International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1739-1745

http://www. irphouse.com

Spontaneous Affect Recognition from Audio-visual Cues using Multi-resolution Analysis

Mayank Kumar Aditia[1], Gyanendra K. Verma[2]

[1,2]NIT Kurukshetra - 136119 (Haryana)

Abstract

It is essential for computer to get affective state of the user for better Human-Computer Interaction. The audio-visual signals play a major role to know the change of human behaviour or affective state of the human being. This paper presents a framework for spontaneous affect recognition from audio-visual cues. To predict spontaneous emotion, the spectral and prosodic features from audio signal and visual features from video signal are extracted followed by Fisher Discriminant Analysis (FDA) in order to select most discriminative features. The multiresolution analysis (MRA) of audio-visual signals is performed using the Discrete Wavelet Transform, a classical transform for analysis of one/two dimensional signals. The spectral and prosodic features from audio signal are also extracted using Mel Frequency Cepstral Coefficient (MFCC). The experiments were performed on eNTERFACE and RML database. The performance of the proposed system was measured in terms of precision, recall and f-measures. Multilayer Perceptron (MLP) and Support Vector Machine (SVM) classifiers are deployed to classify six categories (Anger, Disgust, Fear, Happiness, Sadness and Surprise) of basic emotion. The experimental results prove that our system shows significant performance with 80.5% and above accuracy with fusion of audio-visual cues.

Keywords: Affect Recognition, Discrete Wavelet Transforms, MFCC, SVM and MLP.

Introduction

Affective Computing is a prime research area of Human Computer Interaction (HCI) which combines engineering and computer science with Cognitive Science, Sociology, Physiology, Psychology and many more. In the last few decades most of the research was based on data of spontaneous as well as posed data acquired in the laboratory setting for affect recognition. The different affective states like thinking,

embarrassment or depression can be considered complex affective states and expressed via dozens of anatomically possible facial expression or body gesture.

This work presents the framework to predict emotion from audio-visual modalities. The correlation between different emotions demonstrates significantly improvement in performance. We compared two machine learning techniques (i.e. Support Vector Machine and Multilayer Perceptron) for continuous affect prediction. The rest paper is organized as follow: literature review is described in section 2. The methodology is explained in section 3. Experiments and results are given in section 4 followed by concluding remark in section 5.

Related work

In real life communication information is shared among the persons. And it matters that how easy we can communicate. In human to human communication non-verbal cues are often used like facial expression and variations in voice tones. But unfortunately, we do not have a good human-computer interfaces which can take advantages of these valuable communicative mediums. Emotion plays an important role in human communication to recognize the human feelings.

S. Koelstra et al. [1] proposed an emotion recognition system by using fusion of facial expressions and EEG signals. They we utilize methods for facial expression and EEG signal analysis to investigate the possibilities for multimodal fusion in affect recognition and implicit tagging.

Y. Wang et al. [2] investigated kernel based methods for multimodal information analysis and fusion. They utilized Kernel cross modal factor analysis for modelling the nonlinear relationship between two multidimensional variables. They have also introduced an approach to identify optimal transformations to represent coupled patterns between two different subsets of features.

M. Paleri et al. [3] presented an extensive study on feature selection for automatic audiovisual person independent emotion recognition. They used neural network in order to compare the performance of the emotion recognition system using different features.

An asynchronous feature fusion method for clustering and classification of basic affective states from facial expressions and speech was proposed by M. Mansoorizadeh et al. [4] in order to create a unified hybrid feature space. They claimed that the proposed fusion approach performs better than unimodal face and speech based system. They have also provided comparative results based on synchronous feature and decision level fusion approaches.

D. Datcu et al. [5] presented a fusion model based on Dynamic Bayesian Network (DBN) for emotion recognition from speech and video. They used berlin database for speech and kohn- kanade for facial emotion recognition along with partly eNTERFACE database. A two fold cross validation method has been applied in order to detect six basic emotions.

Proposed Methodology

Recognizing emotion from multimodal cues (audio, video etc.) is an interesting and challenging problem before researchers, for more emphasis on better human computer interaction. A framework has been proposed based on the multi-resolution analysis (MRA) of the audio-visual cues. MRA analyzes the signal at different frequencies with different resolutions.

Pre-processing of video cue.

We have used enterface and RML database for experimentation. The dataset contains the frontal facial image of different subjects with other background objects.We have used face detection utilities of Viola Jones [6] to extract the frontal face. The size of the extracted frame is 720 × 576.

Facial Features Extraction.

