Sign Language to Speech and Text Conversion Using Image Processing And Machine Learning

Prof. Chhaya Narvekar[1], Sayukta Mungekar[2]

[1]Prof. Chhaya Narvekar
Department of Information Technology(Head of the Department)
Xavier Institute of Engineering, Mahim
Mumbai, India

[2]Sayukta Mungekar
Department of Information Technology
Xavier Institute of Engineering, Mahim
Mumbai, India

---***---

Abstract –One of the most precious gifts of nature to the man breed is the ability to express himself by responding to the events occurring in his surroundings. Every normal human being sees, listens and then reacts to the situations by speaking himself out. But there are some less fortunate ones who are deprived of this valuable gift. The deaf and the dumb, rely on some sort of sign language for communicating their feelings to others. In the era of advanced technologies, where computers, laptops and other processor based devices are an integral part of day to day life, efforts are required to be done for making the disables more independent in life. This project consists of image processing and machine learning methods for this purpose. Our aim is to design a human computer interface system that can recognize language of the deaf and dumb accurately. In this paper, vision based hand gesture recognition system have been discussed as hand plays vital communication mode, considering various techniques available for hand tracking, segmentation, feature extraction and classification are referred. Implementation of the project is as, images are captured using webcam and are processed using image processing techniques such as OTSU method and classification of the captured gesture is done by using linear classification method. Here, the captured gestures are stored into folders consisting 120 replicas of the same gesture. Image gesture is captured in the form of histogram.

***Key Words*:**Sign language Recognition, Hand gesture recognition, Image processing, OTSU method, Feature detection, Feature extraction, Naïve Bayes Classifier, Machine Learning

1.INTRODUCTION

Sign Language is a nonverbal method of communication in which gestures are made using hands. Gestures are an integral part of our day to day communication and some expressions are conveyed by gestures only. Rising of eyebrows, shrugging of shoulders, nodding of head are some commonly used gestures. Sign language is a more organized form than gestures. Various commonly used sign languages are ASL (American Sign Language), BSL (British Sign Language) and ISL (Indian Sign Language). There is no one standard form of sign language and it varies from region to region. We have selected a sample sign language with reference to the American Sign Language.Since English is the standard language that is used all over the world for computer keyboards, so our sample sign language is based on English vocabulary.In this project we have used Vision based approach for single handed gestures computed using Image Processing and processed by Naïve Bayes Classification i.e. Linear Classification method.

A. Sign Language

Sign Language is a well-structured language with a phonology, morphology, syntax and grammar. Sign language is a complete natural language that uses different ways of expression for communication in everyday life. Sign Language recognition system transfers the communication from human-human to human-computer interaction. So, there are two main approaches used in the sign language recognition that is Sensor based and Vision based Approach.

Vision Based Approach: In this approach camera takes the image of gesture, extract the main feature and recognizes it. Capturing the image is done by using Image Processing techniques and depending on the image captured the speech and text are obtained using machine learning techniques. The main advantage of this method is that it gives maximum accuracy and robustness.

Sensor Based Approach: In this approach for hand gesture recognition different types of sensors were used and placed on hand, when the hand performs any gesture, the data is recorded and is then further analyzed. Sensor based approach damages the natural motion of hand because of use of external hardware. The major disadvantage is complex gestures cannot be performed using this method.

B. Image Processing

Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be image or characteristics/features associated with that image.

OTSU Method: In computer vision and image processing, is used to perform automatic image thresholding. In the simplest form, the algorithm returns a single intensity threshold that separate pixels into two classes, foreground and background. This threshold is determined by minimizing intra-class intensity variance, or equivalently, by maximizing inter-class variance. Otsu's method is a one-dimensional discrete analog of Fisher's Discriminant Analysis, is related to Jenks optimization method, and is equivalent to a globally optimal k-means.

