

Real Time Conversion of Hand Gestures to Speech using Vision Based Technique

S. G. Mundada, K. Khurana, A. Bagora

Abstract: Sign Language is one of the most common approaches of communication usually used by people having hearing and speech impairment. These languages consist of well-defined set of gestures or pattern and sequence of actions that conveys meaningful words and sentences. The paper presents different algorithms and techniques for automation of single hand gesture detection and recognition using vision based methods. The paper uses basic structure of hand and properties like centroid for detecting the pattern formed by the fingers and thumb and assigning code bits i.e. converting each gesture into a set of 5 digits representation and motion is detected using movement of centroid in each frame. The paper uses techniques like K-means Clustering or Thresholding for background elimination; Convex Hull or a proposed algorithm for peak detection and text to speech API for conversion of words/sentences corresponding to gestures to speech. Combinations of different techniques like thresholding and convex hull or Clustering and proposed algorithm is implemented and results are compared.

Index Terms: K-Means Clustering, Convex Hull, Thresholding, Sign Language, Hand Gestures

I. INTRODUCTION

Sign Language is the non-verbal way of communication usually used by people having hearing impairment. In sign language, gestures are considered to be any specific patterns or movements of the hands, face or body to make certain expressions [10]. Gestures can be expressed with the help of facial expressions, limb movements or any meaningful bodily state [8]. Gesture can be static or dynamic. A static hand gesture configuration consists of a particular pattern of hand which is not moving whereas dynamic hand gesture configuration may contain multiple hand gestures or single hand gesture in motion or multiple hand gestures in motion. The paper aims at automating the process of gesture detection, recognition and converting the gesture to speech using vision-based techniques. The paper focuses on converting dynamic gestures to speech by using K-means clustering algorithm or thresholding for background elimination, convex hull or a proposed algorithm for finding the peaks and using centroid trajectory to find the directions where the hand is moving in the frame. Thus, determining the gesture in the frame and converting the word corresponding to gesture into speech.

Revised Manuscript Received on July 05, 2019

S. G. Mundada, Department of Computer Science Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur

K. Khurana, Department of Computer Science Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur

A. Bagora, Department of Computer Science Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur

II. LITERATURE REVIEW

Hand Recognition: Hand Recognition is done using various techniques. Hand modelling (model-based approach) [3]: this method tries to infer the pose of the palm and joint angles, is ideal for interaction in virtual reality environments. A typical model-based approach may create a 3D model of a hand by using some kinematics parameters and projecting its edges onto a 2D space. Hardware-based method: [5] [10] Glove-based hand gesture recognition methods detects the gesture using the data of the hand position and orientation provided by the glove. S. Oniga & I. Orha [5] used a bracelet that captures the movement of the hand using accelerometers and Field Programmable Gate Array (FPGA) and then modeled, trained and simulated the desired network using Neural Network Toolbox. The method gives accurate results and is expensive. Vision-based methods: [12] This process involves the usage of image processing techniques to identify the gesture, the image is converted to binary. This is done using concepts based on image thresholding techniques. The coordinates of the contour with maximum area enclosed is considered to be the hand producing gesture. The points are identified as tips or valleys based on the magnitude of the angle. The tips are correlated with the centre of the palm to identify the respective gesture. Skin color is mostly used for hand recognition various color models suggested in [13] are cbr color space, RGB color Space, HS (Hue, Saturation value) color space, Normalized RGB & HSV. Trong-Nguyen, Nguyen [3] used skin color, Md Iqbal Quraishi, Krishna Gopal Dhal, J Paul Choudhury [4] used YCbCr color, S. Bhowmick, S. Kumar and A. Kumar [10] have a hybrid HSV+YCbCr colour model for detecting the hand, [3] filling the gaps and mapping to stored dataset. [4] used neural networks for identifying gesture. M. Murugeswari, S. Velucham [7] suggested hand gesture recognition system consists of two stages: the training and testing stage. In training stage build the cluster model and SVM classifier model and in testing stage this models used to recognize the gesture. [14] uses neural network for classification of gesture, [15] uses Softmax classifier for classification; after using background elimination and finding the hand contour [6] Being a newly developed distance measuring hardware, the depth camera gives a depth image that could reflect the 3d feature directly, different parts of the object can be separated. Another method that can be used for hand detection is background subtraction. Rather than detecting the

