NON-FRONTAL VIEW FACIAL EXPRESSION RECOGNITION BASED ON ERGODIC HIDDEN MARKOV MODEL SUPERVECTORS

Hao Tang, Mark Hasegawa-Johnson, Thomas Huang

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Email: {haotang2,huang}@ifp.uiuc.edu, jhasegaw@uiuc.edu

ABSTRACT

Automatic facial expression recognition from non-frontal views is a challenging research topic which has recently started to attract the attention of the research community. In this paper, we propose a novel approach to tackling this problem based on the ergodic hidden Markov model (EHMM) supervector representation of facial images. First, the scale-invariant feature transform (SIFT) feature vectors are extracted from a dense grid of every facial images. Next, an EHMM is trained over all facial images in the training set and is referred to as the universal background model (UBM). The UBM is then maximum a posteriori adapted to each facial image in the training and test sets to produce the image-specific EHMMs. Based on these EHMMs, we derive a supervector representation of the facial images by means of an upper bound approximation of the Kullback-Leibler divergence rate between two EHMMs. Finally, facial expression recognition is performed in the linear discriminant subspace of the EHMM supervectors using the k-nearest-neighbor classification algorithm. Our experiments of recognizing six universal facial expressions over extensive multiview facial images with seven pan angles ($-45^o \sim +45^o$) and five tilt angles ($-30^o \sim +30^o$), which are synthesized from the BU-3DFE facial expression database, show promising results compared to the state of the arts recently reported.

Keywords— Facial expression recognition, hidden Markov model, supervector representation.

1. INTRODUCTION

Automatic facial expression recognition has been a popular research topic in the areas of multimedia, computer vision, and human-computer intelligent interaction [3]. One obvious reason for this is that machine recognition of facial expressions can be potentially applied to a variety of application scenarios in various facets of the society, for example, natural human-computer interaction interfaces, behavioral science study, smart advertising, movies and games, etc. Another reason is that automatic recognition of facial expressions may play a key component role in many other tasks such as face/age/gender recognition (i.e. hard and soft biometrics) in the presence of facial expressions. For a very good general survey on this topic, the readers are advised to refer to [1, 2, 3].

In the past few decades, research on facial expression recognition has mainly been focused on a particular type of facial images, namely those facial images in which the facial pose is constrained to be frontal or near-frontal. While facial expression recognition from frontal or near-frontal facial images is of itself important, the heavily constrained facial pose greatly limits its practical utility. Nevertheless, the problem of facial expression recognition from non-frontal views have been rarely addressed in the literature. The reasons for this situation are manyfold. Compared to facial expression recognition from the frontal or near-frontal view, facial expression recognition from non-frontal views is far more challenging due to the vast intra-class variations introduced by the different facial poses. More importantly, there has been a lack of a multiview facial expression database for the research community, partly due to the many difficulties in constructing one. Without such a database, research on non-frontal view facial expression recognition has been seriously impeded.

Fortunately, the recent development of a 3D facial expression database by Yin at al. at Binghamton University, known as the BU-3DFE database [4], offers an alternative opportunity. Based on the BU-3DFE database, a few researchers have begun to explore this fascinating area of non-frontal view facial expression recognition. They synthesized multiview facial images from the BU-3DFE database by rotating the 3D facial expression models in the database to the desired poses and projecting them onto a 2D image plane. Using the synthesized multiview facial images, Hu et al. [5] investigated the problem of facial expression recognition from non-frontal views with five pan angles, namely 0^o, 30^o, 45^o, 60^o and 90^o, respectively. They combined the "geometric features", defined by the location of 83 facial feature points, and various classifiers such as nearest neighbor and the support vector machine to recognize six universal facial expressions. Zheng et al. [6] studied the same problem with the same five pan angles. Instead of using the "geometric features", they employed the "texture features", defined as the scale-invariant feature transform (SIFT) [7] feature vectors extracted from the sparse location of the 83 facial feature points. They proposed a novel method for feature selection based on minimization of an upper bound of the Bayes error and reduced the dimensionality of the SIFT feature vectors. The reduced-dimensional feature vectors were then classified with the k-nearest-neighbor (KNN) classifier. To the best of our knowledge, these above-mentioned works are pioneer in this particular area of non-frontal view facial expression recognition. However, there are two common pitfalls in both works. One is that they only investigate the non-frontal views of five coarsely-quantized pan angles, which is apparently far from being sufficient for realistic applications. The other is that their methods rely on the localization of the 83 facial feature points which were manually picked in their work. Automatic localization of facial feature points of itself is still an open research issue, especially for non-frontal view facial images. Therefore, these pitfalls have made the practical applicability of their methods very limited.

In this paper, we propose a novel approach to tackling the problem of non-frontal view facial expression recognition based on the ergodic hidden Markov model (EHMM) [8] supervector representation of facial images. Fig. 1 gives a schematic overview of the proposed approach. First, the SIFT feature vectors are extracted from a dense grid of every facial images. Next, an EHMM is trained over all facial images in the training set and is referred to as the universal background model (UBM). The UBM is then maximum a posteriori (MAP)

This work was supported in part by NSF 08-03219.

