Emotional State Recognition Using Facial and Acoustic Features of the Human

G.Saranya¹, G.Mary Amirtha Sagayee ², G.S.Anandha Mala³

¹PG scholar ECE Dept., Parisutham Institute of Technology & Science, Affiliated to Anna University Chennai, Thanjavur, India, saranyaraji91@gmail.com
²Prof &Head ECE Dept., Parisutham Institute of Technology & Science, Affiliated to Anna University Chennai, Thanjavur, India, gmasagayee@gmail.com
³Professor of Computer Science and Engineering, St.Joseph's College of Engineering, Chennai, Tamilnadu, India

Abstract — Emotional state of the human can be detected and analyzed by various aspects such as facial expressions, voice tone and body gestures. The detected expressions are analyzed and interrupted by the system. This method is very useful in Human Computer Interaction. Deblocking is the main step in detecting the various expressions of the human. Expressions are classified using respective classifiers for detecting emotional state using multi modalities. In the proposed approach facial expressions and voice tone are used to detect the emotional state. The key point detection in the image is obtained using Independent Component Analysis (ICA). Principal Component Analysis (PCA) is used for train the images. Similarly for speech Mel-Frequency Cepstral Coefficient (MFCC) and Sub band based Cepstral parameter (SBC) are used for feature extraction from the voice tone. Decision level fusion is used to combine the facial expressions and the emotional speech.

Keywords — Deblocking, ICA, PCA, SBC, MFCC.

1. 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.

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.

Fig. 1. System Design
The overall system design is given in figure 1. The remaining part of this paper is organized as follows: section 2 describes the proposed system, Section 3 explains the facial module and section 4 describes about the voice module. Section 5 gives the experimental results.

2. 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.

Input (video)

Frames

Pre - processing

Expression labeled

Input (video)

Frames

Pre - processing

Expression labeled

Fig. 2. Work flow in training stage

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.

Input (video)

Frames

Face cropped

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 distance 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 detection is 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 part 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.

3. 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 centred 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 it is
given in equation 1.
\[ d(x, y) = \|x - y\|^2 = \sum_{i=1}^{k} (x_i - y_i)^2 \]  
(1)
Minimum distance is calculated for the two image and the
related expression is produced as the output.

4. 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 Subband 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 Guassian mixture model is
used as the classifier.

![Work flow in MFCC](image)

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 in
equation 1.
\[ p = \sum_{i=1}^{M} p_i b_i(\tilde{x}) \]  
(2)
where \( \tilde{x} \) is a D- dimensional random vector , \( b_i(\tilde{x}) \), and
\[ i=1,2,\ldots,M \] are the component densities and
\[ p_i, i=1,2,\ldots,M \] are mixture weights. Each component
density is a D-dimensional Gaussian function of the form is
given in equation 2.
\[ b_i(\tilde{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (\tilde{x} - \mu_i)^T \Sigma_i^{-1} (\tilde{x} - \mu_i) \right\} \]  
(3)
Where \( \mu \) denotes the mean vector and \( \Sigma \) denotes the
covariance vector matrix. The mixture weights satisfy the law
of total probability, \( \sum_{i=1}^{M} p_i = 1 \).

5. FUSION

The multimodal features are extracted and combined using
decision-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.
Decision level fusion combines probability scores or
likelihood values which obtained from separate classifiers to
deide the final decision. OR logic is used for the fusion level
decision of the features from different classifier.

6. 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 a
selected 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.

![Training process](image)

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

![Label file](image)

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. Emotions are tracked ina frame by frame manner
and produces the relevant emotion as a output by performing
multi-level fusion.
The overall performance of the system is evaluated by Precision, Recall and F-score values. The precision value is calculated by fraction of retrieved samples for one emotional class that are relevant. Similarly the recall rate is calculated from the fraction of relevant samples that are correctly recognised. The F-score calculation is performed using the equation 4.

\[
F = \frac{2PR}{P+R}
\]  

(4)

7. 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


