EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Utterance independent bimodal emotion recognition in spontaneous communication

Jianhua Tao*, Shifeng Pan, Minghao Yang, Ya Li, Kaihui Mu and Jianfeng Che

Abstract

Emotion expressions sometimes are mixed with the utterance expression in spontaneous face-to-face communication, which makes difficulties for emotion recognition. This article introduces the methods of reducing the utterance influences in visual parameters for the audio-visual-based emotion recognition. The audio and visual channels are first combined under a Multistream Hidden Markov Model (MHMM). Then, the utterance reduction is finished by finding the residual between the real visual parameters and the outputs of the utterance related visual parameters. This article introduces the Fused Hidden Markov Model Inversion method which is trained in the neutral expressed audio-visual corpus to solve the problem. To reduce the computing complexity the inversion model is further simplified to a Gaussian Mixture Model (GMM) mapping. Compared with traditional bimodal emotion recognition methods (e.g., SVM, CART, Boosting), the utterance reduction method can give better results of emotion recognition. The experiments also show the effectiveness of our emotion recognition system when it was used in a live environment.

Keywords: Bimodal emotion recognition Utterance Independent, Multistream Hidden Markov Model, Fused Hidden Markov Model Inversion

Introduction

The last two decades have seen significant effort devoted to developing methods for automatic human emotion recognition (e.g., [1-15]), which is an attractive research issue due to its great potential in human-computer interactions (HCIs), virtual reality, etc. Although there are a few tentative efforts to detect non-basic emotion states including fatigue (e.g., [16]), and mental states, such as agreeing, concentrated, disagreeing, interested, thinking, confused, and frustration (e.g., [17-20]), most of the existing efforts focus on the some basic emotions due to their universal properties, their marked reference representation in our affective lives, and the availability of the relevant training and test material (e.g., [1,2,7,21]). Many classical machine learning or pattern recognition algorithms were used to infer emotion states. Most of them have used only a single channel (e.g., [[2,8,10,21-28]), for instance, the facial expression (e.g., [2,8,10]) or speech (e.g., [21-28]). As reported in [29], both vocal intonations and facial expressions determine the listener's affective state in up to 93% of cases. Recently, increased attention has been paid to analyzing multimodal information in emotion recognition (e.g., [1,7,9-13,30-34]). However, most of them still use deliberate and often exaggerated facial displays (e.g., [2,5]).

The spontaneous facial expression is always the natural way for the real human to human communication (e.g., [35]). Studies reported in [36-39] investigated explicitly the difference between spontaneous and deliberate facial behavior. In this situation, the facial expressions are sometimes combined with both emotions and expressed utterances [40]. Such problems may sometimes confuse the methods for emotion recognition. For instance, the facial expression of the phoneme "i" might be recognized as a smile. Some efforts have been recently reported on the analysis of spontaneous facial expression data (e.g., [19,20,36-39,41-49]). For instance, Pantic and Rothkrantz [1] and Fasel and Luttin [10] suggested that the facial block near the lips should not be used for emotion recognition. In Zeng et al.'s study [50,51], smoothed facial features are calculated by averaging facial features at consecutive frames to reduce the influence of utterance on facial expression, based on the assumption that the

* Correspondence: jhtao@nlpr.ia.ac.cn
National Laboratory or Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 100190, China

influence of utterance on face features is temporary, and the influence of affect is relatively more persistent. However, these simple averaging facial features may give some error hints for facial expression especially while the utterance is very short, or some paralinguistic features are included. However, most of existing work still simply combines the audio-visual parameters for emotion recognition with model of feature-level fusion or decision-level fusion (e.g., [19,20,45-47]), some of them just focus on getting Action Units (AUs) from facial expression rather than emotion recognition (e.g., [36-39,41-44,48,49]).

In this article, we try to introduce a new utterance-independent method for bimodal emotion recognition in spontaneous communication. At the beginning, a Multi-stream Hidden Markov Model (MHMM) is used to combine the audio-visual features for the emotion recognition. While there is still argument on integrated emotion theory, we focus here on the six basic emotions: "happiness," "surprise," "fear," "anger," "sadness," and "neutral." To do the utterance reduction, the input audio features are classified into two types, content-related features and prosody features. Then the audio-visual mapping from content-related features to facial expression is created. The results of utterance reduction in visual parameters will be finally got by subtracting the audio-visual mapping results from the real facial expressions. We introduce a Fused Hidden Markov Model (HMM) Inversion model to solve the mapping problem. To reduce the computing complexity, the inversion model is further simplified to a Gaussian Mixture Model (GMM) mapping. Furthermore, as inspired by the idea of using averaging facial shapes [51], we also use a dynamic smoothed facial parameters which is tracked as a search in point distribution models (PDMs) [52] to get better visual parameters.

The article makes some detailed experiments and discussion of our methods by comparing with other utterance-dependent methods. The results show that the utterance-independent methods improve the results of the emotion recognition, especially for confusing emotions. Finally, a real-time bimodal emotion recognition system in our live communication environment has been created in our lab. The efficiency of the system is also discussed.

The contributions of this article are concluded as the following points.

(a) Our approach takes the advantages of time series analysis for emotion recognition by combining audio and visual features with a Multi-stream Hidden Markov Model (MHMM) method.

(b) We propose the utterance-independent method to enhance the visual expression parameters for emotion recognition in spontaneous communication by the hybrid of the MHMM and the fused HMM inversion.

(c) To reduce the computing complexity we also propose an alternative utterance reduction model which is based on a GMM and is simplified from the inversion model.

(d) We made detailed performance analysis for the utterance reduction methods and extended them in the live environment, which will have greater potential application as well as higher recognition accuracy.

This article is organized as follows. In "Bimodal fusion with multistream hidden markov model," we use the Multi-stream Hidden Markov Model (MHMM) for audio-visual data fusion in emotion recognition. Unlike traditional couple or fused HMM, the discrete coupling parameters is extended to the continuous observation in our study. Section "Utterance reduction with inversion method" introduces the fused HMM inversion model for the utterance reduction. A two-layer clustering method in visual configuration is further introduced to smooth the visual representations. The simplifying Fused HMM Inversion to GMM mapping is also described in this section followed by the experiments and discussion of our study. The audio-visual parameters and training data is also described here. Different utterance reduction models are discussed. We further compare our study with the typical emotion recognition methods, SVM-based method [13], CART-based method [11], Boosting Method [53], Rule-based decision fusion method [11], and also the methods that use only uni-channel features via extensive experiments. Finally, we conclude our study and discuss future study.