Real Time Conversion of Hand Gestures to Speech using Vision Based Technique

ROI, the background is subtracted by applying clustering algorithms. We use K-means algorithm for the same. Fang et. al. [2] uses extended Adaboost method for hand detection and hand segmentation is done by collecting the color of the hand from neighborhood of features mean position. They further use the scale-space feature detection to detect blob and ridge structures, i.e. palm and finger structures. R. Agrawal & N. Gupta[9] did gesture recognition by separating the palm from the hand by creating a palm mask, i.e. eroding the fingers away from the hand mask, by successive erosion and dilation morphological operations. Then by using convexity defects, the center of the hand and the number of fingers are detected. M. Panwar [1] uses the concept of code bits, with binary values and maps these with the corresponding meaning of the gesture. This concept is extended by using double threshold holding, and generate a sting of numbers, i.e. 0, 0.5 and 1 to map them to the corresponding meaning of the gesture.

III. PROPOSED METHODOLOGY

Proposed technique consists of two phases, Phase I and Phase II. Phase I consists of detecting the static gestures, converting the gesture into its equivalent code bits and mapping Code bits to the word corresponding to these gestures and then the word is spoken using text to speech modules[8].

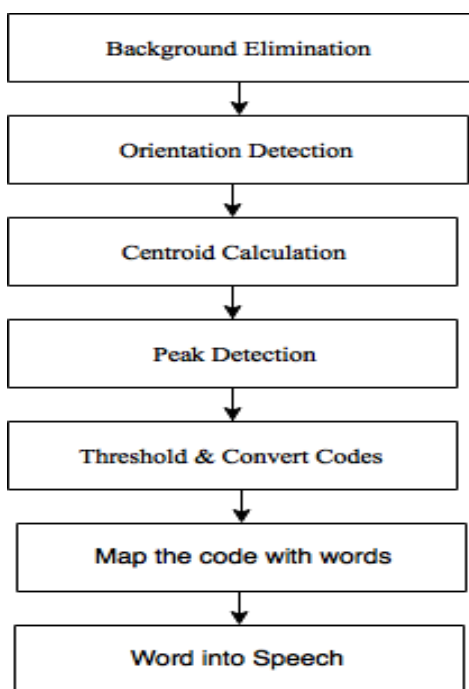


Fig 1: Conversion of static hand gestures to speech

Phase II contains implementation of detection of gestures in a sequence of frames, finding the code bits for each gestures and finding the direction of the gesture in the subsequent frames and mapping these code bits, direction/directions of a gesture or multiple gestures to the word or a sequence of words or sentences.