978-1-4244-7493-6/10/$26.00 ©2010 IEEE

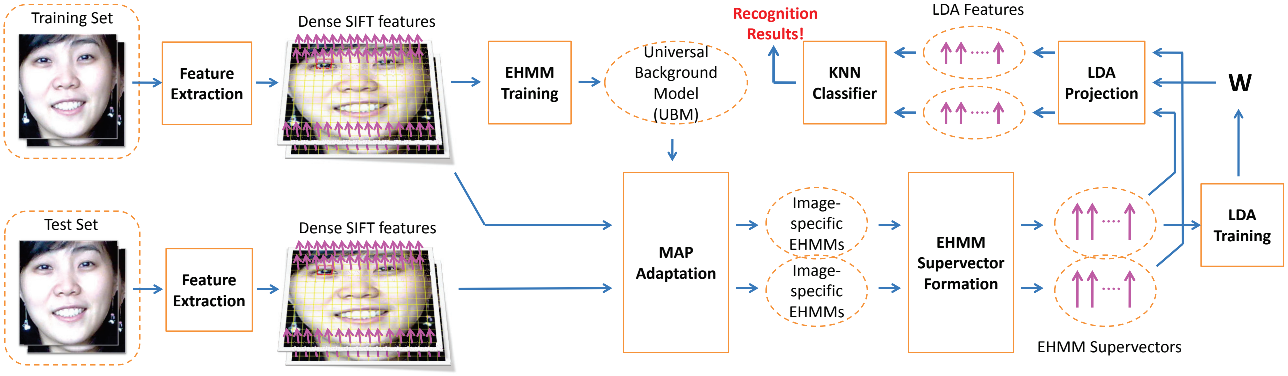


Fig. 1. The schematic overview diagram of the proposed approach.

[9] adapted to each facial image in the training and test sets to produce the image-specific EHMMs. Based on these EHMMs, we derive a supervector representation of the facial images by means of an upper bound approximation of the Kullback-Leibler divergence (KLD) [10] rate between two EHMMs. Finally, facial expression recognition is performed in the linear discriminant subspace of the EHMM supervectors (defined by a linear projection W) using the k-nearest-neighbor (KNN) classification algorithm. We conduct five-fold cross-validation experiments of recognizing six universal facial expressions over extensive multiview facial images with seven pan angles ($-45^o \sim +45^o$) and five tilt angles ($-30^o \sim +30^o$) (i.e. a total of 35 views), which are synthesized from the BU-3DFE facial expression database. Our experiment results are shown to be very promising compared to the state of the arts recently reported. Note that, however, these experiment results should be considered preliminary, as we have not had time to explore all possible design choices and to find the optimal parameters.

The key component of our proposed approach and major contribution of this paper is the novel EHMM supervector representation of facial images, which possesses the following attractive properties:

1. The representation summaries the statistical distribution of the feature vectors compactly and allows the statistical interdependence amongst the feature vectors to be modeled with a systematic underlying structure of first-order Markov chain [11].
2. The representation performs unsupervised segmentation of the facial images implicitly to reveal the local structures of the faces and to allow localized comparison of the images.
3. The representation is in a vector form ready for supervised distance metric learning to further reenforce its discriminative power for classification.

In addition to the above properties, the EHMM supervector representation of facial images is rather generic in nature and may be used with other types of images (e.g. non-facial images) and applied to other image-based recognition tasks such as face/age/gender recognition, although we have found it particularly useful for the task of non-frontal view facial expression recognition. The rationale behind this is that the EHMM supervector representation relaxes the general requirement that the facial images be fairly exactly aligned and that it is robust to occlusions in the facial images.

This paper is organized as follows. Section 2 describes how we synthesize multiview facial images from the BU-3DFE database and how we extract the dense SIFT features from the facial images. Section 3 introduces EHMM modeling and adaptation of the dense SIFT features. Section 4 gives the mathematical derivation of the EHMM supervector representation of facial images in detail. Section 5 presents the experiment results and some discussions, and Section 6 concludes the paper.

2. DATABASE AND FEATURE EXTRACTION

The BU-3DFE database is a 3D facial expression database developed by Yin et al. at Binghamton University. It was designed to sample 3D facial behaviors with different prototypical emotional states. There are a total of 100 subjects in the database among which 56 are female and 44 are male. The subjects are well distributed across different ethnic or racial ancestries, including White, Black, East-Asian, Middle-East Asian, Hispanic Latino, and others. During the recording session, each subject performed six universal facial expressions, namely anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU), and the 3D geometric and texture models of the subject were captured. For a detailed description of the database, the readers may refer to [4].

