Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

Review Paper on Human Emotion Recognition Using Audio-Visual Modalities

Naveen Kumar H N[1], Dr. JagadeeshaS[2]
[1]*Department of E&C, SDMIT Ujire*
[1]*Department of E&C, SDMIT Ujire*

Abstract --- Automatic emotion recognition from voice and face has become a core discipline in machine learning and pattern recognition. From the machine perspective, recognizing the user's emotional state is one of the main requirements for computers to successfully interact with humans. Performance of human emotion recognition system can be improved my combining more input modalities for the interpretation. This paper provides a survey on human emotion recognition by using multimodal signals such as facial and acoustic features. The important issues like collection and availability of testing and training data and challenges are addressed. A brief introduction to existing techniques on human emotion recognition and future enhancements that can be done on emotion recognition system are represented clearly.

Keywords--- Facial Expression Recognition (FER), Hidden Markov Model (HMM), Support Vector Machines (SVM), Human Computer Interaction (HCI).

I. INTRODUCTION

Emotion is a complex psychophysiological experience that results in physical and psychological changes that influence an individual's behavior. Emotion modulates almost all modes of communication such as facial expressions, gesture, posture, and tone of voice, selection of words, respiration and skin temperature. Nonverbal communication is probably less affected than verbal behavior by the censoring of communication. Nonverbal communication may provide feedback on whether someone is listening, is bored, and is getting ready to talk. Speech signal disclose much information about age, gender, race and emotion. Facial expressions are used to analyze human emotion by observing the change of shape and distance of facial regions. Significant amount of information for detecting and interpreting emotion can also be obtained by the body languages of the head. Bodily expressions reveal unconscious attitudes about one's effect. Gaze direction play's its role in determining whether the communication is approach oriented or avoidance oriented.

The research work on human emotion recognition system is highly motivated by the recent scientific findings about the role of emotional abilities in human intelligence and on the way human machine interaction imitates human-human interaction. Today's laptop's or other widely used portable devices supports audio-visual modalities, which is the added advantage. Recognition is most likely to be accurate when it combines multiple modalities, information about the user's context, situation, goal and preferences. A combination of low level features, high level reasoning and natural language processing is likely to provide the best emotion inference.

II. LITERATURE REVIEW

Kartick Subramanian et al [1], proposed Database Independent Human Emotion Recognition with Meta-Cognitive Neuro-Fuzzy Inference System. Their work focused on the generalized case of subject-independent emotion recognition. The emotion recognition system employed is Meta-Cognitive Neuro-Fuzzy Inference System (McFIS). McFIS has two components, a Neuro-fuzzy inference system, which is the cognitive component and a self-regulatory learning mechanism, which is the meta-cognitive component.

The meta-cognitive component monitors the knowledge in the Neuro-fuzzy inference system and decides on what-to-learn, when-to-learn and how to- learn the training samples, efficiently. For each training sample, McFIS decides on whether to delete the sample without being learnt, employ it for addition/ pruning or parameter update or reserve it to be used later. They used JAFFE and TFEID database to evaluate the performance of the algorithm. They used two tests, a 5-fold cross validation study to measure the performance of McFIS on individual databases and an inter-database predictability check, wherein, McFIS trained on JAFFE is tested with TFEID database and vice-versa. The performance of the proposed algorithm compared with standard SVM indicates promising results. The present work fails to evaluate the performance of features extracted based on other techniques, such as curvelet, Gabor based features, etc. A deeper analysis has to be conducted on the transferability of emotion from one database to another.

Sakmongkon Chumkamon et al [2], developed Robot's Eye Expression system for Imitating Human Facial Expression. It is very difficult to perceive and interpret the specific emotion from the eye expressions. The Facial Expression Recognition (FER) system can recognize the basic emotion of the human consequently convey the emotional expression command to the robot's eye expression system for eyes expression of the robot. FER system is implemented using Constrained Local Model (CLM) and Hidden Markov Model (HMM) including the head robot system and the 3D virtual eye. Their results were shown that the proposed system was able to recognize the emotion from the user's face. Consequently, the robot could express the imitating emotion with the expression of the robot eye, which was expressed in the 3D virtual eye that was displayed by the small LCD in the head of the robot. This system can operate the FER and expressing the eye robot around 12 frames per second. It's very challenging to design an independent-person FER system to assure the robot can cooperate with the human in the real world.

