AUDIO-VISUAL BASED EMOTION RECOGNITION USING TRIPLED HIDDEN MARKOV MODEL

Mingli Song, Chun Chen, Mingyu You

College of Computer Science, Zhejiang University
Hangzhou, 310027, P.R. China
{brooksong, chenc, roseymy}@cs.zju.edu.cn

ABSTRACT

Emotion recognition is one of the latest challenges in intelligent human/machine communication. Most of previous work on emotion recognition focused on extracting emotions from visual or audio information separately. A novel approach is presented in this paper to recognize the human emotion which uses both visual and audio from video clips. A tripled Hidden Markov Model is introduced to perform the recognition which allows the state asynchrony of the audio and visual observation sequences while preserving their natural correlation over time. The experimental results show that this approach outperforms only using visual or audio separately.

1. INTRODUCTION

Emotion recognition is becoming more and more important in the intelligent communication between human and computer. Extensive studies have been performed on the relationship between emotion and facial expression or speech since 1970s [1, 5].

Psychological research [4] identified that six principle emotions are universally associated with distinct facial expression. Several approaches [2, 3] were presented to recognize different facial expressions from video frame sequence. In [2, 3], by tracking facial features and measuring the amount of facial movement, they attempt to categorize different facial expressions automatically. However, the computing workload of such approaches is very heavy while the result is not so good with the average accuracy of about 74%.

Much work has also been done for emotion analysis in the field of speech recognition. There are some difficulties in analyzing emotions from speech. The recognition of emotion has been shown more difficult than that of word or phonemes. Technical report [6] gave the accuracy of 63.5% in comparison 90% of word recognition

In our approach, visual and speech are both considered to recognize the emotion from video. The visual information provides the system with features that can not be corrupted by the acoustic noise of the environment. And when the visual information is not good enough to determine the emotion, the audio one can help the system to make the recognition. The emotions are classified into seven basic but representative ones: neutral, joy, anger, surprise, sadness and fear.

It is noticeable that the mouth shape is not necessary reflect the speaker's emotion while speaking. In our approach, facial feature points are divided into two groups: visual speech part and expression one. The latter is important for emotion recognition while the former can perform the auxiliary function. So we factorize these two parts separately to obtain the corresponding two visual vectors.

HMM is a good model to process temporal-space signal. The audio and visual fusion techniques investigated in previous work based on multi-stream HMM assume that audio and video component to have different contribution to the overall observation likelihood. However, it is well known that the acoustic features of speech are delayed from the visual features, and assuming state synchronous models can be inaccurate. The audio visual product HMM can be seen as an extension of the multi-stream HMM that allows for audio-video state asynchrony. In our work, the system models independently the two visual input sequences and the audio ones using three HMMs (Tripled HMM), and combines the likelihood of each observation sequence based on the reliability of each modality. Figure 1 is an overview of our system.

2. VISUAL FEATURE EXTRACTION

2.1 Visual Feature Tracking

The MPEG-4 standard extends FACS to derive a set of Facial Animation Parameters (FAPs). It has been used widely in facial animation for its good compression recently. And it is adequate for basic expression recognition because of its varieties of expressive parameter.

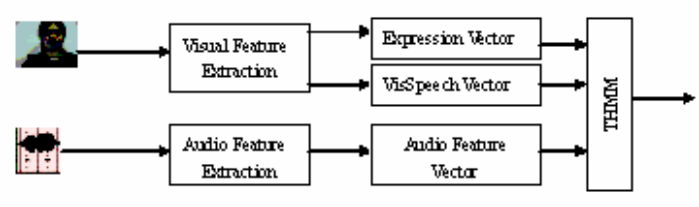

Figure 1: System Overview

In our system, a MPEG-4 compliant feature point set is defined to be tracked in the video frame sequence. An example of tracked image is shown in Figure 2.

As we know, the mouth shape is necessary for speaking and it doesn't reflect the speaker's expression all the time. We usually determine a person's expression through the motion of his/her eye brow, eyelid and cheek. Some facial contour such as the lips and jaw, do not necessarily have the expression information while one is speaking. So in our study, the tracked feature points are divided into two groups experientially. The feature points of the lips and jaw are clarified into visual speech set while the others are processed as expression ones. Active Appearance Model (AAM) [7] is a good method to track the facial feature points in temporal-spatial domain because of its robustness and real time. For tracking the facial feature points mentioned above, we apply this method to track the facial feature points. The image database for the training of the appearance model is obtained using the method in [8]. The images from the sequence are labeled with the feature points manually. Then, the model is trained by using these labeled images. Finally, the facial feature points can be tracked using the trained model.

