International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 4
ISSN: 2088-5334

Multimodal Approach to Emotion Recognition for Enhancing Human Machine Interaction - A Survey

Veni. S [#], Thushara. S[#]

[#] *Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham University*
Amrita School of Engineering, Coimbatore, Tamil Nadu, India
E-mail: thushara.86@gmail.com1, s_veni@cb.amrita.edu2

Abstract **- Emotions are defined as a mental state that occurs instinctively rather than through voluntary effort. They are strong feelings triggered by experiencing the joy, hate, fear, love and is followed by some physiological changes. Emotions play a vital role in social interactions and facilitate the decision making and perception in human being. Emotions are conveyed through speech, facial expression or by physiological signals. There are 6 emotions which are treated as universal emotions: anger, happiness, sadness, disgust, surprise and fear. This paper projects different emotion recognition systems which aim at enhancing the Human-Machine interaction. The techniques and systems used in emotion detection may vary depending on the features inspected. This paper explores them in a descriptive and comparative manner. Further the various applications that adopt these systems to reduce the difficulties in implementing the models in real-time are contemplated. Also, A multimodal system with both speech and facial features is proposed for emotion recognition through which it is possible to obtain an enhanced accuracy compare with the existing systems.**

Keywords**- emotion recognition; human-machine interaction; multimodal; speech features; facial features**

I. INTRODUCTION

For a machine to behave like human being it should have the ability to acquire and spectacle the emotions. Moreover, for having a human-like interaction machines should learn how to identify the faces and to spot the emotions. To achieve an efficient human-computer intelligent interaction (HCII), the natural interaction with the user is very important. The main mediums of interaction in humans are speech, facial expressions and body gestures. Thus the new interactive technology combines the natural sensory modes like sound, touch and sight. This multimodal approach will enhance the understanding level and accomplish in building an efficient emotion recognition system.

In some cases recognizing emotions by computer is not required. For example, automatic teller machine or airplanes have computers that are not meant to recognize the emotions. However the applications like e-learning and humanoid robots, computers are playing a social work like instructor companion or helper, here recognizing the user's emotions will enhance the functionality. By synthesizing the speech it helps to identify the stress level of the user. According to the needs how the human beings are responding to the circumstances is reflected as emotions. When a user approaches a computer with some application specific purpose, how the human computer interface processes the intended goals has a great effect on the emotional state of the user. Such information are used as a feedback to identify whether the goal is met by the user or not [1] [2].

Psychologists and engineers have performed many researches to synthesis the vocal emotions, analyse the facial expressions and gestures to categorize and understand the emotions. Using this knowledge, computers are trained to recognize emotions from videos captured from build in cameras. Since the interaction between computer and humans is such a sizzling topic, different studies are attempted on this. Mainly the studies are concentrated on uni-model approaches for recognizing the basic emotions like anger, happiness, sadness and fear [3], [4], [5], [6], [7]. Human emotion recognition was also performed using EEG waves by Wan Ismail [8].

The drawback of previous studies is that the automatic emotion recognition (AER) is not giving good results in real life situations, where the expressions are not posted or staged. The emotions in staged performance are exaggerated but in real life situations the displayed behaviour is not artificial and to analyse the expressions by the same techniques will not result good output. Thus the researchers started to shift their focus towards spontaneous facial expression display by fusing the audio expression [9]. In addition to it for increasing the performance and robustness researchers started to fuse the facial and acoustic features. The fusion of features can be performed either at the decision level or before

classification. This paper discusses various unimodal approaches in emotion recognition (ER), different feature fusion parameters and limitations by these techniques and need for multimodal system.

Organization of the paper is as follows: Section II describes the Methods used for human insight of emotion recognition system and the related studies. Results and discussions are given in section III. system. Section IV summarizes the challenges of the existing system.

