Hierarchical Bayesian Theme Models for Multi-pose Facial Expression Recognition

Qirong Mao, *Member, IEEE,* Qiyu Rao, Yongbin Yu, and Ming Dong*, *Member, IEEE*

Abstract—As an essential way of human emotional behavior understanding, facial expression recognition (FER) has attracted a great deal of attention in multimedia research. Most of studies are conducted in “lab-controlled” environment, and their real-world performance degenerates greatly due to factors such as head pose variations. In this paper, we propose a pose-based hierarchical Bayesian theme model to address challenging issues in multi-pose FER. Local appearance features and global geometry information are combined in our model to learn an intermediate face representation before recognizing expressions. By sharing a pool of features with various poses, our model provides a unified solution for multi-pose FER, bypassing the separate training and parameter tuning for each pose, and thus is scalable to a large number of poses. Experiments on both benchmark facial expression databases and Internet images show the superior/highly-competitive performance of our system when compared with the current state-of-the-arts.

Index Terms—Face expression recognition, Multi-pose, Hierarchical theme model, Supervised Latent Dirichlet Allocation, Intermediate features

1 INTRODUCTION

As an essential way of human emotional behavior understanding, in the past decades, facial expression recognition (FER) has attracted a great deal of attention in multimedia research. The increasing applications of expression recognition, especially those in Human Computer Interaction (HCI) [22], [39] and affective computing, make it a core component in the next generation of computer system [30], [38], [41], [2], [29].

Recently, advances have been made in automatic FER in terms of face detection, feature extraction and expression classification. Most of these studies are conducted in “lab-controlled” environment, in which the faces captured are usually frontal or near-frontal with only one character and a single scale [49]. The real-world performance (e.g., on internet images or personal photo albums) of these systems degenerates greatly where new challenges arise due to large variations in expressions attributed to poses, identity, scales, etc..

More recently, a handful of methods on multi-pose expression recognition have been proposed [34], [15], [27], [35], [5]. These studies are mainly conducted using deliberately acted multi-pose face images. Models are learned and parameter-tuned separately for different poses [15], [27], failing to explicitly model the relationships between different poses. More specifically, the existing methods on multi-pose FER can be divided into face-shape-based methods (e.g., [3], [53], [18]) and face-shape-free methods (e.g., [27], [16], [11]). Face-shape-based methods rely on 2D/3D face-shape models that are used to decouple image variations caused by changes in facial expressions and head pose. Thus, FER accuracy is highly dependent on how well the shape models are aligned with the image data [18]. Face-shape free methods achieve multi-pose/head-pose-invariance by using pose invariant expression-related facial features extracted from 2D images (e.g., texture and/or geometry-based features extracted from manually marked facial points [49], [43]), or by training the facial expression recognition method pose-wise. However, extracting expression-related facial features independent of head pose is very difficult because the changes in head-pose and facial expressions are nonlinearly coupled in 2D. On the other hand, pose-wise FER requires a large amount of training data in terms of different expressions and poses, which are often not readily available. In addition, the performance of pose-wise face-shape free methods is expected to degenerate when tested on facial images with continuous change in head pose.

Recently, it has been shown that using intermediate features is very helpful for image understanding, image retrieval and object recognition [13], [24]. The

Qirong Mao is with the Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu Province 212013, China. Email: mao_qr@mail.ujs.edu.cn; Phone: 86-0511-88780371; Fax: 86-0511-88780371.
Qiyu Rao is with the Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu Province 212013, China. Email: raoqiyu@mail.ujs.edu.cn; Phone: 86-0511-88780371; Fax: 86-0511-88780371.
Yongbin Yu is with the Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu Province 212013, China. Email: darcy0511@gmail.com; Phone: 86-0511-88780371; Fax: 86-0511-88780371.
**Corresponding author, Ming Dong is with the Department of Computer Science, Wayne State University, Detroit, MI 48202, USA. Email: mdong@cs.wayne.edu; Phone: 1-313-577-0725; Fax: 1-313-577-6868.*

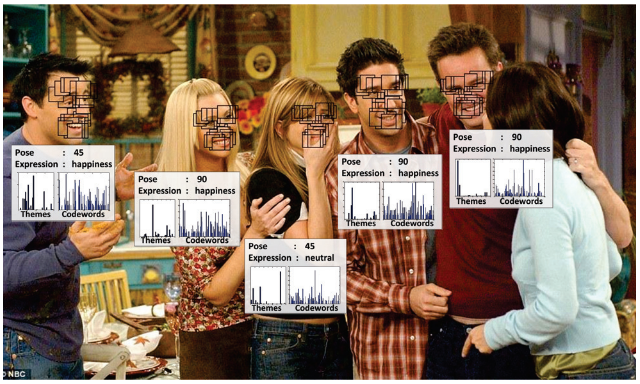


Fig. 1. Multi-pose FER using pose-based hierarchical Bayesian model. Globally optimized landmarks are shown on each detected face. Under each face, the left histogram shows its distribution over the 40 intermediate themes, and the right histogram shows its distribution over 300 codewords.

key is to integrate low-level facial features into an intuitive intermediate representation which can be shared across different poses to improve the learning performance. In [19], [9], hierarchical Bayesian models, e.g., Latent Dirichlet Allocation (LDA), are proposed to learn latent emotional topic features. In [45], Tong et al. proposed a unified probabilistic facial action model based on the dynamic Bayesian network to simultaneously and coherently represent rigid and nonrigid facial motions, their spatial-temporal dependencies, and their image measurements. However, these methods are mainly applied for frontal or near-frontal face images/video frames, and they were not robust to the pose changes.

In this paper, we propose a pose-based hierarchical Bayesian theme model to address the research issues for multi-pose FER (see Fig.1). In our system, automatic face detection and pose estimation are conducted on an input image. Then, features from local patches centered on globally optimized landmarks are used to learn the multi-pose intermediate representation before classifying expressions. From a technique perspective, our model provides a unified solution for multi-pose FER based on the shared pool of features. The major contributions of this paper are:

1) Our pose-based hierarchical Bayesian model provides a unified, shape-free framework to handle various poses in FER and does not require separate training and parameter tuning for each pose.
2) Local appearance features and global geometry information are combined in our model to learn a multi-pose, intermediate facial expression representation. By sharing the pool of features among various poses, our theme model is able to leverage the relationships between different poses, and thus achieve great performance on multi-pose FER.

The rest of the paper is organized as follows. We introduce the related work in Section 2. Section 3 presents our pose-based hierarchical model and Bayesian decision in details. Section 4 describes the facial expression feature extraction. Our experimental results are given in Section 5. Section 6 concludes.

2 RELATED WORK

FER is an active area in multimedia research because of the importance of faces in emotion expression and perception. Most of the existing work on FER studies the expressions of six basic emotions: happiness, sadness, surprise, fear, anger and disgust due to their marked reference representation in our affective lives and the availability of the relevant training and test data [49], [47]. There are also a few tentative efforts to detect non-basic affective states, such as fatigue, boredom, confusion and frustration [49]. At the beginning, studies are mainly based on deliberate and often exaggerated facial display. Later on, efforts have been reported on the automatic analysis of spontaneous facial expression recognition [7], [10]. For a comprehensive survey of the works in expression recognition please refer to [37], [49], [12]. In the following, we first review the work that concentrates on multi-pose/pose-invariant FER, and then discuss the literatures about hierarchical theme models.

2.1 Multi-pose/pose-invariant FER

Recent advances toward automatic multi-pose/head-pose-invariant FER can be classified into face-shape-based approaches and face-shape-free approaches [34]. We first briefly review the face-shape-based approaches, and then focus on the face-shape-free approaches as the method proposed in this paper belongs to the latter category. In [42], Active Appearance Models (AAM) are used to estimate the 3D head pose and locations of characteristic facial points for head-pose-invariant FER. Later on, Ji et al. [53], [45] use 3D face models to decouple rigid head motions and nonrigid muscular motions. A nonlinear mapping function from the 2D shapes of faces at any non-frontal pose to the corresponding 2D frontal face shapes is learned using Gaussian process regression in [1], but it requires the 3D pose of the face as an input to the regression. In general, face-shape-based methods require accurate alignment of the face-shape with the image data, which is challenging under varying facial expressions. Moreover, these methods ignore correlations across different poses.

The second category of approaches toward multi-pose/head-pose-invariant FER are based on 2D face-shape-free models. These methods achieve multi-pose/head-pose-invariance by using pose invariant expression-related facial features extracted from 2D images, or by training the facial expression recognition method pose-wise. According to the feature representation, the 2D-face-shape-free multi-pose/head-pose-invariance approaches can be further divided into shape representation, low-level engineered representation and high-level learning-based representation.

