AUDIO-VISUAL AFFECT RECOGNITION IN ACTIVATION-EVALUATION SPACE

Zhihong Zeng, Zhenqiu Zhang, Brian Pianfetti, Jilin Tu and Thomas S. Huang
University of Illinois at Urbana-Champaign
{zhzeng, zzhang6 , jilintu, huang}@ifp.uiuc.edu, bpianfet@uiuc.edu

ABSTRACT

The ability of a computer to detect and appropriately respond to changes in a user's affective state has significant implications to Human-Computer Interaction (HCI). To more accurately simulate the human ability to assess affects through multi-sensory data, automatic affect recognition should also make use of multimodal data. In this paper, we present our efforts toward audio-visual affect recognition. Based on psychological research, we have chosen affect categories based on an activation-evaluation space which is robust in capturing significant aspects of emotion. We apply the Fisher boosting learning algorithm which can build a strong classifier by combining a small set of weak classification functions. Our experimental results show with 30 Fisher features, the testing error rates of our bimodal affect recognition is about 16% on the evaluation axis and 13% on the activation axis.

1. INTRODUCTION

Changes in a person's affective state play a significant role in perception and decision making during human to human interactions. This fact has inspired the research field of "affective computing" which aims at enabling computers to express and recognize affect [4]. Perhaps the most fundamental applications of affective computing would be in human-computer interaction where the computer could detect and track a user's affective states and initiate communications based on this knowledge, rather than simply responding to a user's commands.

Research presented in this paper is part of an ongoing federally funded project (ITR) (itr.beckman.uiuc.edu) which is to contribute to the development of a multimodal human-computer intelligent interaction (HCII) learning environment. The concepts and tools resulting from this project are applied to an educational context for evaluation. This education based test-bed focuses on the ability of the computing environment to help at-risk middle school students learn math and science concepts through exploration of Lego gears. This project uses a proactive computing learning environment to achieve two goals. The first goal was to keep the children actively engaged in the learning activity. The second goal was to support the exploration of math and science phenomena enabling the children to increase their knowledge. This is done through the recognition of changes in the children's affective states and applying appropriate tutoring strategies.

The psychological study [2] indicated that judging someone's affective states, people mainly rely on facial expressions and vocal intonations. Thus, affect recognition should inherently be the issue of multimodal analysis. The primary aim of this paper is to combine cues from facial expression and vocal modalities to increase the accuracy of affect recognition algorithms.

In this paper, we choose the affect categories in activation-evaluation space, namely active vs. passive states, and positive vs. negative state. These categories are both simple and capable of capturing significant issues in emotion. Based on this affect representation, we apply Fisher Boosting learning method which provides a good framework for feature selection and combination of multiple weak classifiers to achieve good classification performance.

2. RELATED WORK

Researchers from many different disciplines are interested in the possibility of automated affect analysis and recognition. Recent advances in computing power and multimedia technologies are facilitating efforts toward audio-visual affect recognition. According to [3], four papers reported advances of bimodal affect recognition. In addition, there have been three papers of bimodal emotion recognition [1][10][15] recently published in 2004.

All of these above-mentioned studies used the category of six emotions. In addition, [11] considered four HCI-related cognitive states besides six basic emotions. Contrary to the fine categories of emotion representation, this paper explores the coarse categories for automatic affect recognition. That is inspired by the psychological research [5] which represents emotion in the activation-evaluation space. This emotion space is both simple and capable of capturing significant issues in emotion. It has been used in audio-only emotion recognition research [6].

Based on the activation-evaluation space of emotion, we apply Fisher Boosting method for classification which has been successfully used in face localization [10].

