FACIAL EVENT MINING USING COUPLED HIDDEN MARKOV MODELS

Limin Ma, Qiang Zhou, Mehmet Celenk, and David Chelberg

School of Electrical Engineering and Computer Science
Stocker Center,Ohio University, Athens, OH 45701, USA
{limin.ma,qiang.zhou,celenk,chelberg}@ohio.edu

ABSTRACT

Facial event mining is one of the key techniques for automatic human face analysis. It plays an important role in human computer interaction. This paper proposes a new approach to facial event recognition by combining active shape models (ASMs) and coupled hidden Markov models (CHMMs). Based on the assumption that a complex facial event can be decomposed into multiple coupled processes, ASMs are used to track global facial features and to decouple pattern attributes for upper and lower faces separately. These two interacting processes are modeled as a CHMM for training and recognition. Four basic facial events are investigated. Preliminary experiments yield consistent results, that show the significant advantage of CHMMs over conventional HMMs for facial event mining in video.

1. INTRODUCTION

The study of human computer interaction (HCI) deals with the design, evaluation, and implementation of interactive computer systems for human use. The ideal model for HCI is human face-to-face communication. As one form of visual communication, facial event mining mainly involves the process of detecting and classifying facial gestures and expressions. This has found many applications in areas such as face recognition and facial expression analysis, and has received considerable attention in recent years.

In general, automatic facial expression analysis is a complex task, which includes face localization, facial feature tracking, pattern generation and classification. A variety of approaches were proposed based on the deformation of facial features. Active appearance models are utilized in [1] to interpret face images and classify expressions. Lyons et al. [2] represent faces as elastic graphs labeled with 2D Gabor wavelet features and classify high-level attributes using linear discriminant analysis. However, the facial meaning is embedded not only in the nature of facial feature deformation, but also in the temporal evolution of facial attributes. Therefore, many approaches were proposed for facial expression analysis in video. In [3], a control-theoretic method is introduced to extract the spatio-temporal motion-energy representation of facial motions. Zhang and Ji [4] use a multisensory information fusion technique with dynamic Bayesian networks for modeling the temporal behaviors of facial expressions in image sequences. The use of hidden Markov models (HMM) for video analysis has been adopted by computer vision research with increasing interest in recent years. HMMs provide an effective probabilistic framework for modeling stochastic processes such as speech signals and DNA sequences. HMM-based approaches have been exploited in [5][6][7] for modeling the temporal behaviors of facial expressions with features such as optical flow, feature points, and motion units. A comprehensive literature survey concerning the state-of-the-art facial expression analysis can be found in [8].

This paper proposes a new approach to facial event mining by coherent integration of active shape models (ASMs) [9] and coupled hidden Markov models (CHMMs) [10]. In the proposed method, global facial expressions are interpreted as a result of two coupled sub-expression processes involving upper faces and lower faces, respectively. ASMs are used as the front-end unit for global facial feature detection, tracking, and pattern feature decoupling for upper and lower faces. Then, these two interacting stochastic processes are modeled via CHMMs for event classification.

Human faces are constrained by geometrical relationships between facial features. Facial expressions are a result of contractions of facial muscles and characterized by temporally deformed facial characteristics. An active shape model is devised to be parametric and deformable for locating non-rigid objects in cluttered images. ASMs interpret image data in ways which are constrained by the statistical characteristics of the class of objects being modeled. The shape parameters of ASM are a compact representation of shape features of objects with allowable variations. Therefore, ASMs are adopted in this system to accomplish facial feature detection, tracking, and pattern feature decoupling in a unified way. In [11], a similar procedure based on point distribution models (PDM) was proposed for feature extraction. However, their method did not take advantage of the temporal behavior of facial expressions. Our method uses

0-7803-8554-3/04/$20.00

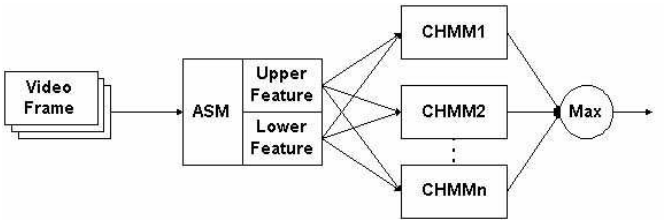

Fig. 1. The structure of proposed facial event mining system

shape parameters and their derivatives directly as feature vectors and employs ASMs to decouple shape features for upper and lower faces.