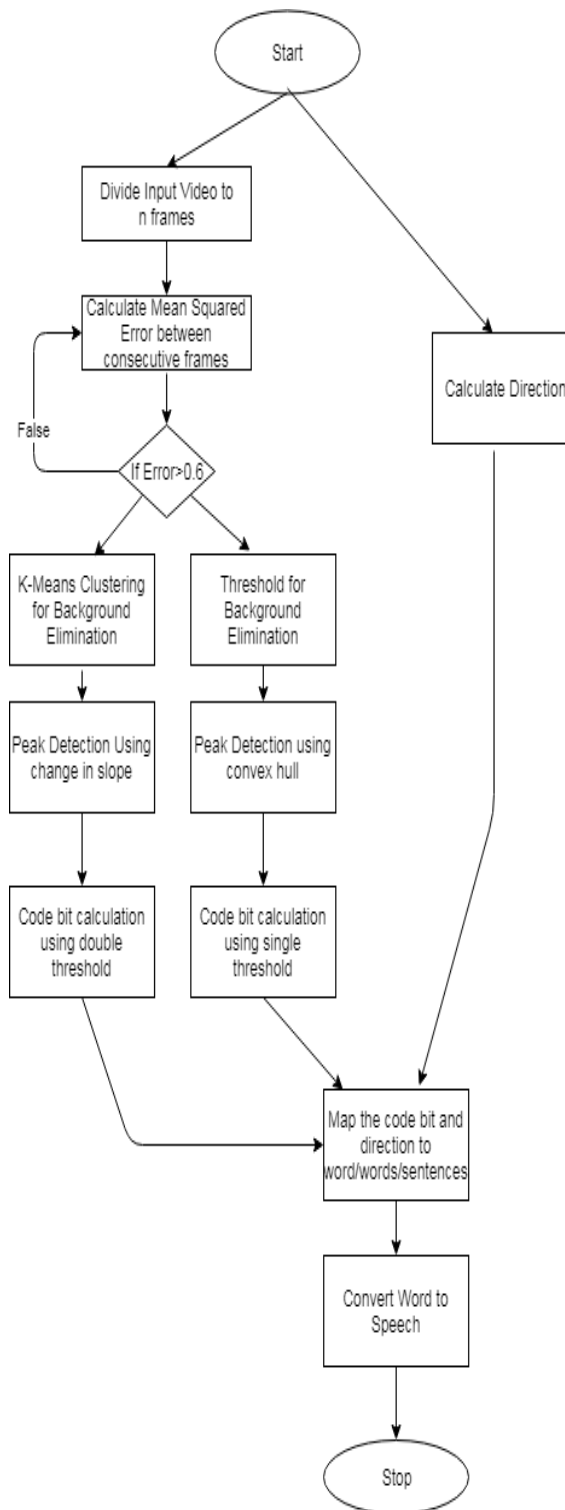


Fig 2: Conversion of dynamic hand gestures to speech

IV. SYSTEM DESIGN

Following are the steps to implement proposed system

1. Conversion of video into frames
2. Detection of gesture and code bits
3. Determination of trajectory of centroid and calculation of Direction
4. Mapping of code bits, direction to words

Real Time Conversion of Hand Gestures to Speech using Vision Based Technique

1. Conversion of video into frames

Video is taken as an input. Each frame has an unique identifier. F_i is calculated as

$$F_i = \text{Frame_id} \% \text{frame_rate}$$

Where,

Frame_id: Unique identifier of each frame

Frame_rate: Number of the frames per second

If $F_i=0$ then

Frame is selected

Else

Frame is not selected

This extracts n frames from the sequence of frames where n is the duration of the video in seconds. Thus, selecting frame at each second. The selected frames are stored and used to find the change in gesture.

2. Detection of gesture and code bits

- a. **Detection of change in gesture:** The stored list of frames is used to find whether the sequence of frames contains only one gesture or multiple gestures. The first frame is passed to the module that determines the code bits of a gesture and then change in gesture is detected by finding the mean squared error between the first frame and the next frame in the stored list of frames. The error is calculated between all the consecutive frames in the list. If the error between two frames is greater than a particular threshold (in this case 60%) then change in gesture is detected, the frame containing the new gesture is passed to the module that converts gesture in a frame to code bits and the code bits are appended to the code bits obtained for the previous gesture this continues till all the selected frames are processed. In case if the sequence of frames contains in only one gesture the error remains below the threshold and no other frame except the first is passed to the module that finds code bits.