0-7803-9332-5/05/$20.00 ©2005 IEEE

3. ACTIVATION-EVALUATION SPACE OF AFFECT

The accuracy of detecting fine emotional states is limited because the differences among them are so subtle. That results in difficulty to make appropriate responses from the computer. In addition, detecting affect states such as joy, anger alone are of little use when applied to learning environments such as the research being conducted in this paper's parent ITR project. An example of this can be in the projects early use of "anger" and "sad" states. In this research, when a subject did not know the answer to a problem but was still working towards a solution his/her affect would be typically classified as being "sad". When a subject would reach a point where he failed to solve the problem and no longer wanted anything to do with the activity he/she would exhibit traits that would be classified as "angry." The problem with this system is that it does not capture the similarities and differences between the affective states that can be used to increase partial accuracy. For example, the "sad" and "angry" are similar because they are both negative emotions; however, they differ in how much a person may be motivated to change their current situation. To help capture the subject's positive or negative emotions as well as their motivation to change the current situation, an activation-evaluation space model was used to [5]. [5] suggests that the majority of emotion misclassifications remain in the same quadrant. That suggests that even misclassification generally convey broad information of emotional state.

In the activation-evaluation space model, the subject's positive or negative evaluations of the situation is plotted along the X axis, while the motivation to change the situation through action, active or passive, is placed on the Y axis. In this four quadrant system, the basic emotions would be located in the periphery of an emotion circle in Figure 1.

In this system, knowing which quadrant the emotion is located can increase the likelihood of an appropriate response by the computer if the responses are focused on the quadrants and not just on the fine emotional state. For example, if the system chooses quadrant III, negative emotion and passive motivation, but the subject is actually in quadrant II, negative emotion and positive motivation, and then the negative emotion aspect of the response would be correct, only the motivational aspect would be wrong. Focusing on the changes in the positive or negative state of the user as well as the likelihood that the user will change his situation can increase the likelihood of at least partially correct affect recognition. Choosing computer responses that also take these two classifications into consideration can also result in producing more appropriate responses.

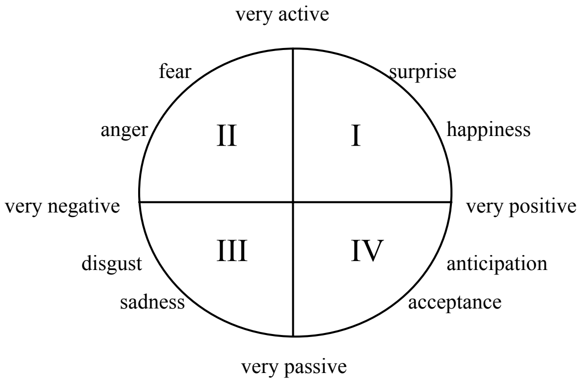

Figure 1. Activation-evaluation space of affect

4. DATABASE

A large-scale database was collected [12] that is customized for multimodal affect recognition research for human-computer interaction applications. In this paper, we choose data of 20 subject on 6 basic emotions (i.e. happiness, sadness, fear, surprise, anger, and disgust) and label them in the activation-evaluation space.

The 20 subjects (10 female and 10 males) in our database consist of graduate and undergraduate students from different disciplines. This set of videos contains subjects with a wide variability in physiognomy. Although the subjects displayed affect expressions on request, the subjects chose how to express each state. They were simply asked to display facial expressions and speak appropriate sentences three times.

5. FEATURE EXTRACTION

In this section, we present the techniques used for extracting the facial features in visual channel and prosodic features in audio channel.

5.1. Facial features

A tracking algorithm called Piecewise Bezier Volume Deformation (PBVD) tracking [9] was applied to extract facial features in our experiment.

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 (frontal view of a neutral facial expression), the 3-D facial mesh model is constructed by manual or automatic selection [10] of landmark facial feature points. Once the model was fitted, the tracker can track head motion and local deformations of the facial features. At the current stage, only local deformations of facial features are used for affect recognition. These deformations are measured in terms of magnitudes of 12 predefined motions of facial features, called Motion Units (MUs), which are shown in Figure 2. The outputs of the face tracker corresponding to 12 MUs are used as facial features for later affect recognition in our experiment.

