Audio-Visual Spontaneous Emotion Recognition

Zhihong Zeng[1], Yuxiao Hu[1], Glenn I. Roisman[1], Zhen Wen[2], Yun Fu[1] and Thomas S. Huang[1]

[1] University of Illinois at Urbana-Champaign, USA
[2] IBM T.J.Watson Research Center, USA
{zhzeng, hu3, yunfu2, huang} @ifp.uiuc.edu, roisman@uiuc.edu, zhenwen@us.ibm.com

Abstract. Automatic multimodal recognition of spontaneous emotional expressions is a largely unexplored and challenging problem. In this paper, we explore audio-visual emotion recognition in a realistic human conversation setting—the Adult Attachment Interview (AAI). Based on the assumption that facial expression and vocal expression are at the same coarse affective states, positive and negative emotion sequences are labeled according to Facial Action Coding System. Facial texture in visual channel and prosody in audio channel are integrated in the framework of Adaboost multi-stream hidden Markov model (AdaMHMM) in which the Adaboost learning scheme is used to build component HMM fusion. Our approach is evaluated in AAI spontaneous emotion recognition experiments.

Keywords: Multimodal Human-Computer Interaction, Affective computing, affect recognition, emotion recognition.

1 Introduction

Human-computer interaction has been a predominantly one-way interaction where a user needs to directly request computer responses. Change in the user's affective state, which play a significant role in perception and decision making during human to human interactions, is inaccessible to computing systems. Emerging technological advances are enabling and inspiring the research field of “affective computing,” which aims at allowing computers to express and recognize affect [1]. The ability to detect and track a user’s affective state has the potential to allow a computing system to initiate communication with a user based on the perceived needs of the user within the context of the user’s actions. In this way, human computer interaction can become more natural, persuasive, and friendly [2-3][45][66].

In the speech recognition community, there is an increasing interest in improving performance of spontaneous speech recognizers by taking into account the influence of emotion on speech [5-7]. The authors in [7] made a systematic comparison of speech recognition under different conditions. Their results show that the influence of emotion is larger than others (i.e. noise, loudness). The studies [5-6] indicated that emotion-sensitive audio-only ASR system improved speech recognition rates noticeably.

Automatic emotion recognition has been attracting attention of researchers from a variety of different disciplines. Another application of automatic emotion recognition is to help people in emotion-related research to improve the processing of emotion data. In recent decades, with the advance of emotion theories [10][11][12] and emotion measurement (e.g. Facial Action Unit System (FACS) [13]), more and more reports of emotion analysis have been conducted in psychology, psychiatry, education, anthropology, neurophysiology [14][15][16]. The emotion-related research includes attachment [17], mother-infant interaction [18], tutoring [19], and psychiatric disorders [20]. All of the above research requires measurement of emotion expressions. At present, this problem was solved mainly by self-reports and observers' judgments of emotion (based on FACS or other labeling schemes). But self-reports and human-based emotion measurements are error-prone, and time consuming. Such limitations influence the rate at which new research can be done. Automatic emotion recognition would reduce dramatically the time it takes to measure emotional states, and improve the reliability of measurement.

In this paper, we explore audio-visual recognition of spontaneous emotions occurring in a realistic human conversation setting—the Adult Attachment Interview (AAI). The AAI is the most widely used and well-validated instrument in developmental research for identifying adult attachment representations. The AAI data in our experiment were collected by the authors in [17] to study links between adults' narratives about their childhood experiences and their emotional expressive, physiological, and self-reported emotion.

Although the ability to recognize a variety of fine-grained emotions is attractive, it may be not practical because the emotional data in the context of realistic conversations is often not sufficient to learning a classifier for a variety of fine-grained emotions. In this paper, we focus on recognizing positive and negative emotions which can be used as a strategy to improve the quality of interface in HCI, and as a measurement in studies conducted in the field of psychology [17]. This work extends our previous work [21][56] that explored separating spontaneous emotional facial expressions from non-emotional facial expressions in order to narrow down data of interest for emotion recognition research.

In the paper, we propose Adaboost multi-stream hidden Markov model (Adaboost MHMM) to integrate audio and visual affective information. In order to capture the richness of facial expression, we use 3D face tracker to extract facial texture images that are then transformed into low dimensional subspace by Locality Preserving Projection (LPP). We use pitch and energy in audio channel to build audio HMM in which some prosody features, like frequency and duration of silence, could have implications. In the audiovisual fusion stage, we treat the component HMM combination as a multi-class classification problem in which the input is the probabilities of HMM components and the output is the target classes, based on the training combination strategy [44]. We use Adaboost learning scheme to build fusion of the component HMMs from audio and visual channels. The framework of our approach is illustrated in Figure 1.

The rest of the paper is organized as follows. In the following section, we briefly describe related work about automatic emotion recognition, focusing on audio-visual emotion recognition. In Section 3 we introduce the AAI data that is used in our spontaneous affective expression analysis. Section 4 introduces a 3D face tracker used to extract facial textures, and Locality Preserving Projection for feature dimension reduction. Section 5 describes prosodic feature extraction in audio channel. In Section 6, we introduce Adaboost multi-stream hidden Markov model for bimodal fusion. Section 7 presents our preliminary experimental results on two AAI subjects to evaluate our method. Finally, we have concluding remarks in Section 8.

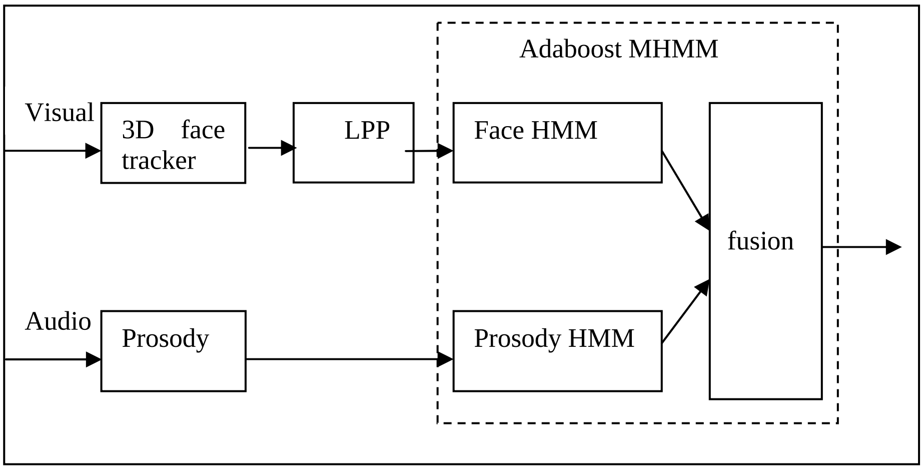


Fig. 1. The audio-visual emotion recognition framework

2 Related Work

The recent appraisal-based emotion theory [12] indicates the importance of the integration of information from different response components (such as facial and vocal expression) to yield a coherent judgment of emotions. In addition, current techniques of both computer vision and audio processing have limitations in realistic applications. For instance, current face trackers are sensitive to head pose, occlusion and lighting changes while audio processing is sensitive to noise and distance between speakers and microphone. Audio and visual channels provide complementary information. Moreover, if one channel fails for some reason, the other channel can still work. Thus, the final fusion performance can be robust.

