Cascaded Multi-level Transformed Dirichlet Process for Multi-pose Facial Expression Recognition

Qirong Mao,Feifei Zhang,Liangjun Wang,Sidian Luo[1] and Ming Dong[2]

[1]*Department of Computer Science and Communication Engineering, Jiangsu University*
[2]*Department of Computer Science, Wayne State University, Detroit, MI 48202, USA*
Email: {mao_qr,susanzhang,ljwang0819}@mail.ujs.edu.cn, mdong@cs.wayne.edu

As an essential way of human emotional behavior understanding, facial expression recognition has been studied extensively in recent years. However, the existing methods of facial expression recognition are typically based on near-frontal face data. High recognition accuracy for multi-pose facial expression recognition continues to be a challenge. In this paper, we present a novel cascaded multi-level Transformed Dirichlet Process (cml-TDP) model for multi-pose facial expression recognition. The top-level structure of the cml-TDP model has been carefully designed to make coarse-to-fine prediction, and the outputs of the model are fused for robust and accurate estimation at each level. There are three primary merits to cml-TDP. First, pose is explicitly introduced into cml-TDP so that separate training and parameter tuning for each pose is not required. Second, cml-TDP describes an image by its detected positions and appearance features to implicitly construct geometric constraints. Third, cml-TDP can learn an intermediate facial expression representation subject to geometric constraints. By sharing the pool of spatially-coherent features over expressions and poses, we provide a scalable solution for multi-pose facial expression recognition. The proposed model has been evaluated on two benchmark databases, BU-3DFE and RAFD, and achieved 79.33% and 75.00% facial expression recognition accuracy on these two datasets respectively, which has outperformed current state-of-the-art facial expression recognition methods.

Keywords: facial expression recognition, multi-level Transformed Dirichlet Process, appearance features, geometric constraints, spatially-coherent 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 attentions in multimedia research. It has tremendous impact to a wide-range of applications, interactive games, online/remote education, entertainment, and intelligent transportation systems, which make FER a core component in the next generation of computer system [1, 2, 3, 4, 5, 6].

In recent years, though great strides have been made in FER, most of these studies are conducted in front or near-front face data with good lighting conditions [7]. The performance of these systems degenerates greatly in multi-pose scenarios, therefore new challenges arise due to large variations on expressions attributed to factors such as pose and illumination. More recently, although a handful of methods on multi-pose FER have been proposed [8, 9, 10, 11, 12], these methods are usually trained multiple models for each specific pose and thus they need parameter-tuned separately for each model [13, 14], which is time consuming. Thus, there is an ever growing need for automated systems that can accurately perform multi-view FER.

For multi-pose FER, one challenge is to learn discriminative, pose robust features from facial images. Another important challenge is to effectively exploit the relationship between different poses in order to facilitate the multi-pose expression classification with a unified model.

Based on how to represent the facial images, multi-pose FER can be divided into three categories: 1) texture-based local features [8, 15, 16], e.g., Local Binary Pattern (LBP), Scale-invariant feature transform (SIFT) and Histogram of Oriented Gradient (Hog), 2) geometry-based global features [10, 17], 3) hybrid features [18, 19]. However, these features are frequently used as disorder and independence with

each other when they represent an image, which are usually extracted from a full face image with irrelevant information. Thus, these features can not construct the relationships between different key regions during the training process. In this paper, we extract texture-based features from key regions, and then we represent a face image with the texture-based features and the corresponding location information of each key region. Hence, in our model, we represent an image by spatially-coherent features.

Classification method for multi-pose FER is another important factor that influences the recognition accuracy and efficiency. According to how they deal with the variations in head-pose and expressions in 2D images, they can be divided into three categories: 1) methods that learn a single classifier with pose-robust features for multi-view FER [20, 21, 22], 2) methods that perform pose normalization before conducting multi-view FER [8, 10], and 3) methods that learn multiple classifiers for each specific poses [14, 13]. However, the main downside of these approaches is that they fail to exploit the relationships among different poses. This, in turn, results in classifiers being less robust for the multi-pose FER task, especially when the number of poses are large, these methods will become more complex.

