Training Combination Strategy of Multi-stream Fused Hidden Markov Model for Audio-visual Affect Recognition

Zhihong Zeng, Yuxiao Hu, Ming Liu, Yun Fu and Thomas S. Huang
University of Illinois at Urbana-Champaign
405 N Mathews Av.
Urbana, IL 61801
{zhzeng, hu3, mingliu1, yunfu2, huang} @ifp.uiuc.edu

ABSTRACT
To simulate the human ability to assess affects, an automatic affect recognition system should make use of multi-sensor information. In the framework of multi-stream fused hidden Markov model (MFHMM), we present a training combination strategy towards audio-visual affect recognition. Different from the weighting combination scheme, our approach is able to use a variety of learning methods to obtain a robust multi-stream fusion result. We evaluate our approach in personal-independent recognition of 11 affective states from 20 subjects. The experimental results suggest that MFHMM outperforms IHMM which assumes the independence among streams, and the training combination strategy has the superiority over the weighting combination under clean and varying audio channel noise condition.

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
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 affects [2]. Perhaps the most fundamental application of affective computing would be 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. In addition, the progress in this area will bring significant impact on the emotion-related studies in education, psychology and psychiatry.

Most of current affect recognition approaches are uni-modal: information for recognition is limited to visual or audio [1]. Multimodal sensory information fusion is a process that enables human ability to assess emotional states robustly and flexibly. The psychological study [4] indicated that people mainly rely on facial expressions and vocal intonations to judge someone's affective states. To simulate the human ability to assess affect, an automatic affect recognition system should also make use of multimodal fusion.

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) [6]. Although there are some audio-visual fusion studies in audio-visual Automatic Speech Recognition (AVASR) literature [6], few studies are found for audio-visual affect recognition [5][11].

Most of current multi-stream combination studies are based on the framework of multi-stream Independent HMM (IHMM) which assumes the independence among different streams. And they 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 [7], stream S/N ratio [7], stream entropy [8], or other reliability measures such as ones in [9].

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 training is not involved.

Recently, multi-stream fused hidden Markov model (MFHMM) was introduced in our work to integrate coupled audio and visual features in multimedia emotion data [5]. 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.; 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MM'06, October 23–27, 2006, Santa Barbara, California, USA.
Copyright 2006 ACM 1-59593-447-2/06/0010...$5.00.

other existing model fusion methods, like the coupled HMM [12] and mixed-memory HMM [13].

Based on the framework of MFHMM, we here present a training combination strategy under which a variety of learning methods can be applied to obtain a robust multi-stream fusion result. Different from the weighting combination whose mechanism is linear, our approach is able to combine fused HMM components in a variety of ways, linear or nonlinear, depending on the learning scheme.

In the MFHMM framework, the different streams are modeled by different fused HMM components which connect the hidden states of one HMM to the observation of other HMMs. Then the probabilities of these fused HMM components are considered as the input of a combination block which outputs the final decision results (target classes). This combination can be considered as a multi-class classification so that various learning methods can be used. We evaluate our framework in personal-independent recognition of 11 affective states from 20 subjects. The experimental results demonstrated that MFHMM outperforms IHMM which assumes the independence among streams, and the proposed training combination strategy has the superiority over the weighting combination under clean and varying audio channel noise condition.

The paper is organized as follows: the next chapter briefly introduces Multi-stream Fused HMM. In Chapter 3, we describe the proposed training strategy to combine Fused HMM components. Chapter 4 describes the database we use to test our algorithms. Chapter 5 is experimental results and Chapter 6 is our conclusion.

2. MULTI-STREAM FUSED HIDDEN MARKOV MODEL (MFHMM)

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

$$\begin{aligned}&\hat{p}(O^{(1)};O^{(2)};\cdots;O^{(n)})\\&= p(O^{(1)})\,p(O^{(2)})\cdots p(O^{(n)})\bullet\frac{p(v^{(1)},v^{(2)},\cdots,v^{(n)})}{p(v^{(1)})\,p(v^{(2)})\cdots p(v^{(n)})}\end{aligned}\qquad(2.1)$$

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

According to the maximum mutual information (MMI) criterion, we can 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 $(i=1,2,\cdots,n)$ set of transform being

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

The *i*-th corresponding fusion model defined by (2.1) yields

$$\begin{aligned}&\hat{p}^{(i)}(O^{(1)};O^{(2)};\cdots;O^{(n)})\\&= p(O^{(1)})\,p(O^{(2)})\cdots p(O^{(n)})\\&\bullet\frac{p(U^{(i)},O^{(1)},\cdots,O^{(i-1)},O^{(i+1)},\cdots,O^{(n)})}{p(U^{(i)})\,p(O^{(1)})\cdots P(O^{(i-1)})P(O^{(i+1)})\cdots p(O^{(n)})}\\&= p(O^{(i)})\,p(O^{(1)},\cdots,O^{(i-1)},O^{(i+1)}\cdots,O^{(n)}\,|\,U^{(i)})\end{aligned}\qquad(2.2)$$

The details of MFHMM, including its learning and inference algorithms, can be found in [5]

3. COMBINATION STRATEGIES

In the MFHMM framework, each stream is modeled by a set of component FHMMs which represent the possible classes. During the recognition process, these component FHMMs are applied to the corresponding streams, and we can obtain a set of probabilities of these component FHMMs. Then all MFHMM outputs are combined to obtain a global decision.