C. Machine Learning

Machine learning(ML) algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Naïve Bayes Classification Method:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes Theorem. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naïve Bayes Classifier is as expressed as following

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

The image processing technique using the camera to capture the image/video. Analysis the data with static images and recognize the image using algorithms and produce sentences in the display, vision based sign language recognition system uses OTSU method and extract the image and eliminate the unwanted background noise. It uses Naïve Bayes Classification method in convolutional neural network to improve the performance of the system. The main advantage of this approach is it's less computational time and fast response in real time applications.

2. RELATED WORK

The man-machine interface plays a very important role in today's life. It is better to implement a model that recognizes user provided patterns and identifies that pattern in the machine. Researchin face recognition, hand detection, speech recognition, and voice recognition and in more areas are increasing rapidly.

Hand gesture recognition is a part of HumanComputer Interaction (HCI), the hand is given as an input and machine should recognize the gesture provided by the user and it should give the output based on the gesture.

Denise Powell [1] gives a case study of two qualified New Zealand Sign Language interpreters working in a post-secondary education setting in New Zealand was undertaken using both qualitative and quantitative methods.

C.W.Ng and S.Ranganath [2]interpret a user's gestures in real-time using hand segmentation to extract binary hand blobs. The shape of blobs is represented using fourier descriptors. This fourier descriptor representation are input to radial-basis function(RBF) networks for postureclassification.

N.Tanibataet al. [3]obtain hand features from a sequence of images. This is done by segmenting and tracking the face and hands using skin colour. The tracking of elbows is done by matching the template of an elbow shape. The hand features like area of hand, direction of hand motion, etc. are extracted and are then input to Hidden Markov Model(HMM).

H.K.Nishiharaet al.[4] (US patent, 2009), generate silhouette images and three-dimensional features of bare hand. Further, classify the input gesture by comparing it with predefined gestures.

Jagdish L. Raheja et al. [5] describes a novel method of fingertips and center of palms detection in dynamic hand gestures generated by either one or both hands without using any kind of sensor or marker. We call it Natural Computing as no sensor, marker or color is used on hands to segment skin in the images and hence user would be able to do operations with natural hand.

Shweta S. Shinde, Rajesh M. Autee and Vitthal K. Bhosale [6] have proposed a method in which the angle and peak calculation approach is used to extract the features of hand gestures by using MATLAB and then they convert the recognized gesture into speech using MATLAB inbuilt command.

3. SYSTEM DESIGN

Computer recognition of sign language is an important research problem for enabling communication with hearing impaired people. This project introduces an efficient and a fast method to convert Sign Language into speech and text. The system does not require the hand to be perfectly aligned to the camera. The project uses image processing system to identify, especially English alphabetic sign language used by the deaf people to communicate. The basic objective of this project is to develop a computer based system that will enable dumb people significantly to communicate with all other people using their natural hand gestures. The idea consisted of designing and building up a system using image processing, machine learning and speech conversion concepts to take visual inputs of sign language's hand gestures and generate easily recognizable form of outputs. Hence the objective of this project is to develop an intelligent system which can act as a translator between the sign language and the spoken language dynamically and can make the communication between people with hearing impairment and normal people both effective and efficient.

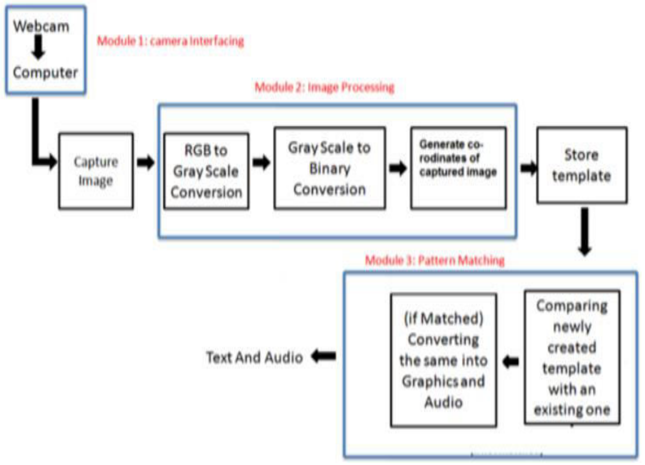

Webcam
Computer
Module 1: camera interfacing
Module 2: Image Processing
Capture Image
RGB to Gray Scale Conversion
Gray Scale to Binary Conversion
Generate co-rodinates of captured image
Store template
Module 3: Pattern Matching
Comparing newly created template with an existing one
(if Matched) Converting the same into Graphics and Audio
Text And Audio