Shape representation relies on facial features such as shape of the face components and/or the coordinates of the facial landmarks. In [17], FER in non-frontal poses is investigated based solely on the coordinate values of 83 facial points. Rudovic et al. [34] proposed a probability-based approach to perform head-pose-invariant FER based on 2D geometric features from 39 facial landmark points. Low-level engineered representations extract local features (e.g., Scale-invariant Feature Transform (SIFT) and Local Binary Patterns (LBP)) and encode them in a transformed image. Low-level engineered features mainly capture skin texture changes such as wrinkles, bulges, and furrows. In [50], a facial image is divided into subregions, and then SIFT descriptors are extracted from each subregion and used as the input to a k-Nearest Neighbors classifier (k-NN). LBP and its variations are systematically evaluated in [27] for FER under different conditions, e.g., image resolution and orientation. In [44], [40], it was shown that a combination of sparse coding and Bag of Features (BOF) achieves good FER results. High-level learning-based representations encode features that are semantically interpretable for FER in a low- to high-level manner. The most well-established paradigms for learning-based representations are s-parse coding [6] and deep learning [32].

2.2 Hierarchical Theme Model

A shortcoming of 2D-face-shape-free multi-pose/head-pose-invariant FER methods is that they perform pose-wise facial expression recognition. That is, feature extraction and classifier training are conducted separately for similar poses. Thus, they can not explicitly model the relationships between different poses and do not scale well to a large number of poses due to the lack of a shared pool of features and a unified classification model [34]. Furthermore, these methods require a large amount of facial expression data per pose in order to train the classifiers. Parameter tuning for each pose model is time-consuming and generally is not applicable in real-world applications.

Hierarchical theme models such as LDA can model the inter- and intra-class structure of feature distributions, in which each image is represented as a finite mixture over an intermediate set of topics. However, topics are typically discovered in an unsupervised fashion and thus have limited use for classification. D-ifferent from unsupervised models which often learn topics hard to interpret, supervised variations of LDA have the ability to control the content of topics and are widely used for text or image classification. The main idea is to incorporate the class label variable into the generative model to enforce content of topics to handle specific classification tasks. Examples in this direction include the classLDA (cLDA) [23], the supervised LDA (sLDA) [4] and the labeled LDA.

More specifically, In cLDA, a class label is introduced as the parent of the topic prior. In this way, each class defines a prior distribution in the topic space, conditioned on which the topic probability vector is sampled. In sLDA, the class variable is conditioned by topics directly. In this paper, we propose to combine cLDA and sLDA into a unified framework that models the variation of both poses and facial expressions. Thus, our model can learn an intermediate FER feature representation shared by different poses.

3 POSE-BASED HIERARCHICAL BAYESIAN MODEL FOR FER

In this paper, we propose a pose-based hierarchical Bayesian theme model for multi-pose FER. The architecture of our system is shown in Fig.2. First, face detection, pose estimation and landmark localization are performed using the tree-based model [52]. Then, faces are modeled as a collection of local patches centered at the landmarks, based on which we construct the codewords and themes, and train our pose-based hierarchical Bayesian model. Finally, FER is performed. In the following, we first present our theme model in details. Feature extraction is discussed in Section 4.

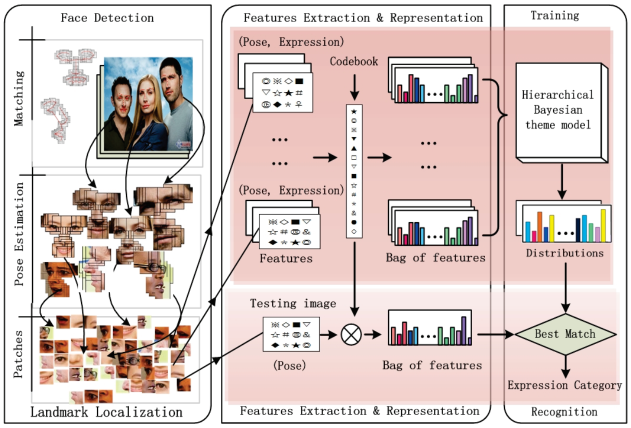


Fig. 2. System architecture

3.1 Pose-based Hierarchical Bayesian Theme Model

Figure 3 is a graphical illustration of our pose-based hierarchical Bayesian theme model. By introducing a variable c for pose, our model is a variation of the supervised Latent Dirichlet Allocation (sLDA) [19]. Pose is explicitly introduced because it is one of the most decisive factors for multi-pose FER and can be estimated with high accuracy. Other environmental factors such as illumination are more complex and hard to be untangled and determined separately.

Specifically, we model a facial image as a collection of local patches. Each patch is represented by a codeword from a large vocabulary of codewords. The goal of learning is to achieve a model that best represents the distribution of these codewords in a category of expressions. In recognition, therefore, we first identify all the codewords in the unknown facial image. Then, we find the expression category model that best fits the distribution of the codewords of the image. These distributions that best represent the distribution of the codewords in each category of expressions are called intermediate features or latent expression themes of expression categories. Next, we present our model by going through the generative process for creating a facial image with a specific expression and pose.

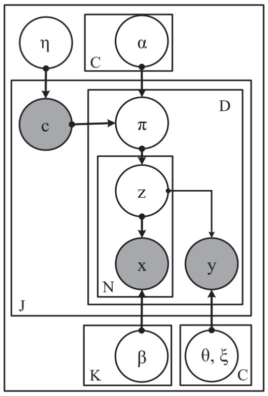


Fig. 3. Pose-based Hierarchical Bayesian theme model for multi-pose FER.

Suppose we have J face images belonging to C poses and Y expressions $\mathcal{J} = \{(I_1, c_1, y_1), ..., (I_J, c_J, y_J)\}$, where each image is labeled with expression $y = \{1, ..., Y\}$ and pose $c = \{1, ..., C\}$. Also assume D is the number of images for each pose. An image I is modeled using a collection of N patches $I = \{x_1, ..., x_N\}$, where a patch x_n is typically represented by the codewords in a codebook. In Fig.3, x, c and y are shaded by common convention to indicate that they are observed variables. Nodes with variables with no shading in the graph are unobserved. Under our theme model, a facial image arises from the following generative process.

1) Choose the pose variable $c \sim p(c|\eta)$ for each face image, where $p(.)$ is the multinomial distribution of poses with C outcomes and parameterized by η. η is a C-dim vector of a multinomial distribution.
2) For this face image with pose c, draw a parameter π from the multinomial distribution $\pi \sim p(\pi|c, \alpha)$ to determine the distribution of the themes, where α is a Dirichlet prior on the training sets and is a matrix of size $C \times K$, where K is the total number of themes and assumed to be known in advance.
3) For each patch x_n in the face image I_i
   a) Choose a latent theme $z_n \sim p(z_n|\pi)$, where $p(.)$ is a latent theme multinomial. z_n is a K dimensional unit vector.
   b) Choose a patch $x_n \sim p(x_n|z_n, \beta)$ with a latent theme z_n, where β is a $K \times T$ matrix, and T is the total number of codewords in the codebook.
4) Draw the facial expression variable $y \sim p(y|z_n, \theta, \xi)$. $p(.)$ is the Gaussian distribution $Gauss(\bar{Z}, \theta, \xi)$ with the parameter θ and ξ, where $\bar{Z} := (1/N)\sum_{n=1}^{N} z_n$ is the mean theme assignment vector. Here, the Gaussian parameters, mean μ and variance σ^2, are equal to $\theta^T \bar{Z}$ and ξ.

Given the parameters $\eta, \alpha, \beta, \theta$ and ξ, the joint probability of poses c, expression y, themes z, patches x can be written as:

$$p(x, z, \pi, c, y|\alpha, \eta, \beta, \theta, \xi) = p(c|\eta)p(\pi|c, \alpha) \\ (\prod_{n=1}^{N} p(z_n|\pi)p(x_n|z_n, \beta))p(y|z_n, \theta, \xi) \tag{1}$$

with

$$p(c|\eta) = Mult(c|\eta), \tag{2}$$

$$p(\pi|c, \alpha) = \prod_{j=1}^{C} Dir(\pi|c, \alpha_{j.})^{\delta(c,j)}, \tag{3}$$

$$p(z_n|\pi) = Mult(z_n|\pi), \tag{4}$$

$$p(x_n|z_n, \beta) = \prod_{k=1}^{K} p(x_n|\beta)^{\delta(z_n^k, 1)}, \tag{5}$$

$$p(y|z_n, \theta, \xi) = \frac{1}{\sqrt{2\pi\xi}} exp\{-\frac{(y - \theta^T \bar{Z})^2}{2\xi}\}, \tag{6}$$

where N is the number of patches in an image; $Mult$ represents the multinomial distribution and Dir denotes the Dirichlet distribution.

In the hierarchical representation of our theme model, the Dirichlet parameter α is at the pose-level, sampled once in the process of generating a pose. The multinomial variable π is at the theme-level, sampled once per face image. Through this parameter, the model can learn the relationships among different poses by sharing the pool of features. The theme variable z and patch x are at the patch-level, sampled every time a patch is generated. Finally, the Gaussian parameters θ and ξ are at the category-level, sampled once in the process of generating an expression.

3.2 Parameter Estimation

The parameters of our model are $\eta, \alpha, \beta, \theta$ and ξ. For convenience, the distribution of $p(c\,|\eta)$ is assumed to be a fixed uniform distribution in which $p(c) = 1/C$. The Dirichlet hyperparameters α can be estimated following the standard procedure of maximum likelihood estimation [33]. Next, we discuss how to estimate β, θ and ξ.