Recently, some papers show that intermediate features are very helpful for image understanding and expression/object recognition [23, 24, 25, 26]. The key idea is to integrate low-level features into an intuitive intermediate representation before classifying or recognizing, which can be shared across different environments (e.g., poses and illumination) to improve the learning performance and help us obtain high recognition rate. Transformed Dirichlet Process (TDP) not only can learn the intermediate features, but can construct geometric constraints between different regions. Furthermore, according to the facial action coding system (FACS) developed by Ekman et al. [27], the facial regions around the eyes, nose and mouth contain much more action units than other regions of a face. Hence, the appearance and geometric features of these facial regions may contribute more to the expression recognition. Third, paper [28] points out that it can enforce the spatial coherency of a model through combining the texture-based features and the corresponding geometry information. Fourth, the key parts segmentation and FER can be finished in a unified TDP model.

Inspired by the intuitions mentioned above, in this paper, we propose a novel graphical model, cml-TDP, for unified multi-pose FER. In our approach, pose is explicitly introduced into a multi-level, spatially-coherent Transformed Dirichlet Process (TDP) model for robust key parts segmentation and expression recognition. Consequently, our model provides a unified solution for multi-pose FER, instead of training multiple models for each specific pose, which will greatly improve training efficiency and avoid parameter-tuned separately for each pose.

The preliminary version of this work was first presented in a shortened form as a conference abstract [29]. The major contributions of this paper can be summarized as follows:

1). Pose is explicitly introduced in cml-TDP to learn a relationship among different views, so that separate training and parameter tuning for each pose is not required. Thus, cml-TDP provides a scalable solution for multi-pose FER.
2). TDP is a hierarchical probabilistic theme model and used to describe the spatial structure of a facial image. In addition, it can learn the latent topic distribution subject to geometric constraints, which is an improvement over the feature independence assumption made in Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) models.
3). A facial image in our model is described by the segmented parts including the positions and corresponding appearance features. The geometric constraints among different facial parts are implicitly encoded in our model and integrated with the local features (represented by codewords) to facilitate multi-pose FER. Thus, cml-TDP can obtain high recognition rate even with large pose variation.
4). cml-TDP provides a cascaded model to make coarse-to-fine multi-pose FER. It integrates parts segmentation and FER in a unified graphical model.

The rest of the paper is organized as follows: Section 2 introduces the related work. In Section 3, we develop cml-TDP, a novel unified model for key parts segmentation and multi-pose FER. Section 4 describes the facial expression feature extraction and representation. The experimental results are illustrated in Section 5. We discuss our results and future directions in Section 6.

2. RELATED WORK

Extensive efforts have been devoted to recognize facial expressions [23, 2, 30, 31, 6, 32]. Most of the existing works for FER are based on the six basic emotions, e.g., happy, sad, disgust, surprise, fear and angry, due to their marked reference representation in our affective lives and the availability of the relevant training and test data [7]. At the beginning, studies were mainly based on frontal or near frontal facial images. Later on, efforts have been reported on the multi-pose facial expressions. Generally, feature representation and expression recognition are the most important parts in FER. In the following, we will review feature representation and FER methods on multi-pose conditions which have tight connections with our work.

Feature representation: Features for FER can be roughly classified into three categories: 1) texture-based local features, 2) geometry-based global features, and 3) hybrid features.