Dragos Datcu et al [3], developed an automatic bi-modal emotion recognition system based on fusion of facial expressions and emotion extraction from speech. The system is bimodal and is based on the fusion of data regarding facial expressions and emotion features that are extracted from the speech. Voila –Jones face detector (open CV) used to detect face from video frames and active appearance model for extracting face shape. SVM used for classification, Optical flow algorithm for computing features needed for the classification of facial expressions. For training and testing Cohn-Kanade database is used. The parameters used for classification consist of the mean, standard deviation, minimum and maximum of the following acoustic features, fundamental frequency (pitch), intensity, F1, F2, F3, F4 and Bandwidth. Fusion model aims at determining the most probable emotion of the subject given the emotion determined in the previous frames. Dynamic Bayesian Network is used for emotion data fusion. The main drawback of the proposed work is higher probability of generating faulty results for the extraction of face shape. Effective face tracking can be used to overcome the above.

Urvashi Agrawal et al [4], proposed Emotion and Gesture Recognition with Soft Computing Tool for Drivers Assistance System in Human Centered Transportation. Facial gesture and emotion are used to identify the driver's attentiveness and in case of less attentiveness the vehicle will be switched to automatic mode. Facial gestures are detected by motion of eyes and lips. The basic emotions like happy, anger, sad and surprise are classified by the different facial expressions. Fuzzy rule based system was used for the better system performance. The Fuzzy logic systems are good at explaining their decisions since they can process imprecise information and they cannot automatically acquire the rules they use to make those decisions.
D'Melloet et al [5], developed a multimodal system which combines conversational cues, gross body language, and facial features to identify the inferred emotion. The recognition results showed that the accuracy of multimodal channels was statistically higher than the best recognized single-channel

model for the fixed emotion expressions. Although the combined channels yielded improvement in recognizing some emotions, they may provide redundant and additive information for other emotions. Carlos Busso et al [6], investigated that emotional content affect the relationship between facial gestures and speech. They showed that facial and acoustic features are interrelated, with higher levels of correlation, which presents significant emotional differences. Hidden Markov Models (HMMs) are used to analyze audio-visual mapping in their work. Limitation of their work is that, they analyzed the facial gestures and speech of a single actress.

Stefanos Kollias et al [7], proposed the best possible techniques for multimodal emotion recognition in HCI applications based on a physiological background. Their work mainly deals with audio and visual emotion analysis, with physiological signal analysis serving as supplementary to these modalities. This work mainly addresses the extraction of emotional features and signs from each modality in separate and integration of the outputs of each modality in order to recognize user's emotional state, by considering emotion models and existing knowledge from both the analysis and synthesis perspective. Their work doesn't address the handling of conflict information conveyed by modalities and optimizing information with high disparity in accuracy.
Alejandro jaimes et al [8], proposed a technique for the automatic classification of unimodal data, bimodal data and multimodal data using Bayesian classifier. The main contribution of their work is the integration of three modalities (Body gesture, Facial expression and Speech information) for the emotion recognition. They compared feature level fusion with decision level fusion and concluded that, feature level fusion is more accurate for emotion recognition. This work fails to address scenarios like, robustness to occlusions, noisy background and head motions. One more important issue is development of methods for multimodal fusion that take into account the mutual relationship between feature sets in different modalities, the correlation between audio-visual information and the amount of information that each modality conveys about the expressed emotion.

Angeliki Metallino et al [9], developed a method for Audio-Visual Emotion Recognition using Gaussian Mixture Models for Face and Voice. Their work aims at improving emotion recognition accuracy by combining facial and vocal modalities in very effective way. Individual modality recognition performances indicate that anger and sadness have comparable accuracies for facial and vocal modalities, while happiness seems to be more accurately transmitted by facial expressions than voice. The neutral state has the lowest performance, possibly due to the vague definition of neutrality. Cheek regions achieve better emotion recognition accuracy compared to other facial regions. Moreover, classifier combination leads to significantly higher performance, which confirms that training detailed single modality classifiers and combining them at a later stage is an effective approach. To improve the performance more detailed modelling of the spatial and temporal relations face and voice modalities can be considered.