II. Material and Method

A. Methods Used for Emotion Expression Recognition

Fundamental component of being human is emotion. It mainly motivates the action by adding meaning and richness to human experience. Many definitions are cited for emotions, kleinginna in 1981 stated the emotion is a reaction to events deemed relevant to the needs, goals, or concern of an individual. Ekman [10] grouped the emotions into different categories, which includes happiness, sadness, anger, fear, disgust and surprise.

From this basic emotion theory we can conclude that most of the ERS process is based on similar emotions. But this emotions fails to cover the real life situations. In order to cover the daily interactions the selection of emotion categories are performed pragmatically. However, sentiments are the properties allotted to the object which motivate the user to perform task which is broadly classified as positive, neutral and negative sentiments [11]. Filko et.al.[12] performed emotion recognition using neural networks. Following sections elaborates the detailed review about the works based on ER.

B. Facial Expression Recognition Studies

Facial expressions impart a great role in ER. Several techniques are implemented to detect the region of face and measuring the displacement of specific points in different emotions. Most common technique used is sign judgment describes the appearance and message judgment concentrates on behaviour. Figure 1 depicts the basic block diagram of emotional recognition system using facial features.

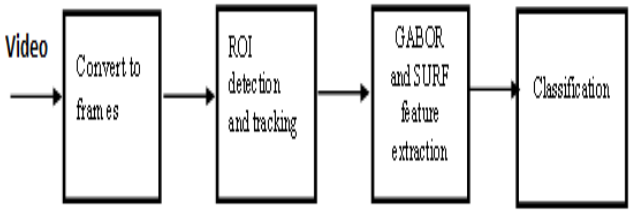

Fig. 1 Basic block diagram of facial ERS

Paul Ekman et.al., in early 1970s, carried an extensive study on facial expression and add up the evidence to support the universality of facial expressions which include happiness, anger, sadness, surprise, fear and disgust. The studies were also performed in different cultures. As a conclusion he proposed that 'display rules' govern the facial expressions in different social context. Facial Action Coding System (FACS) was developed by Ekman and Friesen to recognize the facial expressions. In this method, movements on face are defined by Action Units (AUs) [13]. They explain the relation of each AU with some muscular movements. Each emotion is described by combining different AUs. There are some prescribed rules for this emotion recognition system. Inputs used for this system are mainly still images where the emotions are depicted at its peak. Major disadvantage of this system is its time consuming processing stage. Many studies were based on emotion recognition by putting Ekman's work as inspiration [10] Emotions are categorized by tracking the facial features and measuring the facial movements .

An appearance model was implemented by Lanitis for person identification, gender recognition, and facial emotion recognition [14]. To recover the non-rigid motion, Black and Yacoob used local parameterized models [8]. Finally the parameters were fed into a rule based classifier.

Yacoob and Davis uses optical Flow (OF) to classify six facial emotions [16]. OF is also used by Rosenblum [4] to measure the facial region and then Radial Basis Function (RBF) network is applied on it for classification. Otsuka and Ohya [6] compute the OF and calculate the 2D Fourier transform coefficients which is given to Hidden Marcov Model (HMM) to classify the expressions. This system is capable to recognize one of the six emotions in real time. Static classifiers are adopted by chen for person dependent and person independent result to classify the emotions [14]. Cohen [7] defines two type of classification scheme: static and dynamic. In static system, the structure of Bayesian classifiers is first studied. The input given to the classifier is obtained from the face tracking system which is implemented for each frame in the video. But in dynamic system, a multi-level HMM classifier is implemented which allows the automatic segmentation of arbitrary long sequence to different expression segments. Amir Jamshidnezhad and Md Jan Nordin [17] used quantitative analysis in order to find the most effective features movements between the selected facial feature points.

