Audio-Visual Emotion Recognition in Adult Attachment Interview

Zhihong Zeng, Yuxiao Hu, Glenn I. Roisman, Zhen Wen, Yun Fu and Thomas S. Huang
University of Illinois at Urbana-Champaign
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 affective expressions is a largely unexplored and challenging problem. In this paper, we explore audio-visual emotion recognition in a realistic human conversation setting—Adult Attachment Interview (AAI). Based on the assumption that facial expression and vocal expression be at the same coarse affective states, positive and negative emotion sequences are labeled according to Facial Action Coding System Emotion Codes. Facial texture in visual channel and prosody in audio channel are integrated in the framework of Adaboost multi-stream hidden Markov model (AMHMM) in which Adaboost learning scheme is used to build component HMM fusion. Our approach is evaluated in the preliminary AAI spontaneous emotion recognition experiments.

Categories and Subject Descriptors
H.1.2 [User/Machine Systems]: Human information processing
I.5.4 [Pattern Recognition Applications]: Computer vision, signal processing

General Terms
Algorithms

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 [3]. 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.

Automatic emotion recognition has been attracting attention of researchers from a variety of different disciplines. Another potential application of automatic emotion recognition is to help people in emotion-related research to improve the processing of emotion data.

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 [4] to study links between adults' narratives about their childhood experiences and their emotional expressive, physiological, and self-reported emotion. In their study, the analysis of emotion expressions is currently completed manually using coding systems such as the Facial Action Coding System (FACS) [6]. Automatic emotion recognition can improve the efficiency and objectivity of this psychological research.

While the ability to recognize a variety of fine-grained emotions is attractive, it may be not practical because the emotion data of the realistic conversation 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 [4]. This work extends our previous work [12] which 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 (AMHMM) 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 which are then transformed into low dimensional subspace by Locality Preserving Projection. 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 implication. 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 [13]. 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 the related work about emotion recognition. Section 3 we introduce the AAI data which 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 (AMHMM) framework. Section 7 presents our preliminary experimental results on two AAI subjects to evaluate our method. Finally, we have concluding remarks in Section 8.

![Figure 1. The audio-visual emotion recognition framework.](image)

2. RELATED WORK

In these years, literature about automatic emotion recognition is growing dramatically due to the development of techniques in computer vision, speech analysis and machine learning. However, automatic recognition on emotions occurring on natural communication setting 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 hard to elicit emotions without influencing the results. Manual labeling of spontaneous emotional expressions is very time consuming, error prone, and expensive [16]. This state of affairs leads to a big challenge to spontaneous emotional expression analysis. Due to these difficulties in emotional expression recognition, most of current automatic facial expression studies [1-2] were 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. An overview of databases for emotion classification studies can be founded in [7].The psychological study [17] indicates that the posed nature of the emotions may differ in appearance and timing from corresponding performances in natural settings. Especially, the study [18] indicated that posed smiles were of larger amplitude and has less consistent relation between amplitude and duration than spontaneous smile. In addition, most of current emotion recognizers are evaluated in clear audio-visual input (e.g. high quality visual and audio recording, non-occluded and front portraits) which is different from the natural communication setting. Therefore, the methods based these artificial emotions would perform differently on emotional expressions occurring in natural communication setting.

To simulate the human ability to assess affect, an automatic affect recognition system should also make use of multimodal fusion. However, most of current automatic emotion recognition approaches are uni-modal: information processed by the computer system is limited to either face images or the speech signals. Relatively little work has been done in researching multimodal affect analysis [13-15][23][30].

Recently, there are an increasing number of studies toward spontaneous emotion recognition, including spontaneous facial expression recognition [10-12], spontaneous audio affect recognition [19-20]. One notable study was done by Fragopanagos and Taylor (2005) [9] who explore multimodal emotion recognition on natural emotion data. They use Feeltrace tool [8] to label the emotion data in activation-evaluation space, and use a neural network architecture to handle the fusion of different modalities. In their work, 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. Especially 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.

Comparing 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 FASC 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).
3. MULTIMEDIA EMOTION DATA IN ADULT ATTACHMENT INTERVIEW

The Adult Attachment Interview (AAI) is a semistructured 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 [4].

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 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 amounts of data allow us personal-dependent spontaneous emotion analysis.

In this work, 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 some main 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.

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 [5].

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 [6]. 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.

4. FACIAL EXPRESSIONS

4.1 3D Face Tracker

To handle the arbitrary behavior of subjects in the natural setting, the 3D face tracked is required. The face tracking in our experiments is based on a system called Piecewise Bezier Volume Deformation (PBVD) tracker which was developed in [21].

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. That 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; (c) extracted face texture.

Figure 2: The snapshot of a AAI video. The participant’s face is displayed in the bigger window while the interviewer’s face is in the smaller left-top window.

Figure 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 which is obtained by finding the optimal linear approximations to the eigen-functions of the Laplace Beltrami operator on the manifold [22]. Different from the nonlinear manifold learning techniques, LLP could 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. Different from PCA, LLP 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 [22]. 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 [23] 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 (AMHMM)

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 (AVASR) literature [24], few studies are found for audio-visual affect recognition [9][13-15][23][30].

