Recognizing Emotional State of the Human Using Facial and Acoustic Features

G.Saranya[1]
Post Graduate Student, Dept. of ECE
Parisutham Institute of Technology and Science, Thanjavur.
Affiliated to Anna University, Chennai, India.
Email saranyaraji.91@gmail.com

G.Mary Amirtha Sagayee[2]
Professor & Head, Dept. of ECE
Parisutham Institute of Technology and Science, Thanjavur.
Affiliated to Anna University, Chennai, India.
Email gmasagayee@gmail.com

S.Priyavarsha[3]
Post Graduate Student, Dept. of ECE
Parisutham Institute of Technology and Science, Thanjavur.
Affiliated to Anna University, Chennai, India.
Email priyavarsha.s2@gmail.com

***Abstract*— Emotional state of the human can be analyzed and recognized using various facial expressions and voice tones. The expressions can be detected and interpreted by the system. Deblocking is the main process, which is used to describe the various expressions of the human. Hence the expressions are classified using a classifier. When there is any change in the emotion, this classifier constitutes an emotional code related to the expressions. Various facial expressions such as happy, anger, sad, neutral, disgust have been collected using Independent Component Analysis (ICA). Similarly, various voice tones for the above mentioned expressions are collected. These points are used to detect any changes in the face. In voice sub band based Cepstral parameter (SBC) and Mel-Frequency Cepstral Coefficient (MFCC) are calculated to detect the changes in the emotions. Gaussian Mixture Model (GMM) is used to classify the expressions in the voice tone.**

Keywords—Deblocking; ICA; haaris corner; SBC; MFCC; GMM

I. INTRODUCTION

The study about the development of human system interaction is referred to as affective computing. Human affects (emotions) can be recognized, interpreted, processed and simulate by the systems. The emotional information are detected and recognized to find the emotional state of the human. The sensor such as camera catches facial expressions, body posture and gesture, while a microphone might used to capture the emotional speech. The meaningful patterns are used to recognize the emotional information from the gathered data. The human has an ability to adapt their expressions and spoken style during their conversation with others. To enhance the human system interactions, the study of entrainment on various aspects like pronunciation, tone, speaking rate and various facial expressions in different situations such as happy, anger, disgust, neutral, sad, fear, surprise, and shame becomes essential. Human facial expression recognition has been receiving lot of attention due to its increase in applications. Hence this study reviews the various work on the recognition of human emotional state in the video.

REVIEW ANALYSIS

The sensor takes the image as an input and it processed based on the emotions. Proper training is done over the various expressions and it is stored in the main database. Image comparison is done for the training and testing image to predict the emotional state. Features from the facial images get by the automatic approach which consists of four stages [1].

The human emotional state can also be extracted from the voice. In facial expression recognition the features are placed in the sets and it creates the decision boundaries in the two dimensional image and the expressions can be classified and identified successfully. The emotion state can also deals with the video. For each expression the neutral state expression is consider as the baseline. From the neutral state the other expressions can be classified and resulted.

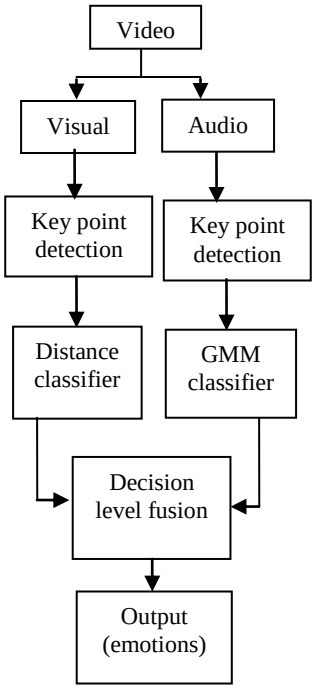

Fig.1. System Design

In this paper the various expressions are trained and stored in the database. During testing process the image under test is compared with the related images in the main database and the result is produced. The training and testing process uses the principal component analysis (PCA) algorithm. The dimensionality in the images can be greatly reduced by using this algorithm. Support vector machine (SVM) [4] is used for classifying the faces. The feature extraction can be done by the haar wavelet in the image. The Euclidean distance for image classification is explained in work of Deepash Raj work [8]. Emotion recognition in speech is explained in K.V.Krishna et.al. Work [9]. It describes the MFCC and wavelet features.

The remaining part of this paper is organized as follows: section 2 describes the proposed system. 3 explain the facial module and 4 describes about the voice module. Section 5 deals with fusion. section 6 gives the experimental results. The overall system design is given in figure 1.