Most common methods for facial expression recognition are based on facial action coding system (FACS) introduced by Ekman and Friesen [7]. Every expression is identified by the action unit (AU). Multi Resolution Analysisis currently most growing technique for feature extraction from the one or two dimensional signals. By using MRA we can analyse signals at multi resolution levels for better feature selection. Essa and Pentland [8] used modular Eigen space for face localization.Since a geometric feature based face analysis is complicated when the appearance of changes due to illumination and pose variations occurs, we decided to use spatiotemporal features to detect small changes in a face. We use the Dollar's method [9] to detect the interest points in image sequences. The method was originally designed to detect small changes in the spatial and temporal domain mostly explored by the human action recognition society. We have extracted spatial features for image sequence using MRA. [14, 15]

Audio Features Extraction.

Different speech features represent different speech information, e.g. emotion, speaker, in highly overlapped manner. These have motivated intensive research of audio emotion recognition in discovering the significant manner of the audio features on specific emotion. Audio features can be classified into 3 groups: vocal tract system, prosodic, and excitation source features. We have extracted wavelet and MFCC features from audio signal to yield feature vector.

MFCC is the Cepstral Coefficients derived from a Mel scale frequency filter bank [10]. In computations of MFCC, audio signal is firstly divided into multiple frames of equal duration. The frames overlap each other to preserve the continuity of the audio signal. Each frame is subsequently multiplied with a Hamming Window so that the continuity of the left and right side of the frame can be increased or maintained. The Hamming Window, w(n) used in MFCC is defined with the equation 1.

$$Mel(n) = 2595 + \log_{10}\left[1 + \left(\frac{n}{700}\right)\right] \quad (1)$$

Feature Level Fusion.

The multiresolution analysis is being performed at different levels in order to obtain

feature bins from different modalities (audio-visual). Then all the features are combined to yield feature bins. Further, the fusion of feature bins is performed, after selection of optimal features for each modality. Finally, the combined feature vector is feed to classifier to predict emotion. The fusion results of two modalities are given in table 1.

Experiments and Results

All the experiments were performed on eNTERFACE [11] and RML [12] emotion database. eNTERFACE database contains 44 subjects; out of them we have selected randomly 8 subjects. RML database contains 8 subjects; from which all of the subjects were used for the experiment. The video frames and audio is extracted from video data of both database. For video feature extraction 80 sample images are selected for each class of the emotion (total 480 sample images with all six emotions) for the audio processing, all the audio sample files are truncated to 2 second long. The dimensionality of audio and visual features set to 26 and 62 respectively. We have used 10-fold cross-validation for evaluation, in which each time 75% of samples are randomly selected for training and the rest for testing. This process is repeated 10 times, and the average of the results is reported.

All the experiments were carried out in Matlab 7.7.0 (R2012), on a 32bit Intel 3.06 GHz processor, with 2 GB RAM. The wavelet transformation was done by using Wavelet toolbox available in Matlab. The SVM-Light software was developed in C using Vapnik's support vector Machine, available at [13].

We have performed multiresolution analysis of signal using discrete wavelet transform. Multiresolution analysis of signal provides more optimal features as the analysis is done with different resolution and scale. Initially results are tested on all of the wavelet families (e.g. Daubechies, Haar, DMeyer etc.). DMeyer (dmey) wavelet family provides good feature, therefore we have used DMeyer function to extract features and formed feature vectors. Min-max normalization, a classical data normalization algorithm is being used to normalize data for classification purpose. The experiments were performed individually by wavelet analysis of each audio and video/images signals. A feature level fusion is also performed in order to fuse the information obtained from both cues i.e. audio and video signals.

Fig.1 and Fig.2 shows the single cue accuracy for audio and video modalities for eNTERFACE and RML database respectively. The results are given in terms of precision, recall and f-measure for six categories of emotions anger, disgust, fear, happiness, sadness and surprise. Looking at figure 1, the highest accuracy is obtained for happy class for audio modality using SVM classifier. Similarly,for video modality SVM classifier outperform compare to MLP.

Table 1 shows the accuracy rate for different emotion category obtained by MLP &SVM classifiers. The results based on fusion of two modalities are also given in table 1. We are getting highest classification results for anger using audio cue (MLP-97.5 %) and (SVM-100 %) on eNTERFACE database. We can conclude that the SVM classifier outperforms to predict emotion from audio-visual signal.