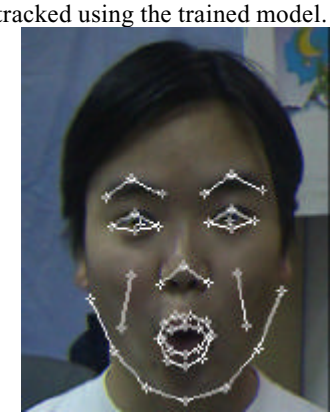

Figure 2 Facial Feature Points Tracking

2.2 Visual Feature Vectorization

For a set of n tracked feature points, $f_1, \cdots f_n$ on the face define the face shape, represented as the vector $\overline{m}$, where $\overline{m} = [f_1^x, f_1^y, \cdots f_n^x, f_n^y]^T$. As mentioned in section 2.1, the facial feature points are grouped into visual speech set and expression set, correspondingly, $\overline{m}$ is split into $\overline{m}^s$ and $\overline{m}^e$, where $\overline{m}^s = [f_1^x, f_1^y, \cdots f_s^x, f_s^y]^T$ and $\overline{m}^e = [f_{s+1}^x, {}^y_{s+1} \cdots f_n^x, f_n^y]^T$. In our case, n=46 and s=18.

With the frames from 1 to C, we can represent the facial data as the facial configuration described above. The input vectors are stacked for the different frames to form two stream vector M^S and M^E where

$$M^S = \left[\overline{m}_1^s, \overline{m}_2^s, \cdots \overline{m}_C^s\right] \quad (1)$$

$$M^E = \left[\overline{m}_1^e, \overline{m}_2^e, \cdots \overline{m}_C^e\right] \quad (2)$$

and they are used as the input of THMM.

4. AUDIO FEATURES FOR EMOTION RECOGNITION

Besides the visual information, the audio data of the observation also reflect the actor/actress's emotion. We extract the acoustic features as used in [9]. For each frame, the following feature parameters are calculated:

Energy features: The first and second derivatives of the logarithm of the mean energy in the frame are used to model the instantaneous value of energy, which reflect both the articulation speed and the dynamic range. And the first and second derivatives of the logarithm of the 8Hz low pass filtered energy in the frame are used to model the syllabic contour of energy.

Pitch features: In order to characterize instantaneous pitch, a simple auto-correlation analysis is performed at every frame. The maximum of the long term auto-correlation is determined and used to form five different parameters: the value of the maximum of the long term auto-correlation, along with its first and second derivatives, and the first and second derivatives of the logarithm of the pitch lag. And the first and second derivatives of the smoothed pitch lag evolution are evaluated in order to estimate the pitch evolution more precisely.

The audio feature vector is formed by 4 energy feature parameters and 7 pitch ones. An 11 dimension vector with these parameters is represented as $\overline{a}^e$, where

$$\overline{a}^e = [e_1, \cdots e_4, p_1 \cdots p_7]. \quad (3)$$

With the frames from 1 to C, an audio input stream vector is

$$A^E = \left[\overline{a}_1^e, \cdots, \overline{a}_C^e \right], \quad (4)$$

while A^E is used as the input of THMM.

5. THE AUDIO-VISUAL MODEL

A novel model for emotion recognition is introduced in our system, which uses a tripled hidden Markov model (THMM). The THMM is a generalized of the HMM suitable for a large variety of multimedia application that integrate two or more streams of data. A Tripled HMM can be seen a collection of HMMs, one for each data stream, where the discrete nodes at time $t-1$ of all the related HMMs. The parameters of a THMM are defined as follows:

$$\pi_0^c(i) = P(q_t^c = t) \quad (5)$$

$$b_t^c(i) = P(O_t^c \mid q_t^c = i) \quad (6)$$

$a_{i|j,k,l}^c = P(q_t^c = i \mid q_{t-1}^0 = j, q_{t-1}^1 = k, q_{t-1}^2 = l)$ where q_t^c is the state of the triple node in the cth stream at time t. In a continuous mixture with Gaussian components, the probabilities of the observed nodes are given by:

$$b_t^c(i) = \sum_{m=1}^{M_i^c} w_{i,m}^c N(O_t^c, \mu_{i,m}^c, U_{i,m}^c) \quad (7)$$

where $\mu_{i,m}^c$ and $U_{i,m}^c$ are the mean and covariance matrix of the ith state of tripled node, and mth component of the associated mixture node in the cth stream. M_i^c is the number of mixtures and the weight $w_{i,m}^c$ represents the conditional probability $P(s_t^c = m \mid q_t^c = i)$ where s_t^c is the component of the mixture node in the cth stream at time t.