II. PROPOSED SYSTEM

The human expressions are trained manually. The work flow is given in figure.2. Human facial expressions are captured through web camera. The captured video is converted into frames by fixing the interval between the frames in a video. The converted frames are stored in the selected location. The images are pre-processed to remove noise and also to enhance the accuracy.

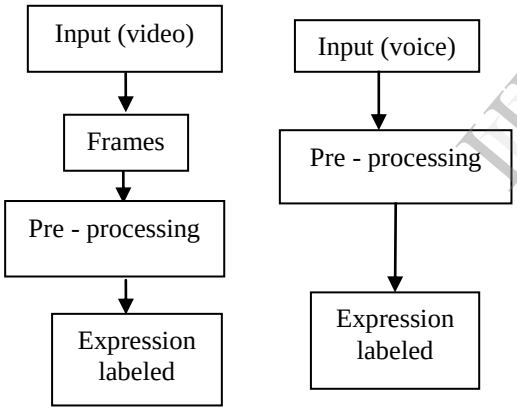

Fig. 2. Work Flow in Training Phase.

Facial part alone is cropped from the image using direct correlation method. This is performed in the pre-processing steps. Steps in Pre-processing are given in figure.3.

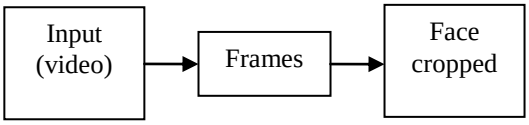

Fig 3. Pre-processing stage for train images.

The steps in the direct correlation technique are given below.

Steps in direct correlation method:

1. Images resize.
2. Intensity value is adjusted.
3. Skin colour extraction.
4. Edge detection.
5. Dilated structure.
6. Fill the holes in edge detected image.
7. Face is cropped.

The score value is calculated for each image by Euclidean distance. To calculate the Euclidean distanced in the image, all images are resized into one common size. Any range of value can be selected. Intensity value for every image is adjusted by fixing the low and high intensity values. The next step is to extract the skin color from the image by converting the RGB image into equivalent image in the YCbCr color space. Edge is detected from the binary image. Edge detection can be performed by various methods such as canny, sobel, prewitt. The corner features are enhanced by forming the morphological structuring elements. The holes in edge detected image are filled to extract the positions in the image. The final step is to remove the background from the image to crop the facial fart alone.

Label file is created to mention the emotions for the images which are in the train dataset. For every image the facial expression is manually created. The dataset must contain various gesture for a single emotion i.e., for happy alone we have to train more than ten image for a person. Similarly for each and every emotion the label file is updated during training phase.

Data set is created by various expressions such as happy, anger, and disgust and sad. The audiovisual dataset is created for various expressions.

III. FACIAL MODULE

The feature vector is calculated for each image in the database. To calculate the feature vector Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are used. The steps involved in PCA are given below.

Steps in PCA

1. Image is converted into matrix format.
2. Rectangular matrix (MxN) is converted into column matrix (Mx1).
3. Mean value is calculated for the entire image.
4. Subtract mean value from the column vector.

The feature vector for the particular image is calculated. Using Principal component the coefficient, score and latent value for the image is created.

Independent Component Analysis is used to extract the maximum information from the multiple visual channels. ICA maximizes the joint entropy and it provide brain-like visual features for the natural image. ICA can also be used for speech separation in the area of speech recognition. ICA is the unsupervised computational and statistical method to discover hidden factors in the data. Steps involved in ICA are given below.

Steps in ICA

1. Centring – make the signals centered into zero.
2. Sphering – make the signals uncorrelated.
3. Rotation – maximization of an object function.

The Gabor wavelet is used to compute the Gabor features of a gray scale image. The wavelet scale, filter orientation, wavelength of small scale filter, scaling factor between successive filters, log Gabor filter transfer function, ratio of angular interval between filter orientations and the standard deviation of the angular Gaussian function, number of standard deviation of the noise energy beyond the threshold point and the polarity values are selected. The feature vector is calculated for the image by calculating the mean squared energy and mean amplitude.

A. *Expression Classification*

The expression is classified by calculating the distance between the feature vectors of the image. Distance classifier is used to classify the expressions. Euclidean distance is used for distance classification. Mean value for neutral expression in the dataset is calculated. Test image is subtracted from the mean neutral to provide the score value for each image.

$$d(x,y) = \|x-y\|^2 = \sum_{i=1}^{k}(x_i - y_i)^2 \quad (1)$$

Minimum distance is calculated for the two images and the related expression is produced as the output.

