Overview on Emotion Recognition System

Ashwini Ann Varghese
Dept. of Computer Science and Engineering
St. Joseph's College of Engineering and Technology
Palai, Kerala, India
ashwiniann91@gmail.com

Jacob P Cherian
Dept. of Computer Science and Engineering
St. Joseph's College of Engineering and Technology
Palai, Kerala, India
jacobpc29@gmail.com

Dr. Jubilant J Kizhakkethottam
Dept. of Computer Science and Engineering
St. Joseph's College of Engineering and Technology
Palai, Kerala, India
jubilantjob@gmail.com

Abstract—Human emotion recognition plays an important role in the interpersonal relationship. The automatic recognition of emotions has been an active research topic from early eras. Therefore, there are several advances made in this field. Emotions are reflected from speech, hand and gestures of the body and through facial expressions. Hence extracting and understanding of emotion has a high importance of the interaction between human and machine communication. This paper describes the advances made in this field and the various approaches used for recognition of emotions. The main objective of the paper is to propose real time implementation of emotion recognition system.

Keywords—Decision level function, Feature level fusion; affective states, Active Appearance Model, Hidden Markov Model, State Sequence ML classifier, Facial Action Encoding.

I. INTRODUCTION

Emotion plays an important role in human life. Interpersonal human communication includes not only language that is spoken, but also non-verbal cues as hand and body gestures, tone of the voice, which are used to express feeling and give feedback and most importantly through facial expression. Human beings express emotions in day to day interactions. Understanding and knowing how to react to people's expression greatly enriches the interaction. The field of psychology has played an important role in understanding human emotion and developing concepts that may aid these HCI technologies [7]. Ekman and Freisn have been pioneers in this area, helping to identify six basic emotions [13](anger, fear, disgust, joy, surprise, sadness) that appear to be universal across humanity [2]. Emotion recognition comes under computer vision. Computer vision seeks to generate intelligent and useful descriptions of visual scenes by performing operations on the signals received from video cameras.

Several approaches to recognize emotions from speech were done. Many efforts were taken to recognize states from vocal information [4]. Some important voice feature vectors have been chosen for recognizing emotions, in which utterance level statistics are calculated. For example; mean, standard deviation, maximum, and minimum of pitch contour and energy in the utterances are mainly used features in this regard [3].

Recognizing emotion using facial expressions is a key element in human communication. Studies on Facial Expressions date back to the early Aristotelian era. The automatic recognition of facial expressions has been an active research topic since the early nineties. The various facial behavior and motions can be parameterized based on muscle actions. This set of parameters can then be used to represent the various facial expressions. There have been two important and successful attempts in the creation of these parameter sets: The Facial Action Coding System (FACS) and The Facial Animation parameters (FAPs) [1].

Most of the work in affective computing does not combine various modalities into one system for the analysis of emotional behavior of a human: different information channels (facial expressions and speech) are considered independently of each other. Further, there are a few ways to integrate information from movement of the body and gestures. In this paper the details about extracting or recognizing emotions from various approaches are carefully studied [5].

Emotion recognition systems find applications in several interesting areas. With the recent advances in robotics, especially humanoid robots, the urgency in the requirement of a robust expression recognition system is evident. Emotion recognition plays a significant role in recognizing one's affect and in turn helps in building meaningful and responsive Human Computer Interface (HCI). Apart from the two main applications, namely robotics and affect sensitive HCI, emotion recognition systems find uses in a host of other domains like Telecommunications, Video Games, Animations, Psychiatry, Automobile Safety, Educational Software, etc.

II. METHODOLOGIES

A. Recognition of facial expression using AAM

Matthew S. Ratliff and Eric Patterson conducted an experiment to recognize the emotion using facial expressions with active appearance models. In their paper a framework for the classification of emotional states, based on still images of the face was done. The technique involved the creation of an active appearance model (AAM) that was