Bimodal fusion with Multistream Hidden Markov Model

The most popular bimodal fusion methods are based on the feature-level fusion [13] or decision-level fusion [11]. The former classifies the bimodal feature vectors combined from audio and visual channels into different emotions directly [13], while the latter makes decisions based on rules after separate acoustic and visual classifications [11]. However, the audio and facial expressions are synchronous at successive times. Some study [24] has proved that the time series analysis methods can improve the robust of such data processing. Thus, we apply the Multi-stream Hidden Markov Model (MHMM) (e.g., [4,51]) for the emotion recognition in our study. In MHMM framework, the composite facial feature from video, acoustic features from audio are treated as two streams, and modeled by two-component HMMs. We use the weighted summation to fuse the results from these component HMMs.

Within all MHMMs, the Fused HMM has been proven as a good model in obtaining the probability between fused audio-visual training pairs (e.g., [4,51]). Given the observed audio-visual parameters, O^a, O^v, and their corresponding HMM, the Fused HMM was

proposed to construct a structure linking the component HMMs together by giving optimal estimation of the joint probability. Taking advantage of the fact that the data from a single sensor can be individually modeled by a HMM, and according to the maximum entropy principle and the maximum mutual information (MMI) criterion, the fusion model yields the following two structures, as shown by [4]:

$$p^{(1)}\left(O^{a};\ O^{v}\right) = p\left(O^{a}\right)\ p\left(O^{v}|\hat{U}^{a}\right) \tag{1}$$

$$p^{(2)}\left(O^{a};\ O^{v}\right) = p\left(O^{v}\right)\ p\left(O^{a}|\hat{U}^{v}\right) \tag{2}$$

where the most possible hidden state sequences $\hat{U}^{a}$ and $\hat{U}^{v}$ are estimated by the Viterbi algorithm. The training process of the Fused HMM includes the following three main steps in general: (a) Two individual HMMs, consisting of visual component HMM and audio component HMM in our study, are trained independently by the EM algorithm; (b) The best hidden state sequences of the HMMs are found using the Viterbi algorithm; (c) The coupling parameters are determined [4]. While training in different emotional corpus, the conditional probability can be used as the probability of the emotion recognition.

In (1), $\hat{U}^{a}$ is asked to be reliably estimated, while in (2) $\hat{U}^{v}$ has to be exactly determined. Previous studies (e.g., [4]) have proven that the first structure will generate more stable results in bimodal emotion recognition because the hidden states of the speech HMM can be estimated more reliably. The coupling parameter in (1) represents the conditional probability distribution of visual observation in visual component HMM, given states in audio component HMM. To use (1) and (2) for audio-visual mapping, we extended the discrete coupling parameters in [4] to the continuous observation as followed

$$b_{j}\left(O^{v}\right) = \sum_{k=1}^{K} C_{jk} N\left(O^{v}|\mu_{jk}, \sum_{jk}\right), \quad 1 \leqslant j \leqslant N, \tag{3}$$

where O^{v} is the visual features being modeled in visual component HMM, and this mixture Gaussian is the visual observation in audio state j. $N(O^{v} \mid \mu_{jk}, \Sigma_{jk})$ is the Gaussian distributed density component related to audio state j, μ_{jk} and Σ_{jk} are the kth mean vector and kth covariance matrix, C_{jk} is the mixture weight, and K is the number of Gaussian functions in the GMM.

Utterance reduction with inversion method

The utterance reduction in visual parameters is trying to find the relationship between the visual parameters and the content-related audio parameters. It can be solved by finding the residual between the real visual parameters and the outputs of the audio-visual mapping which is trained in the neutral expressed audio-visual corpus. (see Figure 1)

Fused HMM inversion model

To find the most possible visual parameters corresponding to content-related speech parameters within the framework of multi-stream HMM, we need to find the best aligned HMM states between two component HMMs. The HMM inversion algorithm was proposed in [4] and applied to the robust speech recognition. Then Choi et al. [54] used HMM inversion in dynamic audio-visual mapping, whose usefulness has been demonstrated in [55]. However, it can only solve the problem of the single HMM. In our study, we extend this work by introducing a Baum-Welch HMM inversion method for multi-stream HMMs.

As shown by Choi and Hwang [54], Xie and Liu [56], and Moon and Hwang [57], the optimal visual counterpart $\hat{O}^{v}$ can be formulated as the optimization of the following object function, $L(O^{v}) = \log P(O^{a}, O^{v}|\lambda^{av})$, given an audio input, where O^{a} is the audio features, and λ^{av} is the parameters of the fused HMM model. The optimization can be found by iteratively maximizing the auxiliary function $Q(\lambda^{av}, \lambda^{av}; O^{a}, O^{v}, \bar{O}^{v})$ based on the Baum-Welch method,

$$\hat{O}^{v} = \arg\max_{\bar{O}^{v}} Q(\lambda^{av}, \lambda^{av}; O^{a}, O^{v}, \bar{O}^{v}) \tag{4}$$

where O^{v} and $\bar{O}^{v}$denote the old and new visual vector sequence, respectively.

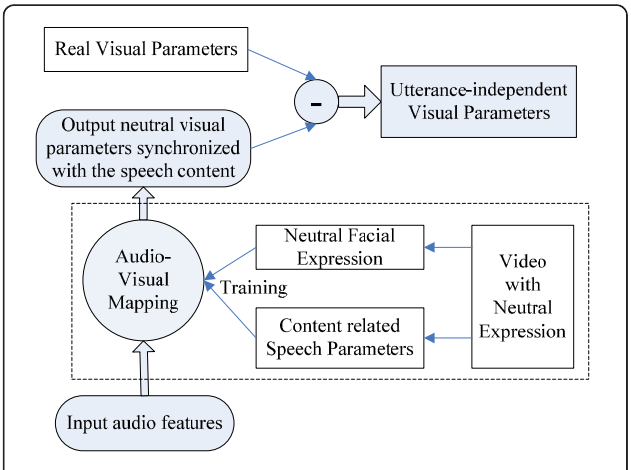

Figure 1 Framework of utterance reduction using audio-visual mapping.

In this study, the fused model can be presented as

$$P(O^a, O^v|\lambda^{av}) = \kappa_1 P(O^a)P(O^v|\hat{U}^a) + \kappa_2 P(O^v|\hat{U}^v) \quad (5)$$

where for constants $\kappa_1 \geq \kappa_2 \geq 0$ with $\kappa_1 + \kappa_2 = 1$, $\kappa_1 > \kappa_2$. It is obvious that the two HMMs will all affect the synthesis result, but have different reliability. It is an easy extension of the presentation in [58].

The objective function can be expressed as

$$\begin{aligned}\arg\max_{\bar{O}^v} L(O^v) &= \arg\max_{\bar{O}^v}\left[\kappa_1 \log P(O^v|\hat{U}^a) + \kappa_2 \log P(O^v)\right] \\ &= \arg\max_{\bar{O}^v}\left[\kappa_1 \sum_{m^{av}} P(O^v, m^{av}|\hat{U}^a, \lambda^{av}) + \kappa_2 \log \sum_{U^v}\sum_{m^{av}} \log P(O^v, U^v, m^v|\lambda^{av})\right]\end{aligned} \quad (6)$$

where m^{av} is the vector $m^{av} = \left\{m^{av}_{\hat{U}^a_1 1}, m^{av}_{\hat{U}^a_2 2}, \ldots m^{av}_{\hat{U}^a_T T}\right\}$, and m^v is the vector $m^v = \left\{m^v_{\hat{U}^v_1 1}, m^v_{\hat{U}^v_2 2}, \ldots m^v_{\hat{U}^v_T T}\right\}$ that indicates the mixture component for each state at each time.