IV. VOICE MODULE

Human emotional state is detected from the voice. The voice signal is recorded through microphone. The key point is detected by collecting the Mel Frequency Cepstral Coefficient (MFCC) and Sub band based Cepstral (SBC). The expression is classified through classifier such as K-nearest Neighbors (KNN), Hidden Markov Model (HMM), Gaussian Mixture Model, Support Vector Machine and Artificial Neural Network. In the proposed approach Gaussian mixture model is used as the classifier.

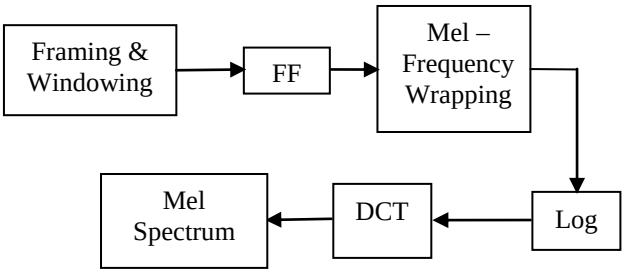

Fig 4. Work flow in MFCC

MFCC is a powerful analytic tool in the field of recognition. MFCC mimic the behavior of human ears by applying the Cepstral analysis. It is computed based on the speech frames. For speech recognition the total number of coefficients used is between nine and thirteen. The work flow in MFCC is given in figure.4. The speech signal is split up into several frames. To avoid the unnatural discontinuities in the signal windowing process is performed. Fast Fourier Transform (FFT) is performed to convert the signal from time domain to the frequency domain. Mel-scale is the scale where the pitch are placed periodic manner. Discrete Cosine Transform (DCT) is performed to convert the signal again to time domain. If the calculated score value is greater than 6.8 then it is considered as the perfect match.

SBC is similar to MFCC instead of FFT it uses wavelet packet transform. SBC parameters are derived from the subband energies. In SBC if the score value calculated is greater than 21.5 then the sample is considered as perfect match.

A. *Expression Classification*

Gaussian Mixture Model is represented as a mixture of the Gaussian densities. GMM is the linear combination of M Gaussians. The equation for the linear combination is given by,

$$p = \left(\frac{\vec{x}}{\lambda}\right) = \sum_{i=1}^{M} p_i b_{i(\vec{x})} \quad (2)$$

where is a D- dimensional random vector $b_{i(x)}$,and i=1,.2,…….M are the component densities and p_i, i=1,2,……..M are mixture weights. Each component density is a D-dimensional Gaussian function of the form

$$b_{i(\vec{x})} = \frac{1}{(2\prod)^{D/2}|\Sigma i|^{1/2}} \exp\{-\frac{1}{2}(\vec{x}-\mu_i)^T \sum_i^{-1}(\vec{x}-\mu_i)\} \quad (3)$$

Where μ denotes the mean vector and Σi ,denotes the covariance vector matrix. The mixture weights satisfy the law of total probability.,

$$\sum_{i=1}^{M} p_i = 1. \quad (4)$$

V. FUSION

The multimodal features are extracted and combined using feature-level fusion. It is the direct fusion method in which feature vectors from the multiple modalities are concatenated to obtain a combined feature vector for a classification task.

VI. EXPERIMENTAL RESULTS

The human expression is recognized from facial expressions and from voice tone. The training dataset is created by saving various expressions made by five persons, 24 images for happy, 12 for sad, 13 for disgust, 11 for neutral and 13 for anger. During recognition process the following steps are followed. The web camera is used to record the human expressions. The various emotions are stored in the desired location as a frame format. The stored frames are compared with the images in the database to produce the results. The training process is given in figure.5.

Fig 5. Training process

The expression for trained image is stored manually in the label file. It is given in figure.6.

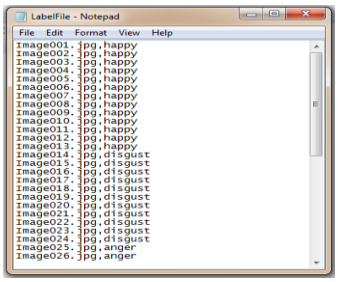

Fig 6. Label file

The trained images are loaded during the testing process. The emotion is tracked to find the changes in the emotions during the testing process. Figure.7 shows the emotional tracking in real time.

