

Audio-visual Affective Expression Recognition through Multi-stream Fused HMM

Zhihong Zeng, *Member, IEEE*, Jilin Tu, Brian Pianfetti, Thomas S. Huang, *Fellow, IEEE*

Abstract—Advances in computer processing power and emerging algorithms are allowing new ways of envisioning Human Computer Interaction. Although the benefit of audio-visual fusion is expected for affect recognition from the psychological and engineering perspectives, most of existing approaches to automatic human affect analysis are uni-modal: information processed by computer system is limited to either face images or the speech signals. This paper focuses on the development of a computing algorithm that uses both audio and visual sensors to detect and track a user's affective state to aid computer decision making. Using our Multi-stream Fused Hidden Markov Model (MFHMM), we analyzed coupled audio and visual streams to detect 4 cognitive states (interest, boredom, frustration and puzzlement) and 7 prototypical emotions (neural, happiness, sadness, anger, disgust, fear and surprise). The MFHMM allows the building of an optimal connection among multiple streams according to the maximum entropy principle and the maximum mutual information criterion. Person-independent experimental results from 20 subjects in 660 sequences show that the MFHMM approach outperforms face-only HMM, pitch-only HMM, energy-only HMM, and independent HMM fusion, under clean and varying audio channel noise condition.

Index Terms—human computing, affective computing, affect recognition, emotion recognition, multimodal fusion, human computer interaction.

I. 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 [28]. One of 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 [23], [26], [7], [8].

Manuscript received May 18, 2007, and revised Oct. 18, 2007. This work was supported in part by Beckman Postdoctoral Fellowship and NSF CCF 04-26627. A short version of this work was published in [45]

Zhihong Zeng, Jilin Tu and Thomas S. Huang are with Beckman Institute, University of Illinois at Urbana-Champaign (UIUC), 405 N Mathews Av., Urbana, 61801, USA. Their emails are {zhzeng,jilintu, huang}@ifp.uiuc.edu.

Brian Pianfetti is with Human Resources Education, College of Education, University of Illinois at Urbana-Champaign, 1310 S. 6th St. MC 708, Champaign, IL 61820, USA (e-mail: bpianfet@uiuc.edu).

Fig. 1 illustrates a prototype of such an affect-sensitive, multimodal computer-aided learning system. The system was built during the NSF ITR project titled “Multimodal Human Computer Interaction: Toward a Proactive Computer”¹. In this learning environment, the user explores Lego gear games by interacting with a computer avatar. Multiple sensors are used to detect and track the user's behavioral cues and his or her task. More specifically, a camera is used to record the user's facial expressions, a set of cameras are used to track eye movements, another camera is used to monitor the progress of the task, and a microphone is used to record the speech signals employed subsequently to recognize the speech and analyze prosody. Multisensory information is then processed and visualized, including the user's emotional state, engagement state, the utilized speech keywords, and the gear state. Based on this information, the avatar offers an appropriate tutoring strategy in this interactive learning environment.

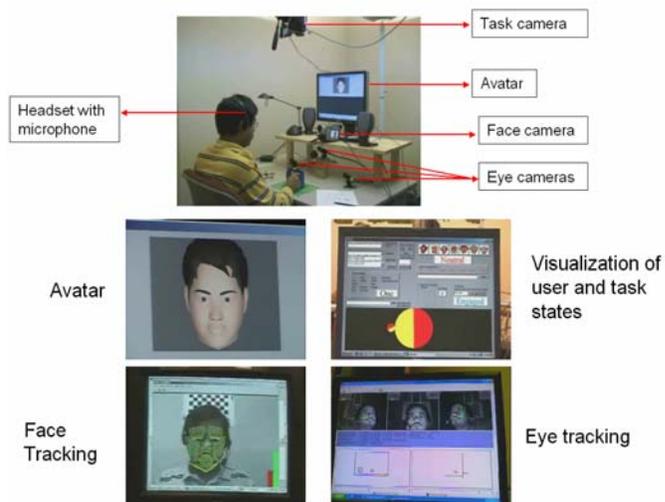


Fig. 1. A prototype of multimodal computer-aided learning system

The psychological studies [19], [1] indicated that judging someone's affective states, people mainly rely on facial expressions and vocal intonations. Another motivation for audio-visual fusion is the fair engineering prospect of improved reliability. More specifically, current techniques for detection and tracking of facial expressions are sensitive to

¹ <http://itr.beckman.uiuc.edu>

head pose, clutter, and variations in lighting conditions, while current techniques for speech processing are sensitive to auditory noise. Due to complementarity and redundancy of the data coming from the two channels, audiovisual human affect recognition is expected to perform more robustly than unimodal methods. Thus, affect recognition should inherently be the issue of multimodal analysis.