In the original MFHMM in [5], weighting scheme is used which sums the weighed log probabilities of component FHMMs across different streams. And the class with maximum of the summation result is regarded as the final recognition result. It is described as follows

$$\begin{aligned}&\hat{p}(O^{(1)};O^{(2)};\cdots;O^{(n)})\\&=\sum_{i=1}^{n}\lambda^{(i)}\hat{p}^{(i)}(O^{(1)};O^{(2)};\cdots;O^{(n)})\end{aligned}\qquad(3.1)$$

where $\sum_{i=1}^{n}\lambda^{(i)}=1$

The weights $\lambda^{(i)}$ of (3.1) is set proportional to the reliabilities of the component HMMs. The weights could be computed from normalized stream recognition rate [7], stream S/N ratio [7], stream entropy [8], and other reliability measures in [9].

Using the above measures as weights is intuitive and reasonable in some way. But it assumes that the combination is linear. This assumption could be invalid in practice. In addition, the outputs from different modalities are not optimally scaled with respect to each other because they deal with different feature spaces. Even it is possible that the weighting combination is worse than individual component performance, as shown in our experiments.

In this paper, we propose the training combination strategy. According to this strategy, this MFHMM combination can be treated as a multi-class classification problem in which the input is the probabilities of FHMM 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 component FHMMs and n the number of streams, this classification contains $s\times n$ input units and s output units, and the parameters of the classifier can be estimated by training. Under this strategy, a variety of leaning methods can be used to build a combination of FHMM components. In our experiments, we choose two simple training combination methods as follows

1) Gaussian-based combination: The distribution of probabilities of component MFHMMs in each target class is learned by assuming that this distribution is a Gaussian density. The decision boundary between any two classes is where the log of the probabilities ratio is zero. The boundaries can be nonlinear if the covariance matrixes of these Gaussian models do not equal.

2) Fisher's combination: the linear discriminant function of probabilities of component MFHMMs between the classes is learned by Fisher's principle, i.e. the between-class variance is maximized and within-class variance is minimized. This multi-class case is implemented by the one-against-all strategy.

4. DATABASE AND FEATURE EXTRACTION

The 20 subjects (10 female and 10 males) in our database [10] consist of graduate and undergraduate students from different disciplines. This set of videos contains subjects with a wide variability in physiognomy. This database consists of performances of 11 affective states, including 7 basic emotions (happiness, sadness, fear, surprise, anger, disgust, and neutral), and 4 cognitive states (interest, boredom, puzzlement and frustration). 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.

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 MFHHs, which are trained on the original clean database audio.

5. EXPERIMENTS

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. The test is repeated 20 times, each time leaving a different person out (leave-one-out cross-validation).

Audio and visual features are extracted from the database using the algorithms reported in [5]. Briefly, in visual channel, the 3D face tracker was applied to obtain 12 facial features around mouth, eyes, and cheeks. Then smoothed facial features are calculated to reduce the influence of speech on facial expression. Due to the different physiognomy of different subjects, each of these 12 facial features is normalized by the corresponding feature mean of the neutral expressions of the same subject. And then they are quantized into 19-size codebook by vector quantization (VQ) to a composite facial feature. In audio channel, we use pitch and energy due to their importance for affect recognition. Some prosody features, like frequency and duration of silence, could have implication in the HMM structure of energy and pitch. Similarly, the energy and pitch are normalized by the corresponding feature mean of the neutral expression sequences of the same subject, and quantized into 19-size codebook by vector quantization respectively.

Table 1. Performance comparison in clean audio condition among uni-stream HMM and multi-stream HMM, IHMM and MFHMM, and weighting and training combination schemes.

