Multi-stream Confidence Analysis for Audio-visual Affect Recognition

Zhihong Zeng, Jilin Tu, Ming Liu and Thomas S. Huang

Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
405 N. Mathews Av., Urbana, IL 61801
{zhzeng, jilintu, mingliu1, huang}@ifp.uiuc.edu

Abstract. Changes in a speaker's emotion are a fundamental component in human communication. Some emotions motivate human actions while others add deeper meaning and richness to human interactions. In this paper, we explore the development of a computing algorithm that uses audio and visual sensors to recognize a speaker's affective state. Within the framework of Multi-stream Hidden Markov Model (MHMM), we analyze audio and visual observations to detect 11 cognitive/emotive states. We investigate the use of individual modality confidence measures as a means of estimating weights when combining likelihoods in the audio-visual decision fusion. Person-independent experimental results from 20 subjects in 660 sequences suggest that the use of stream exponents estimated on training data results in classification accuracy improvement of audio-visual affect recognition.

1 Introduction

The traditional Human Computer Interaction (HCI) system is constructed to emphasize the transmission of explicit messages while ignoring implicit information about the user such as changes in emotional states. However, changes in a speaker's emotion are a fundamental component in human communication. Some emotions motivate human actions while others add deeper meaning and richness to human interactions. Consequently, the traditional HCI that ignores a user's emotional states is only making use of a small portion of the data available in an interaction. This fact has inspired the research field of "emotional computing" [5] which aims at enabling computers to express and recognize emotion. The ability to detect and track a user's affect state has the potential of allowing 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. This enables a computing system to offer relevant information when a user needs help not just when the user requests help. In this way, human computer interaction can become more natural, persuasive, and friendly.

The work in this paper is motivated by the ITR project (itr.beckman.uiuc.edu). The goal of this project is to contribute to the development of multimodal human-computer intelligent interaction environment. An educational learning environment was used as a test-bed to evaluate the ideas and tools resulting from this research.

This test-bed focused on using Lego gears to teach math and science concepts to upper elementary and middle school children. The project focuses on using proactive computing to achieve two ends. First, to help children explore and understand a variety of phenomena ranging from mathematic ratios to advanced concepts of mechanical advantage and least common multiples. The second goal of the project is to support and prolong a student's interest in the activities while also promoting a high level of student engagement. This is accomplished through a multimodal computer learning environment that uses audio-visual sensors to recognize the student's affective states (e.g. interest, boredom, frustration and puzzlement) and to proactively apply appropriate context specific tutoring strategies (e.g. encouragement, transition/guidance, and confirmation). Through these techniques, students explore a variety of math and science concepts in a highly engaged mode of learning.

Multimodal sensory information fusion is a process that enables human ability to assess emotional states robustly and flexibly. To more accurately simulate the human ability to assess affect, an automatic affect recognition system should also make use of multimodal data. In this paper, we present our efforts toward audio-visual affect recognition. Based on the psychological study [9] which indicated people mainly rely on facial expressions and vocal intonations to judge someone's affective states, we focus on the analysis of facial expression in the visual channel, speech prosody in the audio channel, and bimodal fusion.

For integrating audio and visual streams, we applied the multi-stream hidden Markov model (MHMM) [11][12] which can be fused at the parallel architecture. We investigate the use of individual modality confidence measures as a means of estimating weights when combining likelihoods in the audio-visual decision fusion. Our person-independent affect recognition approaches were tested in 660 sequences based on 20 subjects with 11 HCI-related affect states. The experimental results show that the use of stream exponents results in affect classification accuracy improvement of audio-visual affect recognition.

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 [1], only four papers reported advances of bimodal affect recognition. In addition, there have been four papers of bimodal emotion recognition recently published in 2004 and 2005

Among these eight papers, four papers did person-independent audio-visual affect recognition [2-4][14]. Compared with the four reports, we in this paper explore the use of individual modality confidence measures as a means of estimating weights when combining likelihoods in the audio-visual decision fusion. [3-4] applied rule-based methods for combining two modalities. [2] applied the single-modal method in a sequential manner for bimodal recognition. [14] simply uses manual setting of weights in audio-visual fusion.