The auxiliary function can be derived as

$$\Delta l = L(\bar{O}^v) - L(O^v) \geqslant Q(\lambda^{av}, \lambda^{av}; O^a, O^v, \bar{O}^v) + \text{const} \quad (7)$$

So

$$\begin{aligned}Q(\lambda^{av}, \lambda^{av}; O^a, O^v, \bar{O}^v) = \kappa_1 \sum_{l=1}^{M^{av}} h_l \sum_{t=1}^{T} \log c_{\hat{U}^a_t l} + \kappa_1 \sum_{l=1}^{M^{av}} h_l \sum_{t=1}^{T} \log b_{\hat{U}^a_t l}(\bar{o}^v_t) \\ + \kappa_2 \sum_{i=1}^{N}\sum_{l=1}^{M^v}\sum_{t=1}^{T} \log c_{il} \cdot H_{ilt} + \kappa_2 \sum_{i=1}^{N}\sum_{l=1}^{M^v}\sum_{t=1}^{T} \log b_{il}(\bar{o}^v_t) \cdot H_{ilt}\end{aligned} \quad (8)$$

where

$$h_l = \frac{\prod_{t=1}^{T} P(O^v_t, m^{av}_{\hat{U}^a_t t} = l|\hat{U}^a_t, \lambda^{av})}{\sum_{n=1}^{M^{av}} \prod_{t=1}^{T} P(O^v_t, m^{av}_{\hat{U}^a_t t} = l|\hat{U}^a_t, \lambda^{av})} \quad (9)$$

$$H_{ilt} = \frac{P\left(O^v, U^v_t = i, m^v_{U^v_t t} = l|\lambda^{av}\right)}{\sum_{i=1}^{N}\sum_{l=1}^{M^v}\sum_{t=1}^{T} P\left(O^v, U^v_t = i, m^v_{U^v_t t} = l|\lambda^{av}\right)} \quad (10)$$

$$P\left(O^v, U^v_t = i, m^v_{U^v_t t} = l|\lambda^{av}\right) = \sum_{i=1}^{N} \alpha^v_j(t-1) a_{ji} c_{il} b_{il}(O^v_t) \beta^v_i(t) \quad (11)$$

By setting the derivative of $Q(\lambda^{av}, \lambda^{av}; O^a, O^v, \bar{O}^v)$ with respect to $\bar{o}^v_t$ to zeros, i.e., $\frac{\partial Q(\lambda^{av}, \lambda^{av}; O^a, O^v, \bar{O}^v)}{\partial \bar{o}^v_t} = 0$, we can find the re-estimated the $\bar{o}^v_t$

$$\bar{o}^v_t = \frac{\kappa_1 \sum_{l=1}^{M^{av}} h_l \cdot \sum_l^{-1} \cdot \mu_l + \kappa_2 \sum_{i=1}^{N}\sum_{l=1}^{M} H_{ilt} \cdot \sum_l^{-1} \cdot \mu_{il}}{\kappa_1 \sum_{l=1}^{M^{av}} h_l \cdot \sum_l^{-1} + \kappa_2 \sum_{i=1}^{N}\sum_{l=1}^{M} H_{ilt} \cdot \sum_l^{-1}} \quad (12)$$

$\bar{o}^v_t$ is then used for the visual residual computing for utterance reduction.

In our study, we have classified all visual parameters into several visual clusters (see section "Two-layer clustering in visual configuration") and choose a four-state right-left HMM model for each cluster. The visual cluster represents the deformation of the face shape. Based on the time synchronization between audio and visual representation in the neutral audio-visual corpus, the sequences for each clustered visual feature also have their own corresponding audio frames. Then, for each cluster sequence, we also train a three-state right-left HMM model for the audio data. The best hidden state sequences of the audio component HMMs are found using the Viterbi algorithm, while a Gaussian Mixture Model (GMM) is fitted on the visual frame data for each estimated hidden state.

Two-Layer clustering in visual configuration

If we do not control the amount of clusters, we will have a very large number of audio-visual candidates when compared to phoneme-based units. To reduce the computing complexity, therefore, we use a two-layer framework by classifying the corpus into a series of subsets by considering both visual and audio configurations. This two-layer framework is performed by the following steps:

In the first layer, we only classify all audio-visual subsequences into 40 clusters according to the amount of the phoneme set. Each cluster center represents the repertoire of facial specification. Furthermore, each cluster is classified into sub-clusters by the *k*-means method. These sub-clusters constitute the second layer. Then, we can train more Fused HMMs for sub-clusters below the representative Fused HMM.

In the audio visual mapping, we use Fused HMMs of the first layer to select the best cluster. Then all fused HMMs of the second layer within the selected cluster will be further checked to find the best sub-cluster according to the concatenation (smoothing) cost between two visual frames. The target visual output will be got from these selected sub-clusters. The visual output will be more smoothed using the whole subsets, as shown in Figure 2a, compared with b.

Simplifying fused HMM inversion to GMM mapping

While we replace audio HMM states with audio observations in (3), we can find the function (3) will be simply changed to a GMM which combines the audio-visual observations directly,

$$b_j(O^v) = \sum_{k=1}^{K} C_k N(O^v, O^a|\mu^a_k, \mu^v_k, \sum_k^{av}), \quad 1 \leq a \leq M \quad (13)$$

where O^a is the audio observation within a total number M, $\mu_k{}^v$ and $\mu_k{}^a$ are kth mean vectors of visual

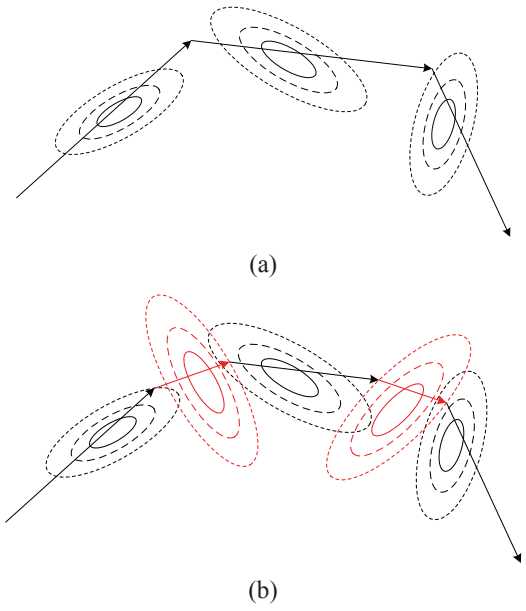

Figure 2 The visual output by using different layers, **(a)** only using the first layer Fused HMMs. **(b)** using the second layer Fused HMMs.

observations and audio observations and Σ_k^{av} are the kth covariance matrix of both audio and visual observations.