Most of current multi-stream combination studies focus on weighting combination scheme with weights proportional to the reliabilities of the component HMMs. The weights could be computed from normalized stream recognition rate [25], stream S/N ratio [25], stream entropy [26], or other reliability measures such as ones in [27].

The weighting combination scheme is intuitive and reasonable in some way. But 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. Even it is 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 [13], 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

\[(x_{i1}, \ldots, x_{in}, y_i), \ldots, (x_{im}, \ldots, x_{m}, y_m)\]

where \(x_j\) 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

\[
\left(P_{10}, P_{11}, \ldots, P_{i0}, P_{i1}, y_i \right), \ldots, \left(P_{n0}, P_{n1}, \ldots, P_{mn}, P_{mn1}, y_m \right)
\]

where \(P_{ij}\) are likelihoods of negative and positive emotions of jth stream of ith sample sequence.

3) Fusion training: based on Adaboost learning scheme [28], 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 [28].

In our experiment, we compare Adaboost MHMM with a traditional weighting combination method named Acc MHMM where the final confidences of positive and negative emotions are the weighted summation of the log likelihoods of component HMMs, and the weights are proportional to the corresponding stream recognition accuracies.

7. EXPERIMENTAL RESULTS

In this section, we present the experimental results of our emotion recognition by using audio-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 1 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 1. Performance Comparison of LPP HMM and PCA HMM

<table>
<thead>
<tr>
<th></th>
<th>Approach</th>
<th>Dimension</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>LPP HMM</td>
<td>5</td>
<td>87.50</td>
</tr>
<tr>
<td></td>
<td>PCA HMM</td>
<td>5</td>
<td>79.69</td>
</tr>
<tr>
<td>Male</td>
<td>LPP HMM</td>
<td>4</td>
<td>84.85</td>
</tr>
<tr>
<td></td>
<td>PCA HMM</td>
<td>5</td>
<td>72.62</td>
</tr>
</tbody>
</table>

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

Figure 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 2 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 affective information than prosody, as the psychological study [29] indicated that during human judgment of emotions, 55% of affective information is from facial expressions and 38% from vocal utterances. 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.
7.3 Audio-visual Fusion

The emotion recognition performance of audio-visual fusion is shown in Table 3. In this table, two combination schemes 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 mean MHMM with the Adaboost learning schemes as described in Section 6.

Table 2. Emotion Recognition of Prosody HMM

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>75.09</td>
</tr>
<tr>
<td>Male</td>
<td>65.15</td>
</tr>
</tbody>
</table>

The results demonstrate that training combination outperforms weighting combination, i.e. Adaboost MHMM are better than Acc MHMM. In addition, Adaboost MHMM is better than uni-stream 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 single modality.

The experimental results of Acc MHMM show that the accuracy-based weighting combination scheme is not good in our application. The accuracy of Acc MHMM is worse than visual LPP HMM for female data, and equals to visual LPP HMM for male data. Thus, although the weighting combination is reasonable and intuitive, it is difficult to obtain the optimal combination because learning is not involved. Even it is possible that the weighting combination is worse than individual component performance, as shown in our experiments.

The confusion matrices of emotion recognition for two subjects are shown in Table 4 and 5. These results demonstrate that negative emotions are more difficult to be recognized than positive emotions. We notice that adult subjects tend to inhibit negative emotion expressions in this interview context. Thus, the negative emotions are shorter and more subtle than positive emotions.

Table 3. Audio-visual Emotion Recognition

<table>
<thead>
<tr>
<th>Bimodal Fusion</th>
<th>Combination scheme</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>Acc MHMM</td>
<td>Weighting</td>
</tr>
<tr>
<td></td>
<td>AdaBoost MHMM</td>
<td>Training</td>
</tr>
<tr>
<td>Male</td>
<td>Acc MHMM</td>
<td>Weighting</td>
</tr>
<tr>
<td></td>
<td>AdaBoost MHMM</td>
<td>Training</td>
</tr>
</tbody>
</table>

Table 5. Confusion Matrix for Male Emotion Recognition

<table>
<thead>
<tr>
<th>Desired</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
</tr>
<tr>
<td>Positive</td>
<td>91.67</td>
</tr>
<tr>
<td>Negative</td>
<td>13.33</td>
</tr>
</tbody>
</table>

8. CONCLUSION

Emotion analysis has been attracting increasing 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.

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-dependent emotion recognition 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 level (positive and negative emotions). Although this assumption is valid at most circumstances, blended emotions can occur when speakers have conflict intension [31]. The exploration of recognition of the blended emotions is our future work.

9. ACKNOWLEDGEMENTS

We like to thank Prof. Maja Pantic and Anton Nijholt for providing valuable comments. The support of Beckman Postdoctoral Fellowship and National Science Foundation: Information Technology Research Grant # 0085980 is gratefully acknowledged.

10. REFERENCES

[2] 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,
Emotional Response During the Adult Attachment Interview, Developmental Psychology, Vol. 40, No. 5, 776-789


[28] Paul A. Viola, Michael J. Jones: Robust Real-Time Face Detection. ICCV 2001


[31] Devillers, L., Abrillan, S. and Martin, J., Representing Real-life Emotions in Audiovisual Data with Non Basic Emotional Patterns and Context Features, Int. Conf. on Affective Computing and Intelligent Interaction, 519-526