TABLE I
COMPARISON OF FACIAL EMOTION RECOGNITION ALGORITHMS [2]

Author	Processing	Classification	Accuracy
Black and Yacoob [15]	Parametric	Rule-based	92%
Yacoob and Davis [16]	Optical flow	Rule-based	95%
Essaet.al [5]	Optical flow	Distance Method	98%
Otsuka [6]	2D FT Optical flow	HMM	93%
Lanitis et.al [14]	Appearance Method	Distance Method	74%
Chen [9]	Appearance Method	Winnow	86%
Cohen et.al [7]	Appearance Method	Bayesian Network	83%

In all these methods, features are extracted from the image first and then fed into a classifier and the output is one of the

emotion categories. This is different from feature extraction from the video images which includes video processing. It mainly falls into two categories: feature based and region based. In feature based approach, features like corners of mouth, eyebrows are detected and tracked. While in region based approach movement in certain areas like eyes, mouth are measured. Many classification algorithms are used to categorize the emotions. Table 1 gives the comparison of different classifiers [2].

There is some confusion in judging the six basic expressions. Ekman reported that anger and disgust are mainly confusing emotions. Fear and surprise also have this problem during judgment. Because of some similar facial actions these emotions get commonly confused [7], [15], [16].

Some researchers uses geometric method for feature extraction. Chang [18] uses 53 facial landmark points to detect and measure the target regions. Pantic [11] uses the characteristic points around eyes, eyebrows, chin to get the feature point. Studies extend to combine both the geometric and appearance based features. By combining features the performance accuracy of the system found to increase. Active appearance model is used to capture both the shape of facial features along with the facial appearance [17]. Piecewise Bezier Volume Deformation Tracker is defined by Huang [20] to extract the appearance and geometric based features from 3D face tracker. For real time emotion recognition this 3D face model is important as it is less controlled and has real time settings. Fusing speech features with facial features will improve the quality of the system which is addressed in the next section.

TABLE II
FACIAL CUES AND EMOTIONS [2]

Emotion	Observed Facial Clues
Surprise	Brows are curved Skin under brows get stretched Horizontal Wrinkles appears on Forehead Eyelids get opened up
Fear	Brows are drawn together Forehead wrinkles are drawn towards centre Lower eyelid is drawn up and upper one get raised Mouth opens Lips get stretched
Disgust	Upper lip get raised Nose get wrinkled Lower lip goes towards upper lips Brows get lowered Cheeks get raised
Anger	Vertical lines between brows Brows lowered Lower and upper lip get tensed Lips pressed towards the centre Eyes get a bulging appearance Nostrils get dilated
Happiness	Teeth get exposed Wrinkles runs down from nose Cheeks get raised Lower eyelids get wrinkles but not tensed
Sadness	Inner corners of lips are drawn up Upper lid is raised Wrinkles below the brow

C. Vocal Emotion Recognition Studies

Another important feature to recognize emotions is served from speech signal. There are explicit (linguistic) messages and implicit (paralinguistic) messages in speech signal. Prosodic features of speech are mainly used by the researchers for acoustic analysis. Table 3 depicts different emotional behaviour along with the common acoustic features. Lexical dictionaries acknowledge the linguistic cues of emotions present in the text. The study is based on mainly language dependent and generalising it is also difficult [16]. Figure 2 depicts the basic block diagram of emotional recognition system using speech signal.

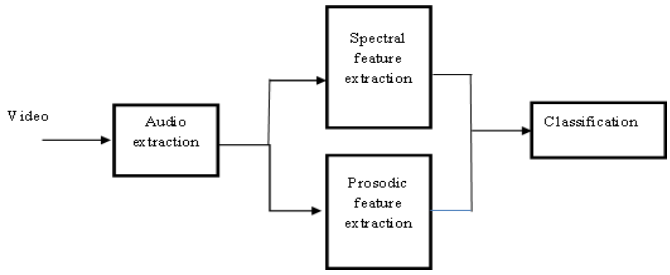

Fig. 2.Basic block diagram of speech ERS

Various kinds of information are embedded in the vocal cord during the utterance of messages. If we consider only the verbal part by disregarding the way in which the messages are spoken, then we might completely misunderstand the meaning of the messages.