For the texture-based local features, it can be further classified into three subgroups: the first one is Gabor wavelet, which is a powerful, but time-consuming. LBP is the second one, which is usually used on arbitrarily gridded sub-regions of images. Third, SIFT descriptor, the most popular texture features as it is scale invariance. Whitehil, et al.[33] adopt the Gabor in GentleBoost and SVMs classifier to realize a practical smile detection. LBP and SIFT features are also extensively used in [34, 18, 35]. For the geometry-based global features [36, 10], 39 landmark points are located from each non-frontal head pose and then normalized to the frontal one through Coupled Scaled Gaussian Process Regression (CSGPR) model. Experiments of the FER are finally carried out based on the normalized 39 landmark points. For the hybrid features, in [37], Gupta et al. propose a hybrid method of feature extraction using Discrete Cosine Transform (DCT), Gabor Filter, Wavelet Transform and Gaussian distribution to improve the recognition rate.

However, the feature representation methods mentioned above can only describe a face image as orderless. They lack the power to describe the spatially coherent images. Recently, many researchers combine the appearance features and geometric features to learn geometric constraints and improve the ability of multi-pose FER [18]. More recently, features learned by machine learning models have attracted lots of attention, such as auto-encoder (AE) and Convolutional Neural Networks (CNNs), which have the ability to automatically extract useful representations from raw data. However, there are few works using them for FER tasks [38, 32, 39]. The major reason is that the labeled facial expression data is too small in current FER datasets, so a feature learning method that has many parameters can easily fall into overfitting when training.

Recognition method: Recognition method is another important factor for FER. Recent advances toward automatic multi-pose FER can be classified into three groups: 1) methods that learn a single classifier for multi-view FER but ignoring the influence caused by poses, 2) methods that perform pose normalization before performing multi-view FER, and 3) methods that learn multiple classifiers for each specific poses.

For the first group, researchers usually learn low-level features which are robust to pose variations on pre-labeled landmark points or a full face image, such as variants of SIFT. Specifically, in [21], authors use the region covariance matrix that obtained by computing the covariance of the SIFT vectors which are extracted from each facial image. However, since the SIFT features are extracted from arbitrary view facial images, they carry much information that are irrelevant to the emotion recognition. The authors have to propose a discriminative feature extraction method to reduce the irrelevance among the features. This can easily lost some valuable information such as features from a not exaggerated expression, e.g., a sightly upper mouth or wrinkling nose, which actually play an important role in FER. For the second group, in literatures [10, 36], 39 landmark points are located from each non-frontal head pose, and then the authors propose a Gaussian process regression model to exploits pair-wise correlations among different poses in order to learn robust mappings from non-frontal to the frontal pose. The performance of these approaches can be easily affected by the errors in the pose normalization step. For the third group [14, 13], view-specific SVMs are learnt for each view. These approaches ignore relationships among different poses, which make them suboptimal for the multi-view FER, especially with the increase of poses.

In addition to the limitations mentioned above, most of the classifiers represent images as orderless collection of local features and assume local pathces of an image are independent with each other. In this paper, we introduce a spatial transformations latent topic model, which could construct a geometric constraints through jointing the appearance features with geometric features, and then it can learn intermediate features. More important, we add the factor of pose in the traditional TDP model.

3. THE GENERATIVE MODEL OF CML-TDP

In this paper, we propose a cascaded multi-level transformed dirichlet process (cml-TDP) model for robust key parts segmentation, and FER. The architecture of our FER system is shown in Figure 1, which cascades two levels of TDP to make coarse-to-fine prediction for key parts segmentation and expression recognition, one level per task. Before passing an image to our system, we first perform face detection and pose estimation using a tree-based part model [40]. After the preprocessing, we apply our cml-TDP for key parts segmentation and expression recognition. Specifically, in the first level (key parts segmentation) of cml-TDP, we detect the key parts (e.g., eyes, nose, mouth) in the facial area using both geometry and SIFT features extracted from Elliptical Interest Regions (EIRs) (see Section 4.2 for details). Then, in the second level, features from the key parts are used to FER by our cml-TDP model.

3.1. Dirichlet Process & HDP & TDP

In this section, we give a brief review of Dirichlet Process (DP) [41], Hierarchical Dirichlet Process (HDP)[42] and Transformed Dirichlet Process (TDP)[43].