In recent years, the literature on automatic emotion recognition has been growing dramatically due to the development of techniques in computer vision, speech analysis, and machine learning. However, automatic recognition on emotions occurring in natural communication settings is a largely unexplored and challenging problem. Authentic emotional expressions are difficult to collect because they are relatively rare and short lived, and filled with subtle context-based changes that make it difficult to elicit emotions without influencing the results. Manual labeling of spontaneous emotional expressions for ground truth is very time consuming, error prone, and expensive [23]. This state of affairs leads to a big challenge for spontaneous emotional expression analysis. Due to these difficulties in emotional expression recognition, most of current automatic facial expression studies have been based on the artificial material of deliberately expressed emotions that were collected by asking the subjects to perform a series of emotional expressions to a camera. The popular artificial facial expression databases are Ekman-Hager data [24], and Kanade-Cohn data [25], Pantic et al.'s data [26], and the JAFFE data [27]. The audio-visual emotion data is Chen-Huang data [28]. An overview of databases for emotion classification studies can be founded in [26][29].The psychological study [30] indicates that the posed nature of the emotions may differ in appearance and timing from corresponding performances in natural settings. More specifically, the study [31] indicated that posed smiles were of larger amplitude and has less consistent relation between amplitude and duration than spontaneous smile. And the study [32] indicated the spontaneous brow actions (AU1, AU2 and AU4 in the Facial Action Coding System) have different characteristics (intensity, during and occurrence order) from corresponding posed brow actions. In addition, most of current emotion recognizers are evaluated in clear and constrained input (e.g., high quality visual and audio recording, non-occluded and front-view face), which is different from the natural communication setting. Therefore, the methods based these artificial emotions would be expected to perform inadequately on emotional expressions occurring in natural communication settings.

Most studies of automatic emotion recognition focus on six basic facial expressions or a subset of them, namely happiness, sadness, anger, fear, surprise, and disgust. The recognition of these basic facial expressions was based on Ekman's extensive study [33] that indicated the universality of human perception of these basic facial expressions in different cultures. However, most of the emotion expressions that occur in our human-human conversation are non-basic emotions [15].

In addition, most of current automatic emotion recognition approaches are uni-modal: information processed by the computer system is limited to either face images [67-73] or speech signals [74-78]. Relatively little work has been done in researching multimodal affect analysis. For extensive survey of automatic emotion analysis done in the recent years, readers are referred to review papers, including [34][35][79] written by Pantic et al. in 2003, 2005 and 2006, [37] by Cowie et al. in 2001, and [36] by Sebe et al. in 2005.

Here we focus on reviewing the efforts toward audio-visual emotion recognition, especially those done in these years, which have not been included in the previous review papers. The first studies for audio-visual emotion recognition include Chen and Huang [38][39], Yoshitomi et al. [40], De Silva and Ng [41]. In that past few years, there are an increasing number of reports investigating the integration of emotion-related information from audio and visual channels in order to improve the recognition performance. They are summarized in Table 1 in which the first column is the reference, the second column is the number of subjects in datasets, the third column is the number of emotional states, the fourth column is classifier, and the fifth column is fusion method (feature-level, model-level, and decision-level), the sixth column is the test method (p: person-dependent; i: person-independent), and last column is recognition accuracy (%). * denotes the missing entry. The data in [44-46][48] included 6 basic emotions (happiness, sadness, fear, disgust, anger, and surprise), 4 cognitive/motivational states (interest, boredom, puzzlement and frustration), and neutral. The data in [51] were collected in a standing car with webcam and an array microphone. In the study [52], the data include six different languages (English, Chinese, Urdu, Punjabi, Persian, and Italian) and subjects were from different races. All of those studies were evaluated on artificial material of deliberately expressed emotions.

Table 1: Properties of studies of Audio-visual emotion recognition on posed emotions

reference	subject	state	classifier	fusion	test	accuracy
Zeng et al. [44]	20	11	MFHMM	model	i	83.64%
Zeng et al. [45]	20	11	MFHMM	model	i	80.61%
Zeng et al. [46]	20	11	MHMM	decision	i	75%
Zeng et al. [47]	20	2	Fisher-boosting	feature	p	>84%
Zeng et al. [48]	20	11	SNoW MHMM	decision	P i	96.30% 72.42%
Song et al. [49]	*	7	THMM	model	*	84.7%
Buss et al. [50]	1	4	SVC	Feature, decision	P	89.1% 89.0%
Hoch et al. [51]	7	3	SVM	decision	p	90.7%
Wang et al. [52]	8	6	LDA	decision	i	82.14%
Go et al. [53]	10	6	LDA K-mean	decision	i	>95%

Because more and more researchers have noticed the potential difference between posed emotion expression and spontaneous emotion expression, there is in the last few years a trend in the emotion recognition community in moving away from posed emotion recognition to spontaneous emotion recognition. These notable studies include visual spontaneous expression recognition, audio spontaneous emotion recognition, and relatively few audio-visual spontaneous emotion recognition. They are summarized in Table 2 in which the first column is the reference, the second column “sub” is the number of subjects in their dataset, the third column “state” is the number of emotional states or facial action units, the fourth column “labeling” is the labeling methods, the fifth column “fea” is the used feature (A: audio; V: visual), the sixth column is the classifier, the seventh column is test method (p: person-dependent; i: person-independent), the last column is the recognition accuracy.