TABLE III
VOICE AND EMOTION [1]

	Fear	Anger	Sadness	Happiness	Disgust
Speech rate	Much faster	Slightly faster	Slightly slower	Faster or slower	Very much slower
Pitch average	Very much higher	Very much higher	Slightly lower	Much higher	Very much lower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Normal	Highest	Lower	Higher	lower
Voice quality	Irregular voicing	Breathy chest tone	Resonant	Breathy blaring	Chest tone
Pitch changes	Normal	Abrupt on chest	Downward inflections	Smooth upward	Wide downward
Articulation	precise	Tense	Slurring	Normal	normal

Studies on speech emotion recognition had started from early 1930s. In most of the studies from traditional to recent works, prosodic features are used to extract information from speech signal. Prosodic features include pitch, intensity and duration of utterances [21]. Spectrum of real emotional

speech is compared with the acted speech by Williams [22] and found many similarities in their spectrograms. Nowadays majority of works are held on investigating the contribution of vocal cord in generation of speech. Five features were extracted from speech and a multilayer neural network is constructed for classifying the emotions [23]. 79.5% accuracy is obtained by using 17 features and different classification algorithms. In his studies, there were 4 categories of emotions. Some research was also performed by comparing the human and machine emotion recognition. Petrushin [24] conducted his study by taking 30 subjects speaking 4 sentences for each emotion categories and the accuracy obtained was around 65%. Some large scale studies were also conducted by using professional actors. 29 feature sets were used for that work. From the studies it was found that sadness and anger are the best recognizable emotions whereas fear, joy and disgust gives the worst result among all the emotions.

Rule-based method was proposed by Chen [9] to classify the input audio data into any one of these categories: happiness, fear, sadness, anger, dislike and surprise. Study consists of two different languages Spanish and Sinhala. Different languages were considered to reduce the linguistic influence. Each speaker spoke 6 different sentences for each emotion classes. Pitch contours and intensity were calculated from the speech signal which is classified by using some predefined rules. Table 4 gives the summary of vocal affects which are listed in relation to neutral voice.

D. Multimodal Approach To Emotion Recognition

Several efforts are made to fuse the facial and acoustic system together. Prosodic and spectral features are extracted from audio signals whereas distances, geometric and appearance features from specific facial points are extracted from visual data. When the modalities are fused together then the performance of the system increases. Audio-vision recognition mainly includes three fusion strategies: feature level, model level and decision level. In feature level fusion different features of unimodal systems are combined to construct a joint vector. Prosodic features of speech signal and geometric feature of video frames are combined at the earlier stages and is given as the input to the classifier [25]. Decision level fusion independently processes each unimodal systems and combines the result at the end.

For speech recognition combining audio and visual data is performed in recent years. When speech waveform appears to be noisy, information derived from the lip movements add up the performance of the system. The speech sounds and lip movements are tightly coupled in speech recognition. But for emotion recognition very little studies were performed on coupling multi modalities.

De Silva [26] projected a rule-based system in which singular classification of audio-visual data is addressed. Each subject is asked to act 12 different emotions by uttering a single English word. Processing of visual and audio data is performed separately. Displacement and velocity of the facial points (mouth and inner corners of eyebrow) are extracted by using OF method. Pitch and pitch contours are estimated from the speech signal. Nearest neighbour method is proposed for facial model whereas acoustic model uses HMM model for classification [27]. For each subject the classification result is plotted and based on these the rules for classification is defined. Busso et.al [28] proposed a system that uses singular classification of basic emotion categories. 5 subjects display 6 emotions 6 times with appropriate vocal expression. This emotion sequences starts and ends with neutral expressions. The single model classification is done in a sequential manner.

TABLE IV
SUMMARY OF HUMAN VOCAL CORD EFFECTS

	Anger	Happiness	Sadness	Fear	Disgust
Speech Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much slower
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy	Blaring	Resonant	irregular	Grumbled

1) Fusing Multimodal Parameters: Main issue in multimodal ER is that the data is processes separately and is only combined at the last stages. But Busso et.al [28] came out with a finding that for accomplishing human like analysis multiple input signals acquired at different sensors cannot be combined in a context free environment. Instead the input data should be made in a joint feature space according to a context-dependent model. The problem arises by using this fusion is their large dimension feature vector, different feature formats and timing. In order to overcome all these drawbacks Pantic et.al [11] developed a tightly packed multimodal data fusion to develop a context-dependent system by using Bayesian interference method [29].