In this paper, we present our efforts toward audio-visual affect recognition. With an HCI application in mind, we expand the number of affective states researched beyond seven prototypical (basic) emotions (neural, happiness, sadness, anger, disgust, fear and surprise) to include four cognitive/motivational states (interest, boredom, frustration and puzzlement). These cognitive/motivational states can give us insight into the progress, strategies, and engagement associated during the course of the interaction. Recognizing these states, the computer is able to proactively apply appropriate tutoring strategies (e.g., encouragement, transition, guidance, and confirmation). Due to the fact that it is quite difficult to obtain sufficient audio-visual spontaneous affect expression material of these fine-grained affective states, we test our algorithm on required displayed affect data from 20 non-actor people. Although the subjects displayed affect expression on request, no instruction was given as to how to perform these affective states. The performed expressions are based on the subject's individual perception of these affective states. Thus, studying these required affect expressions might be an indirect way of studying "real" affect communication.

Although the benefit of audio-visual fusion for affect recognition is expected from engineering and psychological perspectives, our knowledge of how humans achieve this fusion is extremely limited. The neurological studies on fusion of sensory neurons (e.g., [32]) seem to more support early fusion (i.e., feature-level fusion) than late fusion (i.e., decision-level fusion). However, it is an open issue how to construct suitable joint feature vectors composed of features from different modalities with different time scales, different metric levels and different dynamic structures. Due to these difficulties, most researchers choose decision-level fusion that simplifies the fusion problem by assuming the independence among different modalities under certain affective state (e.g., [3], [14], [15], [41], [44]). As a result, decision-level fusion loses the information correlation among multiple modalities. Model-level fusion or hybrid fusion that combines the benefits of both feature-level and decision-level fusion methods may be the best choice for this fusion problem.

In this paper, we present a model-level fusion method named the multi-stream fused hidden Markov model (MFHMM) for audio-visual affect recognition. MFHMM is a generalization of two-stream fused HMM [21] that was originally used for audio-visual speaker verification problem. We provide the generalized formula and implementation of our MFHMM so that the algorithm can handle the recognition problem with more than two feature streams. We discuss some problems different from two-stream case [21] and propose our

solutions, in particular, computation of coupling probability among multiple streams. The success in our audio-visual affect recognition experiment validates our MFHMM that can be extended to other multimodal research fields.

The advantages of MFHMM include: 1) Every feature could be modeled by one component HMM which has optimal connection among other streams according to the maximum entropy principle and a maximum mutual information criterion [21][20]; 2) State transitions of different component HMMs do not necessarily occur at the same time across different streams so that the synchrony constraint among different streams can be relaxed; 3) If one component HMM fails due to some reason, the other HMM can still work. Thus, the final fusion performance will be robust; 4) It achieves a better balance between model complexity and performance than other existing HMM-based model fusion methods, like the coupled HMM [2] and mixed-memory HMM [31].

Our audio-visual affect recognition approach was tested in 660 sequences based on 20 subjects with 11 HCI-related affect states. The person-independent experimental results show that the MFHMM approach outperformed face-only HMM, pitch-only HMM, energy-only HMM, and independent HMM fusion that assumes independence among tightly coupled streams, under clean and varying audio channel noise condition.

This paper is organized as follows. Section II describes related work in affect recognition field. Section III introduces the multimedia emotion databases for our experiment. Section IV presents the computing methods, including facial and audio feature extraction, and audio-visual fusion. Section V presents our experimental results. A summary and closing remarks conclude this paper.

II. RELATED WORK

Automatic affect recognition has attracted more and more attention of researchers in multiple disciplines (e.g., psychology, computer science, linguistics, neuroscience) with a variety of motivations in emotion-related research [11] as well as human computer interaction [23], [26], [7], [8], [24].

Most of the existing efforts studied the expressions of the six basic (prototypical) emotions due to their universal property [10], their marked reference representation in our affective lives, and the availability of the relevant training and test material (e.g., [16]).

In addition, most of them focus on the uni-modal approaches: information processed by the computer system is limited to either face images or the speech signals. For exhaustive surveys of these efforts in the field, readers are referred to the following articles:

- overviews of early work on facial expression analysis: Samal & Iyengar, 1992, [30], and Pantic & Rothkrantz, 2000, [22],
- surveys of techniques for automatic facial muscle action recognition and facial expression analysis: Tian et al., 2005, [36], and Pantic & Bartlett, 2007, [27], and

- overviews of multimodal affect recognition methods: Cowie et al., 2001, [8], Pantic & Rothkrantz, 2003, [23], Pantic et al., 2005, [25], and Zeng et al. [24]

In the comprehensive survey written by Pantic & Rothkrantz in 2003, [23], only four studies [5], [6], [9], [39] were found that were focused on audiovisual affect recognition. Since then, an increasing number of efforts are reported toward this direction.