978-1-4799-1753-2/15/$31.00 ©2015 IEEE

trained on face images from a publicly available database to represent shapes and texture variation key to expression recognition. AAM used a set of feature classification scheme which identifies the six basic emotions. In this approach facial expression was analyzed by the means of the movement of facial muscles. Use FACS parameters to store those values. FACS was used as a framework for classification. The AAM has the ability to aid in initial face-search algorithms and in extracting important information from both the shape and texture (wrinkles, etc.) of the face that may be useful for communicating emotion. So they adopted this technique as a feature extraction method. By giving the computer prior information such as eating habits, stress levels, sleep habits, etc. the ANN predict the emotional state of the user and can change its responses accordingly. In this approach we use the facial expression database known as [9]"FEEDTUM." This database contains still images and video sequences of eighteen test subjects, both male and female, of varying age. Rather than hiring actors to artificially create or mimic emotional responses, this database was developed with the attempt to actually elicit the required emotions. Using a camera mounted on a computer screen, subjects show various movie clips that hopefully trigger an emotional response. The database is organized by category using the six basic emotions. Key areas were chosen to capture the movement of the brow, eyes, mouth, and nasal region as formed by the underlying muscles expected in expression of the face. Once an initial AAM was trained in several subjects, the search function helped automate the labelling process [2].

B. Fully Automatic Recognition of Temporal Phases of Facial Actions

In the paper, Fully Automatic Recognition of Temporal Phases of Facial Actions, proposed by Michel F. Valstar and Maja Pantic the detection of a much larger range of facial behavior that include the classification of more facial expression other than the basic six emotions. The facial behavior is recognized by facial muscle actions which include action units (AUs).AUs help for higher decision making systems like emotion recognition system. Michel F. Valstar and Maja Pantic proposed fully automatic method that allows the recognition of 22 AUs and it also allows storing the temporal characteristics (which include the temporal segments like neutral, onset, offset and apex). For calculating the temporal features the proposed system uses a facial point detector which used Gabor-feature –based booster classifier [10,15] that automatically localize 20 facial fiducially points. The points that are detected from the face are tracked in an entire image sequence. To encode AUs and their temporal activation models, it uses a combination of support vector machines, GentleBoost, and hidden Markov models. The five important steps that are used for the recognition of AUs and their temporal activation models are: registration and smoothening, mid-level parametric representation, facial AU classification, temporal activation models of facial AUs and finally emotion detection [8].

C. Emotion Recognition using Facial Expressions, Speech and Multimodal Information

In the paper emotion recognition using facial expressions, speech and multimodal information recognition of human emotion based on facial expression or speech is done. Only limited system or work has been done to fuse this information. The accuracy and robustness of emotion recognition can be increased by combining these two techniques. This paper explains about the two approaches used to fuse the two modalities: decision level and feature level integration.

The database holds the four emotions to classify: sadness, anger, happiness and neutral state. The database used in this technique is recorded by an actress. By the use of markers on the face, detailed facial motions can be captured with motion capture. The paper reveal that the system based on facial expression gave better performance than the system based on just acoustic information for the emotions considered.

The paper also shows the complimentary of the two modalities and that when these two modalities are combined, the robustness and the performance of the emotion recognition system improve measurably. The methodology used on the system based on speech includes the cues for audio emotion recognition and they are global-level prosodic features such as the statistics of the pitch and the intensity. Therefore, the standard deviations, the ranges, the maximum values, the minimum values and the medians of the pitch and the energy were computed. In addition, the voiced/speech and unvoiced/speech ratio were also estimated.

By the use of sequential backward features selection technique, an 11-dimensional feature vector for each utterance was used as input in the audio emotion recognition system. In the case of a system based on facial expression the spatial data collected from markers in each frame of the video are reduced into a 4-dimensional feature vector per sentence, which is then used as input to the classifier. After the motion data are captured, they are normalized by the following steps: (1) all markers are translated in order to make a nose marker be the local coordinate centres of each frame, (2) a frame with neutral and close-mouth head pose is picked as the reference frame, (3) three rigid markers define a local coordinate origin for each frame, and (4) each frame is rotated to align it with the reference frame. Each data frame is divided into five blocks: eyebrow, low eye, forehead, right cheek and left cheek area. For each block, the 3D coordinate of markers in this block is concatenated together to form a data vector. Principal Component Analysis (PCA) method is used to reduce the number of