Audio-visual fusion is an instance of the general classifier fusion problem. This paper explores decision fusion method in affect recognition application which combines the single-modality classifier outputs to recognize audio-visual affect. Classifier fusion based on their individual decision about the classes of interest is an active area of research with many applications [10][12]. Combination strategies are different in various aspects, such as the architecture used (parallel, cascade, or hierarchical combination), and information level considered at integration (abstract, rank-order, or measurement level). Although examples of most of these categories can be found in audio-visual Automatic Speech Recognition (AVASR) literature [12], few studies are found audio-visual affect recognition. In addition, most of AVASR studies focus on two-stream combination. This paper explores three-stream fusion problem.

3 Database

The datasets used in previous papers [2-4] were small in the number of subjects, and were not related directly to human computer interaction. To overcome these problems, a large-scale database was collected [13]. This database consists of controlled performances of 7 basic emotions (happiness, sadness, fear, surprise, anger, disgust, and neutral), and 4 cognitive states (interest, boredom, puzzlement and frustration).

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. Each subject was required to repeat each state with speech three times. Therefore, for every affective state, there are 3*20=60 video sequences. And there are totally 60*11=660 sequences for 11 affective states. The time of every sequence ranged from 2-6 seconds.

Speech energy was used to determine start and end points of each emotion expression because they easier to detect than those of facial expressions. Once these segments were defined, corresponding points of facial feature sequences were labeled.

4 Facial Feature Extraction

A tracking algorithm called Piecewise Bezier Volume Deformation (PBVD) tracking [6] is 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 [15] 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. These deformations are measured in terms of magnitudes of 12 predefined motions of

facial features around mouth, eyelids and eyebrows, called Motion Units (MUs), which are shown in Figure 1. A facial expression is represented as a linear combination of the 12 Motion Units (MU) in the following formula

$$V = BDP = B[D_1 D_2 ... D_{12}] \begin{bmatrix} p_1 \\ p_2 \\ ... \\ p_{12} \end{bmatrix} \quad (1)$$

where B is constructed by Bezier basis functions, $D_i (i = 1,...,12)$ is the displacement vector of ith MU, and the $p_i (i = 1,...,12)$ is the magnitude of ith MU deformation. The overall motion of the head and face is represented as

$$R(V_0 + BDP) + T \quad (2)$$

where R is the 3D rotation matrix, T is the 3D translation matrix, and V_0 is the initial neutral face model.

The local deformation output P of the tracker is used as facial affective features for affect recognition. The face tracker outputs 30 frames per second.

We notice that the movements of facial features are related to both affective states and content of speech. Thus, smooth facial features [16] are calculated by averaging facial features at consecutive frames to reduce the influence of speech on facial expression, based on the assumption that the influence of speech on face features is temporary, and the influence of affect is relatively more persistent.

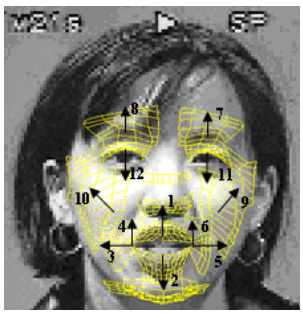

Fig. 1. 12 facial Motion Units

Regarding person-independent affect recognition, facial feature normalization is crucial because every subject has different physiognomy. To express an affect, different subjects will display different magnitudes of 12 MUs. To overcome this difference, the neutral expression for each person has been used as the normalization standard. In detail, for a given subject, the magnitudes of 12 MUs at every frame were normalized by the corresponding feature means of the neutral expression of the same subject.

After the feature vector of each frame is normalized, it is quantized into 19-size codebook by vector quantization (VQ).

5 Audio Feature Extraction

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 [7]. 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 [8] 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.