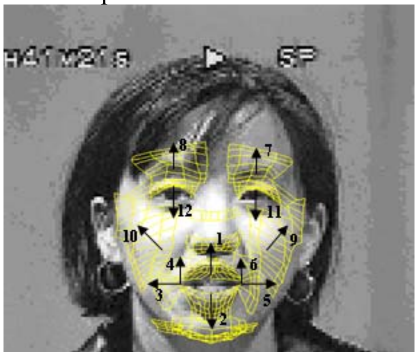

Figure 2. 12 facial Motion Units

5.2. Prosodic features

For audio feature extraction, Entropic Signal Processing System named get_f0, a commercial software package, is used. It implements a fundamental frequency estimation algorithm by using the normalized cross correlation function and dynamic programming [12]. 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 [13] 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.

6. FISHER BOOSTING CLASSIFIER

Based on the affect representation in activation-evaluation space, the affect recognition problem can be divided as two 2-class classification problems: positive and negative states in evaluation level, and active and passive states in activation level.

In order to solve these 2-class classification problems, we apply the Fisher Boosting algorithm which can build a strong classifier by combining a small set of weak classifiers [10]. The study in [10] shows that compared with conventional Adaboosting algorithm, it can build a strong classifier with fewer weak classifiers.

Given a training set of 2-class examples, Fisher boosting learning approach takes a greedy strategy to gradually add the discriminating feature on which a weak classifier is learned. At each stage of the process, the examples are weighted so that those examples misclassified by the previous weak classifier are emphasized. And the weighted Fisher feature is obtained according to Fisher criterion which maximizing the between-class variance and minimizing the within-class variance. Then, the high dimension data are projected to this feature, and the weak classifier is trained. By incrementally learning Fisher features that can best discriminate the data, we can construct an optimal feature set with a small number of Fisher features. The final strong classifier is a weighted linear combination of these weak classifiers, and the weights are inversely proportional to the corresponding training errors.

7. EXPERIMENT

We test our person-dependent affect recognition algorithm on the above-mentioned dataset of 20 subjects (10 females and 10 males). For every subject, the data is divided into two halves. One half of every subject's frames are selected as training data, and the other half as testing data. In the classification evaluation, both testing data and training data are recognized by the trained Fisher Boosting classifier.

We apply Fisher boosting algorithm on the data of the joint features which are constructed by concatenating vectors of audio and visual features. In real-time condition, our face tracker can only achieve 10 frames per second on a P4 1.79GHz. In order to get as many frames as possible, our current approach applies off-line processing. Thus, according to the video rate, the face tracker outputs 30 frames per second. The prosody modality in our experiment can output 90 frames per second in real-time condition. Thus, in constructing fused feature sets, the visual features are upsampled to the audio feature extraction rate (90Hz) by linear interpolation.

In our experiment, Fisher boosting learning algorithm is able to effectively seek discriminating features in 14-dimensional space constructed by 14 affective features (i.e. 2 prosodic feature and 12 facial features). For classification on evaluation axis, the average training and testing error curves of Fisher boosting algorithm are shown in Figure 3. For classification on activation axis, the training and testing error curves are shown in Figure 4.

Figure 3 and Figure 4 illustrate that Fisher Boosting is good at solving the classification overfitting problem. The performance on training data and testing data are close to each other, especially when feature number is small. The curves in Figure 3 show with the increasing number of Fisher features, the training and testing errors of the classifier globally decrease while there is local increase at the beginning. And the changes of the error rates are gradually slower with the increasing number of Fisher features. The curves in Figure 4 have the same performance. The number of 30 features seems to be good point where the training and testing errors are about 0.14 and 0.16 on evaluation axis, and 0.11 and 0.13 on activation axis. These coarse-emotion-category-based performances are greatly above our previous fine-emotion-category-based classification at the frame level [11].

