Talking vs Non-Talking: A Vision Based Approach to Detect Human Speaking Mode

Sadia Afroze
Dept. of Computer Science & Engineering
Chittagong University of Engineering & Technology
Chittagong, Bangladesh
e-mail: sadiacse10@gmail.com

Mohammed Moshiul Hoque
Dept. of Computer Science & Engineering
Chittagong University of Engineering & Technology
Chittagong, Bangladesh
e-mail: mmoshiulh@gmail.com

***Abstract*—Human talking mode detection is an important issue in human-computer interaction. In this work, we propose a method for detecting human talking and non talking mode detection based on supervised machine learning approach. Visual lip information of human is considered as an important clue. Our goal is to develop a method for human talking and non talking mode detection in real time using supervised classification algorithm. We tested our experiment with a single speaker task and compared the results with the previous method. The results show that our approach can obtain a 98.00% accuracy and a fast executed time.**

***Index Terms*—Computer vision; feature extraction; face detection; pattern recognition; evaluation**

I. Introduction

Human speaking mode is a way of communication where human can communicate with other agent like as human, robot,sound sensor device and so on. Human speaking mode detection is an important cue in human-computer interaction (HCI), hypo-vigilance analysis, and fatigue detection. In the recent year, There are many artificial intelligence (AI) system are developing based on human voice command command such as : Arduino Drone, robot , computer game, Smart Speaker and some medical instruments. The voice based system sometime failed to work due the surrounding noise, In that scenario the speaking mode system may improve the service quality of voice based AI system. As the lip movement is a key parameter in interactive communication for talking and non-talking mode detection,so we consider visual based approach for detecting human speaking mode. Detecting human speaking mode can be applicable in several applications. In the case of video conferencing and in the hypo-vigilance analysis, the visual analysis of lip movement is used to aid recognizing the talking state of the human.We can implement our system in driving vehicles for detecting driver's talking mode.As most of the accident occurs due to a conversation with the driver while driving. So, talking or non-talking mode detection can give alert to the driver which may reduce the serious road accident.Various intelligent system work on human voice.But there is problem of noise when those system only consider human voice. If we integrate the vision based speaking mode detection with voice based system then it can overcome the limitation of voice based system. In our work, we detect mouth of human from face region and identify as the lip movement is in talking or non-talking mode. We recognize this mode in real time using a web-cam which will be set in front of a person. First of all, the system detects the human head. After detecting the human head the system extracts the mouth region from the face. The system gives four salient points of mouth region. These feature points, namely, top of the upper lip, bottom of the lower lip, left and right mouth corners. Then the system performs statistical analysis and support vector machine (SVM) to detect the talking and not- talking state of a human face.

II. Related Work

There are several methods which study visual information of leap to recognize the human talking or non talking state. In our work, we recognize human talking and non-talking state in real time. We use visual information of human for leap motion analysis in order to know human mouth state. Bendris et al. [1] presented a method based on lip motion to separate talking and non-talking faces. They used the degree of disorder of pixel directions around the lip to detect the lip motion. They also represent the mouth region as a rectangle because it is very complex to extract the shape of lips due to the quality of appearance. They used Stasm for the facial features detection. But the system is not much reliable with pose changes such as faces looking up or down and expressions such as a mouth wide open or shrink. Khan et al. [2], proposed to detect mouth using a combination of Viola Jones and skin color pixel detection. But this system consumes much time for detecting facial feature and also requires high level of image representation. In the previous work [3],skin-color segmentation filter and edge projection is used for detecting mouth region.But there were some limitations as it could not exactly locate the correct mouth position, especially profile face or light influence. Azim et al. [4], proposed a non-intrusive fatigue detection system based on the video analysis of drivers. In their system, they located the face through Viola-Jones face

978-1-5386-9111-3/19/$31.00 2019 IEEE

detection method to ensure the presence of a driver in the video frame. Then, a mouth window is extracted from the face region, in which lips are searched through spatial fuzzy c-means clustering. But fuzzy c-means is not a deterministic algorithm and it needs large computation time.In the previous work [5], visual lip information from the driver is used for speech recognition in car environment.They used Viola- Jones approach for locating and tracking the driver's lips despite the visual variability of illumination. Most of the work in previous are performed in static image and only detect the mouth region. There is no approach in recognizing human talking mode using mouth contours.In contrast to previous work, we would like to focus on detecting human mouth state means talking and non-talking state of mouth in real time. The geometrical point based lip tracking unable to capture mouth shape where the point is not able to select [6]. Feature extraction is a key term in image processing, the Log-polar Signature [7]is frequency based based feature extraction method where the feature value defend on pixel value. In object and face detection Haar-like features [8] extract the rectangle based feature but not work at low resolution image. Lip tracking or mouth motion tracking is also very important research are in speech recognition [9] and human-robot interaction. Person verication by lip-motion [10] system depend on lip information. In that system lip-motion are used for person verification. There are many research has to be conducted on lip-motion and mouth tracking but each system has some limitations. Now we proposed system where detect the talking or non-talking state using mouth-motion. In our proposed system we select the mouth region using lip key [11] point and extract the feature by histogram of oriented gradients (HOG). The extracted feature feed to the LIBSVM [12] and train with a suitable kernel.