Three fusion strategies (feature-level, decision-level and model-level fusions) are found to be used in the audio-visual affect recognition. Typical examples of feature-level fusion are studies [3], [42] which concatenated the prosodic features and facial features to construct joint feature vectors that are then used to build an affect recognizer. However, the different time scale and metric level of features from different modalities and increasing feature dimension influence the performance of the feature-level fusion. Most of the bimodal affect recognition studies applied decision-level fusion [3], [14], [15], [38], [40], [41], [43], [44], which independently model audio-only and visual-only expressions, then combines these uni-modal recognition results at the end. Since humans display audio and visual expressions in a complementary and redundant manner, the conditional independent assumption of decision-level fusion actually loses the correlation information between audio and visual signals.

Some model-level fusion methods are introduced that can make use of the correlation between audio and visual streams, and relax the requirement of synchronization of these streams. Song et al. [33] presented an extended coupled HMM named tripled HMM to model correlation properties of upper face, lower face and prosodic dynamic behaviors. Fragopanagos & Taylor [12] proposed an artificial neural network with a feedback loop called ANNA to integrate the information from face, prosody and lexical content. Caridakis et al. [13] investigated combining face and prosody expressions by using Relevant Neural Networks.

Most of existing methods are based on deliberately posed emotion expressions, a few exceptional studies are reported toward audio-visual spontaneous emotion recognition. These studies are those of Zeng et al., [43] who used the data collected in psychological research interview (Adult Attachment Interview), and of Fragopanagos and Taylor [12], and Caridakis et al. [13], who used the data collected in Wizard of OZ scenarios. Because their data were not sufficient to build classifiers for fine-grained affective states, they chose to recognize coarse affective states (positive and negative in [43]), or quadrants in evaluation-activation space [13], [12].

The contribution of this paper includes:

First, different from most of existing studies of affect recognition, we explore to detect 4 non-basic emotion (interest, puzzlement, frustration and boredom) as well as basic emotions, considering the importance of these non-basic states in human computer interaction.

Second, we present a general model-level fusion framework named MFHMM to integrate information from multiple

streams according to the maximum entropy principle and the maximum mutual information criterion. We extend the original two-stream fused HMM [21] to the multiple-stream processing, and accordingly propose some general formulas and implementation. The fused HMM have been proved in [21] to have a better balance between model complexity and performance than other existing HMM-based model fusion methods, like the coupled HMM [2] and mixed-memory HMM [31]. Our extensive experimental results of audio-visual affect recognition validate the efficiency and robustness of the MFHMM that can be extended to other multimodal fusion problems.

III. DATABASE

The datasets used in most of existing studies 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 [4]. This database consists of 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 $3 \times 20 = 60$ video sequences. And there are totally $60 \times 11 = 660$ sequences for 11 affective states. The time of every sequence ranged from 2-6 seconds. Table 1 is the statistics of affect expressions in our experimental data.

During labeling, start and end points of each emotion expression were determined by speech energy which is easy to detect. Once these audio segments were defined, corresponding points of facial expressions were labeled.

Table 1: the statistics of affect expressions in our experimental data

Subject	Affective states	Expressions per state per person	Total sequences
20	11	3	660

IV. OUR APPROACH

A. Facial Feature Extraction

A tracking algorithm called Piecewise Bezier Volume Deformation (PBVD) tracking [35] 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 [37] 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 Fig. 2. The outputs of the face tracker corresponding to 12 MUs are used as facial features for later affect recognition in our experiment.

We notice that the movements of facial features are related to both affective states and content of speech. Thus, smooth facial features 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 [44].

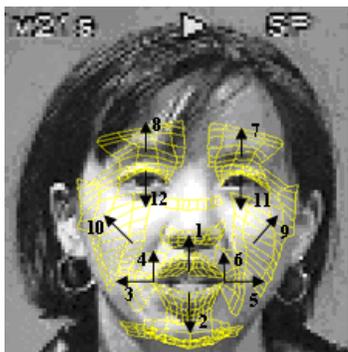


Fig. 2. 12 facial Motion Units

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 magnitude of each MU at every frame was normalized by the corresponding feature mean 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).

B. 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 [34]. 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 [17], [44] 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.

C. Multi-stream Fused HMM (MFHMM)

For integrating coupled audio and visual features, we propose multi-stream fused HMM (MFHMM) which constructs a new structure linking the multiple component HMMs which is optimal according to the maximum entropy principle and maximum mutual information (MMI) criterion. MFHMM is a generalization of two-stream fused HMM introduced by Pan et al. [21] so that MFHMM can be used for the recognition problem with more than two feature streams. For some related theory of fused HMM, the readers are referred to the studies by Pan et al. [20], [21]. In the following paragraphs, we give the multi-stream formulas and algorithms, and discuss the related issues.