Figure 2(a) shows a graphical illustration of DP [41]. It is a Bayesian nonparametric probabilistic

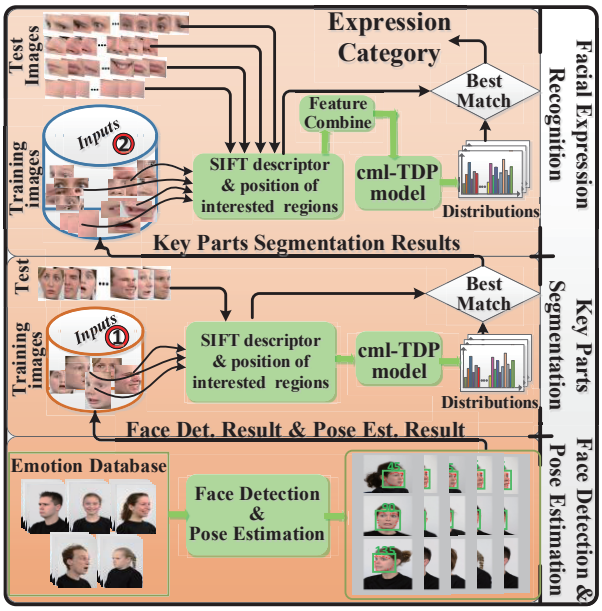


FIGURE 1. System architecture

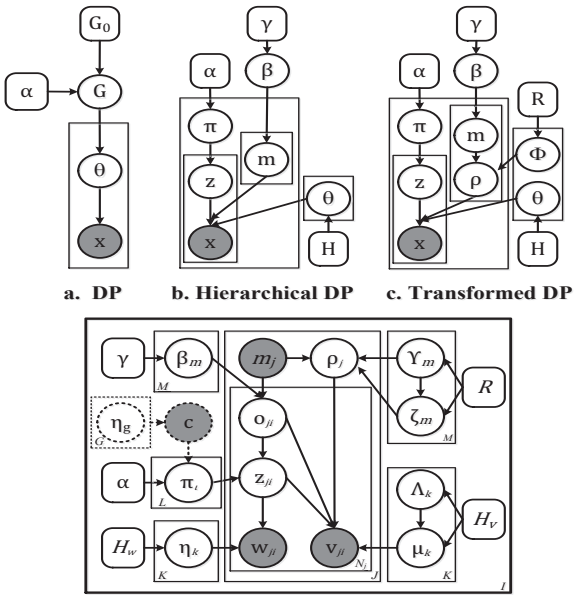


FIGURE 2. Graphical model depiction of DP (a), HDP (b), TDP (c) and cml-TDP (d) models.

model where a Dirichlet random variable θ with k-dimensionality has the property: $\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$. DP describes the distribution of θ with the following probability density:

$$DP(\alpha, \theta) = \frac{\Gamma(\Sigma_{i=1}^{k} \alpha_i)}{\Pi_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1}, \qquad (1)$$

where the parameter α is a k-vector with components $\alpha_i > 1$ and Γ is the Gamma function.

Figure 2(b) shows a graphical illustration of HDP [42]. It is an extension of the DP, mainly used for clustering grouped data. HDP uses a DP for each group of data, with the DPs for all groups sharing a base distribution which is itself drawn from a DP. HDP is also a non-parametric bayesian model that infers the number of latent themes from training data and assumes a hierarchical structure so that data in different groups can share the same themes. However, whether HDP or DP, they all ignore the relative spatial locations of local patches.

Figure 2(c) shows a graphical illustration of TDP [43]. TDP model generalizes HDP by applying a random set of transformations to each clustering center, and therefore the dependencies of the local patches are encoded. Our approach differs from the traditional TDP model in two aspects: 1) we introduce a variable c for pose in TDP to cope with the multi-pose FER, rather than using separated models for different poses; and 2) we build a multi-level structure to provide a unified solution of feature representation and expression recognition.