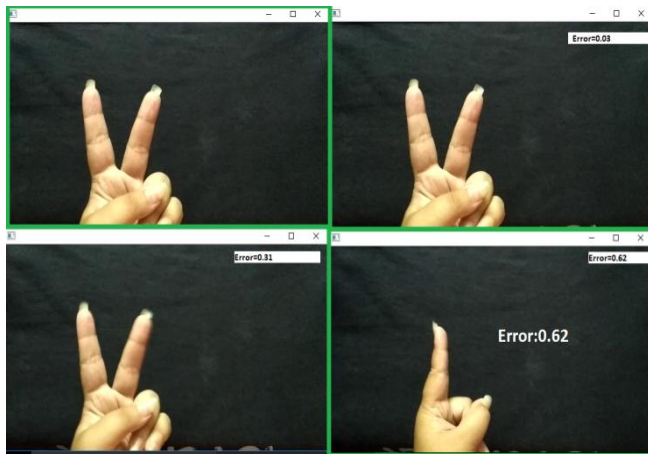


Fig 3: Detection of gesture using mean squared error between frames

As Fig 3 shows a 4 seconds video is converted to 4 frames and the mean squared error is calculated between the frames. First frame is passed to the next module and the error between frames is calculated. Last frame is passed to the module because the error is above threshold (60%) and the change in gesture is detected.

b. Conversion of each gesture to code bits:

1. **Background Elimination:** Each frame contains a particular gesture; the technique of background elimination separates the hand gesture from the background. There are two methods that are used in implementation
 - i. **K-means clustering:** The frame is converted into a list of RGB values and passed to the algorithm. The algorithm forms two clusters by labeling each value in the list either 0 or 1; based on these labels a binary image is formed. If the label is 0 then the RGB values assigned is (0, 0, 0) if label is 1 values assigned is (255,255,255). Thus, the frame is converted into a binary image which separates hand gesture from the background using labels assigned to each pixel in the list.

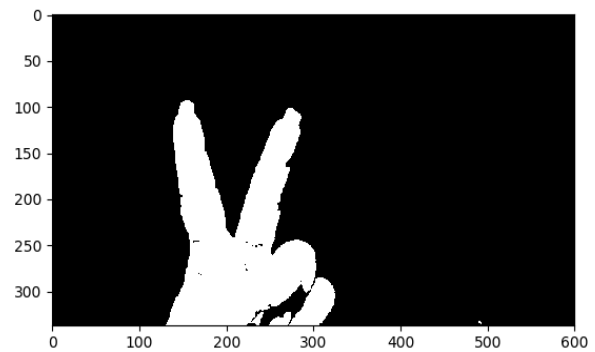


Fig 4: Background Elimination using K-means Clustering

- ii. **Thresholding:** The technique of thresholding is simple and effective; used in image processing to isolate objects from an image, based on a threshold value. The frame is converted to a gray scale image and thresholding is done, the method converts a gray scale image to binary by assigning either 0/1 based on whether the value of the pixel is greater than or less than the threshold value. If the value of the pixel is greater than the threshold value then 1 is assigned; else zero is assigned. This technique thus separates hand region from the background.

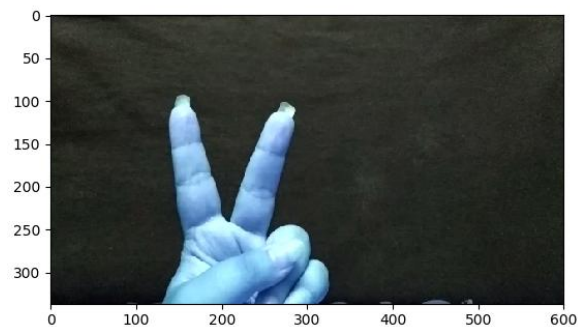


Fig 5: Background Elimination using Thresholding. After getting the binary images we would find the contour corresponding to the hand region.

2. **Peak Detection:** Peaks are the fingertips of the hand which we need to



Real Time Conversion of Hand Gestures to Speech using Vision Based Technique

detect to find whether and which fingers are open or close. To detect the peaks, we use the contour of the hand obtained in the previous step. Two methods are used for finding the peaks.