Facial action coding system (FACS) [12] divides facial expressions into upper and lower facial actions and further subdivides the motion into action units (AUs). Facial expressions are modeled by single AUs or AU combinations. The recognition of AUs in upper and lower faces were investigated in [7] [13]. Oliver et al. [14] developed a system called LAFTER for tracking and recognizing mouth expressions. Therefore, it is reasonable to assume that an expression can be characterized as a composition of the correlated facial actions in upper and lower faces, even though most HMM-based approaches so far interpret faces as a whole. CHMMs are adopted in the proposed method, because it provides an effective mechanism to model multiple interacting processes by introducing conditional probabilities between their hidden state variables. It also has the advantages of HMMs.

Four basic facial events (yawn, neutral, smile and speaking) are investigated in the experiments. The preliminary results justify the premise that facial expressions can be regarded as two interacting processes. It is also observed that CHMMs have significant advantages over conventional HMMs for data fusion and modeling facial events composed of coupled processes. The following sections include a detailed discussion about the methods used and present the experimental results obtained. Conclusions and future research topics are given at the end.

2. METHODS

As shown in Fig. 1, the system is composed of a video input, an ASM-based feature extraction module, and a CHMM-based classifier. ASMs are used for global facial feature tracking and shape feature decoupling for upper and lower faces. Each facial event is decomposed into a set of coupled upper and lower facial actions. They are represented by an observed feature sequence and modeled by a CHMM. In the recognition process, the probability of an observation sequence generated by ASMs is computed for each event model, and the one with the maximum score is chosen as the recognition result.

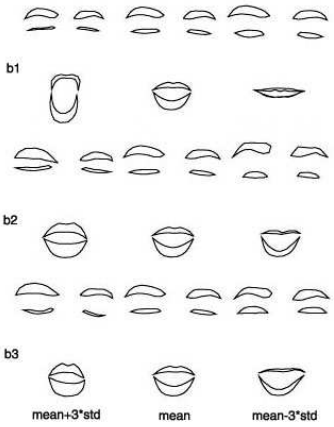

Fig. 2. Point distribution models of facial features

2.1. Active Shape Models

There are two phases in ASMs: training and searching. In the training phase, each face image in the training set is annotated with $n = 104$ landmark points along the boundary of eyes, eyebrows, and mouth, since their deformation largely contributes to facial events. Each face shape is represented by the coordinates of its landmarks, i.e.,

$$\mathbf{x} = \{(x_1, y_1), ..., (x_n, y_n)\}$$

With principle component analysis, an existing shape can be approximated as $\mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{P}\mathbf{b}$, where $\overline{\mathbf{x}}$ is the mean shape, $\mathbf{P}$ is the transformation matrix, and $\mathbf{b} = (b_1, b_2, ..., b_t)^T$ is called the shape parameter. New examples of the shapes can be generated by varying $\mathbf{b}$ within suitable limits; e.g., three times its standard deviation, as shown in Fig. 2. In this way, the model $\mathbf{X}$ in image coordinates is represented as $\mathbf{X} = T_{x_t, y_t, s, \theta}\left(\bar{\mathbf{x}} + \mathbf{P}\mathbf{b}\right)$, where the function $T_{x_t, y_t, s, \theta}$ performs a similarity transform. In order to decouple $\mathbf{b}$, the same procedure is applied to shapes in upper and lower face halves to derive their respective models; i.e.,

$$\mathbf{x}_i \approx \overline{\mathbf{x}}_i + \mathbf{P}_i\mathbf{b}_i, i = upper, lower$$

The dimension of $\mathbf{b}_{upper}$ and $\mathbf{b}_{lower}$ in this research is reduced to 21 and 17, respectively.