The datasets of these studies of spontaneous emotion expressions include human-human interaction and human-to-computer interaction. They are collected from call center [62][61], meeting [61], interview [54][56][58], dialogue system [59], Wizard of OZ scenarios [64][63][60], and kiosk [55].

There are methodological differences between the studies of posed emotion recognition and those of spontaneous emotion recognition. First, as compared to posed emotion recognition where labels are pre-defined, the spontaneous emotion labeling for ground truth is error-prone and time consuming. In Table 2, only the study [55] used the self-reports of subjects as labels, and the other studies used the human observation in which the humans labeled the data based on Facial Action Unit System (FACS), Feeltrace system [65], or ad hoc labeling scheme.

Second, in these studies of spontaneous emotion recognition, only one study [55] tried to recognize the 4 basic emotions (neutral, happiness, surprise and disgust). In the rest studies, the recognized emotion states include coarse emotion states (positive, negative and neutral) [56][61], quadrant states in evaluation-activation space [64][63], or application-dependent states (trouble in [60], annoyance and frustration [62]), and facial action units [54][32]. The main reason could be that there is no sufficient data of basic emotion expressions to train a classifier.

Because these above studies evaluate their algorithms on the different experimental condition (data, labeling, the number of classes, feature set), it is difficult to give their performance rank only based on the accuracies.

Table 2: Properties of studies of Audio-visual spontaneous emotion recognition

reference	sub	state	labeling	fea	classifier	test	accuracy
Barlett et al. [54]	12	16AUs	FACS	V	SVM	i	90.5%
Sebe et al. [55]	28	4	Self-report	V	KNN	*	95.57%
Zeng et al. [56]	2	2	FACS	V	KSVDD	p	79%
Cohn et al. [58]	21	3AUs	FACS	V	Guassian	i	76%
Valstar et al. [32]	*	3AUs	FACS	V	SVM	i	50.4%
Litman et al. [59]	10	3	ad hoc	A	decision tree	p i	66-73%
Batliner et al. [60]	24	2	ad hoc	A	LDA	i	74.2%
Neiberg et al.[61]	*	3	ad hoc	A	GMM	i	85-95%
Ang et al. [62]	*	6	ad hoc	A	decision tree	i	85.4-93.2%(2class)
Fragopanagos et al. [63]	*	4	Feel-trace	AV	Neural Network	*	44-71%
Garidakis et al. [64]	*	4	Feel-trace	AV	Neural network	*	79%

In the studies [64][63] toward audio-visual spontaneous emotion recognition, the authors uses Feeltrace tool [65] to label the data collected in a “Wizard of OZ” scenario. In the study [63], due to considerable variation across four raters, it is difficult to reach a similar assessment with the FeelTrace labels. They observed the difference of the labeling results among four Feeltrace users. Specifically, these Feeltracers judged the emotional states of data by using different modalities. For example, one used facial expressions as the most important cues to make the decision while another used prosody.

Compared with Feeltrace tool mentioned above, FACS could be more objective in labeling and be able to capture the richness and complexity of emotional expressions. Thus, we build our emotion recognizer with FACS labeling in this work. We make the assumption that in our database there is no blended emotions so that the facial expression and prosody belong to same emotional states at the coarse level (i.e. positive and negative emotions).

In addition, different from these two studies above mentioned [64][63], we apply 3D face tracker which is able to capture the wider range of face movement than 2D face tracker. We use facial texture instead of sparse geometrical features in order to capture the richness of subtle facial expressions. For capturing the dynamic structure of emotional expressions and integrating audio and visual streams, we build our recognizer in Adaboost multi-stream hidden Markov model framework in which Adaboost learning scheme is used to build fusion of component HMMs.

3 Data of Adult Attachment Interview

The Adult Attachment Interview (AAI) is a semi-structured interview used to characterize individuals' current state of mind with respect to past parent-child experiences. This protocol requires participants to describe their early relationships with their parents, revisit salient separation episodes, explore instances of perceived childhood rejection, recall encounters with loss, describe aspects of their current relationship with their parents, and discuss salient changes that may have occurred from childhood to maturity [17].

During data collection, remotely controlled, high-resolution (720*480) color video cameras recorded the participants' and interviewer's facial behavior during AAI. Cameras were hidden from participants' view behind a darkened glass on a bookshelf in order not to distract the participant's attention. The snapshot of an AAI video is shown in Figure 2. The participant's face is displayed in the bigger window while the interviewer's face is in the smaller left-top window.

As our first step to explore audio-visual spontaneous emotion recognition, AAI data of two subjects (one female and one male) was used in this study. The video of the female subject lasted 39 minutes, and one of the male lasted 42 minutes. The significant amount of data allowed us personal-dependent spontaneous emotion analysis.

In order to objectively capture the richness and complexity of facial expressions, Facial Action Coding System (FACS) was used to code every facial event that occurred during AAI by two certified coders. Inter-rater reliability was estimated by the ratio of the number of agreements in emotional expression to the total number of agreement and disagreements, yielding for this study a mean agreement ratio of 0.85.

To reduce FACS data further for analysis, we manually grouped combinations of AUs into two coarse emotion categories (i.e., positive and negative emotions) on the basis of an empirically and theoretically derived Facial Action Coding System Emotion Codes which was created by the psychological study [80].

In order to narrow down the inventory to potential useful emotion data for our experiment, we first ignore the emotion occurrences to which these two coders disagree with each other. In order to analyze the emotions occurring in a natural communication setting, we have to face the technique challenges to handle arbitrary head movement. Due to the technique limitation, we filtered out the emotion segments in which hand occluded the face, face turned away more than 40 degree with respect to the optical center, or part of face moved out of camera view. Each emotion sequence starts from and to the emotion intensity scoring scale B (slight) or

C (marked pronounced) defined in [13]. The number of audio-visual emotion expression segments in which subjects displayed emotions using both facial expressions and voice is 67 for female and 70 for male.

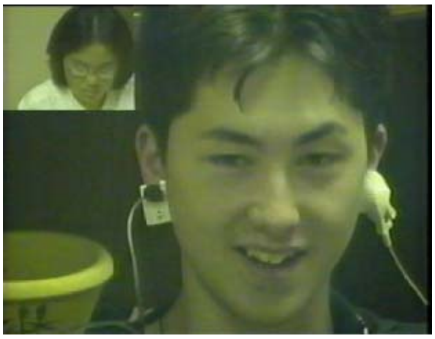


Fig. 2. The snapshot of an AAI video. The participant's face is displayed in the bigger window while the interviewer's face is in the smaller left-top window.