In order to synthesize multiview facial images from the BU-3DFE database, we first rotate every 3D facial expression models in the database by a certain pan angle and tilt angle. The ranges of pan and tilt angles of interest to our study are $\{-45^o, -30^o, -15^o, 0^o, 15^o, 30^o, 45^o\}$ and $\{-30^o, -15^o, 0^o, 15^o, 30^o\}$, respectively. We believe that the combination of these seven pan angles and five tilt angles is able to provide a sufficient level of quantization of the continuous non-frontal views in realistic environments. Then, we render the rotated 3D facial expression models into 2D images using OpenGL [12] with appropriate lighting simulation. Fig. 2 illustrates a sample of the multiview facial images synthesized from the BU-3DFE database in the form of a cube.

For each facial image, we extract a set of dense SIFT features. Specifically, we place a dense grid on a facial image, and extract a 128-dimensional SIFT descriptor at each node of the grid with a fixed scale and orientation. The SIFT descriptor is formed from the histogram of intensity gradients within a neighborhood window of the grid node and is a distinctive feature to represent the texture variation in this local region. In this way, a facial image is encoded by a “bag” of SIFT feature vectors, as shown in Fig. 3(a). Particular to the synthesized multiview facial images in this paper, we perform a further step. Among the extracted SIFT feature vectors, we abandon those with extremely small

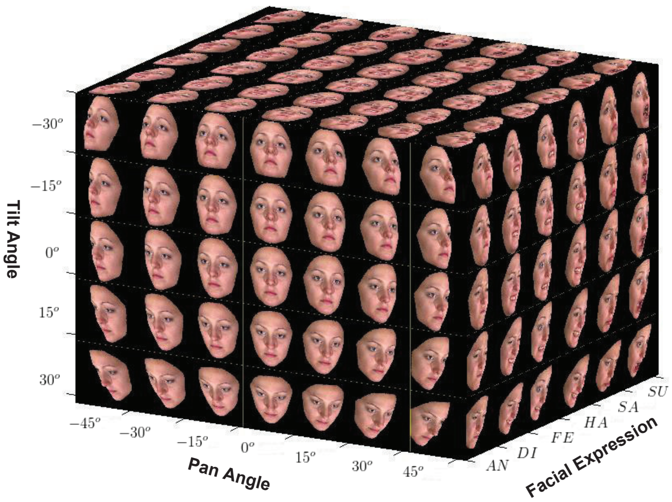
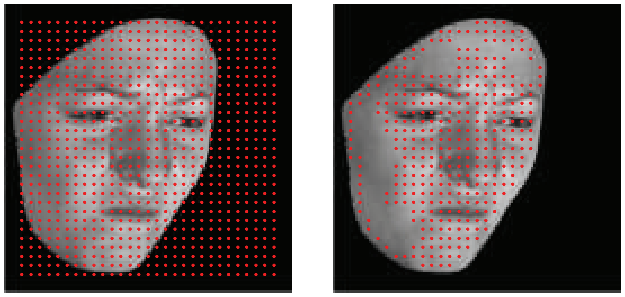


Fig. 2. A sample of the multiview facial images synthesized from the BU-3DFE database in the form of a cube.



(a) (b)

Fig. 3. (a). The SIFT feature vectors are extracted from the grid nodes, shown as red dots; (b) We abandon the SIFT feature vectors with extremely small magnitudes, which correspond to the SIFT feature vectors extracted from the black background in the images as well as those extracted from the low-contrast portion on the face.

magnitudes, which correspond to the SIFT feature vectors extracted from the black background in the images as well as those extracted from the low-contrast portion on the face, as shown in Fig. 3(b).

For each facial image, the extracted SIFT feature vectors are sorted in the order of the grid node location (x, y), with the x coordinate of the location being the fastest changing variable. Note that the ordering here is in fact not important due to the EHMM modeling described in the next section. Thus, an observation sequence, $O = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T\}$, is formed for each facial image, where $\mathbf{o}_t$, $t = 1, 2, \cdots, T$ are the individual SIFT feature vectors (a.k.a. observations) and T is the total number of SIFT feature vectors for the facial image.

3. EHMM MODELING AND ADAPTATION

The hidden Markov model (HMM) [8] is a powerful statistical tool for modeling sequential data. It is a doubly stochastic process consisting of an underlying, hidden, discrete random process which possesses the Markov property (namely a Markov chain) and an observed, discrete, continuous, or mixed discrete-continuous random process which is a probabilistic function of the underlying Markov chain [13]. The likelihood function of an HMM is given by

$$p(O|\lambda) = \sum_{q_1 q_2 \cdots q_T} \left[\pi_{q_1} b_{q_1}(\mathbf{o}_1) \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(\mathbf{o}_t) \right] \qquad (1)$$

An HMM is completely defined by its parameters $\lambda = \{A, B, \Pi\}$. Here, A is the state transition probability matrix whose entries, $a_{ij} = p(q_t = S_j | q_{t-1} = S_i)$, $1 \leq i, j \leq N$, specify the probabilities of transition from state S_i to state S_j at time t. B is the state emission probability matrix whose entries, $b_{jk} = p(o_t = v_k | q_t = S_j)$, $1 \leq j \leq N, 1 \leq k \leq M$, specify the probabilities of emitting an observation symbol v_k given that the model is in state S_j at time t. Π is the initial state probability matrix whose entries, $\pi_i = p(q_1 = S_i)$, $1 \leq i \leq N$, specify the probabilities of the model being initially in state S_i. For the case of continuous observations, the entries of the state emission probability matrix are given by continuous probability density functions, namely $b_j(o_t) = p(o_t | q_t = S_j)$, $1 \leq j \leq N$. One important class of continuous probability density functions widely used for the state emission densities of the continuous-observation HMM is the Gaussian mixture density functions of the form