System	Correlation among streams	Combination scheme	Error rate (%)
Visual HMM	N	N/A	61.36
Pitch HMM	N	N/A	39.54
Energy HMM	N	N/A	34.69
Acc-W IHMM	N	Weighting	27.87
Acc-W MFHMM	Y	Weighting	23.06
Gauss IHMM	N	Training	20.45
Gauss MFHMM	Y	Training	16.36
Fisher IHMM	N	Training	21.67
Fisher MFHMM	Y	Training	17.58

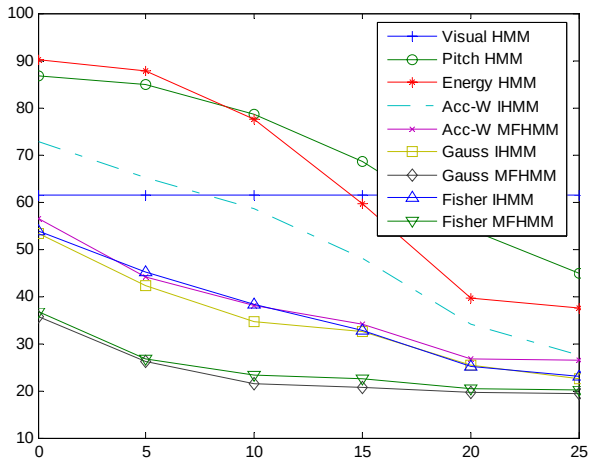

Figure 1. Error rates under various audio SNR present in the audio data

In our experiment, the composite facial feature from video, energy and pitch features from audio are treated as three coupled streams, and modeled by three component HMMs.

We used nine methods to make decisions and compared the recognition results. The performance comparison of these methods in the clear audio condition are shown in Table 1.These methods include uni-stream HMM and multi-stream HMM, IHMM and MFHMM, and weighting and training combination schemes. The uni-stream HMMs are visual-only HMM, pitch-only HMM, energy-only HMM. IHMM is multi-stream HMM which assumes the independence among different stream while MFHMM makes use of the correlation among streams as

described in Section 2. Acc-W means a weighting combination scheme with weights proportional to stream normalized recognition accuracies. Gauss and Fisher mean using training combination schemes based on Gaussian density estimation and Fisher's principle as described in Section 3.

The experimental results under varying level of audio SNR is shown in Figure 1. In Figure 1, 0dB audio SNR means that signal and noise have same energy. In this case, audio still contain non-random information. Thus, the performance of both pitch-only and energy-only modalities is still better than random results.

The results demonstrate that,

1. 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 error rates of Acc-W IHMM in 0 and 5dB audio SNR are higher than visual HMM. That shows that the accuracy-based weighting combination scheme is not good when the performance of some streams is very bad. Thus, although the weighting combination is reasonable and intuitive, it cannot guarantee to obtain a good combination result, especially when the error rates of uni-stream HMMs are low.
2. MFHMM outperforms IHMM. That is because MFHMM takes into account the correlation among different streams while IHMM assumes the independence among streams
3. Training combination outperforms weighting combination, i.e. Gaussian-based and Fisher-based combinations are better than Acc-W combination, in both IHMM and MFHMM systems. By using training process, even some simple training scheme, we are able to find the better combination of HMM components than weighting scheme.

6. CONCLUSION

In this paper, we extend our previous work toward audio-visual affect recognition, and propose to use the training strategy to combine multi-stream FHMMs. The experimental results suggest that MFHMM outperforms IHMM which assumes the independence among streams, and the proposed training combination strategy has the superiority over the weighting combination under clean and varying audio channel noise condition. Worthy of mention, the price of training combination is the additional computational cost resulted from learning. In this paper, we only use two simple training combination schemes, Gaussian and Fisher-based combinations, which demonstrate the noticeable advantage over the weighting combination in the original MFHMM. Under this training combination strategy, we can extend to apply more sophisticated learning methods [3] to improve the recognition performance in the future.

Although we only test our approach on audio-visual affect recognition, our approach is general and can be applied to other multi-stream applications.

7. 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] Picard, R.W., Affective Computing, MIT Press, Cambridge, 1997.

[3] Jain, A.K., Duin, R.P.W. and Mao, J., "Statistical pattern recognition: a review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, 2000, 4-37

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

[5] Zeng, Z., Tu, J., Pianfetti, B., Liu, M., Zhang, T., Zhang, Z., Huang, T.S., Levinson, S.E., Audio-Visual Affect Recognition through Multi-Stream Fused HMM for HCI. CVPR (2) 2005: 967-972

[6] Potamianos, G. , Neti, C. , Gravier, G. , and Garg, A., Automatic Recognition of audio-visual speech: Recent progress and challenges, Proceedings of the IEEE, vol. 91, no. 9, Sep. 2003

[7] Bourlard, H. and Dupont, S., A new ASR approach based on independent processing and recombination of partial frequency bands, ICSLP 1996

[8] Okawa, S., Bocchieri, E. and Potamianos, A., Multi-band Speech Recognition in noisy environments, ICASSP, 1998, 641-644

[9] Garg, A., Potamianos, G., Neti, C. & Huang, T.S., Frame-dependent multi-stream reliability indicators for audio-visual speech recognition, ICASSP, 2003.

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

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

[12] Brand, M. and Oliver, N., Coupled hidden Markov models for complex action recognition, In Proc. Computer Vision Pattern Recognition, 201-206, 1997

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