III. CHALLENGES

3.1. Challenges involved in the selection of database are:

- The subject feels the emotion internally (not acted).
- The subject should be in a real world environment instead of lab environment and emotions should occur spontaneously.
- The subject should not know that he or she is part of an experiment.
- The performance of research work highly dependent on the selection of the database.

3.2. Challenges involved in the speech emotion recognition:

- In case of speech emotion relation between linguistic content and emotion is language dependent. Generalizing from one language to another is difficult to achieve.
- Same utterance may show different emotions.

- Differentiating between various emotions with particular speech features are most useful, but not clear because of the existence of the different sentences, speakers, speaking styles and speaking rates.
- Each emotion may correspond to the different portions of the spoken utterance. It's very difficult to differentiate these portions of utterance. The implicit messages have not yet been fully understood.

3.3. Challenges involved in the facial expression emotion recognition:

- Recognition results dependent on image or video quality. It's very complex to identify the exact facial expression from a blurred facial image.
- Segmentation of a facial image into regions of interest is difficult, particularly when the regions do not have significant differences in their imaging attributes.
- Unlike humans, machines usually do not have visual perception to map facial expression into emotions.

IV. CONCLUSION

This paper presents a brief review of the work carried on human emotion recognition using audio and visual modalities. It's very useful to consider the context information for emotion recognition since emotion is highly dependent on the context. Existing techniques doesn't model the contextual information.

From the exhaustive literature survey it is found that the accuracy can be improved by considering the spatial and temporal relations of the face and voice modalities. It's very challenging to develop methods for multimodal fusion that take into account the mutual relationship between features sets in different in different modalities, and the correlation between audio-visual information. Adjustable weighted segmentation method to determine the final results of emotion recognition by combining the data from facial images and speech signals.

To improve the performance of speech emotion recognition system modulation spectral features can be used. Facial expressions can be accurately recognized by combining appearance and geometric features. It's very challenging to handle conflict information conveyed by modalities. Few emotions like intimacy and anxiety are often expressed nonverbally long before they are expressed verbally. It's very difficult to recognize these emotions.

REFERENCES

[1]. Kartick Subramanian, VenkateshBabuRadhakrishnan and SavithaRamasamy,"Database Independent Human Emotion Recognition with Meta-Cognitive Neuro-Fuzzy Inference System", 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP) Symposium on Cognitive Computing in Information Processing.

[2]. SakmongkonChumkamon, Koike Masato and Eiji Hayashi, "The Robot's Eye Expression for Imitating Human Facial Expression" 978-1-4799-2993-1/14/ ©2014 IEEE.

[3]. DragosDatcu, LeonJ.M. Rothkrantz, "Automatic bi-modal emotion recognition system based on fusion of facial expressions and emotion extraction from speech",978-1-4244-2154-1/08/$25.00 ©2008 IEEE.

[4]. UrvashiAgrawal, ShubhangiGiripunje, Dr.Preeti Bajaj, "Emotion and Gesture Recognition with Soft Computing Tool for Drivers Assistance System in Human Centered Transportation", 2013 IEEE International Conference on Systems, Man, and Cybernetics.

[5]. S.D'Mello, and A.Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," User Modeling and User-Adapted Interaction, vol. 10, pp 147-187, 2010.

[6]. Carlos Busso, Shrikanth S Naratanan, "Interrelation between Speech and facial gestures in Emotional Utterances: A single study", IEEE Transactions on audio, speech and language processing, July 2007.

[7]. StefanosKollias and Kostas Karpouzis, "Multimodal Emotion Recognition and Expressivity Analysis", 0-7803-9332-5/05/2005, IEEE.

[8]. Alejandro Jaimes and NicuSebe, "Multimodal Human Computer Interaction: A Survey", IEEE International workshop on Human Computer Interaction in conjunction with ICCV, Beijing, China, Oct 21, 2005.

[9]. Angeliki Metallinou, Sungbok Lee and Shrikanth Narayanan, *"Audio-Visual Emotion Recognition using Gaussian Mixture Modelsfor Face and Voice",* Tenth IEEE International Symposium on Multimedia 2008.