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{M} c_{jk} N(\mathbf{o}_t | \mu_{jk}, \Sigma_{jk}), \quad 1 \leq j \leq N, 1 \leq k \leq M \qquad (2)$$

where M is the number of Gaussian components, c_{jk} is the k^{th} component weight, and $N(o_t | \mu_{jk}, \Sigma_{jk})$ is a multivariate Gaussian density function with mean vector μ_{jk} and covariance matrix Σ_{jk}.

An ergodic HMM (EHMM) is an HMM with the most generic typology, where all possible transitions between the individual states in the model are allowed. As a result, all the entries in the state transition probability matrix are non-zero. The EHMM can be used to statistically model the data which is non-sequential in nature (e.g. images) by presenting the data observations to the model in sequence.

An EHMM may be trained directly for each facial image with its extracted SIFT feature vectors $O = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T\}$ using the classical Bawm-Welch algorithm [14]. However, since there is only one observation sequence O available for each facial image, training an EHMM in this way will not be robust to noise. Instead, we first train an EHMM over all facial images in the training set, leading to a well-trained EHMM which we call the universal background model (UBM). The UBM is a robust EHMM which captures the global statistical distribution of the SIFT feature vectors across all facial images. Then, for each facial image, we adapt the UBM to the SIFT feature vectors of this particular facial image by the maximum a posterior (MAP) technique [15]. An EHMM obtained for each facial image in this manner is considered to be a robust statistical model for the facial image.

4. DERIVATION OF EHMM SUPERVECTORS

In the previous section, we encode facial images by statistical models (i.e. EHMMs). In this section, we derive a supervector representation of facial images based on these EHMMs. A popular distance measure for two statistical or probabilistic models, $f(\mathbf{x})$ and $g(\mathbf{x})$, is given by the Kullback-Leibler divergence (KLD) between the two models [10]

$$D(f \| g) = \int_{\mathbf{x}} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} d\mathbf{x} \qquad (3)$$

Given two N-state EHMMs with Gaussian mixture emission densities, $\lambda_1 = \{A_1, B_1, \Pi_1\}$ and $\lambda_2 = \{A_2, B_2, \Pi_2\}$, where $A_p = \{\mathbf{a}_i^{[p]}\}_{i=1}^N = \{a_{ij}^{[p]}\}_{i,j=1}^N$, $B_p = \{B_i^{[p]}\}_{i=1}^N = \{\mathbf{c}_i^{[p]} = \{c_{ik}^{[p]}\}_{k=1}^M, \{\mu_{ik}^{[p]}, \Sigma_{ik}^{[p]}\}_{k=1}^M\}_{i=1}^N$, and $\Pi_p = \{\pi_i^{[p]}\}_{i=1}^N$, and the superscript $p = 1, 2$ denotes the model index, a natural extension of the

KLD is the KLD rate (KLDR) [16, 17], defined as

$$R(\lambda_1\|\lambda_2) = \lim_{T->\infty} \frac{1}{T} D(\lambda_1\|\lambda_2) \tag{4}$$

For HMMs, the KLDR does not have a closed-form expression. However, Do [18] shows that there is a simple closed-form expression for a fairly tight upper bound of the KLDR

$$R(\lambda_1\|\lambda_2) \leq \sum_{i=1}^{N} \pi_i^{[1]} \left[D(\mathbf{a}_i^{[1]}\|\mathbf{a}_i^{[2]}) + D(B_i^{[1]}\|B_i^{[2]}) \right] \tag{5}$$

Similarly, there is no closed-form expression for the KLD between two Gaussian mixture models (GMMs). However, there exists an upper bound of the KLD between two GMMs derived from the log-sum inequality [18]

$$\begin{aligned} D(B_i^{[1]}\|B_i^{[2]}) \leq D(\mathbf{c}_i^{[1]}\|\mathbf{c}_i^{[2]}) + \\ \sum_{k=1}^{M} c_{ik}^{[1]} D\left(N(\cdot|\mu_{ik}^{[1]}, \Sigma_{ik}^{[1]}) \| N(\cdot|\mu_{ik}^{[2]}, \Sigma_{ik}^{[2]}) \right) \end{aligned} \tag{6}$$

For two single Gaussian models, however, we can obtain a closed-form solution to the KLD, as follows:

$$\begin{aligned} D\left(N(\cdot|\mu_{ik}^{[1]}, \Sigma_{ik}^{[1]}) \| N(\cdot|\mu_{ik}^{[2]}, \Sigma_{ik}^{[2]}) \right) = \\ \frac{1}{2} \log \frac{\det \Sigma_{ik}^{[2]}}{\det \Sigma_{ik}^{[1]}} + \frac{1}{2} \text{trace}(\Sigma_{ik}^{[2]^{-1}} \Sigma_{ik}^{[1]}) + \\ \frac{1}{2} (\mu_{ik}^{[1]} - \mu_{ik}^{[2]})^T \Sigma_{ik}^{[2]^{-1}} (\mu_{ik}^{[1]} - \mu_{ik}^{[2]}) - \frac{d}{2} \end{aligned} \tag{7}$$

where d is the dimensionality of the observation vector.