III. PROPOSED FRAMEWORK

In the proposed system, we recognize human talking or non-talking mode in real time. To implement that, the system initially detects the human face and select facial key point and from that point list we select the left and right leaf point and select a rectangle for mouth region. We extract feature from the mouth region using HOG (Histogram of Oriented Gradient) feature extraction method. The system builds a train model for detecting mouth state. We consider that a face is in talking mode if its lips are consecutively open and close for a while. So, we consider fifty consecutive frames where there is successive open close state of mouth. Therefore if we have a frame where the lip is open, then we look back the previous frames if these are in frequent open and close state. By combining these two methods we consider human talking state in real time. Fig. 1 Illustrates the schematic diagram of propose system methodology. Details of procedures are described in the following subsections.

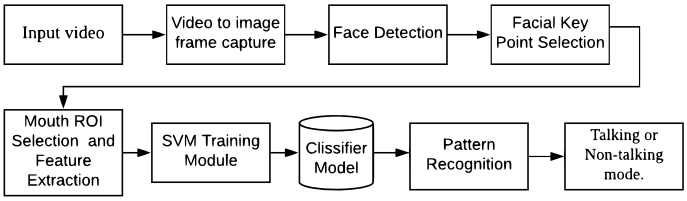

Fig. 1. Schematic diagram of the proposed system.

A. Face Detection and Facial Key Point Selection

We use open source software called dilib [11] for face detection purposes.After face detection we detect the facial land marks using dlib algorithm. Face land marks or key points are used to represent the main regions of the face such as eyes center, eyebrows area, nose tip, mouth region, and jawline. Shape prediction methods are used for facial shape measurement. This software performs facial landmark detection in two steps. At first, the software localizes the face in an image. Then it detects the key facial structures on the face ROI. We are applying a pre-trained HOG model and Linear SVM object detector model for the purpose of face detection. For detecting facial key points a training set of hand-crafted landmarks on an image is used. These images are labeled by human. Mainly the (x, y) coordinate value are annotating by human labour. Now the hand-crafted train data feed to the regression tree based algorithm and the algorithm directly trained the coordinate value using pixel intensity value. In Fig.2. shows the key point with red marked points.

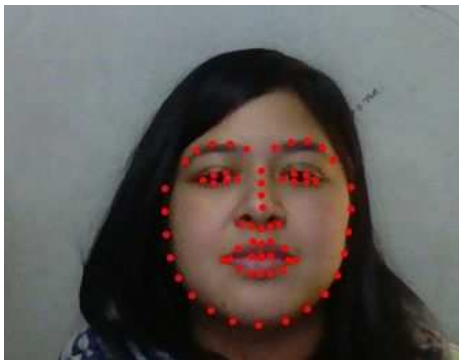

Fig. 2. Facial key point.

For any given image, the pre-trained facial land mark detector estimate the 68 points coordinate value. The value are the key points of any facial structure. Each eye is represented by 6 points (left and right corners, top and bottom eye opening) and each eye brow is represented and each eyebrow is represented by 5 points. The lip contour is described by 20 points. The nose is represented by 9 points and the face contour by 17 points.

B. Lip Region Selection

In Fig.3. represent the detected face with lip region. We detected the human face and select the lip region. We consider four important points from the lip region such as upper and lower points of lip region.These four points prepared the mouth rectangle region.

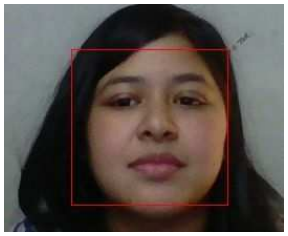

Fig. 3. Lip region selection from detected face rectangle.