6. TRAINING

In our system, an efficient method for the initialization of the maximum likelihood training that uses Viterbi algorithm is derived for the tripled HMM. The Viterbi algorithm for the three streams tripled HMM used in our approach. An extension to a multi-stream tripled HMM is straightforward.

1. Initialization

$$\delta_0(i,j,l) = \pi_0^a(i)\pi_0^e(j)\pi_0^s(l) b_t^a(i) b_t^e(j) b_t^s(l) \quad (8)$$

$$\psi_0(i,j,l) = 0 \quad (9)$$

2. Recursion

$$\delta_t(i,j,l) = \max_{x,y,z} \left\{ \delta_{t-1}(x,y,z) a_{i|x,y,z} a_{j|x,y,z} a_{l|x,y,z} \right\} b_t^a(x) b_t^s(y) b_t^e(z) \quad (10)$$

$$\psi_t(i,j,l) = \underset{x,y,z}{\operatorname{argmax}} \left\{ \delta_{t-1}(x,y,z) a_{i|x,y,z} a_{j|x,y,z} a_{l|x,y,z} \right\} \quad (11)$$

3. Termination

$$P = \max_{i,j,l} \{ \delta_T(i,j,l) \} \quad (12)$$

$$\{ q_T^a, q_T^s, q_T^e \} = \underset{i,j,l}{\operatorname{argmax}} \{ \delta_T(i,j,l) \} \quad (13)$$

4. Backtracking

$$\{ q_t^a, q_t^s, q_t^e \} = \psi_{t+1}(q_{t+1}^a, q_{t+1}^s, q_{t+1}^e) \quad (14)$$

and the segmental K-means algorithm for the tripled HMMs is described as follows:

Step 1: For every training observation sequence r, the data in the stream is uniformly segmented according to the number of states of the tripled nodes and an initial state sequence for the tripled nodes $Q = q_{r,0}^{a,s,e}, \cdots q_{r,t}^{a,s,e}, \cdots, q_{r,T-1}^{a,s,e}$ is obtained. With M_i^c clusters, the mixture segmentation of the data assigned to each state i of the tripled nodes in stream c is obtained using K-means algorithm. Consequently, the sequence of mixture components $S = s_{r,0}^{a,s,e}, \cdots, s_{r,t}^{a,s,e}, \cdots s_{r,T-1}^{a,s,e}$ for the mixtures nodes is obtained.

Step 2: The new parameters of the model are estimated from the segmented data.

$$\mu_{i,m}^{a,s,e} = \frac{\sum_{r,t} \gamma_{r,t}^{a,s,e}(i,m) O_t^{a,s,e}}{\sum_{r,t} \gamma_{r,t}^{a,s,e}(i,m)} \quad (15)$$

$$\sigma_{i,m}^{2\ a,s,e} = \frac{\sum_{r,t} \gamma_{r,t}^{a,s,e}(i,m)(O_t^{a,s,e} - \mu_{i,m}^{a,s,e})(O_t^{a,s,e} - \mu_{i,m}^{a,s,e})^T}{\sum_{r,t} \gamma_{r,t}^{a,s,e}(i,m)} \quad (16)$$

$$w_{i,m}^{a,s,e} = \frac{\sum_{r,t} \gamma_{r,t}^{a,s,e}(i,m)}{\sum_{r,t} \sum_m \gamma_{r,t}^{a,s,e}(i,m)} \quad (17)$$

$$a_{i|x,y,z}^{a,s,e} = \frac{\sum_{r,t} \varepsilon_{r,t}^{a,s,e}(i,x,y,z)}{\sum_{r,t} \sum_x \sum_y \sum_z \varepsilon_{r,t}^{a,s,e}(i,x,y,z)} \quad (18)$$

where

$$\gamma_{r,t}^{a,s,e}(i,m) = \begin{cases} 1, & if \quad q_{r,t}^{a,s,e} = i, s_{r,t}^{a,s,e} = m, \\ 0, & otherwise \end{cases} \quad (19)$$

$$\varepsilon_{r,t}^{a,s,e}(i,x,y,z) = \begin{cases} 1, & if \quad q_{r,t}^{a,s,e} = i, q_{r,t}^a = x, \\ & \quad q_{r,t}^s = y, q_{r,t}^e = z \\ 0, & otherwise \end{cases} \quad (20)$$

Step 3: At consecutive iteration the optimal state sequence **Q** of the tripled nodes is obtained using the Viterbi algorithm. The sequence of mixture component **S** is obtained by selecting at each moment t the mixture $s_{r,t}^{a,s,e}$ such that:

$$s_{r,t}^{a,s,e} = \max_{m=1,\cdots,M_i^{a,s,e}} P(O_t^{a,s,e} \mid q_{r,t}^{a,s,e} = i, m) \quad (21)$$

Step 4: The iterations in Step 2-4 are repeated until the difference between the observation probabilities of the training sequences at consecutive iterations falls below the convergence threshold.