At this point, we may obtain an upper bound of $R(\lambda_1\|\lambda_2)$ from Eqs. 5, 6, and 7. However, the coupled interactions between the mean vectors and covariance matrices make it difficult to construct a vector form representation of the two models out of the upper bound. Therefore, we need to seek an alternative solution. Suppose the EHMMs λ_1 and λ_2 are both MAP adapted from the UBM λ. Since the amount of adaptation data is limited, in order to avoid overfiting [20], it is advantageous to only adapt the mean vectors and covariance matrices of the Gaussian mixture emission densities, and leave the Gaussian mixture component weights, state transition probabilities, and state initial probabilities unchanged. That is, we constrain that $\{\mathbf{c}_i^{[1]}\} = \{\mathbf{c}_i^{[2]}\} = \{\mathbf{c}_i\}$, $A_1 = A_2 = A$, and $\Pi_1 = \Pi_2 = \Pi$, $i = 1, 2, \cdots, N$. We further assume that all the covariance matrices are diagonal, i.e. $\Sigma_{ik} = \text{diag}(\sigma_{ik}^2)$, $i = 1, 2, \cdots, N$, $k = 1, 2, \cdots, M$, which is common practice in GMM modeling [19]. In addition, we consider a symmetric version of the KLDR

$$R_s(\lambda_1\|\lambda_2) = \frac{1}{2} \left[R(\lambda_1\|\lambda_2) + R(\lambda_2\|\lambda_1) \right] \tag{8}$$

Combining Eq. 5 and Eq. 8, we obtain

$$R_s(\lambda_1\|\lambda_2) \leq \sum_{i=1}^{N} \pi_i D_s(B_i^{[1]}\|B_i^{[2]}) \tag{9}$$

where $D_s(\cdot\|\cdot)$ is the symmetric version of the KLD defined in a way similar to Eq. 8, as follows

$$D_s(B_i^{[1]}\|B_i^{[2]}) = \frac{1}{2} \left[D(B_i^{[1]}\|B_i^{[2]}) + D(B_i^{[2]}\|B_i^{[1]}) \right] \tag{10}$$

Note that the terms involving $\mathbf{a}_1$ and $\mathbf{a}_2$ vanish by symmetry, and $\pi_i^{[1]}$ and $\pi_i^{[2]}$ are both replaced by π_i, given the assumed conditions. Campbell [21] shows that the symmetric KLD between two Gaussians can be approximated by

$$\begin{aligned} & D_s\left(N(\cdot|\mu_{ik}^{[1]}, \Sigma_{ik}^{[1]}) \| N(\cdot|\mu_{ik}^{[2]}, \Sigma_{ik}^{[2]}) \right) \\ & \approx \frac{1}{4} \text{trace}\left((\Sigma_{ik}^{[1]} - \Sigma_{ik}^{[2]}) \Sigma_{ik}^{-2} (\Sigma_{ik}^{[1]} - \Sigma_{ik}^{[2]}) \right) + \\ & \frac{1}{4} (\mu_{ik}^{[1]} - \mu_{ik}^{[2]})^T (\Sigma_{ik}^{[1]^{-1}} + \Sigma_{ik}^{[2]^{-1}}) (\mu_{ik}^{[1]} - \mu_{ik}^{[2]}) \\ & = \frac{1}{2} (\sigma_{ik}^{2\,[1]} - \sigma_{ik}^{2\,[2]})^T \frac{1}{2} \Sigma_{ik}^{-2} (\sigma_{ik}^{2\,[1]} - \sigma_{ik}^{2\,[2]}) + \\ & \frac{1}{2} (\mu_{ik}^{[1]} - \mu_{ik}^{[2]})^T (\frac{1}{2} \Sigma_{ik}^{[1]^{-1}} + \frac{1}{2} \Sigma_{ik}^{[2]^{-1}}) (\mu_{ik}^{[1]} - \mu_{ik}^{[2]}) \end{aligned} \tag{11}$$