The lip region height and weight calculated by the bellow equation.

$$W = lip\_right\_x - lip\_left_x + 4 \tag{1}$$

$$H = lip\_top\_y - lip\_bottom\_y + 4 \tag{2}$$

Here W and H represent the lip region width and height. The left and right side of lip position represent by lip_right_x, lip_left_x, lip_top_y and lip_bottom_y. We use these four points to draw a rectangle to show the lip region. Right side of Fig. 3 shows the selected mouth region from a video frame.

C. Feature Extraction

We perform histogram of oriented gradients (HOG) feature descriptor for detecting mouth state. In the HOG feature descriptor, the distribution of directions of gradients is used as features. The feature descriptor converts the image patch to a feature vector. First we calculate the horizontal and vertical gradients to measure the HOG descriptor.Then the magnitude and direction of the gradient is calculated by using the following formula.

$$g = \sqrt{g_x^2 + g_y^2} \tag{3}$$

$$\Theta = \arctan \frac{g_y}{g_x} \tag{4}$$

Eq.3, g is the gradient of the image, gx, and gy mean the x-gradient on vertical lines and the y-gradient on horizontal lines respectively. In Eq.4, is the direction of the gradient. The feature vector fed into Support Vector Machine (SVM) for training.

D. Talking or Non-Talking Mode Selection

To detect the talking mode and non-talking mode, the system initially used a supervised machine learning algorithm. This algorithm decides the current mouth frame is in open state or close state. The system also performs a geometrical analysis of the mouth region. These two techniques were used to decide for detecting the mouth state is in talking or non-talking mode. Finally, the system considered a pattern for detecting human talking mode. The following subsection will elaborate the techniques for detecting human talking mode.

1) SVM Training Module: We used supervised machine learning for classifying the mouth state. Now we train a SVM binary classifier which classify the mouth open or close state. For the training purpose we extract the mouth region feature and feed to the SVM training module.Our extracted HOG feature length from mouth region was 8100.The feature extracted by HOG method.We shows some salient feature in Table.I. In the above table we label $+1$ symbol for open mouth

TABLE I
LIP REGION EXTRACTED HOG FEATURE.

label	f_1	f_2	f_3	f_4	f_5
+1	0.210839	0.496072	0.056212	0.116070	0.150626
-1	0.000440	0.001233	0.006945	0.002553	0.000865

feature vector and -1 for close mouth feature vector.

We train the SVM with different kernel function and the best accuracy obtained by linear kernel method. The SVM objective function aimed is to minimize the loss and maximize the accuracy.

$$Loss(p, \hat{p}) = -(p) * log(\hat{p}) + (1 - y) * log(1 - \hat{y}) \tag{5}$$

Here $Loss(p, \hat{p})$, is cost function which reduces the loss of SVM function, p is a true label and $\hat{p}$is expected label.

2) SVM Testing Module: The SVM trained architecture produce a model which expose by W_{d*f}^T, Here d represent the class number and f represent training feature dimension.

$$f(x^i) = W_{d*f} * x^i + b \tag{6}$$

Where W represents the weight vector, b represents the bias, d represents decision boundary and f represents feature size. We consider 8000 input feature and generate 150 decision boundaries for weight vector. Each decision boundary gives a score and we count maximum expected label from this score.

3) Geometric Analysis: We also consider some points which distances always change on talkative mode. In this regard we consider upper mouth point and lower mouth point because these two points move significantly during talking mode. So we measure the distance between these two points using following Euclidean distance formula.

$$D_i(p_i, q_i) = \sum_{i=1}^{n} |p_i - q_i| \tag{7}$$

Here p_i and q_i is the upper lip point and lower lip point respectively.

4) Pattern Analysis: We consider a pattern for detecting human talking mode in real time. When a person is in talking mode, his mouth will remain open or close consecutively. So, we consider a pattern where mouth is in consecutively open or close state. Therefore if we have a frame where the lip is open, then we look back the previous frames if these are in frequent open and close state. We got 10 frames per second and we consider a pattern in every 5 second.We regard the following rule for pattern analysis.

- rule 1: Two consecutive open mouth followed by close mouth or one/two close mouth followed by open mouth.
- rule 2: 30% Open mouth and 70% close mouth or vice versa.

If the both rules are satisfied for the 5 second video frame then it considered as talking mode otherwise it considered as non-talking mode at the same time then Thus we recognize if the human mouth is in talking mode or not.

E. EXPERIMENTS

To analyze the talkative and non-talkative state accuracy, we emphasize on quantitative approach. We worked with 100 people in the real environment to run this experiment. The average age of participants is 21 years (SD= 4.50).