Obviously, the emotional information in the voice depends on the subject and recording condition. The pitch varies widely from person to person. In general, males speak with a lower pitch than females. Thus, for a given subject, the pitch at every frame is normalized by the pitch mean of the neutral expression sequence of the same subject. The same is done for energy features to normalize amplitude change due to the speaker volume and the distance of a speaker for microphone.

Similarly to the visual feature quantization, the energy and pitch are quantized into 19-size codebook by vector quantization respectively.

6 Decision Fusion for Audio-visual Affect Recognition

The main aim in this paper is to investigate and propose algorithm for the automatic recognition of audio-visual affect recognition. In our case, audio and visual observations are available. Each observation can be used alone to train single-modality statistical classifiers to recognize affective states. However, we hope that combining audio-visual information will give rise to a multi-modal classifier with superior performance to both single-modality ones.

We apply decision fusion method for audio-visual affect recognition which combine the single-modality HMM classifier outputs to recognize audio-visual affect. Specifically, class conditional log-likelihoods from the three classifiers are linearly combined using appropriate weights that explicitly model the reliability of each classifier. Such modeling is very important because discrimination power of the audio and visual streams can vary widely, depending on the acoustic noise in the environment, visual channel degradations, face tracker inaccuracies.

The decision fusion technique for audio-visual affect recognition used in this paper belong to the paradigm of multiple classifier integration using a parallel architecture, adaptive combination weights, and class measurement level information. In our application, the composite facial feature from video, energy and pitch features from audio are treated as three streams, and modeled by three component HMMs. We investigate integration where face-only, energy-only and pitch-only recognizer hypotheses are

rescored by the log-likelihood combination of these streams, which allows complete asynchrony among the three HMMs.

Let us denote the audio-visual observation vector which corresponds to an affective expression by $O = \{O^{(s)}\}$ where $s \in \{v, p, e\}$ representing visual, pitch and energy features respectively. The MHMM models its audio-visual likelihoods of the affective states (classes) as the product of the likelihood of its single-stream components, raised to appropriate stream exponents, namely

$$P[O \mid c] = \prod_{s \in \{v,p,e\}} P(O^{(s)} \mid c)^{\lambda_s} \quad (3)$$

where c denotes the affective states, and λ_s denote the stream exponents (weights), that are non-negative adding up to one, and are a function of the modality. The parameters of the MHMM [11] can be estimated using the EM algorithm. In our case, the stream exponents (weights) capture the confidence of the individual classifiers for our database condition, and are estimated by individual stream component performances (i.e. accuracies) on training data.

7 Experimental Results

In our experiment, the composite facial feature from video, energy and pitch features from audio are treated as three streams, and modeled by three component HMMs with 12 hidden states. The person-independent affect recognition algorithm was tested on 20 subjects (10 females and 10 males). For this test, all of the sequences of one subject are used as the test sequences, and the sequences of the remaining 19 subjects are used as training sequences. Among the training data of 19 subjects, we randomly choose the data of 14 subjects to estimate the parameters of the-single stream component HMMs. Then, the remaining training data of 5 subjects are used to estimate the performance of these component HMMs. Their classification accuracies are linearly mapped to stream exponents in (3) of MHMM which are adding to one. Finally, the test data of one subject are used to test the performance of this MHMM. The procedure is repeated 20 times, each time leaving a different person out (leave-one-out cross-validation). For every affective state, there are 3*20=60 expression sequences. Therefore, there are totally 60*11=660 sequences for 11 affective states.

Besides our above-mentioned audio-visual fusion denoted as MHMM 3, we also applied other five methods in our experiment: 1) face-only HMM. 2) pitch-only HMM. 3) energy-only HMM. 4) MHMM 1: each time 14 of 19 persons on training data are used to train single-stream component HMM , and audio-visual likelihoods are combined without stream exponents λ_s in (3) ; 5)MHMM 2: each time all 19 persons on training data are used to train single-stream component HMM , and audio-visual likelihoods are combined without stream exponents λ_s in (3). The affect recognition results in our experiment are summarized in Table 1.