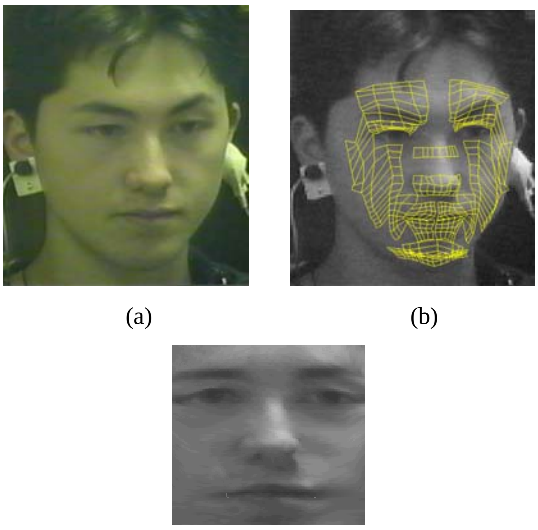
4 Facial Expressions

This section includes 3D face tracker and Locality Preserving Projection which aims to project the high-dimensional images to low dimensional subspace.

4.1 3D Face Tracker

To handle the arbitrary behavior of subjects in the natural setting, it is required to track the 3D face. The face tracking in our experiments is based on a system called Piecewise Bezier Volume Deformation (PBVD) tracker which was developed in [82][81].

This face tracker uses a 3D facial mesh model which is embedded in multiple Bezier volumes. The shape of the mesh can be changed with the movement of the control points in the Bezier volumes, which guarantees the surface patches to be continuous and smooth. In the first video frame, the 3-D facial mesh model is constructed by selection of landmark facial feature points. Once the model was fitted, the tracker can track head motion and local deformations of the facial features by an optical flow method. In this study, we use rigid setting of this tracker to extract facial expression texture. The 3D ridge geometric parameters (3D rotation and 3D translation) determine the registration of each image frame to the face texture map, which is obtained by wrapping the 3D face appearance. Thus, we can derive a sequence of face texture images, which capture the richness and complexity of facial expression. Figure 3 shows a snapshot of the tracking system. Figure 3(a) is the input video frame, and Figure 3(b) is the tracking result where a mesh is used to visualize the geometric motions of the face. The extracted face texture is shown in Figure 3(c).



(a) (b)

(c)

Fig. 3. The 3D face tracker's result. (a) the video frame input; (b) tracking result where a mesh is used to visualize the geometric motions of the face; (c) extracted face texture.

4.2 Locality Preserving Projection

In recent years, computer vision research has witnessed a growing interest in subspace analysis techniques. Before we utilize any classification technique, it is beneficial to first perform dimensionality reduction to project an image into a low dimensional feature space, due to the consideration of learnability and computational efficiency.

Locality Preserving Projection (LLP) is a linear mapping that is obtained by finding the optimal linear approximations to the eigen-functions of the Laplace Beltrami operator on the manifold [83]. As contrasted with nonlinear manifold learning techniques, LLP can be simply applied to any new data point to locate it in the reduced representation manifold subspace, which is suitable for classification application.

Some traditional subspace methods such as PCA aim to preserve the global structure. However, in many real world applications, especially facial expression recognition, the local structure could be more important. In contrast to PCA, LPP finds an embedding that preserves local information, and obtains a subspace that best detects the essential manifold structure.

The details of LPP, including its learning and mapping algorithms, can be found in [83]. The low-dimensional features from LPP are then used to build visual HMM.

5 Vocal Expressions

In our work, we use prosodic features which are related with the way the sentences are spoken. For audio feature extraction, Entropic Signal Processing System named get_f0, a commercial software package, is used. It implements a fundamental frequency estimation algorithm using the normalized cross correlation function and dynamic programming. The program can output the pitch F0 for fundamental frequency estimate, RMS energy for local root mean squared measurements, prob_voice for probability of voicing, and the peak normalized cross-correlation value that was used to determine the output F0. The experimental results in our previous work [48] showed pitch and energy are the most important factors in affect classification. Therefore, in our experiment, we only used these two audio features for affect recognition. Some prosody features, like frequency and duration of silence, could have implication in the HMM structure of energy and pitch.

6 Adaboost Multi-stream Hidden Markov Model

Audio-visual fusion is an instance of the general classifier fusion problem, which is an active area of research with many applications, such as Audio-Visual Automatic Speech Recognition (AVASR). Although there are some audio-visual fusion studies in audio-visual Automatic Speech Recognition literature, few studies are found for audio-visual affect recognition shown in Table 1.

Most of current multi-stream combination studies focus on weighting combination scheme with weights proportional to the reliabilities of the component HMMs. The weights can be computed from normalized stream recognition rate [22], stream S/N ratio [22], stream entropy [8], or other reliability measures such as ones in [9].

The weighting combination scheme is intuitive and reasonable in some ways. However, it is based on the assumption that the combination is linear. This assumption could be invalid in practice. In addition, using the weighting scheme is difficult to obtain the optimal combination because they deal with different feature spaces and different models. It is even possible that the weighting combination is worse than individual component performance, as shown in our experiments.

According to training combination strategy in our previous work [44], the component HMM combination can be treated as a multi-class classification problem in which the input is the probabilities of HMM components and the output is the target classes. This combination mechanism can be linear or nonlinear, depending on learning scheme that we use. In this case, if s represents the number of possible classes and n the number of streams, this classification contains s×n input units and s output units, and the parameters of the classifier can be estimated by training. Under this strategy, we propose Adaboost MHMM in which the Adaboost learning scheme is used to build the fusion of multiple component HMMs.

6.1 Learning Algorithm

Given m training sequences each of which has n streams

$$(x_{11},\cdots,x_{1n},y_1),\cdots,(x_{m1},\cdots,x_{mn},y_m)$$

where x_{ij} is the jth stream of ith sample sequence, and $y_i = 0,1$ for negative and positive emotions in our application. Assume that these n streams can be modeled respectively by n component HMMs. The learning algorithm of the Adaboost MHMM includes three main steps.