features per frame into a 10-dimensional vector for each area, covering more than 99% of the variation. Different emotions appear in separate clusters, so important clues can be extracted from the spatial position of these 10-dimensional features space. In the case of bimodal system proposed in the paper fusing of facial expression and acoustic information is done. The two different approaches were implemented: feature-level fusion, which uses a single classifier with features of both modalities and decision level fusion, which uses a separate classifier for each modality, and the outputs are combined using some criteria. In the first approach, a sequential backward feature selection technique was used to find the features from both modalities that maximize the performance of the classifier. The number of features selected was 10. In the second approach, several criteria were used to combine the posterior probabilities of the mono-modal systems at the decision level: maximum, in which the emotion with greatest posterior probability in both modalities is selected; average, in which the posterior probabilities of each modalities are equally weighted and the maximum is selected; product, in which the posterior probabilities are multiplied and the maximum is selected; and, weight, in which different weights are applied to the different unimodal systems [3].

D. Multimodel Emotion Recognition from Expressive faces body gestures and Speech

Multimodal emotion recognition from expressive faces, body gestures and speech proposed by Ginevra Castellano, et. al [5] a multimodal approach were used in order to recognize the eight emotions. The emotion recognition is done by the integral information from the facial expressions, body movements and speech.

The Bayesian classifier is used for the training and testing of the model. Initially individual classifiers were used for the training of each modality. Then data were fused at a feature level and the decision level. The fusing of multimodal data was increased so much so that the recognition rate of this approach increased 10% with respect to the most successful unimodal system. The eight acted emotion states in this technique are anger, despair, interest, pleasure, sadness, irritation, joy and pride. These states were classified based on the integral information from the body and gestures, facial expressions and speech. In the study of the training and the testing of a model with a Bayesian classifier was done using a multimodal corpus with ten subjects collected during the Third Summer School of the HUMAINE EU-IST project which was held in Genova in September 2006. In this technique after finding the participants from the school technical setting up were done. These were done by the means of two DV cameras (25 fps) which was recorded the actors from the frontal view. One camera recorded the actor's body and the other one was focused on the actor's face. The resolution required for facial features extraction is much larger than the one for body movement detection or hand gestures tracking. This was achieved when one camera was zoomed in the actor's face. Long sleeves and covered neck was preferred since the hand and head detection algorithm was based on colour tracking. For the voice recordings we used a direct-to-disk computer-based system. The speech samples were directly recorded on the hard disk of the computer using sound editing software.

A microphone mounted on the actors' shirt was connected to an HF emitter (wireless system emitter) and the receiver was connected to the sound card using a XLR connector (balanced audio connector for high quality microphones and connections between equipment's). The external sound card included a preamplifier (for two XLR inputs) that was used in order to adjust the input gain and to minimise the impact of signal-to-noise ratio of the recording system. After all these information features should be extracted. Feature extraction can be done by three steps they are face feature extraction, body feature extraction and speech feature extraction. In face feature extraction, first step the face was located, so that approximate facial feature locations could be estimated from the head position and rotation. The face was segmented focusing on the following facial areas: right eye/eyebrow, left eye/eyebrow, mouth and nose. Each of these areas is called feature-candidate areas which contain the features whose boundaries are to be extracted.

Feature extraction was performed for each facial feature, i.e. eyebrows, eyes, nose and mouth, using a multi-cue approach which generates a small number of intermediate feature masks. Feature masks generated for each facial feature were fused together to produce the final mask for that feature. Tracks points in the facial area, we chose to work with MPEG-4 FAPs (Facial Animation Parameters) and not Action Units (AUs). Measurement of FAPs requires the availability of a frame where the subject's expression is found to be neutral.

This frame is called the neutral frame and is manually selected from video sequences to be analysed or interactively provided to the system when initially brought into a specific user's ownership. The whole process was inspired by the equivalent process performed in the acoustic features. For body feature extraction EyesWeb platform were used. Extract five main expressive motion cues, using the EyesWeb Expressive Gesture Processing Library: quantity of motion and contraction index of the velocity, body, fluidity and acceleration of the hand's barycentre. Data were normalised according to the behaviour shown by each actor, considering the maximum and the minimum values of each motion cue in each actor, to compare data from all the subjects. Automatic extraction allows obtaining temporal series of the selected motion cues over time, depending on the video frame rate. Extract the dynamic indicators of the motion of temporal profile: final slope and initial slope, initial and final slope of the peak, maximum