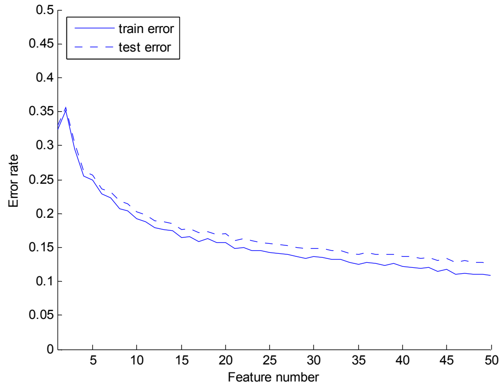

Figure 3: ROC curve of Fisher Boosting on evaluation axis

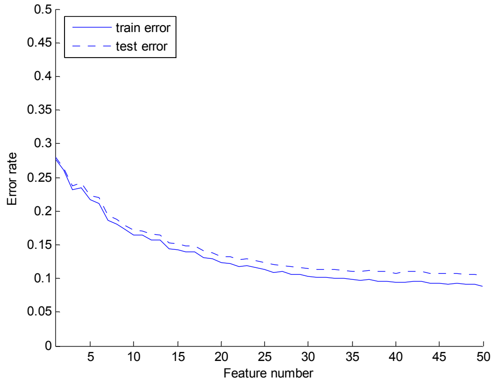

Figure 4: ROC curve of Fisher Boosting on activation axis

8. CONCLUSION

With an automatic affect recognizer, a computer can respond appropriately to the user's affective state rather than simply responding to user commands. In this way, the nature of the computer interactions would become more authentic, persuasive, and meaningful. This type of interaction is the ultimate goal of ITR project where attending to changes in the child's affective states leads to a high level of engagement and knowledge acquisition.

In this paper, we introduce our effort toward multimodal affect recognition. Different from previous bimodal affect recognition studies mentioned in Section 2, we explore the affect recognition on coarse categories in activation-evaluation space which is robust in capturing significant aspects of emotion. We apply Fisher boosting learning algorithm which can build a strong classifier by combining a small set of weak classification functions.

The activation-evaluation space of representing the affective state of a subject also may enable a more continuous tracking between different affect states. This method could allow the tracking of a person's affective state as they are transitioning from one known position in the model to another. Having the ability to account for these in-between states can aid in the prediction of what state the subject may be approaching which would be useful for a tutoring system in helping guide a subject toward desired learning states.

11. REFERENCES

[1] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S., *Analysis of Emotion Recognition Using Facial Expressions Speech and Multimodal Information*, ICMI 2004

[2] Mehrabian, A., *Communication without words, Psychol. Today*, vol.2, no.4, 53-56, 1968

[3] 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

[4] Picard, R.W., *Affective Computing*, MIT Press, Cambridge, 1997.

[5] 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

[6] Cowie, R., Douglas-Cowie, E. and Romano, A., *Changing Emotional Tone in Dialogue and its Prosodic Correlates*. In Proc. ESCA Workshop on Dialogue and Prosody, Netherlands, 1999

[7] Song, M., Bu, J., Chen, C., Li, N., *Audio-Visual Based Emotion Recognition-A New Approach*, CVPR 2004.

[8] Talkin, D., *A Robust Algorithm for Pitch Tracking, in Speech Coding and Synthesis*, Kkeijn, W.B., and Paliwal, K.K., Eds., Amsterdam: Elsevier Science, 1995

[9] Tao, H. and Huang, T.S., *Explanation-based facial motion tracking using a piecewise Bezier volume deformation mode*, CVPR'99, vol.1, pp. 611-617, 1999.

[10] Tu, J., Zhang, Z., Zeng, Z. and Huang, T.S., *Face Localization via Hierarchical Condensation with Fisher Boosting Feature Selection*, In Proc. Computer Vision and Pattern Recognition, 2004.

[11] Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T.S., Roth, D., and Levinson, S., *Bimodal HCI-related Affect Recognition*, ICMI 2004

[12] Chen, L.S, Joint Processing of Audio-Visual Information for the Recognition of Emotional Expressions in Human-Computer Interaction, PhD thesis, UIUC, 2000

[13] Kwon, O.W., Chan, K., Hao, J., Lee, T.W, Emotion Recognition by Speech Signals, EUROSPEECH 2003.