1. n component HMMs are trained independently by the EM algorithm. The model parameters (the initial, transition, and observation probabilities) of individual HMMs are estimated.
2. For each training sequence, likelihoods of these n component HMMs are computed. We obtain
$$(p_{110},p_{111},\cdots,p_{1n0},p_{1n1},y_1),\cdots,(p_{m10},p_{m11},\cdots,p_{mn0},p_{mn1},y_m)$$
where p_{ij0},p_{ij1} are likelihoods of negative and positive emotions of jth stream of ith sample sequence.
3. Fusion training: based on Adaboost learning scheme [4], these estimated likelihoods of n component HMMs are used to construct a strong classifier which is a weighted linear combination of a set of weak classifiers.

6.2 Classification Algorithm

Given a n-stream observation sequence and the model parameters of Adaboost MHMM, the inference algorithm of the Adaboost MHMM includes two main steps.

1. Compute individually likelihoods of positive and negative emotions of n component HMMs.
2. A set of weaker hypotheses are estimated each using likelihood of positive or negative emotion of a single component HMM. The final hypothesis is obtained by weighted linear combination of these hypotheses where the weights are inversely proportional to the corresponding training errors [4].

7 Experimental Results

In this section, we present the experimental results of our emotion recognition by using audio and visual affective information.

The personal-dependent recognition is evaluated on the two subjects (one female and one male) in which the training sequences and test sequences were taken from the same subject. For this test, we apply leave-one-sequence-out cross-validation. For each subject in this test, one sequence among all of emotion expression sequences is used as the test sequence, and the remaining sequences are used as training sequences. This test is repeated, each time leaving a different sequence out.

7.1 Facial Expression Analysis on Locality Preserving Subspace

In this experiment, we evaluate the LPP HMM method which models the facial expressions by using the low-dimensional features in the locality preserving subspace of facial texture images. In addition, the PCA HMM method, which uses the features in the PCA subspace, is also tested to make the performance comparison with LPP HMM. The comparison results are shown in Table 3 for these two subjects. The corresponding facial expression subspaces are called optimal facial expression subspaces for each method. Figure 4 and 5 shows a plot of recognition accuracy of LPP HMM and PCA HMM vs. dimensionality reduction for female and male respectively. It is shown that LPP HMM method largely outperforms PCA HMM. The recognition accuracy of LPP HMM is 87.50% at 5D subspace for female and 84.85% at 4D subspace for male respectively.

Table 3. Performance Comparison of LPP HMM and PCA HMM.

	Approach	Dimension	Accuracy (%)
Female	LPP HMM	5	87.50
	PCA HMM	5	79.69
Male	LPP HMM	4	84.85
	PCA HMM	5	72.62

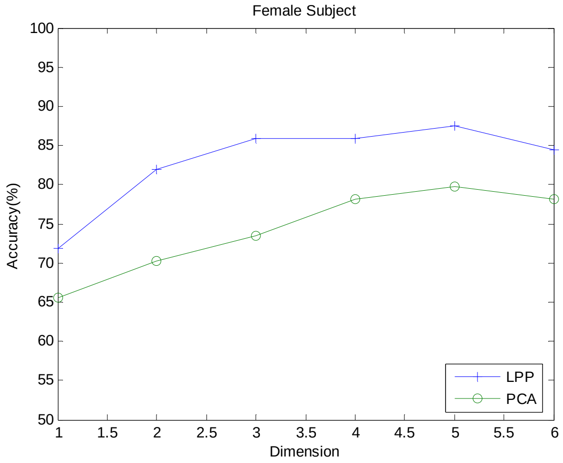


Fig. 4. Facial expression recognition accuracy of LPP HMM and PCA HMM vs. dimensionality reduction on the female emotion data.

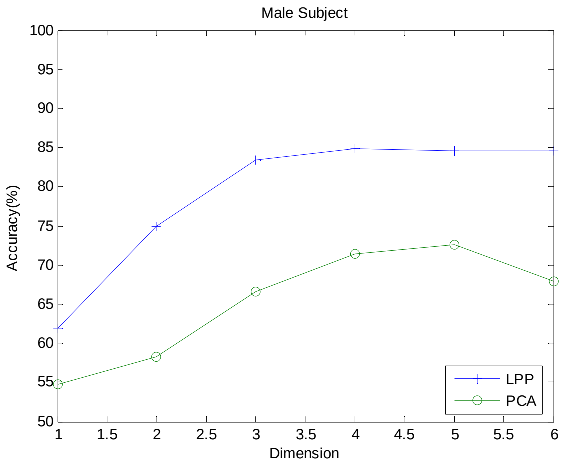


Fig. 5. Facial expression recognition accuracy of LPP HMM and PCA HMM vs. dimensionality reduction on the male emotion data.

7.2 Prosody Expression Analysis

Table 4 shows the experimental results of emotion recognition by using audio HMM. The recognition performance of prosody HMM is better than random, but worse than facial expression recognition. There are two possible reasons why prosodic affective recognition is worse than facial affective recognition. One is that facial expression could provide more reliable affective information than prosody, as the psychological study indicated [15]. The other reason is that we only use information of facial expressions to label our multimedia emotion data. Thus, facial expression recognition is more agreement with human judgment (labels) than prosody expressions.

Table 4. Emotion Recognition of Prosody HMM

Subjects	Accuracy (%)
Female	75.09
Male	65.15

7.3 Audio-visual Fusion

The emotion recognition performance of audio-visual fusion is shown in Table 5. In this table, two combination schemes (weighting and training) are used to fuse the component HMMs from audio and visual channels. Acc MHMM means MHMM with the weighting combination scheme in which the weights are proportional to stream normalized recognition accuracies. Adaboost MHMM means MHMM with the

Adaboost learning schemes as described in Section 6. Because we treat the multi-stream fusion as a multi-class classification problem, there are a variety of methods that can be used to build the fusion. In addition to Adaboost MHMM, we used LDC and KNN (K=3 for female and K=5 for male) to build this audio-visual fusion, which are Ldc MHMM and Knn MHMM in Table 5.

The performance comparison of these fusion methods is as follows:

Adaboost MHMM > Knn MHMM > Acc MHMM > Ldc MHMM

The results demonstrate that training combination outperforms weighting combination, except Ldc MHMM that is a linear fusion. Adaboost MHMM is the best among these four fusion methods.