1) Formulas

Consider n tightly coupled time series $\{O^{(i)}, i = 1, \dots, n\}$. Assume that the series $\{O^{(i)}, i = 1, \dots, n\}$ can be modeled respectively by n HMMs with hidden states $\{U^{(i)}, i = 1, \dots, n\}$. In the multi-stream fused HMM framework, an optimal solution for $p(O^{(1)}; O^{(2)}; \dots; O^{(n)})$ according to the maximum entropy principle [18], [20] is given by

$$\begin{aligned} & \hat{p}(O^{(1)}; O^{(2)}; \dots; O^{(n)}) \\ &= p(O^{(1)})p(O^{(2)}) \dots p(O^{(n)}) \\ & \bullet \frac{p(v^{(1)}, v^{(2)}, \dots, v^{(n)})}{p(v^{(1)})p(v^{(2)}) \dots p(v^{(n)})} \end{aligned} \quad (1)$$

The last term in the equation can be viewed as an enhancement/suppression factor, which absorbs some dependence among $\{O^{(i)}, i = 1, \dots, n\}$. The transforms

$$v^{(i)} = g_i(O^{(i)}) \quad i = 1, 2, \dots, n$$

were introduced so that $p(v^{(1)}, v^{(2)}, \dots, v^{(n)})$ can more easily be calculated than $p(O^{(1)}; O^{(2)}; \dots; O^{(n)})$ and it reflects the statistical dependence among $\{O^{(i)}, i = 1, \dots, n\}$.

According to the maximum mutual information (MMI) criterion, [21] proposed to fuse component HMMs together by connecting the hidden states of one HMM to the observation of another HMM. Thus n sets of transforms can be invoked

with the i -th ($i = 1, 2, \dots, n$) set of transform being

$$v^{(i)} = \begin{cases} U^{(j)} & j = i \\ O^{(j)} & j \neq i \end{cases} \quad (j = 1, 2, \dots, n)$$

The i -th corresponding fusion model defined by Equation (1) yields

$$\begin{aligned} & \hat{p}^{(i)}(O^{(1)}; O^{(2)}; \dots; O^{(n)}) \\ &= p(O^{(1)})p(O^{(2)}) \dots p(O^{(n)}) \\ & \bullet \frac{p(U^{(i)}, O^{(1)}, \dots, O^{(i-1)}, O^{(i+1)}, \dots, O^{(n)})}{p(U^{(i)})p(O^{(1)}) \dots p(O^{(i-1)})p(O^{(i+1)}) \dots p(O^{(n)})} \\ &= p(O^{(i)})p(O^{(1)}, \dots, O^{(i-1)}, O^{(i+1)}, \dots, O^{(n)} | U^{(i)}) \end{aligned}$$

In order to simplify the last term (coupling probability) in the above computation, we use the following conditional independent assumption

$$\begin{aligned} & p(O^{(1)}, \dots, O^{(i-1)}, O^{(i+1)}, \dots, O^{(n)} | U^{(i)}) \\ &= \prod_{j \neq i, j=1}^n p(O^{(j)} | U^{(i)}) \end{aligned}$$

Although this assumption is usually violated in practice, it has a good record in pattern recognition. The reason of the success of this assumption is attributed to the small number of parameters to be estimated. Some complicated algorithms without this assumption are more sensitive to data noise and susceptible to the local maximum.

Thus, the estimate of $p(O^{(1)}; O^{(2)}; \dots; O^{(n)})$ is given by

$$\begin{aligned} & \hat{p}^{(i)}(O^{(1)}; O^{(2)}; \dots; O^{(n)}) \\ &= p(O^{(i)}) \prod_{j \neq i, j=1}^n p(O^{(j)} | U^{(i)}) \end{aligned} \quad (2)$$

The structures defined by Equation (2) are different for different i . Equation (2) emphasizes the dependencies between hidden states $U^{(i)}$ and observations $\{O^{(j)}, j = 1, \dots, n, j \neq i\}$, which requires that $U^{(i)}$ be reliably estimated. In practice, if the n component HMMs have different reliabilities, they may be combined by different weights $\lambda^{(i)}$ ($i = 1, 2, \dots, n$) to get a better result:

$$\begin{aligned} & \hat{p}(O^{(1)}; O^{(2)}; \dots; O^{(n)}) \\ &= \sum_{i=1}^n \lambda^{(i)} \hat{p}^{(i)}(O^{(1)}; O^{(2)}; \dots; O^{(n)}) \end{aligned} \quad (3)$$

where

$$\sum_{i=1}^n \lambda^{(i)} = 1. \quad (4)$$

The weights could be proportional to the reliabilities of the component HMMs. The more reliable one component HMM is, the larger its weight.

2) Learning algorithm

The learning algorithm of the n -stream fused HMM includes three main steps.

- 1) n component HMMs are trained individually by the EM algorithm.

- 2) The best hidden state sequences of the component HMMs are estimated using the Viterbi algorithm
- 3) The coupling parameters among the n HMMs are estimated.

In step 1, the model parameters (the initial, transition, and observation probabilities) of individual HMMs are estimated. And step 2 infers hidden states $U^{(i)}$ ($i = 1, 2, \dots, n$). The details of the EM and the Viterbi algorithms used for solving the above problems can be found in [29].