In the searching phase, a modified iterative fitting algorithm is proposed for image interpretation. First, a neighboring region around each landmark point $\mathbf{X}_i$ is examined to find the most likely target position $\mathbf{X}_i'$ by fitting the current normal profile to a statistical grey-level model learned from the training set. Then the parameters $\mathbf{b}$,x_t, y_t, s, θ are updated to deform $\mathbf{X}$ to $\mathbf{X}'$ with constraints imposed on the shape parameters. This procedure is repeated until it converges. The original local search procedure is actually a greedy algorithm which seeks for the solutions effectively with a good initialization. We propose a strategy to apply

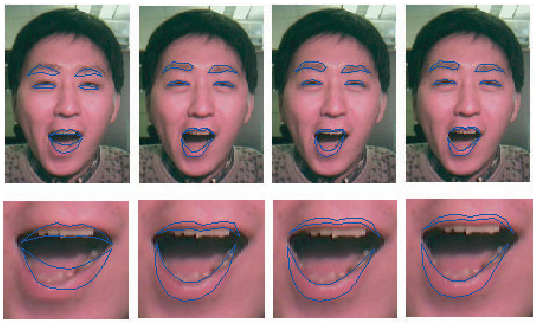

Fig. 3. Iterative search result of ASMs

Fig. 4. Active shape models tracking

a dynamic programming (DP) search in the very first frame and the greedy algorithm in successive frames so that global optimality is guaranteed by DP in case of a poor initialization, and fast solutions are obtained in subsequent frames. An iterative search process is illustrated in Fig. 3, where ASM quickly converges to its true position with an approximate initialization. Tracking is achieved by applying ASM based on the result in previous frames as shown in Fig. 4. Shape parameters are decoupled after the search converges, because the coupling is embedded in global models. Finally, $\mathbf{b}_{upper}$ and $\mathbf{b}_{lower}$ in each frame and their derivatives are used as feature vectors for classification.

2.2. Coupled Hidden Markov Models

A HMM [15] is a powerful probabilistic framework for modeling processes that have varying structure in time. It has

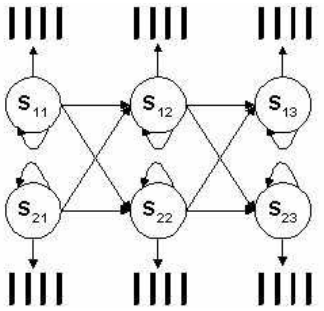

Fig. 5. Topology of coupled hidden Markov models

Bayesian semantics, efficient algorithms for parameter estimation, and the ability to automatically perform dynamic time warping. A HMM, $\lambda = (A, B, \Pi)$, is defined by a set of N discrete states $S = \{a_1, a_2, ..., a_N\}$, a time indexed discrete variable $s_t \in S$, state transition probabilities $A = [P_{s_{t+1}|s_t} = P(s_{t+1} = a_j | s_t = a_i)]$, initial state probabilities Π, and the observation probability density function (PDF) $B = [b_j(o_t)]$ for each state. HMMs are used for classification in such a way that given a sequence of observations $O = \{o_1, o_2, ..., o_T\}$, the most likely state sequence within a model is determined by $S^* = \arg\max_S P(S|O, \lambda)$ which can be efficiently computed by the Viterbi algorithm. As for training, the parameters of HMMs can be optimized via the Baum-Welch reestimation procedure.

CHMMs provide an efficient mechanism for fusing data from multiple sensors and modeling multiple interacting processes by introducing conditional probabilities between their hidden state variables. CHMMs have been used for modeling hand gestures [10], pedestrian interactions [16], and audio-visual speech recognition [17]. A left-to-right CHMM with two chains is illustrated in Fig. 5, where the dynamic interactions are interpreted via two HMMs while the probability of a state in one HMM is dependent not only on its previous state (horizontal state transitions) but also on the previous state in the coupled HMM (diagonal state transitions). The posterior state probability for CHMMs is given by

$$P(S|O, \lambda) = \frac{P_{s_1} b_{s_1}(o_1) P_{s_1'} b_{s_1'}\left(o_1'\right)}{P(O)} \bullet \prod_{t=2}^{T} P_{s_t|s_{t-1}} P_{s_t'|s_{t-1}'} P_{s_t'|s_{t-1}} P_{s_t|s_{t-1}'} b_{s_t}(o_t)\, b_{s_t'}\left(o_t'\right)$$

where superscript denotes the coupled chain. A N-heads dynamic programming is presented in [10] for efficiently computing the posteriors in $O\left(T(2N)^2\right)$ time by relaxing the assumption that every transition must be visited. Here, T denotes the number of observations. This enables fast classification and parameter estimation. The transition matrices are reestimated in a way similar to the conventional HMMs after statistics are collected by the N-heads algorithm[10].