Given a corpus of facial expression images with emotion labels, $D = \{(x_d, y_d)\}_{d=1}^{D}$, we estimate the parameters β, θ and ξ using Expectation Maximization (EM). Specifically, the corpus log-likelihood could be represented as,

$$L(D) = \sum_{d=1}^{D} \log p(x|\alpha, \beta, \theta, \xi, y). \tag{7}$$

In the expectation step (E-step), we approximate the posterior distribution for each facial image using the variable inference algorithm described in the next section. In the maximization step (M-step), the lower bound in Eq. (7) is maximized over all facial images with respect to the model parameters β, θ and ξ by finding the maximum likelihood estimation under expected sufficient statistics [4]. The M-step updates are described below.

Estimating the topics: Given T as the number of codewords, the terms containing $\beta_{1:K}$ are

$$\begin{aligned} L_{[\beta_{1:K}]}(D) = &\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K} \phi_{dni} \log\beta_{k,x_n} \\ &+ \sum_{k=1}^{K} \mu_i (\sum_{t=1}^{T} \beta_{kt} - 1). \end{aligned} \tag{8}$$

Setting $\partial L_{\beta_{1:K}}(D)/\partial\beta_{it} = 0$ leads to

$$\beta_{k,x}^{new} \propto \sum_{d=1}^{D}\sum_{n=1}^{N_d} 1(x_n = t)\phi_{d,n}^{k}. \tag{9}$$

Estimating the GLM parameters: The parameters in the generalized linear model (GLM) are the coefficients θ and the dispersion parameter ξ. The gradient about GLM coefficients θ could be represented as

$$\begin{aligned} \frac{\partial L}{\partial \theta} &= \frac{\partial}{\partial \theta}(\frac{1}{\xi})\sum_{d=1}^{D}\{\theta^T E[\bar{Z}_d] y_d - E[A(\theta^T \bar{Z}_d)]\} \\ &= (\frac{1}{\xi})\{\sum_{d=1}^{D} E[\bar{Z}_d] y_d - \sum_{d=1}^{D} E_d[\mu(\theta^T \bar{Z}_d)\bar{Z}_d]\} \\ &= (\frac{1}{\xi})\{\sum_{d=1}^{D} \bar{\phi} y_d - \sum_{d=1}^{D} E_d[\mu(\theta^T \bar{Z}_d)\bar{Z}_d]\}. \end{aligned} \tag{10}$$

Because the parameters θ and ξ obey a gaussian distribution, setting $\partial L/\partial\theta = 0$ leads to

$$\begin{aligned} &\sum_d E[\bar{Z}_d \bar{Z}_d^T]\theta = E[\bar{Z}]^T y \quad \Rightarrow \\ &\hat{\theta}_{new} \leftarrow (\sum_d E[\bar{Z}_d \bar{Z}_d^T])^{-1} E[\bar{Z}]^T y, \end{aligned} \tag{11}$$

where $E[\bar{Z}] = \bar{\phi} = \frac{1}{N}\sum\limits_{n=1}^{N} \phi_n$, $\mu(\cdot) = E_{GLM}[Y|\cdot]$. The derivative with respect to ξ, evaluated at $\hat{\theta}_{new}$, can be represented as

$$\begin{aligned} &\{\sum_{d=1}^{D} \frac{\partial h(y_d, \xi)/\partial\xi}{h(y_d, \xi)}\} \\ &+ (\frac{1}{\xi})\{\sum_{d=1}^{D} [\theta_{new}^T (E[\bar{Z}_d] y_d) - E[A(\theta_{new}^T \bar{Z}_d)]\}, \end{aligned} \tag{12}$$

where $h(y, \xi)$ is the base measure in GLM with natural parameter y and dispersion parameter ξ. The partial derivative of $\frac{\partial h(y_d,\xi)/\partial\xi}{h(y_d,\xi)}$ is equal to $-\frac{1}{2\xi}$. Using this derivative and definition of $\hat{\theta}_{new}$, the dispersion parameter M-step is exact,

$$\hat{\xi}_{new} \leftarrow \frac{1}{D}\left\{y^T y - y^T E[\bar{Z}]\hat{\theta}_{new}\right\}. \tag{13}$$

3.3 Variational Inference

In Bayesian decision theory, given an unknown image represented as a collection of patches (or codewords), the key inferential problem is to compute the posterior distribution of the latent variables:

$$\begin{aligned} &p(z_n, \pi|x_n, y, c, \alpha, \beta, \theta, \xi) = \\ &\frac{p(\pi|c,\alpha)(\prod_{n=1}^{N} p(z_n|\pi)p(x_n|z_n,\beta))p(y|z_n,\theta,\xi)p(c|\eta)}{\int p(\pi|c,\alpha)\sum\limits_{z_{1:N}}(\prod_{n=1}^{N} p(z_n|\pi)p(x_n|z_n,\beta))p(y|z_n,\theta,\xi)p(c|\eta)d\pi}, \end{aligned} \tag{14}$$

where $\alpha, \beta, \eta, \theta$ and ξ are parameters learned from the training set. As mentioned in Section 3.2, $p(c|\eta)$ is always assumed to be a fixed uniform prior $p(c) = 1/C$. The normalizing value of Eq. (14) is the marginal probability of the observed data which is known as the likelihood. Unfortunately, this marginal probability is computationally intractable because of the coupling between α, β, θ and ξ in the summation over latent themes [4]. Thus, we appeal to variational methods to approximate the posterior. There are a wide variety of approximate inference algorithms which can be considered. In this paper, we adopted mean-field variational algorithm in which Jensen's inequality [4] is used to compute the lower bound of the normalizing value. Our goal is to maximize the log likelihood $\log p(x|\alpha, \beta, \theta, \xi, y)$, which, using Jensen's inequality, is bounded by

$$\begin{aligned} &\log p(x|\alpha, c, \beta, \theta, \xi, y) = \log \int_\pi \sum_z p(x, z, \pi|\alpha, c, \beta, \theta, \xi, y) d\pi \\ &= \log \int_\pi \sum_z p(x, z, \pi|\alpha, c, \beta, \theta, \xi, y)\frac{q(z, \pi|\gamma, \phi)}{q(z, \pi|\gamma, \phi)} d\pi \\ &\geq E[\log p(x, z, \pi|\alpha, c, \beta, \theta, \xi, y)] - E[\log q(z, \pi|\gamma, \phi)]. \end{aligned} \tag{15}$$

If all expectations are taken with respect to $q(z, \pi|\gamma, \phi)$, where γ_i is a K-dimensional Dirichlet parameter vector and each ϕ_{ni} parametrizes a categorical distribution over K elements, the evidence lower bound is achieved as:

$$
\begin{aligned}
&L(x,y|\alpha,c,\beta,\theta,\xi) = E[\log p(\pi|\alpha,c)] \\
&+ \sum_{n=1}^{N} E[\log p(z_n|\pi)] + \sum_{n=1}^{N} E[\log p(x_n|z_n,\beta)] \\
&+ E[\log p(y|z_n,\theta,\xi)] - E[\log q(z,\pi|\gamma,\phi)],
\end{aligned}
\tag{16}
$$

where $q(z,\pi|\gamma,\phi)$ is a distribution of the latent variables. It is introduced to simplify the lower bound and could follow arbitrary variational distribution.

Maximizing the lower bound with respect to γ and ϕ is the same as minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior. This minimization can be implemented via an iterative fixed-point method. The updating rules are given as follows:

$$\gamma^{new} \longleftarrow \alpha_i + \sum_{n=1}^{N} \phi_{ni}, \tag{17}$$

$$
\begin{aligned}
\phi^{new} &\propto exp\{E_q[\log \pi] + E[\log p(x_{ni}|\beta)] \\
&+ (\frac{y_i}{N\xi})\theta - (\frac{1}{2N^2\xi})[2(\theta^T\phi_{-ni})\theta + (\theta \circ \theta)]\},
\end{aligned}
\tag{18}
$$

where i and n are used to index a topic and a patch, respectively, and $\phi_{-mi} := \sum_{ni \neq mi} \phi_{ni}$.

3.4 Classification

Given a new facial image and a fitted model α, β, θ and ξ, we first compute $q(z,\pi)$ and the variational posterior distribution of the latent variables z_n. Then, for Gaussian distributions, since the mapping from the natural parameter to the mean parameter is the identity function, the expression categories of the facial image are given as [4]:

$$E[y|I = \{x_1,...,x_N\}, c, \eta, \alpha, \beta, \theta, \xi] \approx \theta^T E[\bar{Z}]. \tag{19}$$

Expression recognition is then achieved by maximizing the function,

$$
\begin{aligned}
y^* &= \arg \max_{y \in \{1,...,Y\}} \theta_y^T E[\bar{Z}] \\
&= \arg \max_{y \in \{1,...,Y\}} \theta_y^T \bar{\phi}.
\end{aligned}
\tag{20}
$$

Compared with the other models, our model can recognize multi-pose facial expression in a unified model without tuning parameters separately.

4 EXPRESSION FEATURE EXTRACTION

In this section, we present our expression feature extraction method, which is summarized in Fig.4. First, globally optimized landmarks are localized simultaneously with pose estimation through a tree-based part model on a detected face. Then, SIFT or LBP-based features are extracted from the local patches centered at these landmarks and are used to construct a codebook by k-means. Finally, the latent expression themes are learned in the hierarchical Bayesian theme model. Our goal is to represent the appearance variations in facial images, while being robust to pose changes.