value, ratio between the maximum value and the duration of the main peak, mean value, ratio between the mean and the maximum value, ratio between the absolute maximum and the biggest following relative maximum, centroid of energy, distance between maximum value and centroid of energy, symmetry index, shift index of the main peak, number of peaks, number of peaks preceding the main one, ratio between the main peak duration and the whole profile duration. Using this technique we can characterise 80 motion features. In the case of speech feature extraction the set of features based on intensity, pitch, MFCC (Mel Frequency Cepstral Coefficient), Bark spectral bands, voiced segment characteristics and pause length. The full set contains 377 features. The features from the intensity contour and the pitch contour were extracted using a set of 32 statistical features. This set of features was applied both to the pitch and intensity contour and to their derivatives. Here it considers the following 32 features. Maximum, mean and minimum values, sample mode (most frequently occurring value), interquartile range (difference between the 75th and 25th percentiles), kurtosis, the third central sample moment, first (slope) and second coefficients of linear regression, first, second and third coefficients of quadratic regression, percentiles at 2.5 %, 25 %, 50 %, 75 %, and 97.5 %, skewness, standard deviation, variance. Thus, 64 features based on the pitch contour and 64 features based on the intensity contour. we extracted 13 MFCCs using time averaging on time windows, as well as features derived from pitch values and lengths of voiced segments, using a set of 35 features applied to both of them. Finally, extract the feature based on pause (or silence) length and non-pauses lengths (35 each). This paper uses common approach based on a Bayesian classifier (BayesNet) provided by the software Weka which is a free toolbox containing a collection of machine learning algorithms for data mining tasks for comparing the results of unimodal and multimodal systems. The emotion that received the best probability in the three modalities is selected in this technique [11,12].

III. COMPARISONS

Matthew S. Ratliff and Eric Patterson conducted an experiment to recognize the emotion using facial expressions with active appearance models. In their paper a framework for the classification of emotional states, based on still images of the face was done. The technique involved the creation of an active appearance model (AAM) that was trained on face images from a publicly available database to represent shapes and texture variation key to expression recognition. AAM used a set of feature classification scheme which identifies the six basic emotions. In this approach facial expression was analyzed by the means of the movement of facial muscles. Use FACS parameters to store these values. FACS[6] was used as a framework for classification. The AAM has the ability to aid in initial face-search algorithms and in extracting important information from both the shape and texture (wrinkles, etc.) of the face that may be useful for communicating emotion. So they adopted this technique as a feature extraction method. By giving the computer prior information such as eating habits, stress levels, sleep habits, etc. the ANN predict the emotional state of the user and can change its responses accordingly. In this approach we use the facial expression database known as "FEEDTUM." This database contains still images and video sequences of eighteen test subjects, both male and female, of varying age. Rather than hiring actors to artificially create or mimic emotional responses, this database was developed with the attempt to actually elicit the required emotions. Using a camera mounted on a computer screen, subjects show various movie clips that hopefully trigger an emotional response. The database is organized by category using the six basic emotions. Key areas were chosen to capture the movement of the brow, eyes, mouth, and nasal region as formed by the underlying muscles expected in expression of the face. Once an initial AAM was trained in several subjects, the search function helped automate the labelling process [2].

The cons of the above system can be slightly reduced by the system proposed by Michel F. Valstar and Maja Pantic, in which temporal characteristic information is stored which include the temporal segments like neutral, onset, apex and offset information. Apart from the above system this system has a high degree of larger range of facial behavior can be recognized other than the six basic emotions. Accurate fully automatic facial expression analysis will have real world applications. Fully automatic, accurate AU based on generic features is possible in the system. It is possible to detect the four temporal phases of AU with high accuracy. Here multiple databases are used for comparing and testing. The main cons in this system is it can only recognize facial expression if the face is viewed from a pseudo frontal view. If the head has an out of plane rotation greater than 20°, the system will fail. It could not handle occlusion. It is person dependent. That is, it will not fit mask automatically [6].

In the paper, Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information, the robustness and accuracy of emotion recognition system is increased compared to the above two approaches since it uses the combination of facial expression and speech. The system based on facial expression gives better performance than the system based on acoustic information. The robustness and performance of emotion recognition system, improve when these two modularity is combined. Unlike from the system proposed by Michel F. Valstar and Maja Pantic, this system cannot be used in real-time applications. Since actor images are used in database only posed, modulated deliberate images can be taken. For a real time application, we need spontaneous, unmodulated or genuine pictures. Posed expressions are artificial expression that a subject will produce when he or she is asked to do so. Spontaneous expressions are the ones that people give out spontaneously. In this system we use posed expression so it

is not efficient in real time applications [3].