3. EXPERIMENTS

A database including 10 subjects under constant illumination is created for experiments. The Cohn-Kanade database [18] is not appropriate for our research, because it only contains incomplete facial expressions. Four basic facial event classes, namely, yawn, neutral, smile and speaking, are investigated in the experiment. Every subject has two sessions, each of which is composed of all four specific facial events. Video segments vary from 2 to 8 seconds for different kinds of events. In total, there are 20 video segments for each type. Considering the finite size of the database, five subjects are randomly selected for training and the other half are used for testing. Five rounds of random selection are carried out for experiments. Yawn is trained as a CHMM with 5 coupled states, while others are with 3 coupled states. HMMs are also trained with global shape parameters for comparison. The observation PDF for each state is modeled as a Gaussian distribution since only a limited amount of data is available. The average accuracy of both HMMs and CHMMs are listed in Table 1. It is clear from the table that the classification of yawn and speaking has a higher accuracy than that of the other two, and significant improvements (almost 4.5%) are obtained by using CHMMs. This improvement can be attributed to the fact that upper and lower face actions possess closely coupled temporal characteristics for yawn and speaking and their facial features have relatively large dynamic ranges. Neutral and smile interactions have weak coupling and little dynamic changes. As a result, they are likely to be misclassified.

Table 1. Recognition results

Method	Yawn	Neutral	Smile	Speaking
HMM	90%	76%	80%	86%
CHMM	94%	78%	84%	90%

4. CONCLUSIONS AND FUTURE WORK

This paper presents a new approach to facial event mining in video based on ASMs and CHMMs. In the proposed method, the global facial expressions are decomposed into two interacting dynamic processes involving upper faces and lower faces, respectively. ASMs are used for locating and tracking facial features, and decoupling compact visual attributes. CHMMs are adopted for modeling these two interactions and recognizing different events. Experimental results on four classes of basic facial gestures and expressions show the superiority of CHMMs to the conventional HMMs in terms of accuracy, and also justify the assumption that better modeling can be obtained by decomposing a complex facial system into coupled processes. Our future work will investigate asymmetric CHMMs for better modeling, increase the size of database, and experiment with more event classes.

5. REFERENCES

[1] G. Edwards et al., "Interpreting face images using active appearance models," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*, 1998.

[2] M. Lyons et al., "Classifying facial attributes using a 2D Gabor wavelet representation and discriminant analysis," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*, 2000.

[3] I. Essa and A. Pentland, "Coding, analysis, interpretation and recognition of facial expression," *IEEE Trans. on PAMI*, vol. 19, pp. 757–763, 1997.

[4] Y. Zhang and Q. Ji, "Facial expression understanding in image sequences using dynamic and active visual information fusion," in *Proceedings of ICCV*, 2003.

[5] I. Cohen et al., "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, pp. 160–187, 2003.

[6] T. Otsuka and J. Ohya, "Spotting segments displaying facial expression from image sequences using hmm," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*, 1998.

[7] J. Lien et al., "Automated facial expression recognition based on FACS action units," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*, 1998.

[8] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, pp. 259–275, 2003.

[9] T. Cootes et al., "Active shape models-their training and application," *Computer Vision and Image Understading*, vol. 61, pp. 38–59, 1995.

[10] M. Brand, "Coupled hidden markov models for modeling interacting process," in *MIT Media Lab Technical Report 405*, 1997.

[11] C. Huang and Y. Huang, "Facial expression recognition using model-based feature extraction and action parameters classification," *J. Visual Communication and Image Representation*, vol. 8, pp. 278–290, 1997.

[12] P. Ekman and W. Friesen, *Facial action coding system: A technique for the measurement of facial movement*, Consulting Psychologists, 1978.

[13] Y. Tian et al., "Recognizing lower face action units for facial expression analysis," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*, 2000.

[14] N. Oliver et al., "Lafter: A real-time face and lips tracker with facial expression recognition," *Pattern Recognition*, vol. 33, pp. 1369–1382, 2000.

[15] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of IEEE*, vol. 77, pp. 257–286, 1989.

[16] N. M. Oliver et al., "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. on PAMI*, vol. 22, pp. 831–843, 2000.

[17] A. Nefian et al., "A coupled hmm for audio-visual speech recognition," in *Proceedings of ICASSP*, 2002.

[18] T. Kanade et al., "Comprehensive database for facial expression analysis," in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition*, 2000.