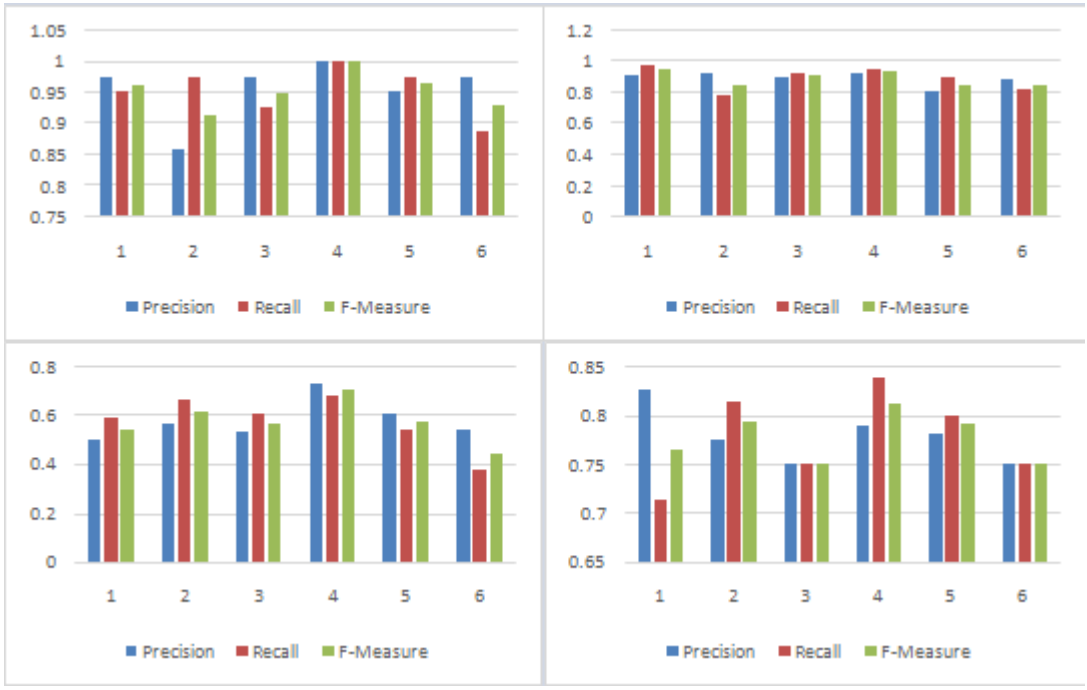

Fig. 1: Single cue results for audio and video modalities over eNTERFACE database (Top: audio results 1. SVM, 2. MLP; Bottom: video results 1. SVM, 2. MLP) (1-Anger, 2-Disgust, 3-Fear, 4-Happiness, 5-Sadness, 6-Surprise)

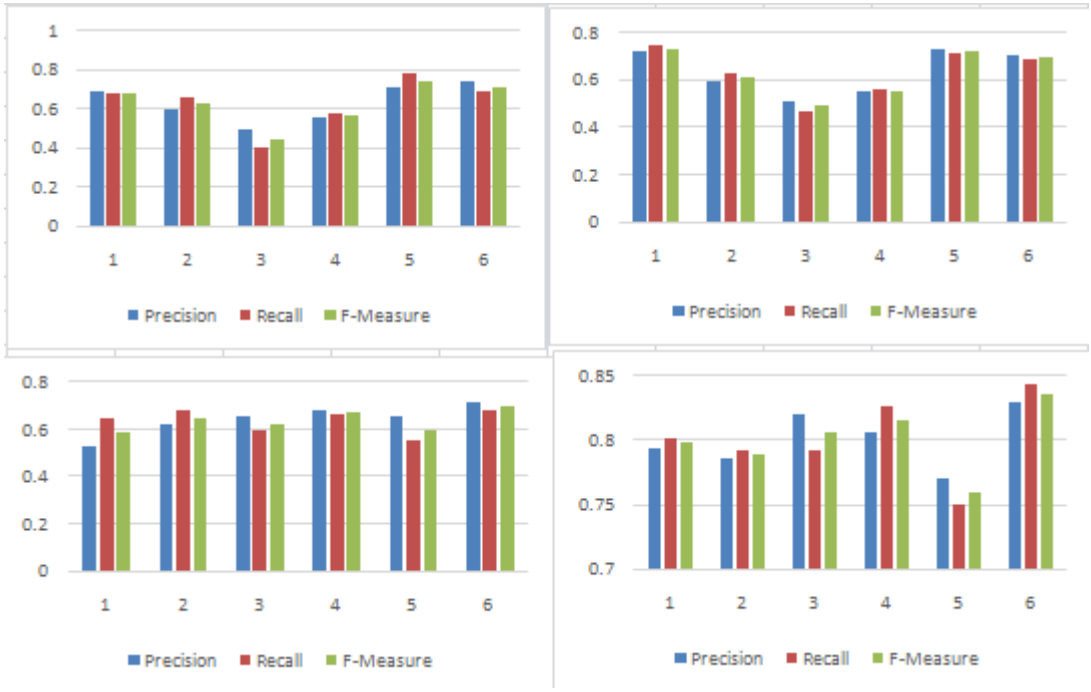

Fig. 2: Single cue results for audio and video modalities overRML database (Top: audio results 1. SVM, 2. MLP; Bottom: video results 1. SVM, 2. MLP) (1-Anger, 2-Disgust, 3-Fear, 4-Happiness, 5-Sadness, 6-Surprise)

Table 1 shows the multiple cues results for both databases. The average accuracy is given for anger, disgust, fear, happiness, sadness and surprise emotion.