The summarization of the performance of different modalities and different fusion methods is illustrated in Figure 6. Results show that Adaboost MHMM and Knn MHMM are better than uni-modal HMM (i.e. visual-only HMM and audio-only HMM). That suggests that multiple modalities (audio and visual modalities) can provide more affective information and have the potential to obtain better recognition performance than a single modality.

In Figure 6, the accuracy of Acc MHMM equals to visual-only HMM for male data but worse than visual-only HMM for female data. Ldc MHMM is worse than visual-only HMM in female and male cases. Both of Acc MHMM and Ldc MHMM are linear bimodal fusion. That suggests that the fusion method play an important role in audio-visual emotion recognition. Although the linear bimodal combination is reasonable and intuitive, it is not guaranteed to obtain the optimal combination at realistic application. Even it is possible that this combination is worse than individual component performance, as shown in our experiments.

The confusion matrixes of emotion recognition for two subjects are shown in Table 6 and 7. These results demonstrate that negative emotions are more difficult to recognize than positive emotions. We noticed that adult subjects tend to inhibit negative emotion expressions in this interactive interview context. Thus, the negative emotions are shorter and more subtle than positive emotions.

Table 5. Audio-visual Emotion Recognition

Bimodal Fusion		Combination scheme	Accuracy (%)
Female	Acc MHMM	Weighting	84.38
	Ldc MHMM	Training	75.00
	Knn MHMM	Training	87.50
	AdaBoost MHMM	Training	90.36
Male	Acc MHMM	Weighting	84.85
	Ldc MHMM	Training	81.82
	Knn MHMM	Training	87.88
	AdaBoost MHMM	Training	89.39

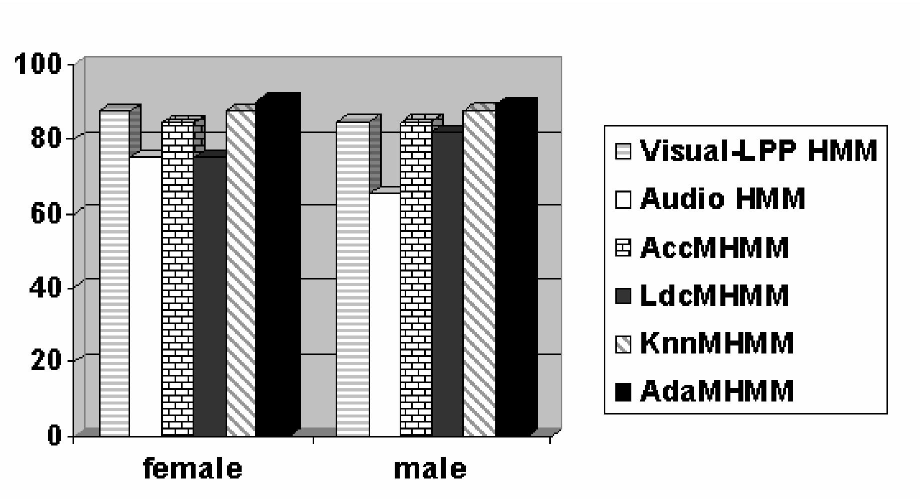


Fig. 4. Performance comparison among different modalities and different fusions

Table 6. Confusion Matrix for Female Emotion Recognition

Female	Detected		
Desired	%	Positive	Negative
	Positive	94.44	5.56
	Negative	10.87	89.13

Table 7. Confusion Matrix for Male Emotion Recognition

Male	Detected		
Desired	%	Positive	Negative
	Positive	91.67	8.33
	Negative	13.33	86.67

8 Conclusion

Emotion analysis has been attracting increased attention of researchers from various disciplines because changes in a speaker’s affective states play a significant role in human communication. Most of current automatic facial expression recognition approaches are based on artificial materials of deliberately expressed emotions and uni-modal methods.

In this paper, we explore audio-visual recognition of spontaneous emotions occurring in Adult Attachment Interview (AAI) in which adults talked about past parent-child experiences. We propose an approach for this realistic application, which includes the audio-visual labeling assumption and Adaboost multi-stream hidden Markov model to integrate facial expression and prosody expression. Our preliminary experimental results from two video of about-40-minute-long AAI suggest the validation of our approach for spontaneous emotion recognition. In the future, our approach in this paper will be evaluated on more AAI data. In addition, we will explore person-independent emotion recognition in which training data and testing data are from different subjects.

Our work is based on the assumption that facial expressions are consistent of vocal expressions at the coarse emotion level (positive and negative emotions). Although this assumption is valid at most circumstances, blended emotions can occur when speakers have conflict intension [23]. The exploration of recognition of the blended emotions is our future work.

Acknowledgments. We like to thank Prof. Maja Pantic and Anton Nijholt for providing valuable comments. The work was supported in part by a Beckman Postdoctoral Fellowship and in part by National Science Foundation Grant CCF 04-26627.

References

1. Picard, R.W., Affective Computing, MIT Press, Cambridge, 1997.
2. Litman, D.J. and Forbes-Riley, K., Predicting Student Emotions in Computer-Human Tutoring Dialogues. In Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), July 2004
3. Kapoor, A. and Picard, R.W., Multimodal Affect Recognition in Learning Environments, ACM Multimedia, 2005, 677-682
4. Viola P. 2004. Robust Real-Time Face Detection. Int. Journal of Computer Vision. 57(2), 137-154
5. Polzin, S.T. and Waibel, A. (1999), Pronunciation Variations in Emotional Speech, Proceedings of the ESCA Workshop, 103-108
6. Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E., Cox, C. (2005), ASR for Emotional Speech: Clarifying the Issues and Enhancing Performance, Neural Networks, 18: 437-444
7. Steeneken, H.J.M. and Hansen, J.H.L. (1999), Speech under stress conditions: Overview of the effect of speech production and on system performance, Int. Conf. on Acoustics, Speech, and Signal Processing, 4:2079-2082
8. Okawa, S., Bocchieri, E. and Potamianos, A., Multi-band Speech Recognition in noisy environments, ICASSP, 1998, 641-644
9. Garg, A., Potamianos, G., Neti, C. & Huang, T.S., Frame-dependent multi-stream reliability indicators for audio-visual speech recognition, ICASSP, 2003.
10. Ekman P, Friesen WV, Ellsworth P. (1972). Emotion in the Human Face. Elmsford, NY: Pergamon.
11. Izard CE. 1971. The face of Emotion. New York: Appleton-Century_Crofts
12. Scherer (2004), Feelings integrate the central representation of appraisal-driven response organization in emotion. In Manstead, A.S.R., Frijda, N.H. & Fischer, A.H. (Eds.),

Feelings and emotions, The Amsterdam symposium (pp. 136-157). Cambridge: Cambridge University Press, 136-157.