When we are considering a highly efficient multimodal system it should be compactable for some imperfect data like noisy data and partial data. To make it compactable, Pantic suggested a method by considering the time-instance versus time-scale dimension of non-verbal communicative signals [11]. In this approach, they considered the previously obtained sample with the current data and computed the statistical prediction. The probability derived from this will give the information about the losses happened due to the malfunctioning or inaccuracy of the input data. Probability graphical models like HMM, Bayesian network, etc., are very much suited for these fusion methodologies. These models can also deal with missing values of features, temporal features and noisy features. For facial ER HMM system provide good results [14]. Fusion of dynamic Bayesian and HMM is used in office activity recognition and event detection from audio visual information. Vocal emotions are predicted from acoustic features extracted from audio tract and facial emotions from the facial features tracked from video data. Visual and audio cues are used to

recognize whether a person speaks or not. Emotions can be recognized even when some information are missing like noisy audio or losing of video track in multimodal systems [29].

Another issue in emotion recognition system is the influence of culture, social vicinity and the current mood in which the observed behavior is encountered [30]. Thus comes the machine learning approach in emotion recognition. Many models are adapting these well-known algorithms and for learning a new model it is possible to use the prior information. i.e, a prior model trained on certain users is used as a starting point of the new model for different users. The time requirement and sensing are the problems of this system.

The main goal of emotion recognition system is to recognize the emotions of a person in natural situations. Forcing someone to smile does not have the real feeling of an authentic smile. Lacking of the actual feelings is the fundamental reason for these artificial expressions. According to Picard [30] five factors influence the data collection.

Posed versus spontaneous: Whether the emotions have to be elicited as a response to the stimulus or subject is asked to produce the emotions.

Lab-setting versus real world: Whether the emotions have to be captured from the laboratory or recorded from the day to day life of the subject.

Expression versus feeling: Whether the importance has to be given to external expressions or to the internal feelings

Open-recording versus hidden-recording: Whether the subject should be aware that he/she is being recorded

Emotion purpose versus other purpose: whether the subject should be aware that he/ she are a part of the experiment.

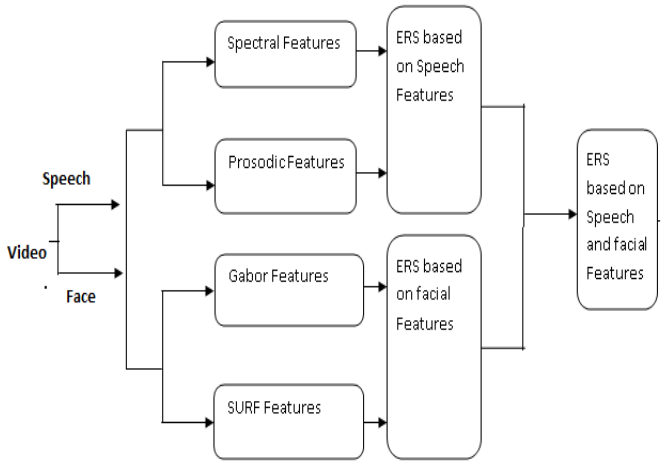

Fig. 2 Proposed Multimodal Emotion Recognition System

2) Proposed System for Fusion of Multimodal System with Speech and Facial Features : Based on the studies performed, it is essential that an efficient ER system is required to improve the quality. One solution is to design a feature fusion based system is proposed and the methodology used is depicted in Figure 2. In this system, the best features of the two systems are fused to get better accuracy. Multimodal system can be built by two ways either by the fusion of the appropriate features or by taking the match scores. It was observed from the experiments that fusing best features outperform the matching score method. For speech, spectral and Prosodic features are used. For face, Gabor and SURF features are used.