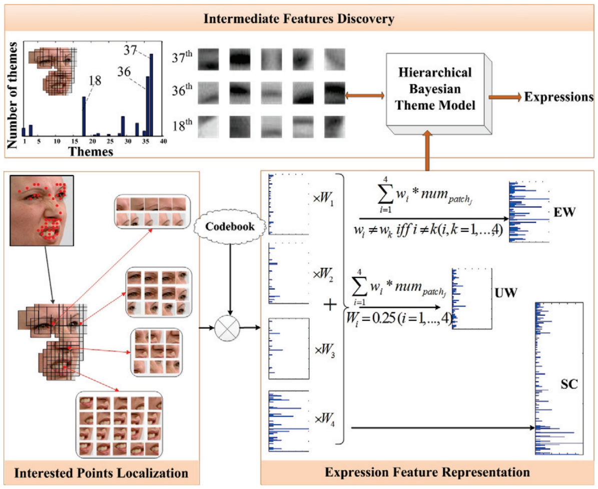

Intermediate Features Discovery
Number of themes
Themes
1
5
10
15
20
25
30
35
40
18
36
37
37th
36th
18th
Hierarchical Bayesian Theme Model
Expressions
Codebook
×W1
×W2
×W3
×W4
+
$\sum_{i=1}^{4} w_i * num_{patch_j}$
$w_i \neq w_k$ iff $i \neq k (i,k=1,...,4)$
EW
$\sum_{i=1}^{4} w_i * num_{patch_j}$
$W_i = 0.25 (i=1,...,4)$
UW
SC
Interested Points Localization
Expression Feature Representation

Fig. 4. The flow chart of expression feature extraction

4.1 Pose Estimation and Landmark Localization

We adopted the tree-based part model in [52] to perform simultaneous pose estimation and landmark localization. This method entails invariance to transformations such as scale, translation and in-plane rotations. Compared with AAM [26], this model captures more of the relevant elastic deformation so that it is more suitable for localizing landmarks for faces with different poses. However, note that the performance gain is achieved with higher computational complexity. The tree-based part model generally runs slower than AAM.

The bottom left panel of Fig.4 shows an example of landmark localization. Specifically, 51 points are located for frontal and near-frontal faces, and 28 points for profile faces. They are usually around the center of eyes and mouths, and along the eyebrows and noses. Compared with dense sampling over the entire face, using patches centered at the landmarks presents a sparse representation of the facial expression, and thus can effectively speed up the FER process. In addition, landmark localization does not need additional time because they are performed during face detection.

4.2 Local Patch Sampling

Local features from patches are more robust to occlusions and spatial variations than global ones. An image patch is first obtained for each landmark and then normalized to 16×16 pixels. To leverage multiscale representation, we tested the following sampling procedures for each patch:

- *Grid*. A patch is sampled with a 2×2 grid, and each cell in the grid is of the size 8×8.
- *Single-scale Dense*. The sampled patch is the same as the original patch, 16×16.

- *Multi-scale Dense*. For a given patch, we constructed a two-level pyramid, a 16×16 region and four 8×8 regions. Features are extracted from each region and then concatenated.
- *DoG Detector*. For a given patch, uncertain regions that are stable and rotationally invariant over the same scales are extracted using the DoG detector [25]. Scale of each region is 8×8.

Finally, Local features including 128-dim SIFT and 243-dim LBP are extracted from each sampled patch.

4.3 Codebook Construction and Encoding

Given the collection of patches from the training images, the codebook of SIFT and LBP descriptors is constructed by k-means algorithm. The codebook is composed of centers of all clusters, usually refereed to as codewords. Notice that face regions, i.e., eyes, mouths, eyebrows and noses, have been automatically detected by the tree-based part model, and they may play different roles in FER depending on the pose. Thus, we propose three encoding methods to construct the codebook: equal weights, unequal weights determined by a grid search, and coding each region separately (see the bottom right panel of Fig.4).

- *EW (Equal Weights)*: The bin of each codeword gives the weighted sum of the number of the closest patches. The four regions are equally weighted at 0.25.
- *UW (Unequal Weights)*: The bin of each codeword gives the weighted sum of the number of the closest patches. The weights of regions are provided by a grid search.
- *SC (Separate Coding)*: We code each of the four regions separately using equal weights. The size of the codebook is quadrupled when compared with EW and UW.

4.4 Intermediate Features

Based on the collected training data, we can build the pose-based hierarchical Bayesian theme model and obtain the distribution of codewords on latent themes and that of themes on each expression and pose. Namely, we can achieve a model that best represents the distribution of codewords over each expression and pose. These distributions are called intermediate features or latent expression themes. They provide a shared pool of expression features in a unified framework, scalable to a large number of poses. For example, we can intuitively understand the intermediate features for disgust expression such as wrinkling nose and raised upper lip (see the top panel of Fig.4). The latent aspects of facial images, hidden behind the bag of features, once discovered, are well suited to reveal the distinctions between facial expressions with various head poses, and thus lead to higher accuracy of recognition.

5 EXPERIMENTS & RESULTS

5.1 Datasets

Our approaches were thoroughly evaluated in subject-dependent experiments on the following datasets: 1) three public multi-pose facial expression databases: the Radboud Faces Database (RAFD) [20], the Karolinska Directed Emotional Faces Database (KDEF), and the Multi-PIE database [14]; 2) one 3D facial expression database: BU-3DFE [48]; 3) one facial expression in the wild dataset: Static Facial Expressions in the wild (SFEW) [8]; and 4) a dataset consisting of images randomly downloaded from the Internet.

The number of facial images in RAFD and KDFE are 8,040 and 4,900, respectively, for five poses and seven basic expressions. We combined RAFD and KDEF as one dataset so that we have sufficient training samples, especially for those pose-wise FER models. The combined dataset contains 12,940 face images covering seven basic expressions (i.e., anger, disgust, fear, happiness, sadness, surprise, and neutral) with five poses $(180°, 135°, 90°, 45°$ and $0°)$. The size of an image is 681 × 1024 in RAFD and 562 × 762 in KDEF. The CMU Multi-PIE face database contains more than 750,000 images of 337 people. Subjects were imaged under 15 view points and 19 illumination conditions while displaying a range of facial expressions.

The 3D facial expression database, namely BU-3DFE [48], has 100 subjects with 3D models and face images. We rendered 2D facial images at four levels of intensity and in five yaw views $(90°, 60°, 45°, 30°,$ and $0°)$ with six facial expressions. In total, 12,000 face images are used for our experiment. The size of an image is 1264 × 931 in BU-3DFE. The SFEW dataset consists of 700 images of 95 subjects, extracted from movies containing facial expressions with various head poses, occlusions and illumination conditions. The images have been labeled in terms of six basic emotion expressions [8].

5.2 Parameter Selection

As the first step of our experiment, we construct a validation set by randomly selecting 2,975 face images from RAFD, 85 images for each expression and pose. This validation set is used to choose the parameters used in our later experiments.

Local patch sampling method: Table 1 shows how different sampling methods and feature descriptors influence the performance of FER. The accuracy here is averaged over all facial expressions and poses. For convenience, we encoded each facial region separately (SC) in this experiment. Clearly, DoG sampling method outperforms all others regardless the choice of using SIFT or LBP descriptors. In addition, SIFT features give higher accuracy than LBP descriptors. So, in our following experiments, we employed DoG to sample local patches and to obtain a multi-scale feature representation.

TABLE 1
Recognition accuracy (reported in %) for different patch sampling methods. Features are encoded with SC. The accuracy reported are averaged over all poses and expressions in RAFD

Descriptor	Grid	Single-scale	Multi-scale	DoG
243-dim LBP	43.05	59.60	61.52	68.53
128-dim SIFT	56.00	60.02	64.44	74.96

Feature Descriptor and Encoding Method: Fig. 5 compares different feature descriptors (i.e., SIFT and LBP) and different encoding methods (i.e., EW, UW and SC) based on recognition accuracy. In EW, the four facial regions (i.e., eyes, mouths, eyebrows and noses) share an equal weight of 0.25. In UW, best weights for regions are grid searched with a step of 0.1, and we adopted the following weights that give the highest overall recognition accuracy: 0.1 for noses, 0.2 for eyebrows, 0.3 for eyes, and 0.4 for mouths. In SC, each region is coded separately. In Fig.5, we report the average accuracy of seven expressions for each pose. Clearly, regardless the choice of SIFT or LBP, SC generally outperforms the other two encoding methods. Between SIFT and LBP, SIFT features encoded with SC achieve a higher recognition accuracy of 74.96%, averaged over all poses. Thus, SIFT and SC are selected as the feature descriptor and encoding method in our following experiments. Also note that better recognition performance is obtained on frontal or near frontal faces, with accuracy of over 80%. This is mainly attributed to the fact that more discriminative features are available under these poses.