3.2. Structure of the cml-TDP

We introduce our cml-TDP model for multi-pose FER in details in this section. Figure 2(d) shows a graphical illustration of our cml-TDP model, in which variable I denotes the different levels, and variable c for poses. Pose is explicitly introduced as an observable variable because it is one of the factors that could influence the results of multi-pose FER. In addition, it can be estimated with high accuracy. Note that the graphical model of each level is the same, shown in Figure 2(d). The only difference between the levels is the input, which precisely prove our model is a faithfully unified model. Next we use the second level (expression recognition level) as an example to present our model in details.

The input to the expression recognition level is the segmented key parts. After segmentation, supposing we have J facial parts (e.g., noses, mouths, eyes) in total, belonging to C poses and M expressions, our model can be described by the following generative process.

1). EIR is the basic unit of an image, and the ith EIR of image j is described by its detected location v_{ji} and discrete appearance feature w_{ji}, typically represented by the codewords in a codebook. That is $EIR_i = (v, w, t)$ with label $t = c \cdot m$, where $c = \{1, ..., C\}$ and $m = \{1, ..., M\}$ are the pose and facial expression labels respectively.
2). The expression transformation ρ_j provides a reference to each of L latent objects o_{ji}, and is used to model their spatial relationship. Specifically, $\rho_j \sim N(\xi_m, \Upsilon_m)$ with normal-inverse-Wishart priors $(\xi, \Upsilon) \sim R$, where N is a Gaussian transformation distribution for each facial expression.
3). For a 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 latent topics z, where α is a Dirichlet prior on the training sets.

4). An image is a collection of N EIRs, denoted by $j = (EIR_1, EIR_2, ..., EIR_N)$.
(1)Choose a latent topic $z \sim p(z|\pi)$, where $p(.)$ is a multinomial distribution. z is a K-dimensional unit vector and K is the number of latent topics.
(2)Choose a $EIR \sim p(EIR|z,\eta)$ with a latent topic z, where η is a $K \times T$ matrix, T is the total number of codewords in the codebook.

In our model, the Dirichlet parameter α is at the pose-levels, sampled once in the process of generating a pose. The multinomial variable π is at the theme-level, sampled once per face image. Through these parameters, the model can learn the relationships among different poses by sharing the pool of features.

Given the expression transformation parameter ρ and η, feature distributions are determined by:

$$\begin{aligned} &w_{ji} \sim \eta_{z_{ji}}, \\ &v_{ji} \sim N(\tau(\mu_{z_{ji}}, \Lambda_{z_{ji}}; \rho_{j\ell})), o_{ji} = \ell, \\ &\tau(\mu, \Lambda; \rho) = (\mu + \rho, \Lambda). \end{aligned} \tag{2}$$

Given the parameters ρ, π, c, o belonging to the ℓth latent object, the features w, v and theme z are computed via a finite mixture model:

$$\begin{aligned} &\mathrm{p}(w_{ji}, v_{ji}, z_{ji} | \rho_j, \pi_\ell, c, o_{ji} = \ell) \\ &= p(c|\eta) \sum_{k=1}^{K} \pi_{\ell k} \eta_k(w_{ji}) N(v_{ji}; \mu_k + \rho_{j\ell}, \Lambda_k) \\ &= p(c|\eta) \sum_{k=1}^{K} p(\pi_{\ell k}|c,\alpha) p(w_{ji}|z_{ji}, \eta_k) p(v_{ji}|\rho_j, \mu_k, \Lambda_k), \end{aligned} \tag{3}$$

with

$$p(c|\eta) = Mult(c|\eta), \tag{4}$$

$$p(\pi_{\ell k}|c,\alpha) = \prod_{r=1}^{C} Dir(\pi_{\ell k}|c,\alpha)^{\delta(c,r)}, \tag{5}$$

$$p(w_{ji}|z_{ji}, \eta_k) = Mult(w_{ji}|z_{ji}, \eta_k), \tag{6}$$

$$p(v_{ji}|\rho_j, \mu_k, \Lambda_k) = \frac{1}{\sqrt{2\pi\Lambda_k}} \exp\{\frac{(v_{ji} - (\mu_k + \rho_{j\ell}))^2}{2\Lambda_k}\}, \tag{7}$$

where K is the number of latent topics; $Mult$ represents the multinomial distribution and Dir denotes the Dirichlet distribution. The overall algorithm is shown in Algorithm 1.