In step 3, the coupling parameters between the n HMMs are determined as follows:

$$\begin{aligned} B^{(i,j)} &= p(O^{(j)} | U^{(i)}) \\ i, j &= 1, 2, \dots, n, i \neq j \end{aligned}$$

3) Inference algorithm

In our application, inference is the process of computing $\hat{p}(O^{(1)}; O^{(2)}; \dots; O^{(n)})$ in (3) given observation sequence $\{O^{(i)}, i = 1, \dots, n\}$ and the model parameters corresponding to each affective state. And the affective state with maximum of $\hat{p}(O^{(1)}; O^{(2)}; \dots; O^{(n)})$ is regarded as the recognition result.

According to (3), we first compute individually according to (2)

$$\begin{aligned} & \hat{p}^{(i)}(O^{(1)}; O^{(2)}; \dots, O^{(n)}) \\ & \quad i = 1, 2, \dots, n \end{aligned}$$

Then their results are combined according to (3) to get

$$\hat{p}(O^{(1)}; O^{(2)}; \dots; O^{(n)})$$

The individual inference algorithm of (2) is derived from the forward inference procedure of traditional HMM [29]. The only difference of our algorithm is that multiple stream observations instead of one-stream observation at time instants are taken into account. In the other words, observation probability

$$p(O^{(i)} | U^{(i)}) \quad (i = 1, 2, \dots, n)$$

in the forward-backward procedure [29] is replaced by

$$\prod_{j=1}^n p(O^{(j)} | U^{(i)}) \quad (i = 1, 2, \dots, n)$$

V. EXPERIMENTAL RESULTS

We tested the MFHMM algorithm in our person-independent affect recognition problem by using leave-one-out cross validation scheme. 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. The test is repeated 20 times, each time leaving a different person out.

In our experiment, the composite facial feature from visual channel, energy and pitch features from audio channel are treated as three tightly coupled streams ($O^{(1)}, O^{(2)}, O^{(3)}$), and modeled by three component HMMs with 12 hidden states. We used the following five methods to make decisions and compared their recognition results:

- 1) face-only HMM;

- 2) pitch-only HMM;
- 3) energy-only HMM;
- 4) independent-HMM (IHMM): assuming $O^{(1)}, O^{(2)}$ and $O^{(3)}$ are independent, it combines the component HMMs by computing:

$$\hat{p}(O^{(1)}; O^{(2)}; O^{(3)}) = p(O^{(1)})p(O^{(2)})p(O^{(3)})$$

5) MFHMM: considering statistical dependence among $O^{(1)}, O^{(2)}$ and $O^{(3)}$, it combines the component HMMs by computing (3). Considering the influence of weights on the performance, we compute the accuracies with different weights. The results are shown in Fig. 3 where the x axis and y axis represent the visual weight and pitch weight individually, and z represents accuracies. The energy weight is automatically fixed after the visual and pitch weights are chosen according to [4]. Fig. 3 shows the smooth change of the performance with weights. And the maximum accuracy (red dot in Fig. 3, 80.61%,) of MFHMM is with 0.2 visual weight, 0.4 pitch weight and 0.4 energy weight, roughly proportional to the reliabilities of corresponding component FHMMs.

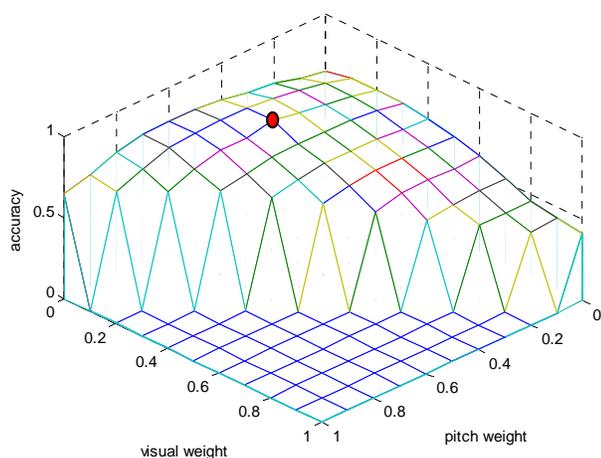


Fig. 3. The MFHMM performance with various weights

In order to make our experiment comparable with the previous basic emotion recognition reports, we also did recognition of 7-state basic emotions besides 11 affective states. The affect recognition results (average accuracies and stand deviations) in our experiment are shown in Fig. 4 for 11-state affect recognition and Fig. 5 for 7-state basic emotion recognition.

Among the five methods mentioned above, face-only HMM gave the poorest performance. The main reason is that speaking decrease the discriminability of facial expressions. Especially, subjects seldom display expressive peaks that are main characteristic for pure facial expressions without speaking. Pitch-only and energy-only HMMs performed better than face-only HMM but worse than IHMM and MFHMM because both of IHMM and MFHMM combine information of face, pitch and energy which provide complementary information for recognition. IHMM gave worse performance

than MFHMM because it assumes independence among $O^{(1)}, O^{(2)}$ and $O^{(3)}$. The performance of MFHMM is best on recognition rate, and the time of its training and inference is only a little more than IHMM.