The Number of Themes and Size of Codebook: Fig. 6(a) plots the recognition rate vs. the different number of themes. Again, the accuracy is averaged over all poses and expressions. Clearly, the highest accuracy is achieved when the number of theme is 40, which we used in our following evaluation. Fig.6(b) illustrates the effects of the codebook size on the recognition accuracy. Notice that the highest recognition accuracy is achieved when the codebook size is 300. Thus, we fix it at 300 in all our following experiments.

5.3 Effect of Landmark Localization and Pose Estimation Error on FER

We use the tree-based part model [52] to estimate the poses and locate the landmarks since it captures more of the relevant elastic deformation compared with the AAM model. The experiments in [52] show that the tree-based part model works best when compared with the multi-AAM and face.com, scoring 91.4% when requiring exact matching, and 99.9% when allowing $\pm 15^o$ error tolerance on the Multi-PIE dataset. For the landmark localization, the tree-based part model also outperforms the state-of-the-art

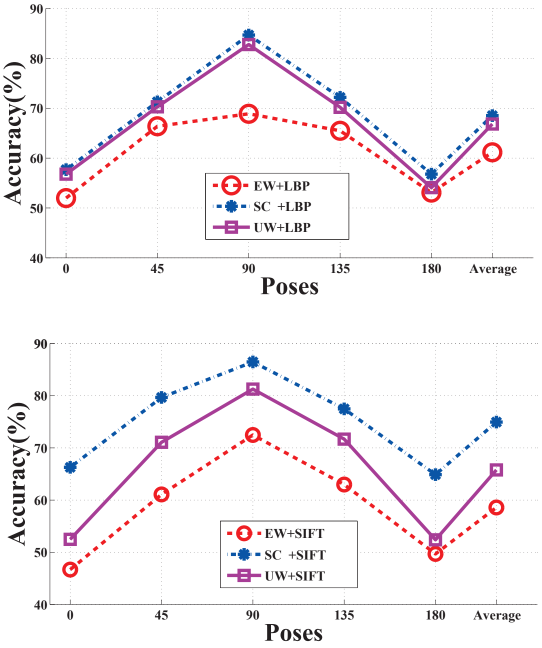
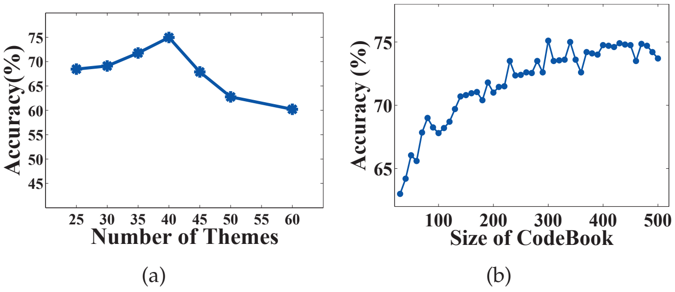


Fig. 5. Performance comparison of feature descriptors (SIFT and LBP) and encoding methods (EW, UW and SC) over five poses $(180°, 135°, 90°, 45°,$ and $0°)$ in RAFD. **Top panel:** Averaged accuracy of LBP with the three encoding methods. **Bottom panel:** Averaged accuracy of SIFT with the three encoding methods.



(a)
(b)

Fig. 6. (a) Number of themes versus average recognition accuracy. (b) Size of codebook versus average recognition accuracy.

Constrained Local Models (CLM) [36] (average error of 4.39 pixels or relative error 2.3% vs. 4.75 pixels or 2.8%) on the Multi-PIE. In this section, experiments are conducted on the validation set of RAFD to analyze how the FER accuracy is affected by erroneous landmarks and poses.

Effect of Landmark Localization Noise on FER: In this experiment, landmark locations of the facial images in the validation set are deliberately corrupted by different levels of noise (randomly selected from the interval) in [-σ, σ], with σ = 0, 2, 5, 8, 15, 20, 30 pixels. The mean FER accuracies and standard deviations of noise-corrupted data with different intervals are reported in Table 2. In each column in Table 2, the results are achieved by five-fold cross-validation on the data corrupted by the same interval of noise, and averaged over all expressions for each pose. The last row of Table 2 gives the accuracy averaged over all poses and expressions for each interval of noise.

Clearly, the highest accuracy averaged over all poses and expressions and the smallest standard deviation are achieved by the clean data. The mean FER accuracy waves when σ is smaller than 8. This is mainly caused by the randomness in noise sampling: the levels of noise are randomly selected in the interval in [-σ, σ], and the intervals are rather small. Note that the smallest standard deviation is always achieved by the noiseless data. When σ is larger than 8, the mean FER accuracy consistently degenerates. Generally, our model is quite tolerable regarding landmark localization errors.

Effect of Pose Estimation Noise on FER: In this experiment, poses of the facial images in the validation set are deliberately corrupted by different levels of noises: $\Delta = 0\%, 2\%, 5\%, 8\%, 10\%, 15\%$ and 20%, where $\Delta = i\%$ indicates that the poses of $i\%$ images in the dataset are randomly changed. The FER accuracies and the standard deviation of noise-corrupted data with different levels of noises are reported in Table 3. In each column in Table 3, the results are achieved by five-fold cross-validation on the data corrupted with the same level of noises, and averaged over all expressions with each pose. The last row of Table 3 shows the accuracy averaged over all poses and expressions.

Clearly, FER accuracy in our model is more sensitive to pose estimation noise than that on landmark locations. Regardless of the pose, the highest accuracy and the smallest standard deviation are always achieved by the noiseless data. This is mainly because pose is explicitly introduced in our hierarchical theme model.

5.4 Performance Evaluation

Latent Themes: Fig. 7 illustrates the latent theme model learned for each expression of the five poses. A small panel in the figure shows the feature distribution over the 40 expression themes, averaged over all the training images with the corresponding expression and pose. Clearly, these distributions vary greatly. In other words, our model can identify latent discriminative features for better multi-pose FER.

Recognition Accuracy on RAFD and KDEF: In order to evaluate our model, we compared its performance with three popular machine learning methods, namely, the multi-SVM [15], DHMM [28] and sLDA model [4], [19].

The multi-SVM model consists of five SVMs, each trained separately with expression images under a specific pose: $180^\circ, 135^\circ, 90^\circ, 45^\circ$, and 0°. The Radial Basis kernel is adopted in each SVM. The parameters $\mathcal{C} \in [-1, 10]$ and $\delta \in [-1, 1.5]$ of each SVM are tuned

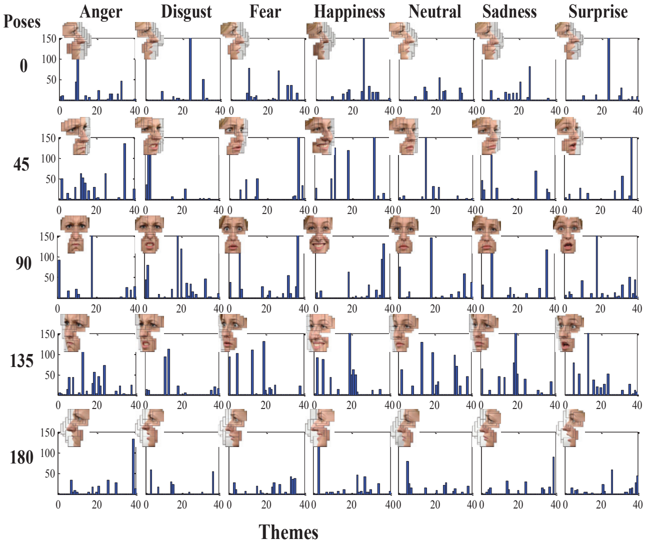


Fig. 7. Theme distributions. Each row represents one pose and each column represents one expression. The panel shows the mean distribution of the 40 latent expression themes on different poses and expressions.

by grid search with a step of 0.1 using the validation set. During recognition, the pose of an input facial image is first estimated, and then the corresponding SVM is used to classify the face into one of the seven expression categories.

Following [28], [21], [46], in DHMM, different parts of a facial image are considered as an observation sequence for FER. Specifically, the observed status of the DHMM model is set as the features of different facial parts in a given facial image, and these observed statuses determine the expression of the face. DHMM model is composed of thirty five DHMMs, each trained separately with images under a specific expression and pose. The number of the observed status in the DHMM model is empirically fixed as 8, and that of the hidden status is set at 90 by a grid search with a step of 10 in the range of $[30, 150]$ using the validation set. The initial values of prior probabilities and state transition probabilities in each DHMM are generated randomly.

Finally, sLDA are trained based on the facial images with expressions regardless of poses. 4,900 facial images are randomly selected from RAFD and KDEF with the seven expressions and five poses (excluding the 2,975 images in the validation set) to train and test our theme model, multi-SVM, DHMM and sLDA. Recognition accuracies (with standard deviation) of the four methods are obtained using five-fold cross-validation and averaged over all expressions and poses in Table 4. In addition, Tables 5 and 6 provide results over each pose and expression.

Clearly, our model outperforms multi-SVM, DHMM and sLDA. The significant accuracy gain over sLDA shows the advantage of explicitly introducing

TABLE 2
FER accuracy and standard deviation with different levels of noise ($[-\sigma, \sigma]$, σ = 0, 2, 5, 8, 15, 20, 30 pixels) added to landmark locations, reported in %. The highest accuracy for each pose is highlighted in bold.