3.3. Gibbs Sampling for cml-TDP model

In this section, we describe a Markov Chain Monte Carlo (MCMC) [44] sampling scheme for cml-TDP.

Algorithm 1 cml-TDP($\{w, v\}$, α, γ, K, t)

Input: appearance feature w, geometry feature v, hyperparameters α, γ, topic number K, label t
Global data: count statistics$\{M_{ml}\}$, $\{N_{lk}\}$, $\{C_{kw}\}$
Output: topic associations, multinomial parameters π, hyperparameter estimates α, γ, expression category m

```
// initialization
for all facial expression images j ∈ [1, J] do
    for all EIR in image j do
        sample latent topic z ~ p(z|π)
        increment feature-topic count {C_kw} to cached statistics
        increment topic-object count {M_ml}, {N_lk} to cached statistics
while not finished do
    for all facial expression images j ∈ [1, J] do
        for all EIR in image j do
            decrement feature-topic count {C_kw}
            decrement topic-object count {M_ml}, {N_lk}
            determine the predictive likelihood f^t_lk(w_ji = w, v_ji) //Eq.8
            sample new objects and topics //Eq.9
            increment feature-topic count {C_kw} to the new topic
            increment topic-object count {M_ml}, {N_lk} to the new object
if converged then
    read out parameter set θ in Eq.10
get expression category by maximizing probability in Eq.11
```

Our goal is to obtain a probability matrix and the theme distribution for every expression. The clustering property can be better understood via the Chinese restaurant franchise representation [45]. It assumes that there is a set of restaurants (each restaurant denoting an image). These restaurants share the same dishes(corresponding to the latent themes). Customers in each restaurant correspond to the EIRs in an image. In our dependent model, we will cluster the dependent EIRs into the same theme. Metaphorically, it means that customers (EIRs from multi-view face data) will order the same dish (latent themes). The dependency information between two EIRs is encoded by the process. Specifically, the sampling process can be described as follows. First, we remove feature (w_{ji}, v_{ji}) from the cached statistics for its current theme and expression category. Second, for each of the $L \cdot K$ pairs of latent key regions categories and themes, we determine the predictive likelihood as

$$f^t_{\ell k}(w_{ji} = w, v_{ji}) = (\frac{C_{kw} + \lambda/T}{\sum_{w'} C_{kw'} + \lambda}) \cdot N(v_{ji} - \rho^{(t-1)}_{jl}; \widehat{\mu}_k, \widehat{\Lambda}_k), \tag{8}$$

where $i \in \{\tau(1), ..., \tau(n_j)\}$, n_j is the number of features in image j, and $\tau(\cdot)$ denotes a random permutation. $v_{ji} - \rho^{(t-1)}_{jl}$ is the location parameter where the superscript t is the index of sampling iterations. $(\widehat{\mu}_k, \widehat{\Lambda}_k)$ is the scale parameter. C_{kw} is the number of times each appearance feature assigned to topic k, and T is the dictionary size through k-means clustering. Third, we sample new objects and topics from the following $L \cdot K$-dim multinomial distribution:

$$\begin{aligned} &(o_{ji}, z_{ji}) \sim \sum_{l=1}^{L} \sum_{K=1}^{k} (M_{m\ell} + \gamma/L), \\ &(\frac{N_{lk} + \alpha/K}{\sum_{k'} N_{lk'} + \alpha}) f_{lk}(w_{ji}, v_{ji}) \delta(o_{ji}, l) \delta(z_{ji}, k), \end{aligned} \tag{9}$$

where $M_{m\ell}$ is the number of the features associated with each latent object, and $N_{\ell k}$ represents the number of the features associated with each latent theme. Fourth, feature (w_{ji}, v_{ji}) is added to the cached statistics for its new object and topic. Finally, we fix (o_j, z_j) and sample a new reference transformation ρ_j^t, and update the features for theme $k = z_{ji}$ accordingly. The training process is stopped at the convergence of MCMC.