13. Ekman P, Friensen WV, Hager JC. 2002. Facial Action Unit System. Published by A Human Face.
14. Ekman P and Rosenberg EL. 2005. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using Facial Action Coding System. 2nd Ed. Oxford University Express.
15. Russell JA, Bachorowski JA and Fernandez-Dols JM. 2003. Facial and Vocal Expressions of Emotion. Annual Review Psychology, 2003, 54:329-49
16. Ekman P. and Oster H. 1979. Facial Expressions of Emotion. Annual Review Psychology. 30:527-54
17. Roisman, G.I., Tsai, J.L., Chiang, K.S.(2004), The Emotional Integration of Childhood Experience: Physiological, Facial Expressive, and Self-reported Emotional Response During the Adult Attachment Interview, Developmental Psychology, Vol. 40, No. 5, 776-789
18. Cohn JF and Tronick EZ. 1988. Mother Infant Interaction: the sequence of dyadic states at three, six and nine months. Development Psychology, 23, 68-77
19. Fried E. 1976. The impact of nonverbal communication of facial affect on children's learning. PhD thesis, Rutgers University, New Brunswick, NJ
20. Ekman P, Matsumoto D, and Friesen WV. 2005. Facial Expression in Affective Disorders. In What the Face Reveals. Edited by Ekman P and Rosenberg EL. 429-439
21. Zeng, Z, Fu, Y., Roisman, G.I., Wen, Z., Hu, Y. and Huang, T.S., One-class classification on spontaneous facial expressions, Automatic Face and Gesture Recognition, 281 – 286, 2006
22. Bourlard, H. and Dupont, S., A new ASR approach based on independent processing and recombination of partial frequency bands, ICSLP 1996
23. Devillers, L., Vidrascu L. and Lamel L., Challenges in real-life emotion annotation and machine learning based detection, Neural Networks, 18(2005), *407-422*
24. Ekman, P., Hager, J.C., Methvin, C.H. and Irwin, W., Ekman-Hager Facial Action Exemplars, unpublished, San Francisco: Human Interaction Laboratory, University of California
25. Kanade, T., Cohn, J., and Tian, Y. (2000), Comprehensive Database for Facial Expression Analysis, In Proceeding of International Conference on Face and Gesture Recognition, 46-53
26. Pantic, M., Valstar, M.F, Rademaker, R. and Maat, L. (2005), Web-based database for facial expression analysis, Int. Conf. on Multimedia and Expo
27. JAFFE: www.mic.atr.co.jp/~mlyons/jaffe.html
28. Chen, L.S, Joint Processing of Audio-Visual Informa-tion for the Recognition of Emotional Expressions in Human-Computer Interaction, PhD thesis, UIUC, 2000
29. Cowie, R., Douglas-Cowie E. and Cox, C., Beyond emotion archetypes: Databases for emotion modelling using neural networks, 18(2005), 371-388
30. Ekman, P. and Rosenberg, E. (Eds.), What the face reveals. NY: Oxford University, 1997
31. Cohn, J.F. and Schmidt, K.L.(2004), The timing of Facial Motion in Posed and Spontaneous Smiles, International Journal of Wavelets, Multiresolution and Information Processing, 2, 1-12
32. Valstar MF, Pantic M, Ambadar Z and Cohn JF. 2006. Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions. Int. Conf. on Multimedia Interfaces. 162-170
33. Ekman, P. (1994), Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique, Psychological Bulletin, 115(2): 268-287
34. Pantic M., Rothkrantz, L.J.M., Toward an affect-sensitive multimodal human-computer interaction, Proceedings of the IEEE, Vol. 91, No. 9, Sept. 2003, 1370-1390

35. Pantic, M., Sebe, N., Cohn, J.F. and Huang, T., Affective Multimodal Human-Computer Interaction, in Proc. ACM Int'l Conf. on Multimedia, November 2005, 669-676
36. Sebe, N., Cohen, I., Gevers, T., and Huang, T.S. (2005), Multimodal Approaches for Emotion Recognition: A Survey, In Proc. Of SPIE-IS&T Electronic Imaging, SPIE Vol 5670: 56-67
37. 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, January 2001, 32-80
38. Chen, L. and Huang, T. S., Emotional expressions in audiovisual human computer interaction, Int. Conf. on Multimedia & Expo 2000, 423-426
39. Chen, L., Huang, T. S., Miyasato, T., and Nakatsu, R., Multimodal human emotion/expression recognition, Int. Conf. on Automatic Face & Gesture Recognition 1998, 396-401
40. De Silva, L. C., and Ng, P. C., Bimodal emotion recognition, Int. Conf. on Automatic Face & Gesture Recognition 2000, 332-335
41. Yoshitomi, Y., Kim, S., Kawano, T., and Kitazoe, T., Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in Proc. ROMAN 2000, 178-183
42. Hoch, S., Althoff, F., McGlaun, G., Rigoll, G., Bimodal fusion of emotional data in an automotive environment, ICASSP, Vol. II, 1085-1088, 2005
43. Wang, Y. and Guan, L., Recognizing human emotion from audiovisual information, ICASSP, Vol. II, 1125-1128
44. Zeng, Z., Hu, Y., Liu, M., Fu, Y. and Huang, T.S., Training Combination Strategy of Multi-stream Fused Hidden Markov Model for Audio-visual Affect Recognition, in Proc. ACM Int'l Conf. on Multimedia, 2005, 65-68
45. Zeng, Z., Tu, J., Pianfetti , P., Liu, M., Zhang, T., et al., Audio-visual Affect Recognition through Multi-stream Fused HMM for HCI, Int. Conf. Computer Vision and Pattern Recognition. 2005: 967-972
46. Zeng, Z., Tu., J., Liu, M., Huang, T.S. (2005), Multi-stream Confidence Analysis for Audio-Visual Affect Recognition, the Int. Conf. on Affective Computing and Intelligent Interaction, 946-971
47. Zeng, Z., Zhang, Z., Pianfetti, B., Tu, J., and Huang, T.S. (2005), Audio-visual Affect Recognition in Activation-evaluation Space, Int. Conf. on Multimedia & Expo, 828-831.
48. Zeng, Z., Tu, J., Liu, M., Huang, T.S., Pianfetti, B., Levinson, S. (2007), Audio-visual Affect Recognition, IEEE Transactions on Multimedia, in press
49. Song, M., Bu, J., Chen, C., and Li, N., Audio-visual based emotion recognition—A new approach, Int. Conf. Computer Vision and Pattern Recognition. 2004, 1020-1025
50. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M. et al., Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. 2004. Int. Conf. Multimodal Interfaces. 205-211
51. Hoch, S., Althoff, F., McGlaun, G., Rigoll, G., Bimodal fusion of emotional data in an automotive environment, ICASSP, Vol. II, 1085-1088, 2005
52. Wang, Y. and Guan, L., Recognizing human emotion from audiovisual information, ICASSP, Vol. II, 1125-1128
53. Go HJ, Kwak KC, Lee DJ, and Chun MG. 2003. Emotion recognition from facial image and speech signal. Int. Conf. of the Society of Instrument and Control Engineers. 2890-2895
54. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J.(2005), Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior, IEEE CVPR'05
55. Sebe, N., Lew, M.S., Cohen, I., Sun, Y., Gevers, T., Huang, T.S.(2004), Authentic Facial Expression Analysis, Int. Conf. on Automatic Face and Gesture Recognition