Fig 1. System Block Diagram

A.Finger Sign Recognition

The finger sign recognition task involves the segmentation of finger sign hand gestures from image sequences. Through the classification of features extracted from these images, sign gesture recognition can be achieved. Since a perfect method of segmenting skin color objects from images with complex backgrounds has not yet been proposed, recent studies on finger sign recognition make use of different methodologies. First the segmentation of hands by skin color detection methods and background modeling. Then, Histogram of Oriented Gradient descriptors are used to classify hand features. It incorporate motion descriptors into skin color based segmentation to improve the accuracy of hand segmentation. GUI makes use of human past behavioral patterns in parallel with skin color segmentation to achieve better hand segmentation.

B. Finger Sign Synthesis

The fingerspelling synthesis can be seen as a part of the sign language synthesis. Sign language synthesis can be used in two forms.

C. Speech Synthesis

Speech synthesis is the artificial production of human speech. Speech synthesis (also called text-to-speech (TTS) system converts normal orthographic text into speech translating symbolic linguistic representations like phonetic transcriptions into speech.

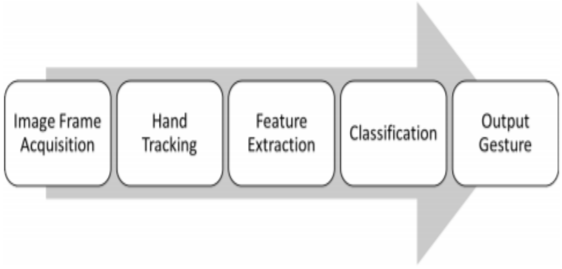


Fig 2. System Flow

The motto of this project is to design and implement a system that can translate finger sign to speech, by using recognition and synthesis techniques for each modality. Such a system will enable communication with the hearing impaired when no other modality is available.

4. IMPLEMENTATION METHEDOLOGY

The system consists of four modules. Image is captured through the webcam. The camera is mounted on top of system facing towards the wall with neutral background. Firstly, the captured Colored image is converted into the gray scale image. The calculated coordinates are then stored into the database in the form of template. The templates of newly created coordinates are compared with the existing one. If comparison leads to success then the same will be converted into audio and textual form. The system works in two different mode i.e. training mode and operational mode. Training mode is part of machine learning where we are training our system to accomplish the task for which it is implemented. Thus, the objective of this project is to develop a system which can act as a translator between the sign language and the spoken language dynamically and can make the communication between people with hearing impairment and normal people both effective and efficient.

A. Camera Initialization and Orientation

The Camera Interface block is the hardware block that interfaces and provides a standard output that can be used for subsequent image processing.

OpenCV: OpenCV is an Open source computer vision library is used as an interface between the user and a machine. OpenCV comes with many versions supports many languages like C, Python, C++.

Camera Orientation: It is important to carefully choose the direction in which the camera points to permit an easy choice of background. The two realistic options are to point the camera towards a wall or towards the floor (or desktop). However since the lighting was single overhead bulb, light intensity would be higher and shadowing effects least if the camera was pointed downwards.

Camera Specification: We are using Intex Night Vision 16 MP Webcam. The Intex 16 MP Webcam comes with Night Vision feature. This webcam gives clear video imaging and can work even in the darkness. The night vision gives good images in the dark also. The upper portion of the webcam is Movable depend on the need. The resolution of captured image is 640x480 having frame rate up to 30fps and The Format of image is RGB 24, 1420.