3.4. Bayesian Inference

In order to perform FER, we need to construct a probability matrix for each facial expressions m. Let θ_i^m denote the mixture components, which is decided by the multinomial distribution η_k for the appearance features and the Gaussian distribution for the location features $N(\mu_k + \rho_j, \Lambda_k)$. Specifically, the probability matrix can be represented by $p(w_j, v_j \,|\theta_i^m)$. Assuming that an unknown testing image j has N local EIRs, represented by features $(w_n, v_n), n = \{1, ..., N\}$, the probability $p(j\,|m)$ is calculated as,

$$\begin{aligned} p(j\,|m) &= \prod_{n=1:N} p(\mathrm{w}_n, v_n\,|m) \\ &= \prod_{n=1:N} (\sum_i p(w_n, v_n\,|\theta_i^m) p(\theta_i^m\,|m)). \end{aligned} \tag{10}$$

Expression recognition is then achieved by maximizing the probability,

$$m = \arg\max_m p(j\,|m). \tag{11}$$

So far the FER is implemented. Compared with other models, our model can recognize multi-pose facial expression in a unified model without tuning parameters separated.

4. EXPRESSION FEATURE EXTRACTION

4.1. Face Detection and Pose Estimation

We adopted the tree-based part model in [40] to perform simultaneous face detection and pose estimation. This method entails invariance to transformations such as scale, translation and in-plane rotations. Compared with Active Appearance Models (AAMs) [46], this model captures more of the relevant elastic deformation so that it is more suitable for face detection with different poses. However, note that the improvement of performance is achieved with higher computational complexity. The tree-based part model generally runs slower than AAM. Specifically, the face detection and pose estimation time of the tree-based part model is about 28s for a facial image with 512*428 pixel, while it is nearly real-time for AAM. The bottom left panel of Figure 3 shows some examples of face detection and pose estimation. The pose parameter only need to estimate in this process and it will be a label in the segmentation and recognition level.

4.2. Spatial-coherent Feature Encoding

The detailed feature extraction process is summarized in Figure 3. It shows that for each input image, we first perform face detection and pose estimation, then detect the corner-like structure by finding pixels with significant second derivatives. Through the Laplacian of Gaussian operator, we can detect a characteristic scale for each corner. After this, we find candidate edges via a canny detector, and link them into segments broken at points of high curvature. These lines then form the major axes of EIRs which is shown in the region detection part of Figure 3. EIR is the basic unit of an image, from which 128-dimensional SIFT features are extracted and the center location v are getting simultaneously. Given the collection of EIRs from the training images, the codebook of SIFT descriptors is constructed by k-means algorithm. The codebook is composed of central EIRs of all clusters, usually refereed to as codewords. The appearance feature w of each EIR is represented by the codeword in a codebook having the minimal Euclidean distance, and then, the ith EIR of image j is described by the detected location v_{ji} and appearance feature w_{ji}. That is $EIR_i = (v, w, t)$ with label $t = c \cdot m$, where $c = \{1, ..., C\}$ and $m = \{1, ..., M\}$ are the pose and facial expression labels, respectively. Finally, each image is represented as the collection of all the EIRs in it. That is $image_j = \{(w_{j1}, v_{j1}, t_{j1}), (w_{j1}, v_{j1}, t_{j1}), ..., (w_{jn}, v_{jn}, t_{jn})\}$ with n EIRs.

Through this way, the Feature Vector with Geometric Constraints (FVGC) of each EIRs is constructed. Based on the collected training data, we build the cml-TDP model and obtain the distribution of codewords on latent themes and that of themes on each expression and pose.