TABLE II
TRAINING MODULE DATA SUMMARY

#Participants	Video duration(sc)	#open mouth	#close mouth
10	5250	25450	27050

To evaluate the system, we told the participants about our experiment and also described to them what they should do. Participants interacted with the system one by one.To build the training model we collect mouth open-close picture from several participants. Table II summarize the statistics of data used for training module where we represent the total number of open and close mouth for 10 person. The number of frame per second (fps) was 10. We tested our proposed system in different lighting condition and different person. We also consider different head pose and different expressions of a person to detect human talking or non-talking mode. Table III represents the data sets summary which we used for testing purpose.

TABLE III
TESTING MODULE DATA SUMMARY

#Participants	Video duration(sc)	#total frame
10	4250	42500

The system observed their talking and non-talking mode from the video. We also collect some news reading video to measure our system accuracy.

IV. EVALUATION MEASURES

We experiment our proposed system using the following statistical methods such as precision, recall, F1-measure, and overall accuracy. **Precision:** In the domain of computer vision , precision represents the ratio of correctly predict the open or closed lip that are relevant to the given image.

$$\textbf{Precision} = \frac{True_p}{True_p + False_p} \tag{8}$$

Recall: In the research area of computer vision, recall represents the ratio of the relevant to lip state that are correctly predicted.

$$\textbf{Recall} = \frac{True_p}{True_p + False_n} \tag{9}$$

$$\textbf{Accuracy} = \frac{True_p + True_n}{True_p + False_p + True_n + False_n} \tag{10}$$

F_1 **- measure:** The measurement of combining recall value with precision value and their harmonic mean which represents by the following equation.

$$\mathbf{F_1 - measure} = \frac{2 * Precision * Recall}{Recall + Precision} \tag{11}$$

Where $True_p$ = the system and actual label is the talking mode, $False_n$ = actual label is non-talking mode and system detect talking mode, $True_n$ = actual label is non-talking mode and the system detect non-talking mode, $False_p$ = actual label is talking mode and system detect non-talking mode.

V. RESULTS

In order to evaluate the proposed system we collect the ten person data with different lighting condition. The support is the total number of sample in the experiments.

In Table IV the consecutive 50 frame are selection for talking or non-talking mode, That means the system provide 10 frame in each second and after 5 second interval we check the talking or non-talking mode.

TABLE IV
SUMMARY OF THE TALKING AND NON-TALKING MODE

#Person	$True_p$	$True_n$	$False_p$	$False_n$	Support
p1	250	130	8	12	400
p2	352	134	6	8	500
p3	340	245	7	8	600
p4	364	326	4	6	700
p5	426	364	7	3	800
p6	380	505	6	9	900
p7	550	430	8	12	1000
p8	560	530	5	5	1100
p9	660	535	3	2	1200
p10	760	525	9	6	1300
total	4642	3724	63	71	8500

In this table the person eight (p8) gain the maximum accuracy due to training and testing data set collect almost same lighting environment. The minimum $True_p$ and $True_n$ value gain by person one (p1) due to lighting and facial expression.

The true negative rate is is mean the actual negative result. In our system the true negative rate is lower with respect to total number of support. The training and testing

module included both the true negative and true positive contentious video .

In Table V The maximum precision and recall value achieved

TABLE V
SUMMARY OF THE ANALYSIS

#Person	Precision	Recall	F_1-score	#Frame	Support
p1	0.97	0.95	0.96	2000	400
p2	0.98	0.97	0.97	2500	500
p3	0.97	0.97	0.97	3000	600
p4	0.98	0.99	0.97	3500	700
p5	0.98	0.99	0.99	4000	800
p6	0.98	0.98	0.99	4500	900
p7	0.99	0.99	0.99	5000	1000
p8	0.99	0.99	0.99	5500	1100
p9	0.96	0.96	0.96	6000	1200
p10	0.93	0.98	0.96	6500	1300
Avg./total	0.97	0.97	0.97	42500	8500

by the person eight (p8) and minimum value from person one (p1). The person one (p1) and person eight data set contain both the pose and lighting variation, but the person eight is only female which data not include in train module. In Fig. 4 show the precision vs recall curve. The x-axis contain the recall value and y-axis contain the precision value. The

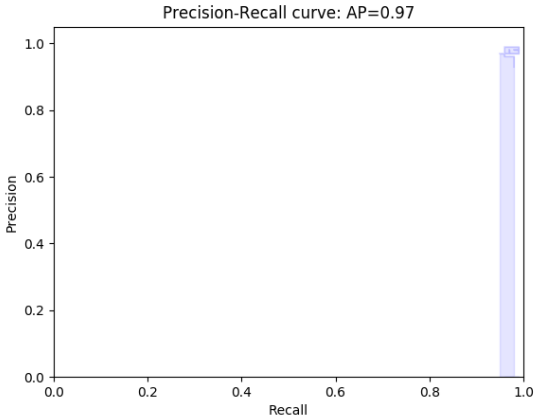

Fig. 4. Recall vs Precision.

average precision and recall value is to 0.97. Most of the precision value above the 0.93 and most of the recall value above the 0.95.