The GMM conversion will reduce the computing complexities compared with inversion method, however, it weakens the time-series analysis in the audio-visual processing by simply replacing HMM states with real audio observations. After the GMMs are trained by the EM method, the optimal estimate of neutral facial deformation ($\hat{\hat{O}}^v$) given by the content related speech parameters (O^a) can be obtained according to the transform function of conditional expectation,

$$\hat{\hat{O}}^v = E\left\{\hat{\hat{O}}^v/O^a\right\} = \sum_{k=1}^{K} p_k(O^a)\left[\mu_k^v + \sum_k^{av}\left(\sum_k^a\right)^{-1}(O^a - \mu_k^a)\right] \quad (14)$$

$$\hat{\hat{O}}^v = E\left\{\hat{\hat{O}}^v/O^a\right\} = \sum_{k=1}^{K} p_k(O^a)\left[\mu_k^v + \sum_k^{av}\left(\sum_k^a\right)^{-1}(O^a - \mu_k^a)\right]$$

where $\sum_k^a$ is the covariance matrix in audio vector space, $p_k(O^a)$ is the probability that the given audio observation belongs to the mixture component (Figure 3)

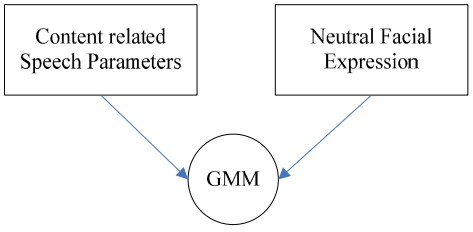

Figure 3 Framework of utterance reduction by using GMM.

$$p_k(O^a) = \frac{w_k N(O^a; \mu_k^a, \sum_k^a)}{\sum_{k=1}^{K} w_k N(O^a; \mu_k^a, \sum_k^a)} \quad (15)$$

Experiments and discussion

Visual parameters

The usually extracted facial features are either geometric features such as the shapes of the facial components (eyes, mouth, etc.) (e.g., [52]) and the location of facial salient points (corners of the eyes, mouth, etc.) (e.g., [59]) or appearance features (e.g., [43,44,60]), representing the facial texture using Fisher's linear discriminant analysis (FDA) [61], principal component analysis (PCA) (e.g., [62,63]), independent component analysis (ICA) [64], and Gabor wavelets [65], Haar features [66], spatial ratio face template [67], or manifold subspace [41].

In this article, we do not want to argue which parameters are the best for the recognition, but only focus on the modality fusing method and the utterance reduction method. We then choose the geometric features by using 20 salient facial points (see Figure 4) including six brow corners and mid-points (p_1, p_2, p_3, p_4, p_5, and p_6), eight eye corners and mid-points (p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13}, and p_{14}), two nostrils (p_{15} and p_{16}), and four mouth corners and mid-points (p_{17}, p_{18} p_{19}, and p_{20}) to represent the facial shape. This representation is a trade-off between the modeling capacity of the facial expressive structure and the efficiency of feature extraction.

To better describe facial expression information, we divide the facial shape into two regions, the upper and lower regions. In the upper region, we have $\varphi_2 = \frac{1}{2}(\theta + \gamma)$, $\varphi_2 = \frac{1}{2}(\theta + \gamma)$, $d_1 = \frac{1}{2}(|\overline{p_3p_9}| + |\overline{p_4p_{13}}|)$, where θ, γ, α, and β are angles defined in Figure 4. We also define two directions X and Y, which are collinear with the vectors $\overline{p_3p_4}$ and $\overline{p_{18}p_{20}}$, respectively. $|\overline{p_3p_9}|$ and $|\overline{p_4p_{13}}|$ are the distances of vectors $\overline{p_3p_9}$ and $\overline{p_4p_{13}}$.

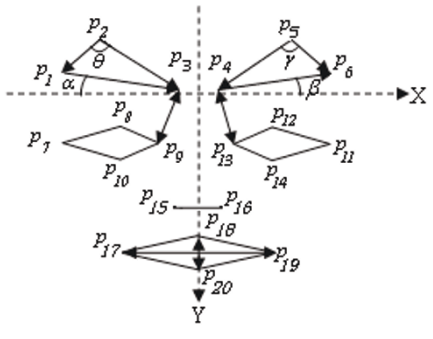

Figure 4 Face shape represented by some salient facial feature points.

In the lower region, we have $d_2 = \overline{|p_{17}p_{19}|}$, $d_3 = \overline{|p_{18}p_{20}|}$, where $\overline{|p_{17}p_{19}|}$ and $\overline{|p_{18}p_{20}|}$ are the distances of vectors $\overline{|p_{17}p_{19}}$ and $\overline{|p_{18}p_{20}}$.

As inspired by the idea of using averaging facial shapes [51], we also use a dynamic smoothing smoothed facial parameters which are tracked as a search in point distribution models (PDMs) [52].

In PDMs, each facial shape is approximately represented by a linear combination of basic variations as

$$X = \bar{X} + P\eta, \tag{16}$$

where $\bar{X}$is the mean facial shape

$$\bar{X} = \frac{1}{L}\sum_{i=1}^{L}\hat{X}_i \tag{17}$$

$_P$ is a matrix

$$P = (p_1, \ldots p_n, \ldots p_{2N}), \tag{18}$$

for which columns p_n, $n \in \{1,2,K,2N\}$, denote all facial variation directions. η is the PDM representation of the facial shape,

$$\eta = (b_1, \ldots b_n, \ldots b_{2N}), \tag{19}$$

where $b_n$indicates how much variation is exhibited for each direction.

The method for calculation of P and η has been reported in [52]. With the suggestion that the utterance expression makes a kind of random variations in facial expression, this mean facial shape $\bar{X}$ can be considered as the smoothed facial expression for utterance reduction. However, to get the dynamic facial features in time sequences, we segment the whole facial utterance into several small periods. For each period, we get a mean facial shape of the PDM and concatenate these mean facial shapes together for the emotion recognition (Figure 5).

Audio parameters

With our existing research study on audio parameters useful for emotion speech classification [22] and speech-driven talkinghead [68], we have got that the prosody parameters including F0, speed, energy, etc., have a good "resolving power" for emotion expression while some spectrum parameters including MFCCs [69] have strong influence on the utterance expression in face. To simplify the study, we only use the MFCCs to reduce utterance expression in face.