Substituting Eq. 11 into Eq. 9, we have

$$\begin{aligned} R_s(\lambda_1\|\lambda_2) \leq & \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \pi_i c_{ik} (\sigma_{ik}^{2\,[1]} - \sigma_{ik}^{2\,[2]})^T \frac{1}{2} \Sigma_{ik}^{-2} (\sigma_{ik}^{2\,[1]} - \sigma_{ik}^{2\,[2]}) + \\ & \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \pi_i c_{ik} (\mu_{ik}^{[1]} - \mu_{ik}^{[2]})^T (\frac{1}{2} \Sigma_{ik}^{[1]^{-1}} + \frac{1}{2} \Sigma_{ik}^{[2]^{-1}}) (\mu_{ik}^{[1]} - \mu_{ik}^{[2]}) \\ \approx & \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \pi_i c_{ik} (\sigma_{ik}^{2\,[1]} - \sigma_{ik}^{2\,[2]})^T \frac{1}{2} \Sigma_{ik}^{-2} (\sigma_{ik}^{2\,[1]} - \sigma_{ik}^{2\,[2]}) + \\ & \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \pi_i c_{ik} (\mu_{ik}^{[1]} - \mu_{ik}^{[2]})^T \Sigma_{ik}^{-1} (\mu_{ik}^{[1]} - \mu_{ik}^{[2]}) \end{aligned} \tag{12}$$

Here, the terms involving $\mathbf{c}_1$ and $\mathbf{c}_2$ vanish by symmetry, $c_{ik}^{[1]}$ and $c_{ik}^{[2]}$ are both replaced by c_{ik}, and we have made a further approximation by replacing the average of the two inverse adapted covariance matrices with the inverse covariance matrix of the UBM. Through some linear algebra, we can re-write Eq. 12 as

$$\begin{aligned} R_s(\lambda_1\|\lambda_2) \leq & \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \| \sqrt{\pi_i c_{ik}/2} \Sigma_{ik}^{-1} \sigma_{ik}^{2\,[1]} - \sqrt{\pi_i c_{ik}/2} \Sigma_{ik}^{-1} \sigma_{ik}^{2\,[2]} \|^2 \\ & + \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \| \sqrt{\pi_i c_{ik} \Sigma_{ik}^{-1}} \mu_{ik}^{[1]} - \sqrt{\pi_i c_{ik} \Sigma_{ik}^{-1}} \mu_{ik}^{[2]} \|^2 \end{aligned} \tag{13}$$

where $\|\cdot\|$ denotes the L_2 norm or Euclidean distance.

Thus, if we approximate the symmetric KLDR between two EHMMs λ_1 and λ_2 by the derived upper bound given in Eq. 13, and form two supervectors as follows

$$\mathbf{s}_1 = \left[\sqrt{\pi_i c_{ik} \Sigma_{ik}^{-1}} \mu_{ik}^{[1]}; \sqrt{\pi_i c_{ik}/2} \Sigma_{ik}^{-1} \sigma_{ik}^{2\,[1]} \right]_{i=1 k=1}^{N \quad M} \tag{14}$$

$$\mathbf{s}_2 = \left[\sqrt{\pi_i c_{ik} \Sigma_{ik}^{-1}} \mu_{ik}^{[2]}; \sqrt{\pi_i c_{ik}/2} \Sigma_{ik}^{-1} \sigma_{ik}^{2\,[2]} \right]_{i=1 k=1}^{N \quad M} \tag{15}$$

then the distance between the two supervectors $\mathbf{s}_1$ and $\mathbf{s}_2$ in the Euclidean space is equivalent to the symmetric KLDR between the two corresponding EHMMs λ_1 and λ_2 (up to a constant scale $\frac{1}{2}$).

In the form of an EHMM supervector, a facial image is represented as a single data point in the Euclidean space. One significant advantage of the EHMM supervector representation is that it offers the precious opportunity to perform optimal distance metric learning (e.g., linear discriminant analysis or LDA [22]) in the Euclidean space prior to classification.

5. EXPERIMENTS

To demonstrate the effectiveness of the proposed approach for non-frontal view facial expression recognition, we conduct experiments over extensive multiview facial images that we synthesize from the BU-3DFE database as described in Section 2. The projected facial images have an original resolution of 512×512 pixels. To speedup the feature extraction process, we downsample the facial images to 128×128 pixels and convert them into grayscale images. The 128-D SIFT feature vectors are then extracted from a dense grid with 4×4 pixel spacing on every grayscale facial image with a fixed scale (12 pixels) and orientation (0^o). The 100 subjects in the database are partitioned into five groups each of which consists of 20 subjects. We consider all the six universal facial expressions, namely, anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU). Thus, each group consists of 20(subjects) $\times$ 7(pan angles) $\times$ 5(tilt angles) $\times$ 6(facial expressions) $=$ 4200 facial images. For the classification purpose, we adopt a "universal" approach in which the classifier is trained with facial images of all views and tested with facial images of any view. Such a "universal" approach is practically very useful as it does not require the facial pose of a test facial image to be known a priori or estimated. In order to obtain confident results, we perform five-fold cross validations on the database. At each fold, four out of the five groups are used for training and the remaining group is used for test. The average overall facial expression recognition rate, as well as the average recognition rates for different views and the average recognition rates for different facial expressions, are reported.

At each fold, all the facial images in the training set are used to train the UBM (3 states, 64-component Gaussian mixture emission densities), which is then MAP adapted to every facial image in both training and test sets to produce the image-specific EHMMs. The image-specific EHMMs are then used to construct the corresponding EHMM supervectors as described in Section 4. The EHMM supervectors in the training set are then used to learn a linear discriminant analysis (LDA) [22] subspace, into which all the original EHMM supervectors in both training and test sets are projected. In the LDA subspace, the KNN classifier ($k = 20$) is employed to perform facial expression recognition.