B. Image Acquisition

Image: An image is defined as a two-dimensional function, F(x,y), where x and y are spatial coordinates, and the amplitude of F at any pair of coordinates (x,y) is called the intensity of that image at that point. When (x,y), and amplitude values of F are finite, we call it a digital image. In other words, an image can be defined by a two-dimensional array specifically arranged in rows and columns.Digital Image is composed of a finite number of elements, each of which elements have a particular value at a particular location. These elements are referred to as picture elements, image elements and pixels. A Pixel is most widely used to denote the elements of a Digital Image.

Image Matrix: Images are represented in rows and columns we have the following syntax in which images are represented:

$$f(x,y) = \begin{bmatrix} f(0,0) & f(0,1) & f(0,2) & \ldots & f(0,N-1) \\ f(1,0) & f(1,1) & f(1,2) & \ldots & f(1,N-1) \\ \vdots & \vdots & \vdots & & \vdots \\ f(M-1,0) & f(M-1,1) & f(M-1,2) & \ldots & f(M-1,N-1) \end{bmatrix}$$

The right side of this equation is digital image by definition. Every element of this matrix is called image element , picture element , or pixel.

Image Acquisition: The first step of Image Acquisition is of acquiring an image during runtime through integrated camera and while acquiring these images will be stored in the directory after they are captured and the recently captured image will be acquired and that image will be compared with images stored for specific letter in the database and the comparison will give the gesture that was done and the translated text for the following gesture. The images will be captured through basic code of opening a web cam through OPENCV and then capturing the image through frames per second which will be stored in another directory where all the inputs images are stored in another directory and the recent captured image is picked up and the comparison with given set of images are made.

Capture Histogram: Histogram solves many problems related to images. It is possible to differentiate between poorly exposed images and perfect images with the histogram. So generating histogram from the captured image is necessary.

C. Image Recognition and Conversion

Hand Contour: Finding hand contour is important to find palm of the hand and it is based on OpenCV method cv2.contourarea(). It finds contour area of hand to further feed to model.

RGB Colour Recognition: Basically, any color image is a combination of red, green, blue colors. An important trade-off when implementing a computer vision system is 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).Although using intensity alone (black and white) reduces the amount of data to analyze and therefore decreases processor load it also makes differentiating skin and markers from the background much harder (since black and white data exhibits less variation than colour data). Therefore it was decided to use colour differentiation. Further maximum and minimum HSL pixel colour values of a small test area of skin were manually calculated. These HSL ranges were then used to detect skin pixels in a subsequent frame (detection was indicated by a change of pixel colour to white). Mathematical Expression: Colour Calibration

In order to automatically calculate the colour ranges (1), an area of the screen was demarcated for calibration (2). It was then a simple matter to position the hand or marker (color rings) within this area and then scan it to find the maximum and minimum RGB values of the ranges(3). A formal description of the initial calibration method is as follows: The image is a 2D array of pixels:

$$\bar{I}(x,y) = \begin{pmatrix} r(x,y) \\ g(x,y) \\ b(x,y) \end{pmatrix} \qquad \ldots\ldots\ldots\ldots\ldots(1)$$

The calibration area is set of 2D points:

$$\bar{J} = \{\bar{x}_1 \ldots \bar{x}_n\}$$
$$\text{where } \bar{x}_i = (x,y) \qquad \ldots\ldots\ldots(2)$$

The color ranges can then be defined for this area:

$$\begin{aligned} r_{max} &= \max_{\bar{x} \in \bar{J}} r(\bar{x}) & r_{min} &= \min_{\bar{x} \in \bar{J}} r(\bar{x}) \\ g_{max} &= \max_{\bar{x} \in \bar{J}} g(\bar{x}) & g_{min} &= \min_{\bar{x} \in \bar{J}} g(\bar{x}) \\ b_{max} &= \max_{\bar{x} \in \bar{J}} b(\bar{x}) & b_{min} &= \min_{\bar{x} \in \bar{J}} b(\bar{x}) \end{aligned} \qquad \ldots\ldots\ldots\ldots(3)$$

