and the brain. It is remarkable the apparatus has developed to enable not only the speech production but also serves other purposes such as breathing or eating [6]. It was discovered that various specific areas in the brain are considered and regarded to be of prime importance for speech and language. These are called the speech centers - damages to any of these areas causes disruption to speech.

The vocal tract and vocal cord play a major role in speech production.

The vocal tract consists of several organs and muscles which are regularly monitored and carefully controlled by the speech centers. The precise controlling is achieved by internal feedback in the brain. As an example auditory feedback helps us to ensure that we are producing the correct speech sounds and that they are of the correct intensity for the environment. Speech sounds are produced when air is exhaled from the lungs and causes either vibration of vocal cord or turbulence at some point of contraction in the vocal tract. The shape of the vocal tract influences the sound harmonics. The way in which the vocal cord is vibrated and the shape of the vocal tract is varied in order to produce a range of speech sounds with which we are familiar.

The vocal cord is situated in larynx called the Adams apple. It is the source for speech production in humans. It generates two kinds of speech sounds these are voiced and unvoiced. The frequency of vibration of the cord is determined by several factors; the tension exerted by the muscle, it's mass and it's length. These factors vary between sexes and according to age. The 25 vibration of vocal cord produces harmonics - the amplitude of the harmonics decrease with increasing frequency.

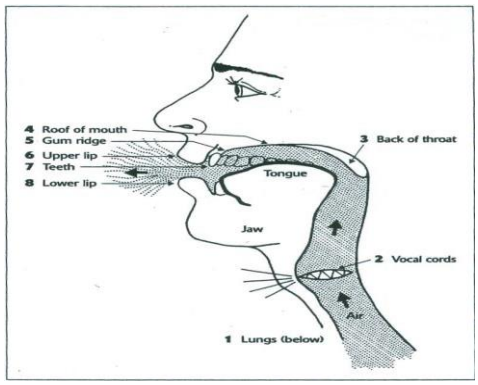

Fig. 9. Speech processing in human body

Procedure:

- A person is made to sit in front of a camera.
- The movements are recorded or captured as a video and the image is read.
- During the recording face part is detected using the algorithms and then the lips are detected. As a rectangle is drawn around the face indicating that it is face part.
- Another rectangle is drawn around the lips, again indicating that they are lips.
- The face and lips is indicated by the change in color of rectangles. Used to detect them
- Then the lip samples are extracted and then compared from those in the data base and then the referred word is displayed.

So, for this, Haar features shown in below figure are used. They are just like similar to our convolution kernel. Each feature is a single value obtained by subtracting sum of pixels under white rectangle from sum of pixels under black rectangle.

OpenCV comes with a trainer as well as detector. If one wants to train their own classifier for any object like plane, car etc. you can use OpenCV to create it. Its complete details are

given here Cascade Classifier training.

Fig. 10. Haar feature detection

Here detection is performed by OpenCV as it already contains many pre-trained classifiers for eyes, face, smile etc. Those XML files are then placed by storing in OpenCV/data/haarcascades/ folder. Let's create face and lip detector with OpenCV.

Now, load the necessary XML classifiers. Then the input image or video is loaded in gray scale mode.

Next face is found in the image. It should return the positions of detected faces when faces are found i.e. as Rect(x,y,w,h).

- Once locations are found, ROI is created for the face and then lip detection is applied on this ROI (since lips are always on the face).
- If no lips are detected, then the user is alerted to be under the proper lightening conditions.
- And the extracted samples of lips are being compared with the samples in the data base.
- If the samples get matched, then the corresponding string from the data base is displayed.

4. Algorithm

A. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm. This can be used for classification and regression challenges. However, it is commonly used in classification problems. In this type of algorithm, each data item acts as a point in n-dimensional space where n is number of features you have with the value of each feature being the value of a particular coordinate are plotted. Then, classification by finding the hyper-plane that differentiate the two classes very well is performed.

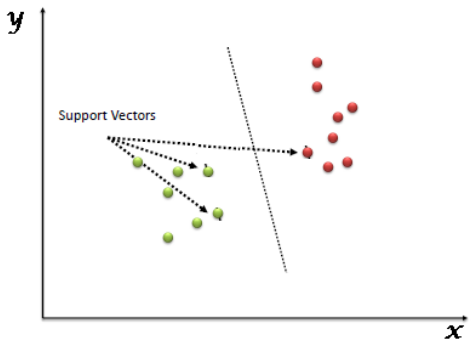

Fig. 11. One class Hyper plane plot

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes hyper-plane/ line.