- i. **Convex Hull Method:** This method is used to find the smallest convex polygon containing all the points from the frame. The method finds the local maxima's in the contour of the hand which represents the fingertips of the hand.

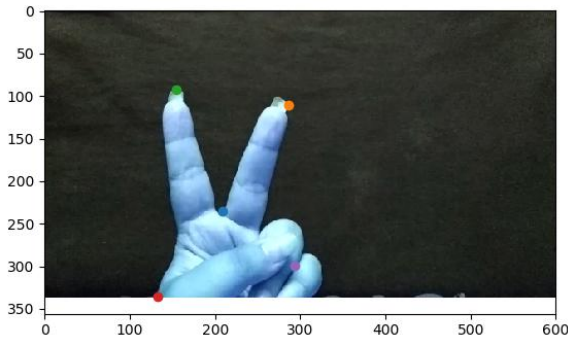


Fig 6: Peak Detection using convex hull method

- ii. **Finding Peaks by using slope:** In this method we calculate the slope between the consecutive points in the contours. If the slope changes from positive to negative sharply and is negative for some iterations then the point of change is marked as peak.

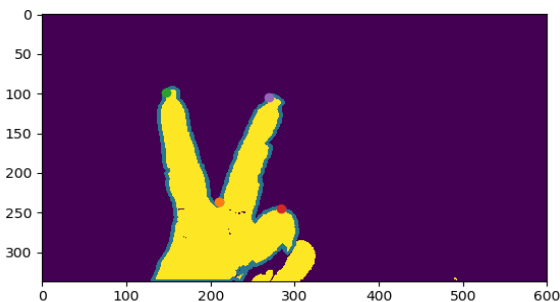


Fig 6: Peak Detection using proposed algorithm method

Let (x_i, y_i) , (x_{i+1}, y_{i+1}) be points on the contour then, slope is calculated by,

$$S = (y_{i+1} - y_i) / (x_{i+1} - x_i) \quad (1)$$

For detecting fingertips slope calculated is used. Slope increases initially and then sharply turns negative indicating local maxima. Hence, the point at which the slope turns from positive to negative is stored as peak.

3. **Centroid Calculation:** Centroid of the hand region can be calculated using the image moment. Image moment is a certain particular weighted average of the image pixels' intensities is calculated using the formula in Opencv; an Image Processing library,

$$M_{ij} = \sum \sum x_i y_j I(x, y) \quad (2)$$

where M_{ij} is image moment, $I(x, y)$ is the intensity at coordinate (x, y) .

Moments are used to find various properties of the region of interest such as centroid, area.

To calculate centroid's x and y coordinates formula used in Opencv is,

$$cX = (M["m10"] / M["m00"]) \quad (3)$$

$$cY = (M["m01"] / M["m00"]) \quad (4)$$

$$\text{Centroid} = (cX, cY) \quad (5)$$

where cX is the x coordinate of the centroid and cY is the y coordinate.

4. **Assigning Code bits:** The distance between the peaks and centroid is calculated. Let the distance between centroid and peak be 'd'. Code bits can be assigned using two methods.

Single Thresholding: In this method a threshold is used to assign a code bit. Let T be the threshold

If $d \geq T$ then
code bit = 1

Else
code bit = 0

This method is used in combination of with convex hull method used for peak detection as convex hull detects only open or closed fingers. Code bit takes a value 1 if finger is open using single threshold and if finger is closed then it takes a value 0.

Double Thresholding: In this method two thresholds are used. Let the two thresholds be Upper Threshold U_t , Lower Threshold L_t

If $d > U_t$ then:

Code-bit = 1

Else If $d < L_t$ then:

Code-bit = 0

Else $L_t \leq d \leq U_t$ then:

Code-bit = 0.5

Two thresholds divide the hand region in three parts if the finger is closed it falls beyond the lower threshold, if it is open distance is more than the upper threshold, if the finger is half folded then the distance falls in the range between the thresholds. Thus, by increasing number of thresholds we get more information about the state of the fingers.