A formal description of skin detection is then as follows,

The skin pixels are those pixels(r, g, b) such that,

$$\begin{aligned} &((r \geq r_{min}) \& (r \leq r_{max})) \& \\ &((g \geq g_{min}) \& (g \leq g_{max})) \& \\ &((b \geq b_{min}) \& (b \leq b_{max})) \end{aligned}$$

Call this Predicate S(r, g, b)

The set of all skin pixels location is then:

$$L = \{\bar{x} \mid S(r(\bar{x}), g(\bar{x}), b(\bar{x})) = 1\}$$

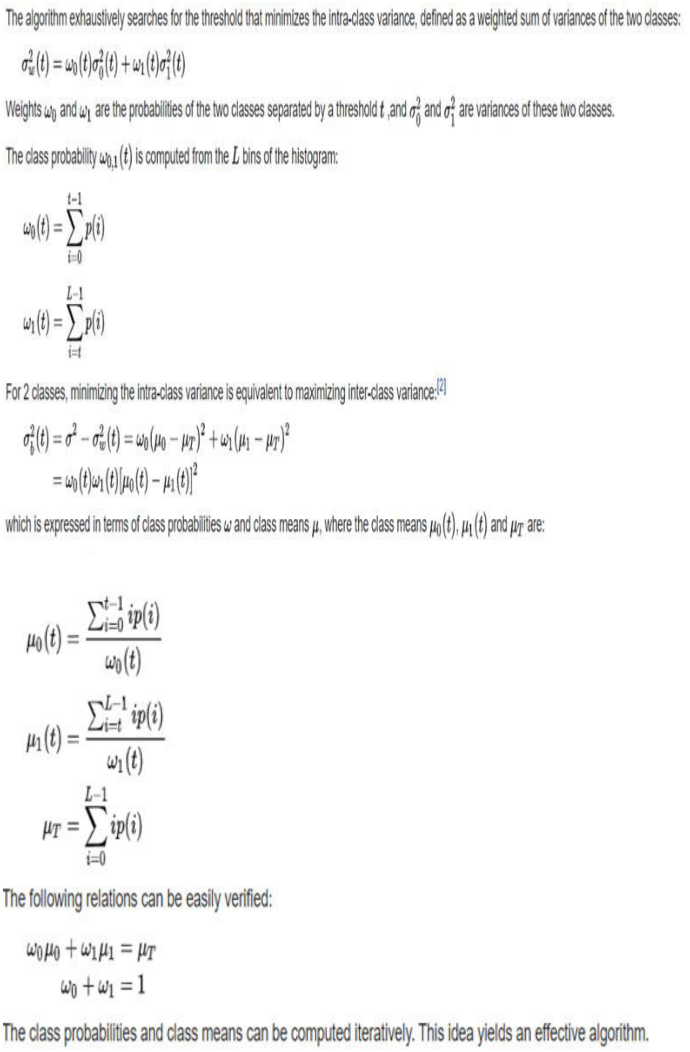
Using this method skin pixels were detected at the rate of 15fps on 2.00 GHz laptop

Color image to Binary image conversion: To convert any color to a grayscale representation of its luminance, first one must obtain the values of its red, green, and blue (RGB) primaries. Grayscale or grayscale digital image is an image in which the value of each pixel is a single sample, that is, it carries only intensity information. Images of this sort, also known as black and white, are composed exclusively of shades of gray, varying from black at the weakest intensity to white at the strongest. A binary image is a digital image that has only two possible values for each pixel. Typically the two colors used for a binary image are black and white though any two colors can be used. The color used for the object in the image is the foreground color while the rest of the image is the background colour. Until now a simple RGB bounding box has been used in the classification of the skin and marker pixels.