Fig 7. Emotional tracking

VII. CONCLUSION

The visual features of human emotional states are recognized to improve the performance of human system recognition during non-verbal communication. It is useful in human machine interaction. To get the efficient result the entrainment over various expressions are performed. The image which is under test is compared with the images in main database. The result will be produced by comparing and retrieving the related expressions from the main database. Time consumption for testing phase is more. The delay will be large. In the future work the delay can be reduced by using various techniques, where the dimensionality to save and retrieve the image will be greatly reduced.

REFERENCES

[1] Bellakhdhar, Kais Loukil, Mohamed *Svm Classification For Face Recognition*, Faten Journal of intelligent computing volume 3 Number 4 December 2012.

[2] Carlos Busso, Zhigang Deng , Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, Shrikanth Narayanan , Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information, Emotion Research Group, Speech Analysis and Interpretation Lab Integrated Media Systems Center, University of Southern California, Los Angeles.

[3] Ce Zhan, Wanqing Li, Philip Ogunbona, and Farzad., *"Real-Time Facial Feature Point Extraction"*, Safaei University of Wollongong, Wollongong, NSW 2522, Australia. Zhan, F. (2007). Pacific- Rim Conference on Multimedia (pp. 88-97). Germany: Springer.

[4] Deepesh raj, A Real Time Face Recognition System Using PCA And Various Diatance Classifiers., Spring 2011.

[5] Faten Bellakhdhar, Kais Loukil, Mohamed ABID, computer embedded system, University of Sfax 2012. SVM classification for face recognition, Journal of intelligent computing volume 3 Number 4 December.

[6] G.U.Kharat, S.V. Dudul, 2009 Emotion Recognition from facial expression using neural networks, Human-computer systems interaction advances in intelligent and soft computimg.

[7] Hua Gu Guangda Su Cheng Du Department of Electronic Engineering, Feature Points Extraction from Faces Research Institute of Image and Graphics, Tsinghua University, Beijing, China. Image and vision computing NZ.

[8] Ira Cohen, Ashuto,sh Garg, Thomas S. Huang, Emotion Recognition from Facial Expressions using Multilevel HMM, Beckman Institute for Advanced Science and TechnologyThe University of Illinois at Urbana-Champaign.

[9] Jui-Chen Wu, Yung-Sheng Chen, and ICheng Chang., *An Automatic Approach to Facial Feature extraction for 3-D Face Modeling, ,* IAENG International Journal of Computer Science, 33:2, IJCS_33_2_1, 24 May 2007.

[10] K.V.Krishna., *Emotion Recognition In Speech Using MFCC And Wavelet Features.,.,* 2013 3rd IEEE International Advance Computing Conference (IACC).

[11] L.S.Chen. Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction. PhD thesis, University of Illinois at Urbana-Champaign, Dept. of Electrical Engineering, 2000.

[12] Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh A., Busso,C., Deng, Z., Lee, S., Narayanan, S.S. Emotion Recognition based on Phoneme Classes. To appear in Proc. ICSLP'04, 2004.

[13] Mase K. Recognition of facial expression from optical flow. IEICE Transc., E. 74(10):3474–3483, 0ctober 1991.

[14] P.Ekman and W.V. Friesen, Facial action coding system: Investigator's Guide. Consulting PsychologistsPress, Palo Alto, CA, 1978.

[15] Priya Metri1, Jayshree Ghorpade and Ayesha Butalia,*Facial Emotion Recognition Using Context Based Multimodal Approach*", Int. J. Emerg. Sci., 2(1), 171-182,March 2012 ISSN: 2222-4254 © IJES 171, Pune.

[16] Qiuxia wu, Zhiyong Wang, Feiqi Deng, Zheru Chi, David Dagan Feng, *Realistic Human Action Recognition With Multimodal Feature Selection And Fusion.* IEEE transactions on systems, man, and cybernetics: systems, VOL.43, NO, 4, July 2013.

[17] T.Kanade,T.Kanade, J.F. Cohn, and Y. Tian. Comprehesive database for facial expression analysis. In Proc. Of 4rd Intl Conf. Automatic Face and Gesture Rec., pages 46–53, 2000.

[18] vSoroosh Mariooryad, Carlos Busso., Exploring Cross-Modality Affective Reactions for Audiovisual Emotion., IEEE Transactions On Affective Computing, Vol. 4, No. 2, April-June.

[19] Yoshitomi, Y., Sung-Ill Kim, Kawano, T., Kilazoe, T. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. Robot and Human Interactive Communication, 2000. RO-MAN 2000. Proceedings. 9th IEEE International Workshop on, 27-29.