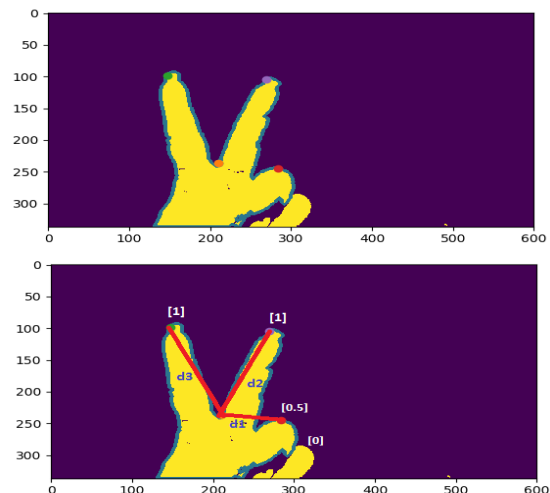


Fig 7: Finding code bits

As shown in Fig 7 the distance between the peak and centroid are calculated using Euclidean Distance.



Real Time Conversion of Hand Gestures to Speech using Vision Based Technique

The above algorithm of double and single thresholding are applied and code bits are calculated. In the Fig 7, If single thresholding is applied then d1 and d2 are above threshold hence assigned a code bit “1”, d3 and d4 is less than 1 threshold hence the code bit is “0”. Using double thresholding code bit for Fig 7 is [1,1,0,0].

If double thresholding is applied then d1 and d2 are above upper threshold hence assigned a code bit “1”, d3 falls between upper and lower threshold hence code bit assigned is “0.5” and d4 is less than lower threshold hence the code bit is “0”. Using double thresholding code bit for Fig 7 is [1,1,0.5,0].

3. Finding trajectory of centroid and calculating direction: Motion of centroid of the hand is used to find the direction in which the hand moves throughout the video. The algorithm runs for each frame in the video finds the centroid of the frame and compares the centroid frame j with the centroid of the frame $j+1$. Change in position of centroid is calculated as

$$dx = Cx_{j+1} - Cx_j \quad (6)$$

$$dy = Cy_{j+1} - Cy_j \quad (7)$$

where, Cx_i , Cy_i are the x and y coordinates of the centroid in the frame i .

If there is significant change in any of the direction x or y then the direction is assigned using the magnitude of the change and the direction of change.

If the change is in only x-direction:

If the magnitude is positive (>0) then:

Direction=Right

else

Direction=Left

If the change is in only y-direction:

If the magnitude is positive (>0) then:

Direction=Up

else

Direction=Down

If there is significant change in both x and y directions then:

Both directions are appended

The direction is temporarily stored and compared with the direction calculated in the next pair of frames .If the hand moves in a particular direction throughout the video then the direction is not changed and contains only one direction.

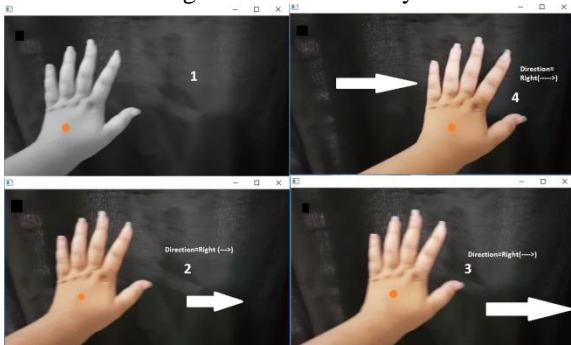


Fig 8: Finding direction using centroid of the hand.

If other direction is encountered then the new direction is appended thus forming a sequence of directions.