III. Results and Discussions

As said by Go et.al [29], emotional skill plays a crucial role in intelligence. Studies reveal the importance of emotional abilities in finding a way to human-machine interaction. As a result many researchers are conducting in this field to develop machine intelligence. Emotion controls many modes of human communications like gestures, facial expression, postures, voice tone, respiration, skin temperature. Expressions can even change the meaning of the messages. Most visible emotion communication tends to be the face but according to the studies by Picard [30] the expressions on face can be controlled to great extend compared to voice or some other modes. The recognition is likely to be accurate when multiple modalities, user's context information, goal and preferences are included [2]. Combining the high-level features, low-level features and natural processing of language provide best emotion recognition.

In the proposed system, both prosodic and spectral features are combined to get the speech ER system. The accuracy by using the polynomial kernel is 84.21%. Similarly, face features alone gives accuracy as 90.47%. The proposed multimodal system gives 94.73% accuracy compared with the existing system with accuracy using 93.62% using morphological and spectral features.

IV. Conclusions

The current system is only dealing with the profile view of face image that have appreciable resolution and lighting conditions. However it is not the case with real time system. In real time, disturbances like hand occlusion movement and low resolution might be present which can decrease the throughput of the system [21]. Next challenge is how unfailingly extract the paralinguistic features and linguistic features from the acoustic channel. Now the studies are mainly done by extracting the prosodic information from the speech. But the results are not up to the mark. Thus to increase the efficiency of the system the linguistic information like repetition, correction, syntactic information has to be done [32]. Fusion of modalities still appears to be problem in multimodal recognitions. Challenge is really in creating the feature set combining the features from different models in different time scale, different dynamic structures and different metric level [31]. The advanced machine learning techniques like deep learning neural networks are expected to produce more accuracy for these types of fusion techniques [33].

References

[1] K. Sreenivsa Rao and Shashidhar G. Koolagudi, "Recognition of emotions from video using acoustic and facial features," International journal on Signal Image and Video Processing, Vol. 9, Issue. 5, pp. 1029-1045, July 2015.

[2] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J.G , "Emotion Recognition in Human-