Poses	σ = 0	σ = 2	σ = 5	σ = 8	σ = 15	σ = 20	σ = 30
0°	**66.28**±4.28	62.86±4.56	65.43±4.73	63.14±7.42	58.86±4.18	57.14±4.52	53.14±4.45
45°	82.86±4.11	83.48±4.23	81.14±4.73	**84.57**±6.96	75.71±6.39	76.86±6.09	72.00±6.17
90°	**84.57**±4.18	83.14±4.64	85.14±4.93	82.57±4.18	77.43±5.10	78.85±4.73	76.29±4.94
135°	79.71±3.47	78.17±4.45	78.29±5.14	**80.57**±6.98	78.57±5.50	74.85±3.90	72.57±6.22
180°	**66.57**±4.43	61.43±5.03	64.57±4.77	63.14±4.77	61.20±6.48	56.00±4.73	54.29±4.44
Mean	**76.00**	73.82	74.91	74.80	70.35	68.74	65.56

TABLE 3
FER accuracy and standard deviation with different levels of noise ($\Delta = 0\%, 2\%, 5\%, 8\%, 10\%, 15\%$ and 20%) added to poses, reported in $\%$. The highest accuracy for each pose is highlighted in bold.

Poses	0%	2%	5%	8%	10%	15%	20%
0°	**66.28**±4.28	65.71±4.30	64.00±5.38	64.00±4.55	63.14±4.38	64.29±5.03	59.57±4.92
45°	**82.86**±4.11	82.57±4.90	79.71±4.73	80.86±6.36	82.86±4.52	80.71±6.35	77.57±7.48
90°	**84.57**±4.18	84.00±4.66	81.14±4.28	81.71±5.53	82.00±5.61	80.71±5.29	77.57±5.71
135°	**79.71**±3.47	79.14±3.55	78.00±3.94	77.71±5.76	76.29±5.48	77.00±5.18	75.29±5.12
180°	**66.57**±4.43	66.29±5.31	64.57±4.46	62.86±5.50	64.00±6.29	60.28±5.64	58.43±5.93
Mean	**76.0**	75.54	73.49	73.43	73.66	72.80	69.89

TABLE 4
Overall performance comparison among our model, multi-SVM, DHMM and sLDA. The highest one is highlighted in bold

Database	Our model	Multi-SVM	sLDA	DHMM
RAFD & KDEF	**74.96**±2.28	67.84±2.31	59.06±2.96	61.86±2.32
BU-3DFE	**80.47**±1.79	74.07±2.01	64.67±2.66	62.72±2.35

TABLE 6
Recognition accuracy and standard deviation comparison across five poses among our model, multi-SVM, DHMM and sLDA on RAFD and KDEF, and the highest one for each pose is highlighted in bold.

	0°	45°	90°	135°	180°
Multi-SVM	56.83±1.21	73.89±0.82	79.72±0.73	77.81±1.03	50.95±0.98
sLDA	49.68±1.01	65.39±0.92	69.98±0.54	63.73±0.74	46.51±1.17
DHMM	50.75±0.88	65.56±0.76	74.36±0.56	66.37±0.35	52.28±0.97
Our model	**64.03**±0.52	**77.26**±0.65	**85.60**±0.11	**80.65**±0.23	**67.60**±0.91

poses in the hierarchical Bayesian model. In addition, our unified model avoids the separate training and parameter tuning in multi-SVM and DHMM, and thus highly scalable to the large number of poses seen in multi-character images.

The confusion table in Fig.8(a) provides details of the performance of our model on each facial expression. The values in the main diagonal give the recognition accuracy of each expression, averaged over all poses. A closer look at the table reveals that, among the seven expressions, anger, disgust, happiness and surprise are easier to be recognized with accuracy of over 80%. Other expressions have a lower recognition rate, with the lowest at 55.7% for fear. The averaged accuracy over all expressions is 75.0%.

Recognition Accuracy on BU-3DFE: In this experiment, we compared FER accuracy of our model with multi-SVM, DHMM and sLDA model on BU-3DFE. Recognition accuracies (with standard deviation) of the four methods are obtained using five-fold cross-validation and averaged over all expressions, poses, and intensity levels in Table 4. The confusion table in Fig.8(b) provides details of the performance of our model on each facial expression of BU-3DFE. A closer look at the table reveals that, like the results on RAFD and KDEF, among the six expressions, anger is easier to be recognized, while fear and sadness are more difficult to be recognized. In addition, the recognition accuracy of fear is higher than that in Fig.8(a) since samples of fear at intensity level 3 and 4 in BU-3DFE are more exaggerated than those in RAFD and KDEF.

Moreover, we also compared the recognition accuracy of our method with the current state-of-the-art results reported in [27], [51], [50] on BU-3DFE for head pose invariant FER. Details regarding each method, e.g., feature extraction, classification algorithms, pose numbers and the expressions adopted, are summarized in Table 7. All results are obtained using object-independent ten-fold cross-validation. Specifically, we randomly divide the 100 subjects into a training set with 90 subjects and a testing set with 10 subjects. Both training and testing sets are over all six expressions, four intensities, and five views of each subject, resulting in 10,800 training facial images and 1,200 testing facial images in total. Overall, our method gets the highest recognition accuracy of 79.11% on BU-

TABLE 5
Recognition accuracy and standard deviation across seven expressions achieved by our model, multi-SVM, DHMM and sLDA on RAFD and KDEF. The highest accuracy for each expression is highlighted in bold.

	Happiness	Sadness	Anger	Surprise	Disgust	Fear	Neutral
Multi-SVM	82.14±0.53	68.26±0.95	78.56±0.86	80.04±0.98	63.38±0.74	**56.02**±1.05	46.50±1.08
sLDA	70.60±0.33	51.01±0.51	72.81±0.62	53.01±0.89	68.75±0.68	31.00±0.78	65.21±1.12
DHMM	78.10±0.85	63.92±1.23	74.68±0.77	59.23±1.99	70.27±0.94	34.79±1.71	52.01±1.13
Our model	**88.25**±0.23	**71.41**±0.43	**80.41**±0.56	**80.65**±0.88	**82.70**±0.65	55.65±0.71	**65.35**±0.91

TABLE 7
Comparison with the state-of-the-art results on BU-3DFE. The highest accuracy is highlighted in bold.

Method	Classifier	Features	Poses	Expressions		Recognition Rates(%)
				number	levels	
Moore et al. [27]	svm	lbp^{u^2}	5	6	1,2,3,4	58.4
Moore et al. [27]	svm	$lbp^{u^{ms}}$	5	6	1,2,3,4	65.0
Moore et al. [27]	svm	lgbp	5	6	1,2,3,4	68.0
Moore et al. [27]	svm	lgbp/lbp^{ms}	5	6	1,2,3,4	71.1
Zheng et al. [51]	knn	sparse sift	5	6	1,2,3,4	78.5
Zheng [50]	GSRRR	lbp^{u^2}	5	6	1,2,3,4	66.0
Zheng [50]	GSRRR	sparse sift	5	6	1,2,3,4	78.9
Wenming Zheng [50]	GSRRR	83 landmark points	5	6	1,2,3,4	71.4
Our model	Pose-LDA	sift	5	6	1,2,3,4	**79.11**

TABLE 8
Performance comparison between our model and DS-GPLVM on Multi-PIE, reported in %. The highest average accuracy is highlighted in bold.

Poses	30°	15°	0	−15°	−30°	Mean
DS-GPLVM	90.11±0.028	89.97±0.023	82.42±0.018	96.96±0.012	93.55±0.019	**90.60**
Our model	89.30±0.028	88.71±0.022	88.70±0.014	92.30±0.012	92.21±0.010	90.24

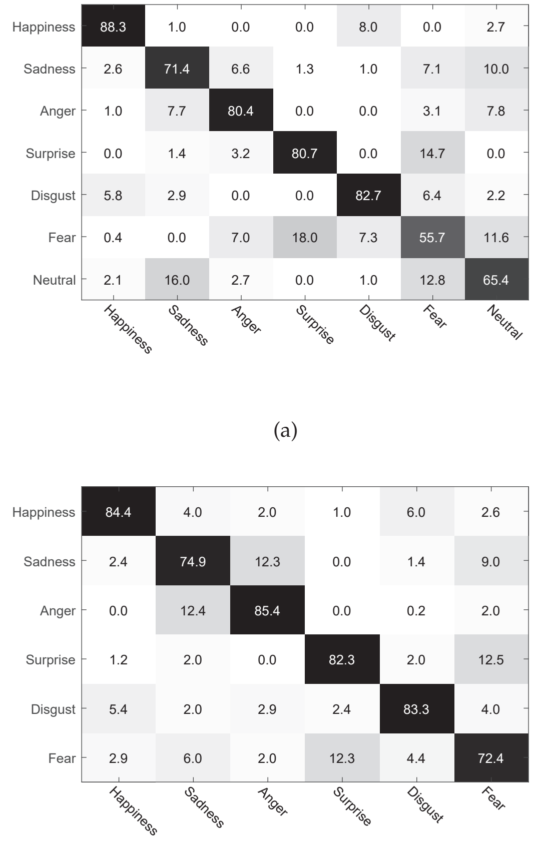
3DFE.