4. Mapping code bits, direction to words:

Calculation of code bits and direction can be done simultaneously by running two different threads one for calculating code bits and other for calculating the direction in which the hand moves throughout the video it may be unidirectional or a sequence of directions. By using the code bits and directions the algorithm maps both code bit and direction to corresponding word or sequence of words or sentences as output. The output would then be passed to Text to speech module/API to convert it into speech.

V. RESULTS

Table 1 shows the results obtained by using k-means clustering for background elimination , finding peaks through slope, calculation of code bits and mapping of code bits to word for static gestures .The input is different images that contain static one hand gestures and the output produced by the algorithm is shown in the table.

Input Image	Gesture	Code	Word
		11100	Coin
		1 0.5 0.5 1 0	Rock and Roll
		0 0 1 1 0	Peace
		0.5 0.5 1 1 1	Vehicle

TABLE 1[8]

DEMONSTRATION OF THE GESTURE RECOGNITION MODULE OF THE ALGORITHM

Description of each gesture:

1. The first gesture has all the fingers and thumb open moving in one direction i.e. right in the video which means “swipe”.

2. The second gesture has only one finger open and fist is closed; the hand moves in circular direction along y-axis in the video which means “always”.

3. The third gesture has two fingers open and the hand taps at one end moves to right and then



Real Time Conversion of Hand Gestures to Speech using Vision Based Technique

tap again in the video which means “Hard of Hearing”.

4. The fourth gesture has thumb and last fingers open the motion is similar as above and it means “Oh! I see”

5. The fifth gesture contains a sequence of gesture .The first gesture have two fingers open that means “See You” and then the hand moves right and gesture changes i.e. only one finger is open which means “Later”.

Table 2 shows the results obtained for dynamic hand gestures using k-means clustering and proposed algorithm of finding peaks using the slope .Video is passed as an input where in the gestures are not static ,but are moving in subsequent frames . Change in gestures and motion of the hand throughout the video is calculated using the proposed algorithm in the paper. Results show the code bits and direction corresponding to each gesture and the word spoken.




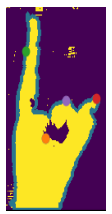
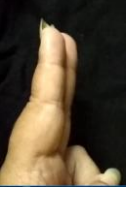
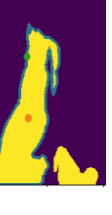




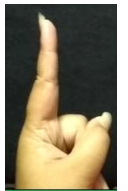

Input Frame	Gesture Detected	Code bits	Direction	Words Sentences
		11111	Right	Swipe
		110.50 0	Circular	Always
		11000	Down Right Down	Hard Of Hearing
		10001	Down Right Down	Oh! I see
		11000	Down Right Down	See You Later
		10000		

TABLE 2 DEMONSTRATION OF DYNAMIC HAND GESTURE RECOGNITION MODULE USING CLUSTERING AND CHANGE IN SLOPE

Table 3 shows the results obtained for dynamic hand gestures using thresholding and convex hull for finding peaks.

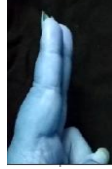







Grayscale Frame	Peak Detection	Code bits	Direction	Words Sentence
		11000	Down Right Down	Hard Of Hearing
		10001	Down Left Down Left	Oh! I see
		11000	Down Right Down	See You Later
		10000		

TABLE 3 DEMONSTRATION OF DYNAMIC HAND GESTURE RECOGNITION MODULE USING THRESHOLDING AND CONVEX HULL

VI. COMPARATIVE ANALYSIS

In the paper, basically two methods for hand gesture recognition are mentioned. First method is for static hand gesture recognition and second one is for dynamic hand gesture recognition.

First one uses clustering for background elimination and changes in slope for peak detection and the other method uses Thresholding and convex hull for Image segmentation and peak detection respectively. Each technique has its pros & cons.

Technique	Advantages	Drawbacks
Thresholding	<ul style="list-style-type: none"> Simple, Fast 	<ul style="list-style-type: none"> Sensitive to intensity of light
K-means Clustering	<ul style="list-style-type: none"> Independent of Image intensity. Forms clusters at run time. 	<ul style="list-style-type: none"> Takes background as Region of Interest (ROI).