Table 1. Average affect recognition rates in our experiment

	Face	Pitch	Energy	MHMM 1	MHMM 2	MHMM 3
Accuracy	0.39	0.60	0.69	0.70	0.72	0.75

Among the six methods mentioned above, face-only HMM gave the poorest performance. The main reason is that speaking influences facial expressions. Especially, subjects seldom display expressive peaks which are main characteristic for pure facial expressions without speaking. Pitch-only and energy-only HMMs performed better than face-only HMM but worse than MHMM because MHMM combine information of face, pitch and energy which provide complementary information for recognition. Among MHMM methods, MHMM1 gave worst performance because it only used the training data of 14 subjects and does not use stream exponents in the fusion stage. MHMM2 has the better recognition rate than MHMM1 because MHMM2 use more data for training. MHMM3 has the best performance because it uses stream exponents estimated on training data of 5 subjects. That suggests that with the limited training data, it is better to divide the training data for estimating single-stream component HMMs and for estimating stream reliability respectively in decision fusion than using all training data for single-stream component HMMs like MHMM2. It is important in the case where single-stream component HMM performances differ largely.

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. To accomplish this end, this paper applies an audio-visual fusion method for person-independent affect recognition.

We investigate the use of single modality confidence measures as a means of estimating weights when combining likelihoods in the audio-visual decision fusion. Person-independent experimental results from 20 subjects in 660 sequences show that the use of stream exponents estimated on training data results in affect classification accuracy improvement in audio-visual affect recognition.

Multimodal recognition of human affective states is a largely unexplored and challenging problem. The elicited nature of the affects performed in our database has the potential to differ from corresponding performances in natural settings. The next stage in the evaluation of this algorithm will be attempting to detect these affect states in human interactions where the states are performed naturally.

Acknowledgement

We like to thank Dr. Lawrence Chen for collecting the valuable data in this paper for audio-visual affect recognition. This work was supported by National Science Foundation (Information Technology Research Grant# 0085980) and Beckman postdoctoral fellowship funding.

References

1. 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
2. Chen, L. and Huang, T. S., Emotional expressions in audiovisual human computer interaction, Int. Conf. on Multimedia & Expo 2000, 423-426
3. 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
4. De Silva, L. C., and Ng, P. C., Bimodal emotion recognition, Int. Conf. on Automatic Face & Gesture Recognition 2000, 332-335
5. Picard, R.W., Affective Computing, MIT Press, Cambridge, 1997.
6. 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
7. 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
8. Kwon, O.W., Chan, K., Hao, J., Lee, T.W, Emotion Recognition by Speech Signals, EUROSPEECH 2003.
9. Mehrabian, A., Communication without words, Psychol. Today, vol.2, no.4, 53-56, 1968
10. Jain, A.J., Duin, R.P.W., and Mao, J., Statistical Pattern Recognition: A Review. IEEE PAMI, 4-37, Vol.22, No.1, January, 2000.
11. Bourlard, H. and Dupont, S., A New ASR Approach Based on Independent Processing And Recombination of Partial Frequency Bands, 1996
12. G. Potamianos, C. Neti, G. Gravier, and A. Garg, Automatic Recognition of audio-visual speech: Recent progress and challenges, Proceedings of the IEEE, vol. 91, no. 9, Sep. 2003
13. Chen, L.S, Joint Processing of Audio-Visual Informa-tion for the Recognition of Emotional Expressions in Human-Computer Interaction, PhD thesis, UIUC, 2000
14. Zeng, Z., Tu, J., Pianfetti, B., Liu, M., Zhang, T., Zhang, Z., Huang, T.S., Levinson, S., Audio-visual Affect Recognition through Multi-stream Fused HMM for HCI, CVPR 2005
15. 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.
16. 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