The experiment results are shown in Table 1. The rightmost column represents the average recognition error rates for different views (a total of 35 views), the bottom row represents the average recognition error rates for different facial expressions (a total of six universal facial expressions), and the bottom-right corner cell represents the average overall recognition error rate. Fig. 4 shows the average confusion matrix of the six universal facial expressions.

Our experiment results are promising compared to the state of the arts recently reported. In the work of Hu et al. [5], the best average overall facial expression recognition error rate was reported to be 33.5% with the help of a support vector machine (SVM) classifier. In the work of Zheng et al. [6], they reported a reduced overall facial expression recognition error rate of 21.7% on the same database under the same experimental settings. However, note that the experimental settings of both of their works only involve non-frontal views of five pan angles (as compared to 35 combinations of seven pan angles and five tilt angles in the experimental settings of our work). Also note that both of their works completely rely on the location of 83 facial feature points in all facial images, which have to be labeled manually. The impractical manual labeling of facial feature point locations seriously limits the applicability of their works. On the contrary, our proposed approach is fully automatic, requiring neither facial alignment nor facial feature point localization. In our experiments, we achieve an average overall facial expression recognition error rate of 24.7%, which is significantly lower than that of Hu et al.'s work and comparable to that of Zheng et al's work, but under far more challenging and useful experimental settings. In addition, we bring to the attention of the readers that our experiment results in this paper should be considered preliminary, as we have not had time to explore all possible design choices and to find the optimal parameters for the experiments which could possibly lead to even better experiment results.

Table 1. Experiment results in terms of recognition error rates. The leftmost column indicates the different views (pan and tilt angles x, y in degrees), and the top row indicates the different facial expressions.

%	AN	DI	FE	HA	SA	SU	Ave.
$-45, -30$	21.0	34.0	48.0	10.0	45.0	5.0	27.2
$-45, -15$	23.0	21.0	41.0	7.0	47.0	10.0	24.8
$-45, +0$	34.0	19.0	35.0	17.0	32.0	7.0	24.0
$-45, +15$	34.0	12.0	37.0	19.0	40.0	4.0	24.3
$-45, +30$	40.0	18.0	30.0	14.0	41.0	9.0	25.3
$-30, -30$	25.0	20.0	33.0	14.0	31.0	8.0	21.8
$-30, -15$	30.0	24.0	54.0	22.0	37.0	8.0	29.2
$-30, +0$	26.0	20.0	53.0	11.0	38.0	7.0	25.8
$-30, +15$	20.0	19.0	41.0	7.0	43.0	6.0	22.7
$-30, +30$	37.0	12.0	35.0	7.0	37.0	5.0	22.2
$-15, -30$	37.0	13.0	43.0	11.0	44.0	4.0	25.3
$-15, -15$	28.0	19.0	35.0	13.0	42.0	6.0	23.8
$-15, +0$	25.0	18.0	37.0	14.0	38.0	8.0	23.3
$-15, +15$	33.0	22.0	52.0	19.0	41.0	5.0	28.7
$-15, +30$	28.0	29.0	38.0	12.0	33.0	6.0	24.3
$+0, -30$	26.0	21.0	42.0	11.0	35.0	6.0	23.5
$+0, -15$	36.0	17.0	46.0	11.0	32.0	6.0	24.7
$+0, +0$	39.0	18.0	38.0	18.0	34.0	3.0	25.0
$+0, +15$	27.0	15.0	32.0	10.0	42.0	5.0	21.8
$+0, +30$	30.0	18.0	38.0	10.0	40.0	6.0	23.7
$+15, -30$	35.0	23.0	45.0	12.0	33.0	6.0	25.7
$+15, -15$	37.0	22.0	49.0	13.0	26.0	6.0	25.5
$+15, +0$	37.0	22.0	49.0	11.0	30.0	9.0	26.3
$+15, +15$	43.0	14.0	41.0	13.0	22.0	7.0	23.3
$+15, +30$	37.0	13.0	31.0	8.0	41.0	5.0	22.5
$+30, -30$	37.0	12.0	30.0	10.0	39.0	6.0	22.3
$+30, -15$	29.0	16.0	39.0	12.0	35.0	4.0	22.5
$+30, +0$	27.0	24.0	46.0	14.0	42.0	6.0	26.5
$+30, +15$	32.0	31.0	43.0	15.0	28.0	6.0	25.8
$+30, +30$	47.0	17.0	45.0	7.0	25.0	6.0	24.5
$+45, -30$	45.0	14.0	28.0	6.0	32.0	5.0	**21.7**
$+45, -15$	36.0	23.0	34.0	18.0	41.0	9.0	26.8
$+45, +0$	42.0	21.0	38.0	11.0	41.0	9.0	27.0
$+45, +15$	28.0	13.0	44.0	10.0	42.0	5.0	23.7
$+45, +30$	24.0	33.0	44.0	15.0	46.0	9.0	28.5
Ave.	32.4	19.6	40.4	12.3	37.0	**6.3**	**24.7**

We note that our approach yields better performance for the three facial expression categories - disgust, happiness, and surprise than for the other three facial expression categories - anger, fear, and sadness. Especially, the performance for surprise is noticeably the best among all (6.3% average recognition error rate). The performance across different views is, however, comparable to one another. This observation strongly supports that our approach is robust to the varying views of the facial images.