A. Sample Output of Talking and Non-Talking Mode

In Fig.5. shows the talking mode captured image from video. The SVM module provide the lip motion and pattern analysis module provided the talking and non talking mode selection. The lip motion selection accuracy almost 99.50% but the pattern analysis module accuracy around 97.00%.

In Fig.6. represented the non-talking mode image sequence which are collected from non-talking mode video. When the non-talking model selected from the proposed algorithm the

Fig. 5. Talking mode.

Fig. 6. Non-Talking mode.

pattern are provided consecutively zero value. The non talking mode accuracy almost 98.5% which is better from talking mode. Moreover for detecting a talking or non-talking mode the system need 1.02 second.

B. Comparison with Existing Approaches

To evaluate the effectiveness, we compare the proposed system with existing approaches [8] . Table VI represents the comparison with existing system performance. This results shows that our proposed system performs better than the previous system in term of accuracy. From Table VI shows

TABLE VI
PERFORMANCE COMPARISON

Method	#Training Frame	#Testing Frame	Accuracy(%)
Log-polar + SVM [8]	900	537	97.4
Proposed	52500	42500	98.00

that the number of frame in testing and training module is higher than the previous system [8].

VI. CONCLUSION

Our main concern was to develop a system that can recognize talkative and non-talkative state in real time. In our work, we addressed human mouth state based on leap motion detection. This system achieved 98.00% accuracy which is better with respect to previous work. Our proposed system experiment with a single person with well lighted condition and the captured video contain similar edge and similar expression. For future research, we will extend our framework for other forms of multiple person. Moreover, we will also include more person with more data to improve the overall performance of the system. The proposed system should be more accurate if we include different facial expression, age and color person data in training section.

REFERENCES

[1] M. Bendris, D. Charlet and G. Chollet, Lip activity detection for Talking Faces Classification in TV-Content, The 3rd International Conference on Machine Vision, 2010.

[2] I. Khan, H. Abdullah and M.S.B. Zainal,Efficient Eyes and Mouth Detection Algorithm Using Combination of Viola Jones and Skin Color Pixel Detection, International Journal of Engineering and Applied Sciences, Vol. 3, No. 4, 2013.

[3] H. Huang and Y.Ching Lin, An Efficient Mouth Detection Based on Face Localization and Edge Projection, International Journal of Computer Theory and Engineering, Vol. 5, No. 3, June 2013.
[4] T.Azim, M.Jaffar and A.Mirza, Fully Automated Real Time Fatigue Detection of Drivers Through Fuzzy Expert Systems, International Conference on Information Science, Signal Processing and their Applications,2010.
[5] R. Navarathna, P. Lucey, D. Dean, C. Fookes and S. Sridharan, Lip Detection for Audio-Visual Speech Recognition in Car Environment, 10^{th} International Conference on Information Science, Signal Processing and their Applications, 2010.
[6] N.Eveno, A. Caplier, and P-Y Coulon, Automatic and Accurate Lip Tracking, IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no.5, pp. 706-715, May 2004.
[7] C. Bouvier, A. Benoit1, A. Caplier1 and P. Y. Coulon, Open or Closed Mouth State Detection: Static Supervised Classification Based on Log-polar Signature, 2009.
[8] L. Wang, Wang and J. Xu, Lip Detection and Tracking Using Variance Based Haar-like Features and Kalman filter, Fifth International Conference on Frontier of Computer Science and Technology , 2010.
[9] K.Saenko,K.Livescu,M.Siracusa,K.Wilson,J.Glass, and T. Darrell, Visual speech recognition with loosely synchronized feature streams, Tenth IEEE International Conference on Computer Vision(ICCV), 2005.
[10] M. I. Faraj and J. Bigun, Person verication by lip-motion, in Computer Vision and Pattern Recognition Workshop on Biometrics, Piscataway, New York, June 2006, pp. 3744.
[11] D. E. King, Dlib-ml: A Machine Learning Toolkit, Journal of Machine Learning Research, vol. 10, pp. 1755-1758, 2009.
[12] C.C. Chang and C.J. Lin, LIBSVM : a library for support vector machines, ACM Transactions on Intelligent Systems and Technology,2011.