Table1: Multiple cue results for ENTERFACE and RML database (A-Anger, D-Disgust, F-Fear, H-Happiness, S-Sadness, Sp-Surprise)

			A	D	F	H	S	Sp
ENTERFACE DATABASE	MLP	Audio	97.5	77.5	91.25	95	88.75	81.25
		Video	71.25	81.25	75	83.75	80	75
		Fusion	91.25	83.75	90	91.25	91.25	82.5
	SVM	Audio	95	97.5	92.5	100	97.5	88.75
		Video	58.75	66.25	60	67.25	53.75	37.5
		Fusion	82.5	82.5	90	81.25	76.25	68.75
RML DATABASE	MLP	Audio	74.17	62.50	46.67	55.83	70.83	68.33
		Video	80	79.17	79.17	82.50	75	84.17
		Fusion	79.17	80	73.33	75.83	85.83	83.33
	SVM	Audio	67.50	65.83	40	57.50	77.50	68.33
		Video	65	68.33	59.17	66.67	55	68.33
		Fusion	80.83	81.67	68.33	74.17	82.50	80.83

Conclusion and Future Work

In this study, we have proposed a framework for emotion recognition based on the multi-resolution analysis of audio-visual data. It also concludes that Wavelet transforms works very well for emotion recognition from multimodal data. We have evaluated the proposed emotion model with standard emotion datasets i.e. eNTERFACE and RML. The experimental results obtained from multimodal fusion are very promising. Further, we wish to test other features for real time emotion recognition to better human-machine interaction.

References

[1] Koelstra S, Muhl C, Soleymani M, Yazdani A, Lee J S, Ebrahimi T, Pun T, Nijholt A, Patras I. *DEAP: A Database for Emotion Analysis Using Physiological Signals.* IEEE Trans. Affective Computing 2012; 3(1): 1831.

[2] Wang Y, Ling G, Anastasios N V.*Kernel Cross-Modal Factor Analysis for Information Fusion with Application to Bimodal Emotion Recognition.* Multimedia, IEEE Transactions 2012; 14(3): 597-607.

[3] Paleari M, Chellali R, Huet B. *Features for multimodal emotion recognition: An extensive study.* Cybernetics and Intelligent Systems (CIS), IEEE Conference On. IEEE 2010.

[4] Mansoorizadeh M, Charkari N M. *Multimodal information fusion application to human emotion recognition from face and speech.* Multimedia Tools and Applications 2010; 49(2): 277-297.
[5] Datcu D, Rothkrantz L. *Semantic audio-visual data fusion for automatic emotion recognition.*Euromedia2008.
[6] Viola P, Jones J. *Robust Real-Time Face Detection.* Int. J. of Computer Vision 2004; 57(2): 137–154.
[7] Ekman P, Friesen W. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978.
[8] Essa I A, Darrell T, Pentland A. *Tracking Facial Motion.* Proceedings of the IEEE Workshop on No rigid and Articulate Motion 1994: 36-42.
[9] Dollar P, Rabaud V, Cottrell G, Belongie S. *Behaviour recognition via sparse spatio-temporal features.* VS-PETS 2005: 65– 72.
[10] Sahidullah Md, Saha G. *Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition.* Speech Communication 2012; 54 (4): 543–565.
[11] Martin O, Kotsia I, Macq B, Pitas I. *The eNTERFACE 05 Audio-Visual Emotion Database*. Proc. of the First IEEE Workshop on Multimedia Database Management, Atlanta, April 2006.
[12] Wang Y, Guan L. *Recognizing human emotional state from audiovisual signals*. IEEE Trans. Multimedia 2008; 10(5): 936–946.
[13] Cortes C, Vapnik V. *Support-vector networks.* Machine learning 1995; 20(3):273-297.
[14] Aldroubi A, Unser M. *Families of multiresolution and wavelet spaces with optimal properties.* Numerical Functional Analysis and Optimization 1993, 14(5-6), 417-446. 1993
[15] Heil C E, Walnut D F. *Continuous and discrete wavelet transforms.* SIAM review 1989; 31(4): 628-666.