Recognition Accuracy on Multi-PIE: In this experiment, we compared our model with the current state-of-the-art results reported on Multi-PIE dataset, which are obtained using a Discriminative Shared Gaussian Process Latent Variable Model (DS-GPLVM) [11]. DS-GPLVM first learns a discriminative manifold shared by multiple views of a facial expression, and then facial expression classification is performed in the expression manifold, either in the view-invariant manner (using only a single view of the expression) or in the multi-view manner (using multiple views of the expression). We trained our model using the same experiment setting employed in [11]. That is, we performed five-fold cross-validation on images of 270 subjects, depicting six acted facial expressions of Neutral, Disgust, Surprise, Smile, Scream and Squint, captured at pan angles $30°, 15°, 0°, -15°$ and $-30°$ (1,531 images per pose).

The comparison results (FER accuracy and standard deviation) are provided in Table 8, which shows that our model can achieve highly competitive results with DS-GPLVM. Compared with the manifold learned in DS-GPLVM, our model can discover intuitive intermediate face representations or latent expression themes for better interpretation of the FER results. In addition, expression recognition is achieved in our model by computing the posterior distribution of the latent variables. No separate classifier (e.g., kNN in DS-GPLVM) is needed.

Recognition Accuracy on SFEW: In this experiment, we conducted FER on SFEW, a facial expression in the wild dataset, and compared our model with VGG-Face [31], a state-of-the-art face model based on a 16-layer convolutional neural network. As the number of images in SFEW is very small, both our model and VGG-Face are trained using the Multi-PIE dataset and tested on SFEW. Note that there are only four common emotions between Multi-PIE and SFEW, i.e., happiness, disgust, neutral and surprise. So, the testing set contains 327 images in SFEW with these emotion labels. Also note that VGG-Face was pre-trained with a large dataset containing 2.6 million facial images spanning more than 2.6K identities, and Multi-PIE was only used to fine-tune the learned weights. The performance of VGG-Face and our model on SFEW are reported in Table 9 for each facial expression. Apparently, our method has a more consistent performance over different facial expressions and outperforms VGG-Face on the mean recognition accuracy.

Recognition Accuracy on Internet Images: 248 Random Internet images are downloaded to further evaluate the performance of the four methods. From



(a)

(b)

Fig. 8. (a) Confusion table on RAFD and KDEF by our model. The average recognition rate is 75.0%. (b) Confusion table on BU-3DFE by our model. The average recognition rate is 80.5%.

TABLE 9
Performance comparison between our model and VGG-Face on SFEW, reported in %. The highest average accuracy is highlighted in bold.

Methods	Disgust	Happiness	Neutral	Surprise	Mean
VGG-Face	63.59	17.24	28.92	56.95	41.68
Our model	43.59	43.01	41.67	50.60	**44.72**

these 248 random Internet images, we select 338 facial images (one downloaded image may include multiple faces) with expression and label them by the validation vote of five non-professional persons. Table 10 shows the recognition accuracies of four methods for each expression. Clearly, our model achieves the highest accuracy compared with other three methods for all expressions. Specifically, the average accuracy of our model is 6.65% higher than that of the second best method: multi-SVM. Note that compared with the ”lab-controlled” environment, the tree-based part model will miss some faces in the unconstrained environment. In these cases, we will count the miss detections as wrong expressions, and therefore, it will negatively affect the expression recognition performance.

TABLE 10
Performance (reported in %) of four state-of-the-art methods on random downloaded facial images. In the first column, the number in the parenthesis is the total number of facial images with an expression. In other columns, the number in the parenthesis is the total number of facial images recognized correctly by the corresponding method . The highest accuracy for each expression is highlighted in bold.

Expressions	Our model	Multi-SVM	sLDA	DHMM
Happiness(56)	**62.50(35)**	55.36(31)	42.86(24)	33.93(19)
Sadness(45)	**49.00(22)**	42.22(19)	37.78(17)	33.33(15)
Anger(51)	**51.02(26)**	45.10(23)	37.25(19)	35.29(18)
Surprise(45)	**68.89(31)**	62.22(28)	51.11(23)	22.22(10)
Disgust(48)	**75.00(36)**	62.50(30)	31.25(15)	43.75(21)
Fear(45)	**35.33(15)**	31.56(16)	24.44(11)	28.89(13)
Neutral(48)	**47.92(23)**	44.08(25)	41.67(20)	45.83(22)
Average	**55.66**	49.01	34.75	38.05

Fig. 9 provides several examples, in which both detected faces (white boxes) and recognized expressions by the four methods are shown with different colors. The recognition results of our model, multi-SVM, sLDA and DHMM for each face are listed respectively from left to right in the white box under the facial image. Again, the expression label of each facial image is achieved by the validation vote of five non-professional persons, which is shown above the facial image. Clearly, our model outperforms other three methods. As long as no heavy occlusion, most faces can be accurately detected, and the expressions can be correctly recognized by our model. The processing time for an internet image varies greatly with the image resolution and the number of faces in the image.

6 SUMMARY & DISCUSSION

In this paper, we present a pose-based hierarchical Bayesian theme model for multi-pose FER. Local appearance features and global geometry information are combined in our model to learn an intermediate face representation before recognizing expressions. By sharing a pool of features with various poses, our unified model can perform FER without separate training and parameter tuning for each pose, and thus is scalable in real-world applications with a large number of poses. Experiments on both benchmark facial expression databases and Internet images show the superior/highly-competitive performance of our system when compared with the current state-of-the-arts. In the future, we plan to collect and label a large number of natural facial images so that our model can be trained with more poses.

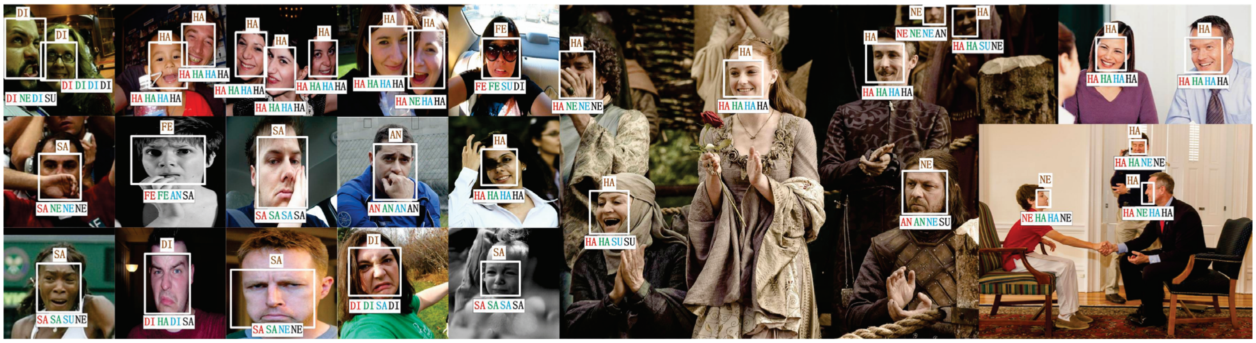


Fig. 9. The comparison examples of FER by four state-of-the-art methods (i. e., our model, multi-SVM, sLDA and DHMM). The images are downloaded from Internet randomly. The first row of the left panel shows the results with face overlapping and faces with glasses. The second row of the left panel shows successful recognition with mouth or chin occlusion. The third row shows the results on different poses and scales. The bigger image in the middle demonstrates our results on faces with complex background, such as multi-person, different face sizes and occlusions. The right bottom image shows the results on profile faces. In the figure, HA: happiness; SA: sadness; DI: disgust; AN: anger; SU: surprise; FE: fear; NE: neutral.

7 Acknowledgement

Work by Qirong Mao was partially supported by the National Nature Science Foundation of China under Grants 61272211 and 61672267, the Six Talent Peaks Foundation of Jiangsu Province under Grant DZXX-027 and the General Financial Grant from the China Postdoctoral Science Foundation 2015M570413. Work by Ming Dong was partially supported by US National Science Foundation (NSF) under grant CNS-1637312, and by Ford Motor Company University Reserach Program under grant 2015-9186R.

References

[1] A. Asthana, M. J. Jones, T. K. Marks, K. H. Tieu, and R. Goecke. Pose normalization via learned 2D warping for fully automatic face recognition. In *Proceedings of the British Machine Vision Conference*, pages 1–11, Sep. 2011.

[2] D. Beant, K. Rafal, and I. S. et. al. Culture and facial expressions: A case study with a speech interface. *Human-Computer Interaction*, 6947:392–404, 2011.

[3] M. J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal Computer Vision*, 25:23–48, 1997.

[4] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing Systems*, 20:121–128, 2008.

[5] R. A. Calvo and S. D. Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[6] S. Cotter. Sparse representation for accurate classification of corrupted and occluded facial expressions. In *Proceedings of IEEE International Conference Acoustic, Speech Signal Process*, pages 838–841, Dallas, Texas, USA, Jul. 2010.

[7] A. Dhall. Expression analysis in the wild: from individual to groups. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 325–328, Dallas, Texas, USA, Apr. 2013.