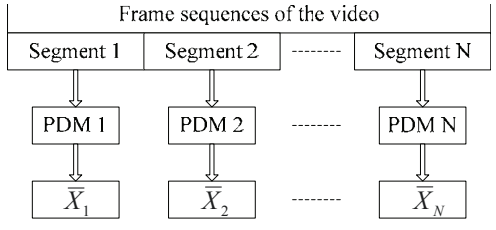

Figure 5 Framework of utterance reduction by PDMs.

Training database

Most of current spontaneous emotion recognition system used datasets which were collected in the following data-elicitation scenarios: human-human conversation (e.g., [20,60,70-73]). In our study, the training database was collected from 30 subjects (15 males and 15 females) in National Laboratory of Pattern Recognition (NLPR). In each time, one of them was asked to sit in the noise reduction environment, and to talk to us for about 2 h with exaggerated expression during conversation, like drama actors/actresses. They were simply asked to display facial expressions and speak in natural way. After recording of all speakers, the data was labeled by three annotators with "happiness," "sadness," "anger," "fear," "surprise," and "neutral" in piece by piece. We selected 400 sentences for each emotion (about 1.8-h data) for the training. The SPTK toolkit [69] and the AAM method [74] were used to get the audio and visual features. For each emotion state, 90% of the data are used for training while others are used for testing. The Fused HMM inversion and GMM training are based on the whole training set of the neutral videos (Figure 6).

To make the study comparable with others, we also use Belfast Naturalistic Database for testing. Some samples of Belfast corpus are shown in Figure 7.

The results of emotion recognition based on our three utterance reduction methods

Figure 8 shows the results of emotion recognition in NLPR's emotional corpus by the method of Fused HMM (MHMM), and the utterance reduction methods which are combined with Fused HMM Inversion (MHMM + Inv), GMM (MHMM + GMM), and PDM (MHMM + PDM). From the results, it is clear that the utterance reduction methods can improve the emotion recognition results than that without utterance reduction models.

We can find HMM-inversion-based method is better than GMM-based method. Using the HMM state or the center of the visual clusters as the outputs of visual parameters, the HMM inversion can simulate the

Figure 6 Samples selected from NLPR Emotional Database.

Figure 7 Spontaneous data selected from Belfast Naturalistic Database.

detailed facial deformation while speaking. In our previous study, we even use it for the system of speech-driven facial animation [68]. In our study, it gets the better utterance reduction results than GMM-based method which may give an over-smoothing visual parameter outputs. The results of "neutral" and "fear" in GMM-based method are even worse than that without utterance reduction method. The results confirm the report in [75], which proved the over-smoothing problem while using GMM for conversion problems.

The results by only using PDM model and MHMM are not so good, compared with two other utterance reduction methods in thearticle. However, it is slightly better than that without utterance reduction methods. As the facial expression of utterance presentation cannot be considered as the random visual variation, the average face shape based on PDM simplifies the problem. Especially if the same phonemes are repeated frequently in a short period, the PDM mean face shape still consists of utterance information which may be easily confused with some emotion states. This confusing more happens between the phoneme "a" and "surprise," or "i" and "angry." Thus, this kind of improvement of only using PDM is poor.

In the experiments, we also made an interesting test by combine the PDM with HMM inversion and GMM. We first use HMM inversion to reduce the influence of utterance after the visual tracking. Then the PDM method is used for the further smoothing of facial deformation. This is really helpful because we always get the random variation after we calculate the residue between the real input visual parameters and the outputs from audio-visual conversion models. Results in Figure 8 confirm our proposal. The recognition accuracies are improved and emotion confusions are decreased.

The further tests were also made based on the Belfast Naturalistic Database. Due to the different emotional presentation styles, only four emotion states, "happiness," "sadness,", "anger," and "surprise" are selected from the Belfast database for the experiment. The results are shown in Figure 9.

From the results, we can find that the conclusion we got from NLPR's emotional corpus is also suitable for Belfast Naturalistic Database, however, most of the emotion recognition rates are lower than that from NLPR's corpus. Major reason is that the NLPR's corpus is a kind of Posed corpus. The speakers were asked to sit in front of the camera and were not allowed to do the complicated action, e.g., looking around, nodding, etc. The speech is also recorded in noise reduction environment. Compared with the Belfast corpus which is more spontaneous and contains more actions, the emotion recognition results on NLPR's corpus are higher than that from Belfast corpus. The difficulty of facial expression tracking might be another reason to cause lower emotion recognition rate in Belfast database.

Comparisons with uni-modal methods

To compare with methods using uni-modal parameters, we performed experiments in which parameters extracted from a single audio or visual channel were inputted into a HMM emotion recognition approach. The testing results from NLPR database are shown in Figure 10.

Compared to the two methods using the uni-modal parameters individually, the results confirm that the compensation between the two channels in the bimodal method improves the performance of emotion recognition.

Comparison with other bimodal fusion methods

To make the further comparison with other studies, we repeated four typical methods, the SVM method [13], CART method [11], Boosting method [53], and the rule-

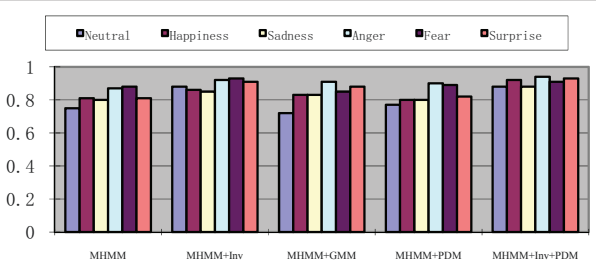

Figure 8 The comparison among MHMM, MHMM + Inv and MHMM + GMM methods based on NLPR Emotional Database.

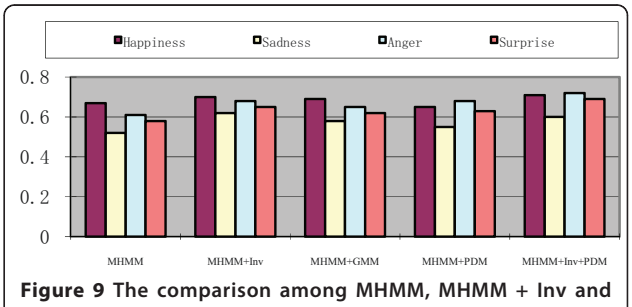

Figure 9 The comparison among MHMM, MHMM + Inv and MHMM + GMM methods based on Belfast Naturalistic Database.

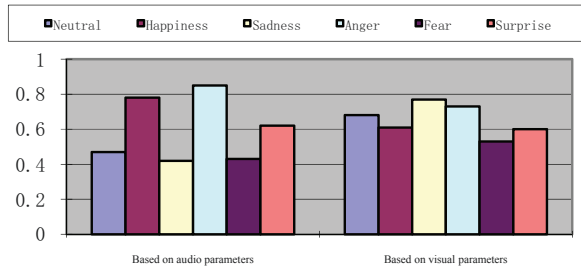

Figure 10 The comparison among the emotion recognition results only based on audio or visual parameters.

based decision fusion method [11]. The results from NLPR emotional database are shown in Figure 11.