In the paper 'Multimodal emotion recognition from expressive faces, body gestures and speech' proposed by Ginevra Castellano, Loic Kessous, and George Caridakis the multimodal approach gave an improvement for more than 10% with respect to the most successful unimodal system. The fusion performed by the feature level has better results than the one performed at the decision level. The main cons in these approaches are long sleeve and covered neck should be needed for the actor since the hand and head selection algorithm is based on colour trading. Uniform background was needed in order to make the background subtraction process easier. In order to reduce the occlusion on face expression some prerequisites such as they should lack the eyeglasses, beards and moustaches [5].

IV. CONCLUSION

The main objective of this paper was to analyse the strengths and weaknesses of various techniques adopted in the emotion recognition system. The inter personal communication between a human being and a computer can be increased rapidly by combining all the pros. Most of the real time issues can be improved by using this recognition system. The emotion recognition system can play a vital role in the real world applications [14] such as HCI, humanoid robotics, security, games etc.

From the study done in the specified papers various techniques and methodologies were identified. So a suggestion can be made for an improved integrated system which includes the advantages of contributed by each system.

REFERENCES

[1] Bettadapura, Vinay. "Face expression recognition and analysis: the state of the art." *arXiv preprint arXiv:1203.6722* (2012).

[2] Ratliff, Matthew S., and Eric Patterson. "Emotion recognition using facial expressions with active appearance models." In *Proceedings of the Third IASTED International Conference on Human Computer Interaction,(Innsbruck, Austria)*, pp. 138-143. 2008.

[3] Busso, Carlos, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. "Analysis of emotion recognition using facial expressions, speech and multimodal information." In *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 205-211. ACM, 2004.

[4] Emerich, Simina, Eugen Lupu, and Anca Apatean. "Emotions recognition by speech and facial expressions analysis." In *Proceedings of the 17th European Signal Processing Conference (EUSIPCO'09)*, pp. 1617-1621. 2009.

[5] Caridakis, George, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis. "Multimodal emotion recognition from expressive faces, body gestures and speech." In *Artificial intelligence and innovations 2007: From theory to applications*, pp. 375-388. Springer US, 2007.

[6] P. Ekman and W.V. Friesen, *"Manual for the Facial Action Coding System,"*Consulting Psychologists Press,1977.

[7] Litman, Diane J., and Kate Forbes-Riley. "Predicting student emotions in computer-human tutoring dialogues." In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 351. Association for Computational Linguistics, 2004.

[8] Valstar, Michel F., and Maja Pantic. "Fully automatic recognition of the temporal phases of facial actions." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 42, no. 1 (2012): 28-43.

[9] Christian Wallraven, Heinrich H. B¨ulthoff, Douglas W. Cunningham, Jan Fischer, and Dirk Bartz*."Evaluation of real-world and computer-generated stylized facial expressions."*ACM Transactions on Applied Perception, volume 4, page 16, New York, NY, USA, 2007. ACM.

[10] E. Holden and R. Owens, "Automatic facial point detection," in *Proc. Asian Conf. Comput. Vis.*, 2002, pp. 731–736.

[11] Scherer, K.R. and Ellgring, H.: Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? Emotion 7(1).

[12] Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: The state of the art. IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(12):1424–1445 (2000)

[13] H. Kobayashi and F. Hara, ªRecognition of Six Basic Facial Expressions and Their Strength by Neural Network,º Proc. Int'lWorkshop Robot and Human Comm., pp. 381-386, 1992.

[14] A. Samal and P.A. Iyengar, ªAutomatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey,º Pattern Recognition, vol. 25, no. 1, pp. 65-77, 1992.

[15] M.J. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," *Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition,* pp. 200-205, Japan, April 1998.

[16] MPEG Video and SNHC, "Text of ISO/IEC FDIS 14 496-3: Audio," in *Atlantic City MPEG Mtg.*, Oct. 1998, Doc. ISO/MPEG N2503.