[8] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expressionanalysis in tough conditions: Data, evaluation protocol and benchmark. In *Proceedings of International Conference Computuer Vision Workshops*, pages 2106–2112, Barcelona, S-pain, Nov. 2011.

[9] A. Dhall, J. Joshi, I. Radwan, and R. Goecke. Finding happiest moments in a social context. In *Proceedings of Asian Conference on Computer Vision (ACCV), Lecture Notes in Computer Science*, volume 7725, pages 613–626, Daejeon, Korea, Nov. 2012.

[10] H. Dibeklioglu, A. A. Salah, and T. Gevers. Recognition of genuine smiles. *IEEE Transactions on Multimedia*, 17(3):279 – 294, 2015.

[11] S. Eleftheriadis, O. Rudovic, and M. Pantic. Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition. *IEEE Transactions on Image Processing*, 24(1):189–204, 2015.

[12] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1):259–275, 2003.

[13] I. J. Goodfellow, A. Courville, and Y. Bengio. Scaling up spike-and-slab models for unsupervised feature learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1902–1914, 2013.

[14] R. Gross, I. Matthews, J. Cohna, T. Kanade, and S. Baker. Multi-pie. *Image and Visison Computing*, 28(5):807–813, 2010.

[15] N. Hesse, T. Gehrig, H. Gao, and H. K. Ekenel. Multi-view facial expression recognition using local appearance features. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 3533–3536, Tsukuba Science City, JAPAN, Nov. 2012.

[16] Q. Hu, X. Peng, P. Yang, F. Yang, and D. N. Metaxas. Robust multi-pose facial expression recognition. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR)*, pages 1782–1787, Stockholm, Sweden, Aug. 2014.

[17] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang. Multi-view facial expression recognition. *Automatic Face and Gesture Recognition*, pages 1–6, 2008.

[18] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. *International Journal Computer Vision*, 83(2):178–194, 2009.

[19] P. Lade, V. N. Balasubramanian, and S. Panchanathan. Probabilistic topic models for human emotion analysis. In *Proceedings of Neural Information Processing Society Workshop on Topic Models (NIPS)*, Lake Tahoe, Nevada, US, Dec. 2013.

[20] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, and S. T. Hawk. Ad van knippenberg: Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, 24(8):1377–1388, 2010.

[21] H. S. Le and H. Li. Recognizing frontal face images using Hidden Markov Models with one training image per person. In *Proceedings of the 17th International Conference on Pattern*

Recognition (ICPR), volume 1, pages 318–321, 2004.

[22] C. Lee-Johnson and D. Carnegie. Mobile robot navigation modulated by artificial emotions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(2):469–480, 2010.

[23] F. Li and P. Pietro. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 524–531, San Diego, California, USA, Jun. 2005.

[24] Z. C. Li and J. H. Tang. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Transactions on Multimedia*, 17(11):1989–1999, 2015.

[25] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on Computer vision (ICCV)*, pages 1150–1157, Kerkyra, Corfu, Greece, Sep. 1999.

[26] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[27] S. Moore and R. Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541–558, 2011.

[28] J. A. Ojo and S. A. Adeniran. One-sample face recognition using HMM model of fiducial areas. *International Journal of Image Processing*, 5(1):58–68, 2010.

[29] L. Pang, S. Zhu, and C. W. Ngo. Deep multimodal learning for affective analysis and retrieval. *IEEE Transactions on Multimedia*, 17(11):2008–2020, 2015.

[30] J. W. Park, W. H. Kim, W. H. Lee, and M. J. Chung. Artificial emotion generation based on personality, mood, and emotion for life-like facial expressions of robots. *Human-Computer Interaction*, 332:223–233, 2010.

[31] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of British Machine Vision Conference*, pages 1–6, 2015.

[32] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza. Disentangling factors of variation for facial expression recognition. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 808–822, Florence, Italy, Oct. 2012.

[33] G. Ronning. Maximum likelihood estimation of Dirichlet distributions. *Journal of statistical computation and simulation*, 32(4):215–221, 1989.

[34] O. Rudovic, M. Pantic, and I. Y. Patras. Coupled Gaussian process for pose-invariant facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1357–1369, 2013.

[35] O. Rudovic, I. Patras, and M. Pantic. Regression-based multi-view facial expression recognition. In *Proceedings of the 20th International Conference of Pattern Recognition (ICPR)*, pages 4121–4124, Beijing, China, Aug. 2010.

[36] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, pages 200–215, 2011.

[37] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1132, 2015.

[38] B. Sergio, H. Isabelle, C. Eva, and B. Sandra. Recognizing emotions from video in a continuous 2D space. *Human-Computer Interaction*, 6949:600–603, 2011.

[39] S. E. Shepstone, Z. Tan, and S. H. Jensen. Using audio-derived affective offset to enhance TV recommendation. *IEEE Transactions on Multimedia*, 16(7):1999–2010, 2014.

[40] K. Sikka, T. Wu, J. Susskind, and M. Bartlett. Exploring bag of words architectures in the facial expression domain. In *Proceedings of European Conference on Computer Vision (ECCV). Workshops and Demonstrations*, pages 250–259, Frience, Italy, Oct. 2012.

[41] W. Steffen, S. Stefan, and S. M. et. al. Multimodal emotion classification in naturalistic user behavior. *Human-Computer Interaction*, 6763:603–611, 2011.

[42] J. Sung and D. Kim. Real-time facial expression recognition using STAAM and layered GDA classifier. *Image and Vision Computing*, 27(9):1313–1325, 2009.

[43] U. Tariq, K. Lin, Z. Li, and X. Zhou. Recognizing emotions from an ensemble of features. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):1017–1026, 2012.

[44] U. Tariq, J. Yang, and T. S. Huang. Multi-view facial expression recognition analysis with generic sparse coding feature. In *proceedings of European Conference on Computer Vision (ECCV), Workshops and Demonstrations*, pages 578–588, Frienze, Italy, Oct. 2012.

[45] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):258–273, 2010.

[46] N. S. Vu and A. Caplier. Patch-based similarity HMMs for face recognition with a single reference image. In *Proceedings of 20th International Conference on Pattern Recognition (ICPR)*, pages 1204–1207, 2010.

[47] S. Yang and B. Bhanu. Understanding discrete facial expressions in video using an emotion avatar image. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):980–992, 2012.

[48] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3D facial expression database for facial behavior research. In *Proceedings of the 7th international conference on Automatic face and gesture recognition*, pages 211–216, Southampton, UK, Apr. 2006.

[49] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[50] W. Zheng. Multi-view facial expression recognition based on group sparse reduced-rank regression. *IEEE Transactions on Affective Computing*, 5(1):71–85, 2014.

[51] W. Zheng, T. Hao, Z. Lin, and T. S. Huang. A novel approach to expression recognition from non-frontal face images. In *Proceedings of IEEE 12th International Conference on Computer Vision (ICCV)*, pages 1901–1908, Kyoto, Japan, Sep. 2009.

[52] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886, Providence, Rhode, Island, Jun. 2012.

[53] Z. Zhu and Q. Ji. Robust real-time face pose and facial expression recovery. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 681–688, New York, USA, June 2006.

Qirong Mao received her MS and PhD degree from Jiangsu University, Zhenjiang, P. R. China in 2002 and 2009, both in computer application technology. She is currently an associate professor of the School of Computer Science and Communication Engineering, Jiangsu University. Her research interests include affective computing, pattern recognition, and multimedia analysis. Her research is supported by the National Science Foundation of China (NSFC), Jiangsu province, and the Education Department of Jiangsu province. She has published over 40 technical articles, some of them in premium journals and conferences such as ACM Multimedia. She is a member of the IEEE.

Qiyu Rao received his BS degree in automation (numerical control technology) from Nanjing Institute of Technology, Nanjing, P. R. China in 2014. He is currently a MS candidate in computer science and technology with the School of Computer Science and Communication Engineering at Jiangsu University. His research interests include affect computing and deep learning.

Yongbin Yu received his MS degree in computer science and technology from Jiangsu University, Zhenjiang, P. R. China in 2015. He is currently a researcher in FiberHome StarrySky Co. LTD. of Nanjing, China. His research interests include affect computing and pattern recognition.

Ming Dong received his BS degree from Shanghai Jiao Tong University, Shanghai, P.R. China in 1995 and his PhD degree from the University of Cincinnati, Ohio, in 2001, both in electrical engineering. He is currently an associate professor of Computer Science and the Director of the Machine Vision and Pattern Recognition Laboratory. His research interests include pattern recognition, data mining, and multimedia analysis. His research is supported by the US National Science Foundation (NSF) and State of Michigan. He has published over 90 technical articles, many in premium journals and conferences such as IEEE Trans. on Pattern Analysis and Machine Intelligence, IEEE Trans. on Visualization and Computer Graphics, IEEE Trans. on Neural Networks, IEEE Trans. on Computers, IEEE Trans. on Knowledge and Data Engineering, IEEE ICDM, IEEE CVPR, ACM Multimedia, and WWW. He was an associate editor of IEEE Trans. on Neural Networks, Pattern Analysis and Applications (Springer) and was on the editorial board of International Journal of Semantic Web and Information Systems. He also serves as a program committee member for many related conferences. He is a member of the IEEE.