Although the SVM and boosting methods are the fine classifiers, their results are slightly poorer than our MHMM + Inv method (see Figure 11). This is a clear demonstration to show the importance of including the time serials in bimodal emotion recognition. By integrating the utterance reduction from audio to visual parameters in a reasonable way, a more efficient emotion recognition system is able to be developed.

Figure 11 shows that rule-based decision fusion algorithm is the worst method of emotion recognition tested in our data. As different emotions may be expressed in different ways, a fixed modality-specific dominance measured by some rules for all people or emotions is not enough.

Emotion recognition in paralinguistic expression

It is also very interesting to know emotion recognition results while the subjects only speak only one or two emotion-related paralinguistic words, e.g., "[A]", "[x ⊠]", "[ən]", etc. Among them, "[A]" might be used for "surprise" expression, "[x ⊠]" is a typical "happy" mood, "[ən]" could be related to "angry". However, the expressions are various among different subjects. The expression of paralinguistic words gives us a hard problem. Do our utterance-independent models also work for these problems?

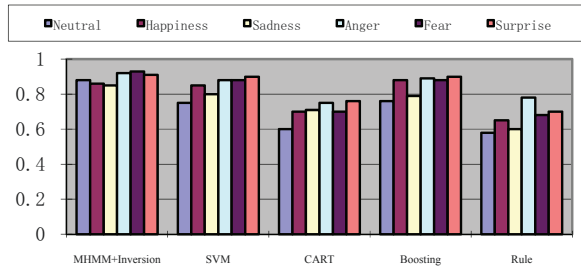

Figure 11 The comparison among MHMM + Inv, SVM, CART, Boosting, rule-based decision fusion methods.

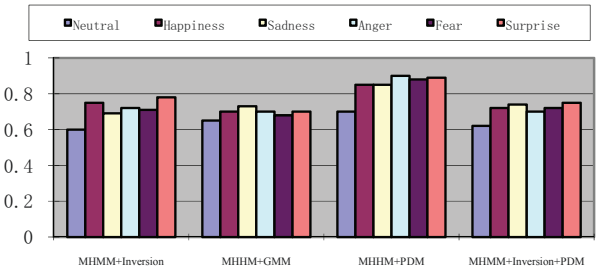

Figure 12 The comparing emotion recognition of paralinguistic words among MHHM + Inv, MHMM + GMM, MHHM + PDM, and MHMM + Inv + PDM methods.

We selected 121 emotional sentences from NLPR database which consist of these paralinguistic words for testing and the results are shown in Figures 12 and 13. There are also some situations that the emotions are influenced by some modal words. But these problems are out of discussion in this article.

From Figures 12 and 13, unfortunately, we found the MHMM + Inv and MHMM + GMM methods do not give the good results as we expect. It tells that emotions sometimes can hardly be separated from speech content while in paralinguistic expressions. However, we find the hybrid method of MHMM and PDM give the best results among all methods. In general, the fused model which integrates the time sequences still works better than other fusion methods even in paralinguistic expressions. And, the smoothed facial shapes with PDM method can always improve the recognition accuracy.

Tests of time delay

To use the methods in real applications, we calculated the time delay of the major models and list them in Table 1. The two indicators are the average emotion recognition rate for the whole database (mean accuracy) and the average running time per image (in millisecond). It shows that the MHMM + Inv method can get the best average emotion recognition rate, while the time consuming of this method is also compared with others.

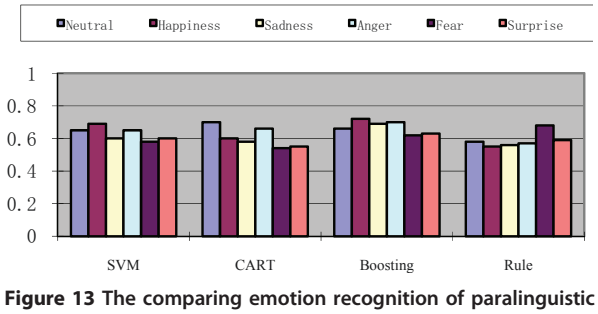

Figure 13 The comparing emotion recognition of paralinguistic words among SVM, CART, Boosting, and rule-based decision fusion methods.

Table 1 Time analysis of our systems compared with the SVM-based system

Method	SVM	MHMM	MHMM + Inv	MHMM + GMM	MHMM + Inv + PDM
Mean accuracy	0.786	0.825	0.892	0.852	0.91
Time (ms)	0.32	0.112	0.154	0.210	0.456

Conclusion and future study

This article presented a framework using MHMM for bimodal emotion recognition. Six different emotions are classified by integrating both audio and visual input channels in communication. Within this framework, the article introduces an utterance reduction method to improve the quality of visual parameters in emotion recognition by introducing the Fused HMM inversion model. To reduce the computing complexity the Inversion model can be further simplified to a GMM. The PDM is also introduced to smooth the visual tracking results.

We took several experiments to discuss our methods. The final results show that the hybrid method which consists of MHMM, HMM inversion, and PDM work best in most of cases except some emotions expressed by paralinguistic words. In paralinguistic expression, the method combining both MHMM and PDM works best. Compared with previous bimodal emotion recognition methods, e.g., SVM, CART, Boosting, and rule-based decision fusion methods, our methods can give the better emotion recognition results.

As the current research still focuses on the six basic emotions, in the future, more databases with spontaneous expressions will be recorded. Fused emotions, e. g., "painful", etc., will be added. Some dataset will be collected from TV directly. Additionally, we will pay more attention on classifications for more paralinguistic information in spontaneous conversation.

Abbreviations

FDA: Fisher's linear discriminant analysis; GMM: Gaussian Mixture Model; HCIs: human-computer interactions; ICA: independent component analysis; MHMM: Multistream Hidden Markov Model; PDMs: point distribution models; PDMs: point distribution models; PCA: principal component analysis.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (grants 60575032, 60873160, and 90820303) and the 863 Program (Grant 2009AA01Z320).

Competing interests

The authors declare that they have no competing interests.