D. Thresholding

Thresholding is the simple method of image segmentation. In this method we convert the RGB image to Binary image. Binary image is digital image and has only two values (0 or 1). For each pixel typically two colors are used black and white though any two colors can be used. Here, the background pixels are converted into black color pixels and pixels containing our area of interest are converted into white color pixels. It is nothing but the preprocessing. For thresholding we are using the OTSU thresholding method-

Mathematical Expression: OTSU Method



The algorithm exhaustively searches for the threshold that minimizes the intra-class variance, defined as a weighted sum of variances of the two classes:

$$\sigma_w^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t)$$

Weights ω_0 and ω_1 are the probabilities of the two classes separated by a threshold t ,and σ_0^2 and σ_1^2 are variances of these two classes.

The class probability $\omega_{0,1}(t)$ is computed from the L bins of the histogram:

$$\omega_0(t) = \sum_{i=0}^{t-1} p(i)$$

$$\omega_1(t) = \sum_{i=t}^{L-1} p(i)$$

For 2 classes, minimizing the intra-class variance is equivalent to maximizing inter-class variance:[2]

$$\sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = \omega_0(\mu_0 - \mu_T)^2 + \omega_1(\mu_1 - \mu_T)^2$$
$$= \omega_0(t)\omega_1(t)\left[\mu_0(t) - \mu_1(t)\right]^2$$

which is expressed in terms of class probabilities ω and class means μ, where the class means $\mu_0(t)$, $\mu_1(t)$ and μ_T are:

$$\mu_0(t) = \frac{\sum_{i=0}^{t-1} ip(i)}{\omega_0(t)}$$

$$\mu_1(t) = \frac{\sum_{i=t}^{L-1} ip(i)}{\omega_1(t)}$$

$$\mu_T = \sum_{i=0}^{L-1} ip(i)$$

The following relations can be easily verified:

$$\omega_0\mu_0 + \omega_1\mu_1 = \mu_T$$
$$\omega_0 + \omega_1 = 1$$

The class probabilities and class means can be computed iteratively. This idea yields an effective algorithm.

E. Feature Extraction and Orientation Detection

For any of the object there are many features, interesting points on the object, which can be extracted to provide a “feature” description of the object. SIFT image features gives a set of features of an object which are not affected by many of the complications experienced in other methods, like object scaling and rotation. SIFT approach, for generation of image feature, takes a picture and transform it into a “big collection of local feature vectors”. Each of the feature vectors never changes to any of scaling, rotation or translation of the image. It will take the input of hand movement in any of the form or any orientation the gesture will be detected by the described section of feature extraction as the SIFT algorithm also includes the orientation assignment procedure.

F. Gesture Recognition

Finally when the whole process is complete the application will be then converted into its recognized character or alphabet from the gesture which might be helpful to be understood in layman’s language. The following process contain passing out the 1- dimensional array of 26 character corresponding to alphabets has been passed where the image number stored in database is provided in the array. For this purpose we are using the Naïve Bayes Classifier that follows Linear Classification.

Mathematical Expression: Naïve Bayes Classification

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.
P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.
P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.
P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

G. Text to Speech

When the character gets selected based on recognized sign using speech conversion, respective text is converted to speech. This part comes under Artificial Intelligence where once the template matching operation becomes successful the matched image is then translated into text and audio format. For this purpose, predefined methods are used for conversion.

3. CONCLUSIONS

It’s well known that inability to speak and hear is one major challenge for human race. The system will provide an interface that can easily communicate with deaf people by Sign Language Recognition. The system is not only can apply in family environment, but also can apply in public. For the

Social use this system is very helpful for deaf and dumb people. We will build simple gesture recognizer based on OpenCV toolkit and integrated it into Visionary framework. Our project works towards bridging the gap by introducing an inexpensive computer in the communication path so that the sign language can be automatically captured, recognized and translated to speech for the benefit of blind people. In this project, we have used Naïve Bayes Classification method that provides maximum accuracy by less noise distortion and creating multiple replicas of each gesture. The webcam used is equally a powerful device and the histogram captures the necessary image strategically. In the other direction, speech must be analyzed and converted to either sign or textual display on the screen for the benefit of the hearing impaired. Also for people with partial voice disabilities the speech recognition system will do the further enhancement in speech systems for the disable people.