56. Zeng, Z., Fu, Y., Roisman, G.I., Wen, Z., Hu, Y., and Huang, T.S. (2006). Spontaneous Emotional Facial Expression Detection. Journal of Multimedia, 1(5): 1-8.
57. Valstar MF, Pantic M and Ambadar Z, and Cohn JF. 2006. Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions. Int. Conf. on Multimodal Interfaces. 162-170
58. Cohn JF, Reed LI, Ambadar Z, Xiao J, and Moriyama T. 2004. Automatic Analysis and recognition of brow actions and head motion in spontaneous facial behavior. Int. Conf. on Systems, Man & Cybernetics, 1, 610-616
59. Litman, D.J. and Forbes-Riley, K., Predicting Student Emotions in Computer-Human Tutoring Dialogues. In Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), July 2004
60. Batliner A, Fischer K, Hubera R, Spilkera J and Noth E. 2003. How to find trouble in communication. Speech Communication, Vol. 40, 117-143.
61. Neiberg D, Elenius K, and Laskowski K. 2006. Emotion Recognition in Spontaneous Speech Using GMM. Int. Conf. on Spoken Language Processing, 809-812
62. Ang J, Dhillon R, Krupski A, et al. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog, ICSLP.
63. Fragopanagos, F. and Taylor, J.G., Emotion recognition in human-computer interaction, Neural Networks, 18 (2005) 389-405
64. Garidakis, G., Malatesta, L., Kessous, L., Amir, N., Paouzaiou, A. and Karpouzis, K.. 2006. Modeling Naturalistic Affective States via Facial and Vocal Expression Recognition. Int. Conf. on Multimodal Interfaces. 146-154
65. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'Feeltrace': an instrument for recording perceived emotion in real time. Proceedings of the ISCA Workshop on Speech and Emotion, 19–24
66. Maat L and Pantic M. 2006. Gaze-X: Adaptive Affective Multimodal Interface for Single-User Office Scenarios. Int. Conf. on Multimodal Interfaces. 171-178
67. Lanitis, A., Taylor, C. and Cootes, T. (1995), A Unified Approach to Coding and Interpreting Face Images, in Proc. International Conf. on Computer Vision, 368-373
68. Black, M. and Yacoob, Y.(1995), Tracking and Recognizing Rigid and Non-rigid Facial Motions Using Local Parametric Models of Image Motion, in Proc. Int. Conf. on Computer Vision, 374-381
69. Rosenblum, M., Yacoob, Y., and Davis, L. (1996), Human Expression Recognition from Motion Using a Radial Basis Function Network Architecture, IEEE Trans. On Neural Network, 7(5):1121-1138
70. Essa, I. and Pentland, A. (1997), Coding, Analysis, Interpretation, and Recognition of Facial Expressions, IEEE Trans. On Pattern Analysis and Machine Intelligence, 19(7): 757-767
71. Cohen, L., Sebe, N., Garg, A., Chen, L., and Huang, T. (2003), Facial expression recognition from video sequences: Temporal and static modeling, Computer Vision and Image Understanding, 91(1-2):160-187
72. Tian, Y., Kanade, T., Cohn, J.F. (2001), Recognizing Action Units for Facial Expression Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2): 97-115
73. Pantic M and Patras I. 2006. 'Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences', IEEE Transactions on Systems, Man and Cybernetics - Part B, vol. 36, no. 2, pp. 433-449
74. Kwon, O.W., Chan, K., Hao, J., Lee, T.W (2003), Emotion Recognition by Speech Signals, EUROSPEECH.
75. Polzin, Thomas (1999), Detecting Verbal and Non-verbal cues in the communication of emotion, PhD thesis, Carnegie Mellon University
76. Amir, N. and Ron, S. (1998), Toward Automatic Classification of Emotions in Speech, in Proc. ICSLP, 555-558

77. Dellaert, F., Polzin, T., and Waibel, A. (1996), Recognizing Emotion in Speech, In Proc. ICSLP, 1970-1973
78. Petrushin, V.A. (2000), Emotion Recognition in Speech Signal, In Proc. ICSLP, 222-225
79. Pantic M, Pentland A, Nijholt A and Huang TS. 2006. Human Computing and Machine Understanding of Human Behavior: A Survey. Int. Conf. Multimodal Interfaces. 233-238
80. Huang, D. (1999), Physiological, subjective, and behavioral Responses of Chinese American and European Americans during moments of peak emotional intensity, honor Bachelor thesis, Psychology, University of Minnesota.
81. Tao, H. and Huang, T.S., Explanation-based facial motion tracking using a piecewise Bezier volume deformation mode, IEEE CVPR'99, vol.1, pp. 611-617, 1999
82. Wen Z and Huang T. 2003. Capturing Subtle Facial Motions in 3D Face Tracking. Intl. Conf. on Computer Vision (ICCV). 1343-1350
83. He, X., Yan, S., Hu, Y., and Zhang, H, Learning a Locality Preserving Subspace for Visual Recognition, Int. Conf. on Computer Vision, 2003