Received: 2 August 2010 Accepted: 13 May 2011
Published: 13 May 2011

References

1. M Pantic, LJM Rothkrantz, Toward an affect-sensitive multimodal human-computer interaction. Proc IEEE. **91**(9):1370–1390 (2003). doi:10.1109/JPROC.2003.817122
2. M Pantic, LJM Rothkrantz, Automatic analysis of facial expressions: the state of the art. IEEE Trans PAMI. **22**(12):1424–1445 (2000). doi:10.1109/34.895976
3. Z Zeng, J Tu, M Liu, TS Huang, B Pianfetti, D Roth, S Levinson, Audio visual affect recognition. IEEE Trans Multimedia. **9**(2):424–428 (2007)
4. Z Zeng, J Tu, P Pianfetti, M Liu, T Zhang, Z Zhang, TS Huang, S Levinson, Audio visual affect recognition through multi stream fused HMM for HCI. Proceedings of the International Conference on Computer Vision and Pattern Recognition. 967–972 (2005)
5. Z Zeng, M Pantic, GI Roisman, TS Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans PAMI. **31**(1):39–58 (2009)
6. R Cowie, E Douglas-Cowie, Emotion recognition in human-computer interaction. IEEE Signal Process Mag. 33–80 (2001)
7. LS Chen, TS Huang, T Miyasato, R Nakatsu, Multimodal human emotion/expression recognition. Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition. 366–371 (1998)
8. B Fasel, J Luttin, Automatic facial expression analysis: a survey. Pattern Recog. **36**(1):259–275 (2003). doi:10.1016/S0031-3203(02)00052-3
9. M Song, J Bu, C Chen, N Li, Audio-visual based emotion recognition: a new approach. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1020–1025 (2004)
10. C Busso, Z Deng, S Yildirim., *et al*, Analysis of emotion recognition using facial expressions, speech and multimodal information. Proceedings of the 6th International Conference on Multimedia Interfaces. 205–211 (2004)
11. D Silva, T Miyasato, R Nakatsu, Facial emotion recognition using multi-modal information. Proceedings of the International Conference on Information and Communications Security. 397–401 (1997)
12. C Liyanage, D Silva, CN Pei, Bimodal emotion recognition. Proceedings of the Forth IEEE International Conference on Automatic Face and Gesture Recognition. 332–335 (2000)
13. CY Chen, YK Huang, P Cook, Visual/acoustic emotion recognition. Proceedings of the International Conference on Multimedia and Expo. 1468–1471 (2005)
14. T Balomenos, A Raouzaiou, S Ioannou, A Drosopoulos, K Karpouzis, S Kollias, Emotion analysis in man-machine interaction systems. LNCS. **3361**, 318–328 (2005)
15. A Jaimes, N Sebe, Multimodal human computer interaction: a survey. Proceedings of the Workshop on Human Computer Interaction in conjunction with ICCV. (2005)
16. Q Ji, P Lan, C Looney, A probabilistic framework for modeling and real-time monitoring human fatigue. IEEE Trans SMC A. **36**(5):862–875 (2006)
17. AB Ashraf, S Lucey, JF Cohn, T Chen, Z Ambadar, KM Prkachin, PE Solomon, The painful face: pain expression recognition using active appearance models. Proceedings of the International Conference on Multimodal Interfaces. 9–14P (2007)
18. A Kapoor, W Burleson, RW Picard, Automatic prediction of frustration. Proc Int J Hum Comput Stud. **65**(8):724–736 (2007). doi:10.1016/j.ijhcs.2007.02.003
19. K Karpouzis, G Caridakis, L Kessous, N Amir, A Raouzaiou, L Malatesta, S Kollias, Modeling naturalistic affective states via facial, vocal, and bodily expression recognition, Artifical Intelligence for Human Computing, Lecture notes in artificial intelligence. (Springer, Berlin, 2007)**4451**, , pp. 91–112
20. GC Littlewort, MS Bartlett, K Lee, Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. Proceedings of the International Conference on Multimodal Interfaces. 15–21 (2007)
21. F Dellaert, T Polzin, A Waibel, Recognizing emotion in speech. *Proceedings of the International Conference on Spoken Language Processing*. (Philadelphia, PA, 1996), pp. 1970–1973
22. JH Tao, YG Kang, Features importance analysis for emotion speech classification. Proceedings of the 1st International Conference on Affective Computing and Intelligence Interaction. 449–457 (2005)
23. CM Lee, S Yildirim, M Bulut, A Kazemzadeh, C Busso, ZG Deng, S Lee, S Narayanan, Emotion recognition based on phoneme classes. Proceedings of the International Conference on Spoken Language Processing. 889–892 (2004)
24. B Schuller, G Rigoll, M Lang, Hidden Markov model based speech emotion recognition. Proc ICASSP. **2**, 1–4 (2003)