The more details of comparison of the five methods of 11-state affect recognition are presented in Table 2 that lists the recognition rate of each affect using each method. In face-only HMM recognition results in Table 2, the 3 cognitive states (frustration, interest and boredom) has the lowest recognition rates. That suggests that it is difficult to judge these subtle cognitive states if only using information of facial expression. On the other hand, the audio expressions (pitch and energy) of these states provide complementary information that results to increase of discriminability from other states. The confusion matrices of the MFHMM method for 11 affect recognition are presented in Table 3.

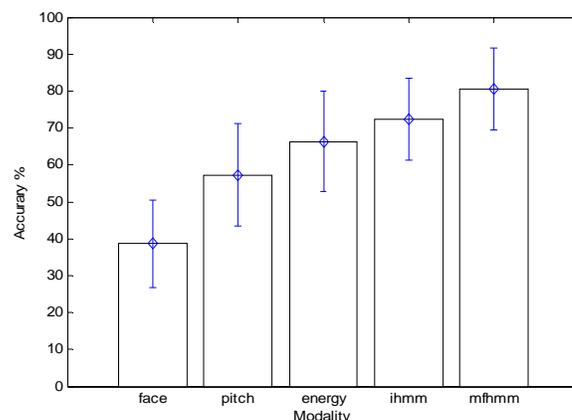


Fig. 4: Average accuracies and stand deviations of recognition of 11 affective states (7 basic emotion and 4 cognitive states)

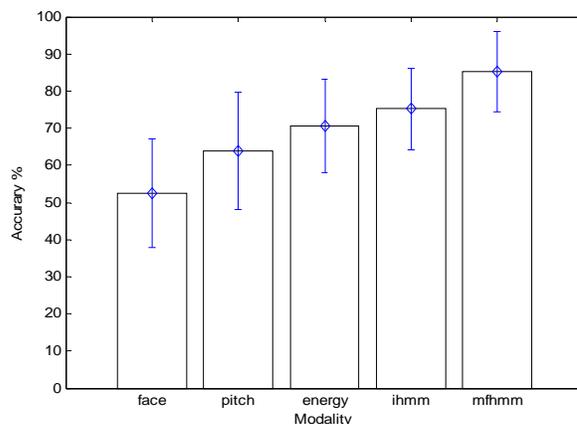


Fig. 5. average accuracies and stand deviations of recognition of 7 basic emotions

To test the performance of our algorithm under varying stream reliability conditions, we artificially add additive white noise to the database audio, at various levels of signal-to-noise ratio (SNR). Such noise is added to the test set of the database only, thus creating a mismatch to the audio-only HMMs, which are trained on the original clean database audio. Thus the performance of visual-only HMM keeps constant in the

experiments.

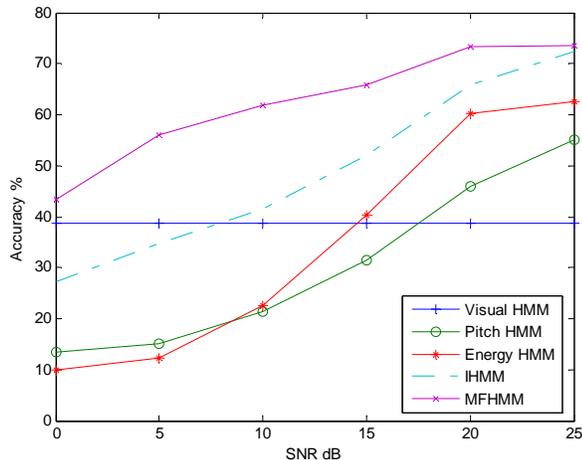


Fig. 6. Accuracies of different methods under various audio SNR conditions

The results under various audio SNR conditions are shown in Fig. 6. They demonstrate that audio-visual fusion outperforms uni-stream methods at most cases, i.e. both of IHMM and MFHMM are better than visual HMM, pitch HMM and energy HMM. The exceptions are that the accuracies of IHMM in 0 and 5dB audio SNR are lower than visual HMM. That shows that the IHMM combination scheme cannot achieve better performance than individual modality when the performance of certain individual streams is very bad. On the other hand, the performance of MFHMM is still little higher than the visual-only HMM in 0dB audio SNR. Thus, MFHMM is more robust to process noisy data than IHMM

VI. 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 where attending to changes in the user's affective states leads to a high level of engagement and knowledge acquisition in a computer-aided learning system. To accomplish this end, we investigate audio-visual affect recognition in this paper. Specifically, we explore to detect 4 cognitive/motivational states as well as 7 basic emotions due to the importance of these cognitive/motivational states in HCI.