Real Time Conversion of Hand Gestures to Speech using Vision Based Technique

Convex Hull	<ul style="list-style-type: none"> • Easy to implement • Every point in the contour needn't be accessed 	<ul style="list-style-type: none"> • Does Not detect fingers that are half folded.
Peak Detection using slope	<ul style="list-style-type: none"> • Detects more number of fingertips as compared to convex hull. • Also detects half folded fingers 	<ul style="list-style-type: none"> • Need to access each point in the contour of ROI.

TABLE 4 :
COMPARISON OF TECHNIQUES USED IN THE
PROPOSED ALGORITHM

Both the methods work better when compared with traditional methods in terms of accuracy.

VII. CONCLUSION

Different threads are used for detecting gestures and finding direction. Thus, time taken for the process is minimal. The limitation of the method is it works for only single hand gestures and the dataset used is small. Machine Learning and Artificial Intelligence algorithm can be used to extend the algorithm for larger, real-time datasets.

REFERENCES

1. M. Panwar (Centre for Development of Advanced Computing, Noida), 'Hand Gesture Recognition based on Shape Parameters'
2. Y. Fang et. al. 2007, 'A REAL-TIME HAND GESTURE RECOGNITION METHOD'
3. T. Nguyen & H. Huynh, 'Static Hand Gesture Recognition Using Artificial Neural Network', Journal of Image and Graphics, Volume 1, No.1, March, 2013
4. M. Quraishi et. al., 'A Novel Human Hand Finger Gesture Recognition Using Machine Learning', 2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing
5. S. Oniga & I. Orha, 'Intelligent Human-Machine Interface Using Hand Gestures Recognition'
6. L. Chen et. al., 'A Survey on Hand Gesture Recognition', 2013 International Conference on Computer Sciences and Applications
7. M.Murugeswari(PG Scholar, Communication Systems, Anna University,Tamil Nadu) ,S.Veluchamy (Assistant Professor, Communication Systems, Anna University,Tamil Nadu), 'Hand Gesture Recognition system for Real-Time Application', 2014 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)
8. R. Agrawal & N. Gupta, 'Real Time Hand Gesture Recognition for Human Computer Interaction', 2016 IEEE 6th International Conference on Advanced Computing
9. Sourav Bhowmick et. al, 'Hand Gesture Recognition of English Alphabets using Artificial Neural Network', 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)
10. Ayushi Shrivastava, Radhika Agrawal, Nidhi Budhraj, Ruchi Panpaliya, S. G. Mundada, 'A Technique for Hand Gesture Recognition on Real Time Basis',2018, International Journal of Computer Applications,(Volume 181, Issue 9)
11. Prashanth Suresh, Niraj Vasudevan, Nilesh Ananthanarayanan, 'Computer-aided Interpreter for Hearing and Speech Impaired', 2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks
12. Prof.R.R.Itkarkar, Dr. Anil V Nandi, 'Hand Gesture to Speech Conversion using Matlab', 4th ICCNT 2013 July 4-6, 2013, Tiruchengode, India

13. Sujeet D.Gawande, Prof. Nitin .R. Chopde, 'Neural Network based Hand Gesture Recognition', International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-2, Issue-3)
14. Meiping Tao, Li Ma, 'A Hand Gesture Recognition Model Based on Semi-supervised Learning', 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics

AUTHORS PROFILE



Mrs. Shyamal Mundada is working as an Assistant Professor in Shri Ramdeobaba College of Engineering and Management, Nagpur. Her area of interest include Image Processing, Machine Learning and Real Time Scheduling.



Mrs. Khushboo Khurana is working as an Assistant Professor in Shri Ramdeobaba College of Engineering and Management, Nagpur. Her area of interests include Image, Video Processing and Big Data Analytics.