6. EMOTION RECOGNITION

The emotion recognition is carried out through the computation of Viterbi algorithm for the input vector stream of modeled basic emotion database. The parameters of the THMM corresponding to the basic emotion are obtained after training. In the recognition stage, the influence of the audio and visual streams is weighted based on the relative reliability of the audio and visual features for different levels of the acoustic noise. With the increasing of SNR, the weight of the stream rises. Formally, the observation probability at time t for the observation vector $O_t^{a,s,e}$ is

$$\tilde{b}_t^{a,s,e}(i) = b_t(O_t^{a,s,e} \mid q_t^{a,s,e})^{\alpha_{a,s,e}} \quad (22)$$

where $\alpha_a + \alpha_s + \alpha_e = 1$, $\alpha_s = \alpha_e$ and $\alpha_a, \alpha_s, \alpha_e \geq 0$ are the exponent of the audio and two visual streams. The values of α_a, α_s and α_e corresponding to a specific acoustic SNR are obtained experimentally to maximize the average recognition rate.

Each emotion THHM is trained with 100 samples firstly. And Table 1 shows the confusion matrix identified emotions for the 684 samples of the test database. The result is appreciable with the accuracy higher than 85%, which is the best comparing with that of the past system.

	S	J	A	F	D	T	N	Tot.
Surprise	138	6	3	6	3	0	0	156
Joy	8	89	6	1	1	0	0	105
Angry	0	2	90	5	3	0	0	100
Fear	2	0	0	96	1	2	1	102
Disgust	0	2	2	1	75	1	1	82
Sadness	0	0	0	0	0	80	9	89
Neutral	0	0	0	0	0	0	50	50
Total	148	99	101	109	83	83	61	684

Table 1:Confusion matrix of the emotion recognition Columns represent the emotion elected in first choice for samples belonging to the emotion of each row, where A stands for anger, S for surprise, J for joy, F for Fear, D for disgust, T for sadness and N for neutral.

7. CONCLUSION

In this paper, an audio-visual emotion recognition system that uses a three-stream tripled HMM to model the audio and video observation sequences. Unlike the HMM, the THMM allows for asynchrony in the audio and visual states, while preserving the natural dependency of the audio and video signals. Furthermore, with the tripled HMM, the audio and the two visual sequences are treated separately and is no need for the concatenation of the observation that is often a challenging problem. The advantage of the THMM is confirmed by the experiment results. This model can be applied to a variety of human/machine system.

Future work includes, first of all, improvement of current model. Besides this, a larger research on synthesis of facial animation with emotion will be carried out.

8. ACKNOWLEGDMENT

This work is partly supported by NSFC grants 60203013

9. REFERENCES

[1] P. Ekman and W.V. Friesen. *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, 1978.

[2] Ira Cohen, Nicu Sebe, Fabio G. Cozman, Marcelo C. Cirelo, Thomas S. Huang. "Learning Bayesian Network Classifiers for Facial Expression Recognition with both Labeled and Unlabeled data". IEEE Conference on Computer Vision and Pattern Recognition, 2003.

[3] Ira Cohen, Nicu Sebe, Larry Chen, Ashutosh Garg, Thomas S. Huang, "Facial Expression Recognition from Video Sequences Temporal and Static Modeling" , *Computer Vision and Image Understanding*, Special Issue on Face Recognition, Volume 91, Issues 1-2, Pages 160-187, July-August 2003.

[4] P. Ekman. "Strong evidence for universals in facial expressions: A reply to Russelll's mistaken critique". *Psychological Bulletin*, 115(2):268-287, 1994.

[5] F. Dellaert, T. Polzin, and A. Waibel. "Recognizing emotion in speech"., In Proc. ICSLP 1996, Oct 1996.

[6] V. Petrushin. "Emotion recognition in speech signals: Experimental study, development, and application", In Proc. ICSLP 2000, 2000.

[7] T. F. Cootes, G. J. Edwards and C. J. Taylor. " Active Appearance Models", In Proc. European Conference on Computer Vision 1998 (H. Burkhardt & B. Neumann Ed.s), Vol. 2, pp. 484-498, Springer, 1998.

[8] M. Covell, C. Brgler, " Eigen-Points", In Proc. IEEE Int. Conf. on Image Processing, 1996.

[9] Albino Nogueiras, Asunció n Moreno, Antonio Bonafonte, José B. Mariño, " Speech Emotion Recognition Using Hidden Markov Models", European Conference on Speech Communication and Technology, 2001.