For integrating tightly coupled audio-visual streams, we present the multi-stream fused HMM that is able to build optimal connection among multiple streams according to the maximum entropy principle and the maximum mutual information criterion. Experimental results from analyzing 11 affect states of 20 subjects suggests that the MFHMM outperformed face-only HMM, pitch-only HMM, energy-only HMM, and independent HMM fusion which assumes independence among tightly coupled streams, under various

audio-channel noise condition. The MFHMM is a general fusion method, and can be extended to other multimodal research fields.

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.

REFERENCES

- [1] Ambady, N., Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, Vol. 111, No. 2, 256-274
- [2] Brand, M. and Oliver, N.(1997), Coupled hidden Markov models for complex action recognition, In *Proc. Computer Vision Pattern Recognition*, 201-206
- [3] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S. (2004), Analysis of Emotion Recognition Using Facial Expressions Speech and Multimodal Information, *Int. Conf. Multimodal Interfaces*. 205-211
- [4] Chen, L.S (2000), Joint Processing of Audio-Visual Information for the Recognition of Emotional Expressions in Human-Computer Interaction, PhD thesis, UIUC
- [5] Chen, L. and Huang, T. S. (2000), Emotional expressions in audiovisual human computer interaction, *Int. Conf. on Multimedia & Expo 2000*, 423-426
- [6] Chen, L., Huang, T. S., Miyasato, T., and Nakatsu, R. (1998), Multimodal human emotion/expression recognition, *Int. Conf. on Automatic Face & Gesture Recognition*, 396-401
- [7] Cohn, J.F. (2006), *Foundations of Human Computing: Facial Expression and Emotion*, *Int. Conf. on Multimodal Interfaces*, 233-238
- [8] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J.G. (2001), Emotion Recognition in Human-Computer Interaction, *IEEE Signal Processing Magazine*, January, 32-80
- [9] De Silva, L. C., and Ng, P. C.(2000), Bimodal emotion recognition, *Int. Conf. on Automatic Face & Gesture Recognition 2000*, 332-335
- [10] Ekman P. (1972). Universals and cultural differences in facial expressions of emotion. *Nebr. Symp. Motiv.* 1971, 207-283
- [11] 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
- [12] Fragopanagos, F. and Taylor, J.G. (2005), Emotion recognition in human-computer interaction, *Neural Networks*, 18: 389-405
- [13] Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Paouzaoui, A. and Karpouzis, K. (2006). Modeling Naturalistic Affective States via Facial and Vocal Expression Recognition. *Int. Conf. on Multimodal Interfaces*. 146-154
- [14] 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
- [15] Hoch, S., Althoff, F., McGlaun, G., Rigoll, G. (2005), Bimodal fusion of emotional data in an automotive environment, *ICASSP*, Vol. II, 1085-1088, 2005
- [16] 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
- [17] Kwon, O.W., Chan, K., Hao, J., Lee, T.W (2003), Emotion Recognition by Speech Signals, *EUROSPEECH 2003*.
- [18] Luttrell, S.P. (1989). The use of Bayesian and entropic methods in neural network theory. In: *Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers, Boston, 363-370
- [19] Mehrabian (1968), A., *Communication without words*, *Psychol. Today*, vol.2, no.4, 53-56, 1968
- [20] Pan, H., Liang, Z.P. and Huang, T.S. (2001). Estimation of the joint probability of multisensory signals, *Pattern Recognition Letter*, 22: 1432-1437
- [21] Pan, H., Levinson S., Huang, T.S., and Liang, Z.P. (2004), A fused Hidden Markov Model With Application to Bimodal Speech Processing, *IEEE Transaction on Signal Processing*, Vol.52, No.3, 573-581, March

[22] Pantic M and Rothkrantz L J M (2000). Automatic analysis of facial expressions—the state of the art. IEEE PAMI, Vol.22, No.12, 1424-1445

[23] Pantic M., Rothkrantz, L.J.M. (2003), Toward an affect-sensitive multimodal human-computer interaction, Proceedings of the IEEE, Vol. 91, No. 9, Sept., 1370-1390

[24] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2007). A survey of affect recognition methods: audio, visual, and spontaneous expressions. Int. Conf. on Multimodal Interfaces. In press

[25] Pantic, M., Sebe, N., Cohn, J.F. and Huang, T. (2005), Affective Multimodal Human-Computer Interaction, in Proc. ACM Int'l Conf. on Multimedia, 669-676

[26] Pantic, M., Pentland, A., Nijholt, A., and Huang, T.S. (2006), Human Computing and Machine Understanding of Human Behavior: A Survey, Int. Conf. on Multimodal Interfaces, 239-248

[27] Pantic M and Bartlett M S (2007). Machine analysis of facial expressions. In Face Recognition, Kurihara K, Ed. Vienna, Austria: Advanced Robotics System, 327-366

[28] Picard, R.W. (1997), Affective Computing, MIT Press, Cambridge, 1997.