6. CONCLUSION

In this paper, we propose a novel approach to non-frontal view facial expression recognition based on the EHMM supervector representation of facial images. The EHMM supervector representation is demonstrated to possess a set of attractive properties and is particularly effective for the task of non-frontal view facial expression recognition. Our experiments over extensive multiview facial images with seven pan angles ($-45^o \sim +45^o$) and five tilt angles ($-30^o \sim +30^o$), which are synthesized from the BU-3DFE facial expression database, have

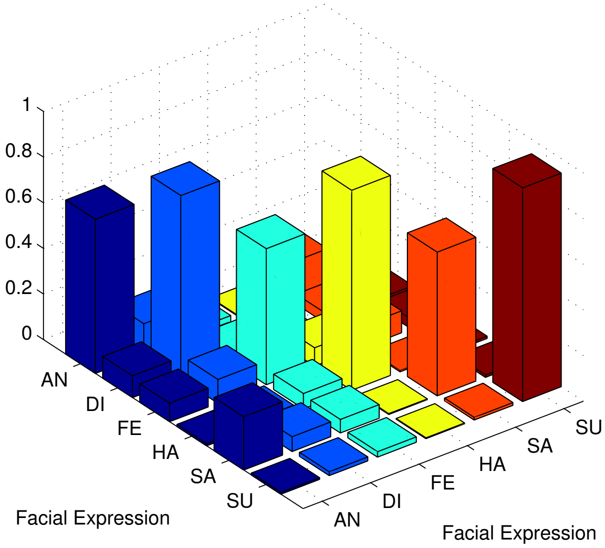


Fig. 4. The average confusion matrix of six universal facial expressions, namely anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU).

shown promising results compared to the state of the arts recently reported. Yet, the EHMM supervector representation is rather generic in nature and may be used with other types of images (e.g. non-facial images) and applied to other image-based recognition tasks such as face/age/gender recognition. One direction of our future work will be to further explore the EHMM supervector representation for new applications.

7. REFERENCES

[1] B. Fasel, and J. Luettin, "Automatic Facial Expression Analysis: A Survey," Pattern Recognition, 36:259–275, 1999.

[2] M. Pantic, and L.J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: the State of the Art," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424–1445, 2000.

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

[4] L. Yin, X. Wei, Y. Sun, J. Wang, and M.J. Rosato, "A 3D facial expression database for facial behavior research," Proceedings of 7th International Conference on Automatic Face and Gesture Recognition, pp. 211–216, 2006.

[5] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, and T.S. Huang, "A study of non-frontal-view facial expressions recognition," Proceedings of International Conference on Pattern Recognition, pp. 1–4, 2008.

[6] W. Zheng, H. Tang, Z. Lin, and T.S. Huang, "A novel approach to expression recognition from non-frontal face images," Proceedings of IEEE International Conference on Computer Vision, pp. 1901–1908, 2009.

[7] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[8] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," Proceedings of the IEEE, 77 (2): 257–286, 1989.

[9] J.-L. Gauvain, and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech and Audio Processing, (2):291–298, 2004.

[10] S. Kullback, and R. Leibler, "On Information and Sufficiency," Ann. Math. Statist., vol. 22, no. 1, pp. 79–86, 1951.

[11] J.R. Norris, Markov Chains, Cambridge series in Statistical and Probabilistic Mathematics, 1999.

[12] OpenGL: The Industry's Foundation for High Performance Graphics, http://www.opengl.org/.

[13] Stephen Levinson, Mathematical Models for Speech Technology, John Wiley & Sons, Ltd, 2005.

[14] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," Ann. Math. Statist., vol. 41, no. 1, pp. 164–171, 1970.

[15] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, no. 1-3, 2000.

[16] T.M. Cover, and J.A. Thomas, Elements of Information Theory, 1st ed., John Wiley & Sons, Inc., New York, NY, 1991.

[17] Z. Rached, F. Alalaji, and L.L. Campbell, "The Kullback-Leibler divergence rate between Markov sources," IEEE Trans. Info. Thy., vol. 50, no. 5, pp. 917–921, 2004.

[18] M. Do, and M. Vetterli, "Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models," IEEE Signal Processing Letters, vol. 10, no. 4, pp. 115118, April 2003.

[19] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, vol. 17, iss. 1-2, pp. 91–108, 1005.

[20] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. JohnWiley & Sons, Inc., 2nd edition, 2001.

[21] W.M. Campbell, "A Covariance Kernel for SVM Language Recognition," ICASSP 2008, Las Vegas, NV, 2008.

[22] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.