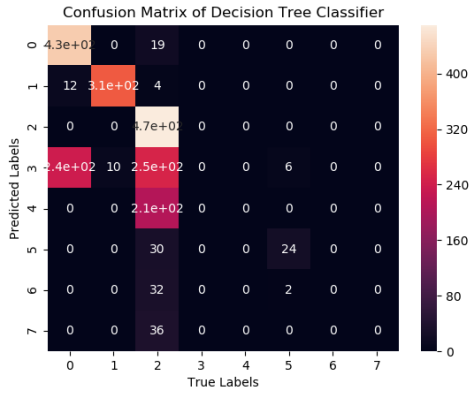

Fig. 12. Confusion matrix of Decision Tree Classifier

A confusion matrix is a technique for summarizing the performance of a classification algorithm.

Classification accuracy alone can be misleading if you have more than two classes in your dataset or in each class if you have an unequal number of observations.

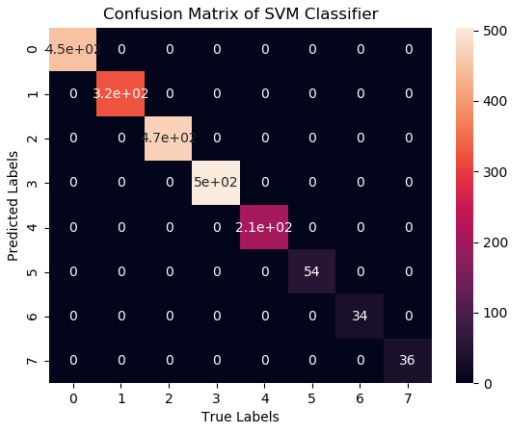

Fig. 13. Confusion matrix of SVM classifier

Calculating a confusion matrix can result a good idea of what your classification model is getting correct or right and what types of errors it is making.

5. Output

A. Training Samples given to SVM

A training dataset is a dataset of examples used for learning, that is to fit the parameters (e.g. weights) of a classifier.

Many approaches which search through training data for empirical relationships results to overfit the data, meaning that they can exploit apparent relationships and identify in the training data that cannot be hold in normal.

The inputs are provided to the system. The provided inputs are matched with the database samples to find whether the given sample is in data base or not. Some of the provided inputs.

Fig. 14. Training samples given to SVM

B. Test samples

A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset. Dataset if a model fit the training dataset which also fits the test dataset as well, then minimal overfitting takes place. Also if better fitting of the training dataset takes place against the test dataset then it points to overfitting.

A test set is therefore a set of examples used only to assess the performance or that can be generalization of a fully specified classifier.

Fig. 15. Test samples given to SVM

Output:

When a face is shown in front of the web cam, the dimensions of the face is captured by the 24x24 window and the lip region is detected and the rest is ruled out. Then, different type of gestures are recognized and displayed as text on the screen.

6. Conclusion and Further work

The face recognition system using OpenCV using Python is designed and the system is verified on the collected image database. The data base is collected for 5 words i.e. 20 samples for each word. As the gesture is identified by the system with the help of database samples then the corresponding word is displayed. But when the gesture is not identified the nearer reference samples are taken into consideration and corresponding words are displayed. Further, the system can be trained for number of words and also the system can be implemented using hardware.

References

[1] Kanchan Dabre, Surekha Dholay "Machine Learning Model for Sign Language Interpretation using Webcam Images", International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014.

[2] Amro Mukhtar Hassan, Ashraf Haitham Bushra, Osama Amer Hamed, L.M. Ahmed, "Designing a verbal deaf talker system using mouth gestures", International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2018.

[3] Siddhartha Pratim Dasa, Anjan Kumar Talukdarb, Kandarpa Kumar Sarmac, "Sign Language Recognition using Facial Expression", Procedia Computer Science, 58, 2015.

[4] N. Sriram, M. Nithiyanandham, "A Hand Gesture Recognition Based Communication System for Silent Speakers", International Conference on Human Computer Interactions (ICHCI), 2013.

[5] Rafiqul Zaman khan, "Hand Gesture Recognition", International Journal of Artificial Intelligence & application, vol. 3, Issue 4, pp. 15-18, 2012.

[6] P. John Hubert & M. S. Sheeba, "Lip and Head Gesture Recognition based on PC Interface," in the Proceedings of IEEE, vol. 5, Issue 18, pp. 59-71, 2015.

[7] Manjunatha M. B, Pradeep Kumar B. P, Santhosh, "Survey paper on Hand Gesture Recognisation", IJAREEIE, vol. 3. Issue 4, pp. 177-182, 2014.

[8] S. Ganesh, Saravana Kumar, "A Novel Voice Recognition System for Dumb People", Journal of Theoretical & Applied Information Technology, vol. 53, Issue 5, pp. 1306-1388, 2013.

[9] T. Chen, "Audiovisual speech processing. Lip reading and lip synchronization," in IEEE Signal Processing Magazine, vol.18, pp. 9-21. 2001.