[29] Rabiner, L.R. (1989), A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of IEEE, Vol.77, No.2, February

[30] Samal A and Iyengar P A (1992). Automatic recognition and analysis of human faces and facial expressions: A survey. Pattern Recognition, Vol. 25, No.1, 65-77

[31] Saul, L.k. and Jordan, M.I. (1999), Mixed memory Markov model: Decomposing complex stochastic processes as mixture of simpler ones, Machine Learning, Vol.37, 75-88, Oct. 1999

[32] Stein, B., Meredith, M.A. (1993). The Merging of Senses. MIT Press, Cambridge, USA

[33] Song, M., Bu, J., Chen, C., and Li, N (2004)., Audio-visual based emotion recognition—A new approach, Int. Conf. Computer Vision and Pattern Recognition, 1020-1025

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

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

[36] Tian Y L, Kanade T and Cohn J F (2005). Facial expression analysis. In: Handbook of Face Recognition, Li S Z and Jain A K (Eds.), Springer, New York, USA, 247-276

[37] Tu, J., Zhang, Z., Zeng, Z. and Huang, T.S. (2004), Face Localization via Hierarchical Condensation with Fisher Boosting Feature Selection, In Proc. Computer Vision and Pattern Recognition, 719-724.

[38] Wang, Y. and Guan, L. (2005), Recognizing human emotion from audiovisual information, ICASSP, Vol. II, 1125-1128

[39] Yoshitomi, Y., Kim, S., Kawano, T., and Kitazoe, T. (2000), Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in Proc. ROMAN 2000, 178-183

[40] Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T.S., Roth, D., and Levinson, S. (2004), Bimodal HCI-related Affect Recognition, Int. Conf. on Multimodal Interfaces, 137-143

[41] 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

[42] 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.

[43] Zeng Z, Hu Y, Roisman G I, Wen Z, Fu Y and Huang T S (2006): Audio-visual emotion recognition in adult attachment interview. Int. Conf. Multimodal Interfaces: 139-145

[44] Zeng, Z., Tu, J., Liu, M., Huang, T.S., Pianfetti, B., Roth D. and Levinson, S. (2007), Audio-visual Affect Recognition, IEEE Transactions on Multimedia, Vol. 9, No.2, February

[45] Zeng, Z., Tu, J., Pianfetti, B., Liu, M., Zhang, T., Zhang, Z., Huang, T.S., and Levinson, S. (2005), Audio-visual Emotion Recognition through Multi-stream Fused HMM for HCI Applications, the Conf. on Computer Vision and Pattern Recognition (CVPR), (2) 967-972.

Table 2. Accuracy comparison of face-only HMM, pitch-only HMM, energy-only HMM, IHMM and MFHMM for 11-state affect recognition

	neutral	happy	sad	angry	disgust	surprise	fear	frustrated	puzzle	interest	bore
face	0.45	0.40	0.32	0.58	0.43	0.43	0.52	0.18	0.42	0.25	0.27
pitch	0.92	0.42	0.55	0.63	0.68	0.42	0.62	0.35	0.72	0.62	0.38
energy	0.92	0.53	0.33	0.67	0.85	0.60	0.72	0.63	0.78	0.58	0.68
IHMM	0.97	0.53	0.67	0.77	0.90	0.57	0.72	0.58	0.82	0.82	0.63
MFHMM	0.98	0.70	0.68	0.82	0.88	0.78	0.78	0.75	0.85	0.85	0.78

Table 3. MFHMM confusion matrix for 11-state affect recognition

MFHMM		Detected										
		neut	happ	sad	ang	dis	surp	fear	frus	puzz	inter	bore
Desired	neutral	0.98	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	happy	0.02	0.70	0.02	0.03	0.07	0.07	0.05	0.00	0.00	0.00	0.05
	sad	0.07	0.00	0.68	0.00	0.00	0.02	0.00	0.03	0.07	0.02	0.12
	angry	0.00	0.00	0.00	0.82	0.05	0.00	0.08	0.00	0.03	0.02	0.00
	disgust	0.00	0.00	0.00	0.00	0.88	0.00	0.10	0.00	0.00	0.00	0.02
	surprise	0.00	0.00	0.02	0.00	0.02	0.78	0.03	0.03	0.08	0.02	0.02
	fear	0.00	0.02	0.02	0.05	0.00	0.03	0.78	0.02	0.03	0.03	0.02
	frustrated	0.02	0.00	0.02	0.00	0.02	0.02	0.02	0.75	0.02	0.13	0.02
	puzzle	0.02	0.00	0.02	0.00	0.00	0.10	0.00	0.02	0.85	0.00	0.00
	interest	0.00	0.00	0.02	0.00	0.00	0.00	0.05	0.07	0.02	0.85	0.00
	bore	0.03	0.00	0.05	0.00	0.07	0.00	0.03	0.00	0.00	0.03	0.78