4. RESULT

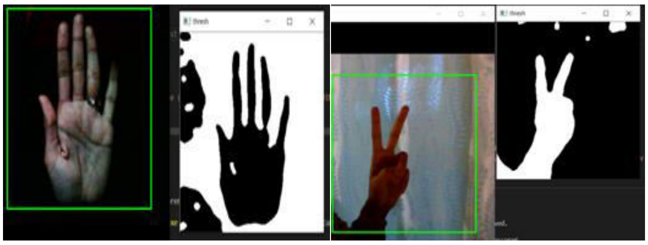


Fig (a). Creating a gesture

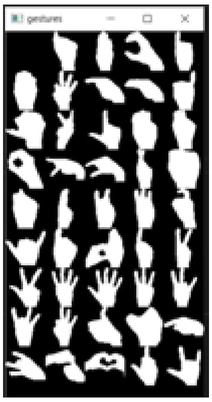


Fig (b). Displaying all created gestures

5. FUTURE SCOPE

Motivations to carry out further research in order to develop enhanced version of the proposed system. System would be able to communicate in both directions i.e. It will have the capability to translate normal languages to hand gestures successfully. The image processing part of the system will also be modified to work with every possible environment.A challenge will be to recognize signs that involve motion.

The project involves distinguishing among different alphabets of English language. Future work may include recognition of all the English alphabets & numbers. Further, we may move on to recognizing words, from as large a dictionary as possible, from Indian Sign Language.

ACKNOWLEDGEMENT

I would like to thank Fr (Dr). John Rose S.J. (Director of Xavier Institute of Engineering,Mahim,Mumbai) for providing me with such an environment so as to achieve goals of my project and supporting me constantly.I express my sincere gratitude to our Honorable Principal Mr. Dr. Y.D.Venkatesh for encouragement and facilities provided to me.I would like to place on record my deep sense of gratitude to Prof. ChhayaNarvekar, Head of Department Of Information Technology, Xavier Institute of Engineering, Mahim, Mumbai, for her generous guidance help and useful suggestions.With deep sense of gratitude I acknowledge the guidance of our project guide Prof. ChhayaNarvekar. The time-to-time assistance and encouragement by her has played an important role in the development of our project.

REFERENCES

[1] Denise Powell, 'A Case Study of Two Sign Language Interpreters Working in Post-Secondary Education in New Zealand', International Journal of Teaching and Learning in Higher Education,,2013 Volume 25, Number 3, 297-304.

[2] C.W.Ng and S.Ranganath , 'Static Hand Gesture Recognition System using Convolutional Neural Networks', IJSRD - International Journal for Scientific Research &Development,Vol. 6, Issue 01, 2018 ,ISSN (online): 2321-0613.

[3]N.Tanibata, 'Automated Extraction of Signs from Continuous Sign Language Sentences using Iterated Conditional Modes', Journal of Engineering and Applied Sciences, August 2014.

[4] H. K. Nishihara, 'Recognition of American Sign Language using Image Processing and Machine Learning', International Journal of Computer Science and Mobile Computing, Vol. 8,Issue 3,March 2019,pp:352-357.

[5] Jagdish L. Raheja, 'Android based Portable Hand Sign Recognition System', International Journal of Computer Trends and Technology (IJCTT) V40(3):165-171, October 2015. ISSN:1651-1656.

[6] Shweta S.Shinde, Rajesh M. Autee and Vitthal K. Bhosale, 'Sign Language Recognition for Deaf & Dumb', International Journal of Advanced Research in Computer Science and Software Engineering 3(9), September - 2013, pp. 103-106.