Computer Interaction," IEEE Signal Processing Magazine, pp. 32-80, Aug 2002.
[3] Simon Dobrisek, Rok Gajsek, France Mihelic, Nikola Pavesic, and Vitomir Struc, "Towards Efficient Multi-Modal Emotion Recognition, "International Journal on Advanced Robotic System, Vol. 6, No. 1, January 2015.
[4] M. Rosenblum, Y. Yacoob, and L. Davis, "Human expression recognition from motion using a radial basis function network architecture," IEEE Trans. on Neural Network Vol.7, Issue.5, pp. 1121–1138, 1996.
[5] Essa. I. A and A.P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," IEEE Trans. on Pattern Analysis and Machine Intelligence Vol.19, No.7, pp. 757–763, 1997.
[6] T. Otsuka and J. Ohya, "Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences," Proc. International Conf. on Image Processing, pp. 546 - 549, 1997.
[7] I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," International Journal on Computer Vision and Image Understanding, Vol.91, Issue 1-2, pp. 160–187, 2003.
[8] W.O.A.S. Wan Ismail, M. Hanif, S.B.Mohamed, Noraini Hamzah, and Zairi Ismael Rizman, "Human Emotion Detection via Brain waves Study by Using Electroencephalogram (EEG)," International Journal Advance Science Engineering Information Technology (IJASEIT), Vol..6, No. 6, 2016.
[9] L. Chen, "Joint processing of audio-visual information for the recognition of emotional expressions in human computer interaction," Thesis, University of Illinois at Urbana-Champaign, 2000.
[10] P. Ekman and W. Friesen, Facial Action Coding System: Investigator's Guide, Consulting Psychologists Press, 1978.
[11] Pantic and Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," IEEE Transactions on. System, Man, Cybernatics society, Vol. 36, Issue.2, 2006
[12] Filko. D and Martinovic "Emotion recognition system by a neural network based facial expression analysis," AUTOMATIKA 54, 2013.
[13] Ekman, Paul, Freisen, Wallace V, Ancoli, Sonia, "Facial signs of emotional experience," Journal of Personality and Social Psychology, Vol. 39, Issue.6, pp.1125-1134, 1980
[14] A. Lanitis, C. J. Taylor, and T. Cootes, "A unified approach to coding and interpreting face images, " Proc. International Conf. on Computer Vision, pp. 368–373, 1995.
[15] M. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion," Proc. International Conference on Computer Vision, pp. 374 -381, 1995.
[16] Y. Yacoob and L. S. Davis, "Recognizing human facial expressions from long image sequences using optical flow," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.18, Issue.6, pp. 636 - 642, 1996.
[17] Amir Jamshidnezhad, and Md Jan Nordin, "A Classifier Model based on the Features Quantitative Analysis for Facial Expression Recognition, "International Journal Advance Science Engineering Information Technology (IJASEIT), " Vol. 1, No.4, pp. 391-394, 2011
[18] Chang Y, Hu C, Feris R and Turk M, "Manifold based analysis of facial expression. J. Image and Vision Computing, Vol. 24. No.6, 605-614, 2006
[19] Mina Navran and Nasrollah Moghadam Charkari, 'Fusion of Feature Sets for Facial Expression Recognition," IEEE Transactions on Telecommunication, Vol. 62, No. 6, 2014
[20] L. Huang, F. Thawn, and L. Didaci,"Bimodal emotion recognition by man and machine," Pattern Recognition, Vol. 42, No. 11, pp. 2807–2817, Nov. 2009.
[21] Y. Sagisaka, N. Campbell, and N. Higuchi, "Computing Prosody", Springer-Verlag, New York, NY, 1997.
[22] C. Williams and K. Stevens, "Emotions and speech: Some acoustical correlates, " Journal of the Acoustic Society of America Vol. 52, Issue 4B, 2005
[23] C. Chiu, Y. Chang, and Y. Lai, "The analysis and recognition of human vocal emotions," in Proc. International Computer Symposium, pp. 83-88, 1994.
[24] V.A. Petrushin,"How well can people and computers recognize emotions in speech," Proc. AAAI Fall Symposium, pp. 141–145, 1998.
[25] VN. Vapnik V, "Overview of Statistical Learning Theory", IEEE Transactions on Neural Networks, Vol.10, Issue.5, 1999.
[26] L. De Silva and P. Ng, "Bimodal emotion recognition," Proc. Automatic Face and Gesture Recognition, pp. 332 - 335, 2000
[27] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," IEEE Trans. on Signal Processing Vol.39, Issue.1, pp. 40-48, 1991.
[28] C. Busso, S. Lee and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," IEEE Transactions on Audio, Speech and Language Processing, Vol. 17, Issue. 4. 2009.
[29] Hj. Go, KC. Kwak, DJ. Lee DJ, and MG.Chun ," Emotion recognition from facial image and speech signal," International Conference of the Society of Instrument and Control Engineers (SICE), 2003.
[30] R. Picard, E. Vyzas, and J. Healey, Toward machine emotional intelligence: Analysis of affective physio-logical state, IEEE Trans. on Pattern Analysis and Machine Intelligence 23(10), pp. 1175–1191, 2001
[31] Hoch, S., Althoff, F., McGlaun, G., Rigoll, G , "Bimodal fusion of emotional data in an automotive environment," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05), 2005.
[32] D.Govind and SRM. Prasanna, "Epoch extraction from emotional speech," Proc. International Conference on Signal Processing and Communications (SPCOM), 2012.
[33] S. Poria, E. Cambria and Gelbukh, A," Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," Proc. Conference on Empirical Methods in Natural Language, pages 2539–2544, 2015.