25. D Roy, A Pentland, Automatic spoken affect classification and analysis. Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition. 363–367 (1996)
26. N Campbell, Perception of affect in speech - towards an automatic processing of paralinguistic information in spoken conversation. *Proceedings of the International Conference on Spoken Language Processing*. (Jeju, 2004), pp. 881–884
27. C Gobl, AN Chasaide, The role of voice quality in communicating emotion, mood and attitude. Speech Commun. **40**, 189–212 (2003). doi:10.1016/S0167-6393(02)00082-1
28. R Tato, R Santos, R Kompe, JM Pardo, Emotional space improves emotion recognition. *Proceedings of the International Conference on Spoken Language Processing*. (Denver, CO, 2002), pp. 2029–2032
29. A Mehrabian, Communication without words. Psychol Today. **2**(4):53–56 (1968)
30. LS Chen, Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction, PhD thesis, UIUC. (2000)
31. L Chen, TS Huang, T Miyasato, R Nakatsu, Multimodal human emotion/expression recognition. Proceedings of the International Conference on Automatic Face and Gesture Recognition. 396–401 (1998)
32. BE Stein, MA Meredith, *The Merging of the Senses*. (MIT Press, Cambridge, MA, 1993)
33. Q Summerfield, Some preliminaries to a comprehensive account of audio-visual speech perception. in *Hearing by Eye*. (Lawrence Erlbaum Associates, Hillsdale, NJ, 1987), pp. 3–51
34. J Robert-Ribes, JL Schwartz, P Escudier, A comparison of models for fusion of the auditory and visual sensors in speech perception. Artif Intell Rev. **9**(4-5):323–346 (1995). doi:10.1007/BF00849043
35. G Caridakis, L Malatesta, L Kessous, N Amir, A Raouzaiou1, K Karpouzis1, Modeling naturalistic affective states via facial and vocal expression recognition. Proceedings of the International Conference on Multimodal Interfaces. 146–154 (2006)
36. P Viola, M Jones, Robust real-time face detection. Int J Comput Vision. **57**(2):137–154 (2004)
37. MS Bartlett, G Littlewort, I Fasel, JR Movellan, Real time face detection and facial expression recognition: development and application to human computer interaction. Proceedings of the CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction. 53 (2003)
38. P Michel, RE Kaliouby, Real time facial expression recognition in video using support vector machines. Proceedings of the International Conference on Multimodal Interfaces. (2003)
39. Y Wang, H Zhou, B Wu, C Huang, Real time facial expression recognition with AdaBoost. Proceedings of the International Conference on Pattern Recognition. 926–929 (2004)
40. TS Huang, L Chen, H Tao, Bimodal emotion recognition by man and machine. *Proceedings of ATR Workshop on Virtual Communication Environments*. (Japan, 1998)
41. Y Tian, T Kanade, JF Cohn, Recognizing action units for facial expression analysis. IEEE Tans PAMI. **23**(2):97–115 (2001). doi:10.1109/34.908962
42. IA Essa, AP Pentland, Facial expression recognition using a dynamic model and motion energy. Proceedings of the 5th International Conference on Computer Vision. 360–367 (1995)
43. MS Bartlett, G Littlewort, B Braathen, TJ Sejnowski, JR Movellan, A prototype for automatic recognition of spontaneous facial actions. Adv Neural Inf Process Syst. **15**, 1271–1278 (2003)
44. MS Bartlett, G Littlewort, M Frank, C Lainscsek, I Fasel, J Movellan, Fully automatic facial action recognition in spontaneous behavior. Proceedings of the International Conference on Automatic Face and Gesture Recognition. 223–230 (2006)
45. L Devillers, I Vasilescu, Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. Proceedings of the International Conference on Spoken Language Processing. 801–804 (2006)
46. L Devillers, L Vidrascu, L Lamel, Challenges in real-life emotion annotation and machine learning based detection. Neural Netw. **18**, 407–422 (2005). doi:10.1016/j.neunet.2005.03.007
47. S Hoch, F Althoff, G McGlaun, G Rigoll, Bimodal fusion of emotional data in an automotive environment. Proceedings of the ICASSP. 1085–1088 (2005)
48. M Pantic, I Patras, Dynamics of facial expression: recognition of facial actions and their temporal segments form face profile image sequences. IEEE Trans SMC B. **36**(2):433–449 (2006)
49. M Pantic, LJM Rothkrantz, Facial action recognition for facial expression analysis from static face images. IEEE Trans SMC B. **34**(3):1449–1461 (2004)
50. Z Zeng, J Tu, M Liu, T Zhang, N Rizzolo, Z Zhang, TS Huang, D Roth, S Levinson, Bimodal HCI-related affect recognition. Proceedings of the International Conference on Multimodal Interfaces. 137–143 (2004)
51. Z Zeng, J Tu, M Liu, TS Huang, Multi-stream confidence analysis for audio-visual affect recognition. Proceedings of the International Conference on Affective Computing and Intelligent Interaction. (2005)
52. CL Huang, YM Huang, Facial expression recognition using model-based feature extraction and action parameters classification. J Visual Commun Image Represent. **8**(3):278–290 (1997). doi:10.1006/jvci.1997.0359
53. RE Schapire, Y Singer, Improved boosting algorithms using confidence-rated prediction. Mach Learn. **37**, 297–336 (1999). doi:10.1023/A:1007614523901
54. K Choi, JN Hwang, Baum-Welch HMM inversion for reliable audio-to-visual conversion. Proceedings of the IEEE International Workshop Multimedia Signal Processing. 175–180 (1999)
55. SL Fu, R Gutierrez-Osuna, A Esposito, PK Kakumanu, ON Garcia, Audio/visual mapping with cross-modal hidden Markov models. IEEE Trans Multimedia. **7**(2):243–252 (2005)
56. L Xie, ZQ Liu, Speech animation using coupled hidden Markov models. Proceedings of the 18th International Conference on Pattern Recognition. 1128–1131 (2006)
57. SY Moon, JN Hwang, Robust speech recognition based on joint model and feature space optimization of hidden Markov model. IEEE Trans Neural Netw. **8**(2):194–204 (1997). doi:10.1109/72.557656
58. H Pan, S Levinson, TS Huang, ZP Liang, A fused hidden Markov model with application to bimodal speech processing. IEEE Trans Signal Process. **52**(3):573–581 (2004). doi:10.1109/TSP.2003.822353
59. MJ Black, Y Yacoob, Recognizing facial expressions in image sequences using local parameterized models of image motion. Proc Int J Comput Vision. **25**(1):23–48 (1997). doi:10.1023/A:1007977618277
60. MS Bartlett, G Littlewort, M Frank, C Lainscsek, I Fasel, J Movellan, Recognizing facial expression: machine learning and application to spontaneous behavior. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. 568–573 (2005)
61. GJ Edwards, TF Cootes, CJ Taylor, Face recognition using active appearance models. Proceedings of the European Conference on Computer Vision. **2**, 681–695 (1998)
62. A Andrew, J Calder, M Burton, A principal component analysis of facial expression. Vision Res. **41**, 179–208 (2001)
63. M Tipping, C Bishop, Probabilistic principal component analysis, *Technical Report NCRG/97/010, Neural Computing Research Group*. (Aston University, Birmingham, UK, 1997)
64. The Birmingham Cognition and Affect Project,http://www.cs.bham.ac.uk/%7Eaxs/cogaff.html
65. MJ Lyons, S Akamatsu, M Kamachi, J Gyoba, Coding facial expressions with Gabor Wavelets. Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition. 200–205 (1998)
66. J Whitehill, CW Omlin, Haar features for FACS AU recognition. Proceedings of the International Conference on Automatic Face and Gesture Recognition. 217–222 (2006)
67. K Anderson, PW McOwan, A real-time automated system for recognition of human facial expressions. IEEE Trans SMC B. **36**(1):96–105 (2006)
68. JH Tao, L Xin, PR Yin, Realistic visual speech synthesis based on hybrid concatenation method. IEEE Trans ASLP. **17**(3):469–477 (2009)
69. http://sp-tk.sourceforge.net/
70. JF Cohn, LI Reed, Z Ambadar, J Xiao, T Moriyama, Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. Proceedings of the International Conference on Systems, Man & Cybernetics. 610–616 (2004)
71. JF Cohn, KL Schmidt, The timing of facial motion in posed and spontaneous smiles. Int J Wavelets Multiresolution Inf Process. **2**, 1–12 (2004). doi:10.1142/S0219691304000317
72. R Cowie, E Douglas-Cowie, C Cox, Beyond emotion archetypes: databases for emotion modeling using neural networks. Neural Netw. **18**, 371–388 (2005). doi:10.1016/j.neunet.2005.03.002

73. E Douglas-Cowie, N Campbell, R Cowie, P Roach, Emotional speech: towards a new generation of database. Speech Commun. **40**, 33–60 (2003). doi:10.1016/S0167-6393(02)00070-5
74. S Lucey, AB Ashraf, JF Cohn, Investigating spontaneous facial action recognition through AAM representations of the face, in *Face Recognition*. (I-Tech Education and Publishing, Vienna, Austria, 2007), pp. 275–286
75. YG Kang, ZW Shuang, JH Tao, A hybrid GMM and codebook mapping method for spectral conversion. Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction. 303–310 (2005)

doi:10.1186/1687-6180-2011-4
Cite this article as: Tao *et al.*: Utterance independent bimodal emotion recognition in spontaneous communication. *EURASIP Journal on Advances in Signal Processing* 2011 **2011**:4.