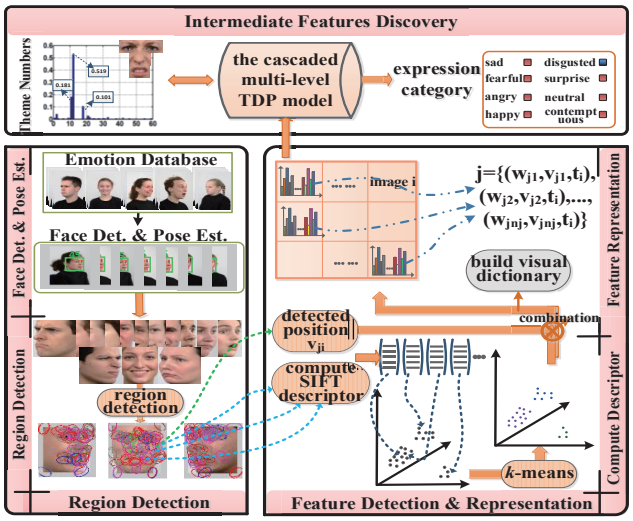


FIGURE 3. The flow chart of expression feature extraction

4.3. Intermediate Features

Based on the collected training data, we can build the multi-level probabilistic 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 spatially-coherent expression features in a unified framework, scalable to a large number of poses and expressions. 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 Figure 3). The latent aspects of facial images, hidden behind the feature vector, once discovered, are well suited to reveal the distinctions between multi-pose facial expressions, and thus lead to higher accuracy of recognition.

5. EXPERIMENTS & RESULTS

5.1. Datasets

We evaluated our cml-TDP on two public facial expression datasets: 1) one 3D facial expression database: BU-3DFE [47]; and 2) s public multi-pose facial expression databases: the Radboud Faces Database (RAFD) [48].

1) BU-3DFE: The 3D facial expression database, namely BU-3DFE, has 100 subjects with 3D models and face images. The 100 subjects include undergraduates, graduates and faculty from the State University of New York Binghamton. Age ranges from 18 years to 70 years old. The database consists of 60% female and 40% male with a variety of ethnicity (White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino). Besides, this dataset contains images depicting seven facial expressions of anger, disgust, fear, happiness, sadness, surprise and neutral. With the exception of the neutral expression, each of the six prototypic expressions includes four levels of intensity.

Implementation details for BU-3DFE: In our experiments, we first render 2D facial images from the 3D models at the four levels of intensity, and six universal facial expressions, i.e., anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA) and surprise (SU). To investigate the multi-pose expression recognition issue, we render the facial images with 5 yaw views $(90°, 60°, 45°, 30°, 0°)$. Consequently, we have $100 \times 6 \times 5 \times 4 = 12,000$ facial images in total for our experiments.

We randomly divide the 100 subjects into a training dataset with 80 subjects and a testing dataset with 20 subjects, such that there are no overlap between the training subjects and the testing subjects. As a result, the training dataset comprises 9,600 facial images whereas the testing one comprises 2,400 facial images. To determine the hyperparameters, we randomly select 1,920 facial images from the training dataset to form an independent validation set. Ten independent trials of experiments are conducted in the experiments.

2) RAFD: RAFD is a public facial expression database with displayed expressions, gaze direction, and varied head orientation, which can be downloaded freely on the website. Specifically, it contains a set of pictures of 67 actors (including Caucasian males and females, Caucasian children, both girls and boys, and Moroccan Dutch males) displaying eight different emotions (i.e., anger, disgust, fear, happiness, sadness, surprise, contemptuous and neutral) with three different gaze directions (right, frontal, and left) and five different poses $(180°, 135°, 90°, 45°$ and $0°)$, which have 8,040 face images at all.

Implementation details for RAFD: In our experiments, we consider all of the eight emotions with three gaze directions and three yaw views $(135°, 90°, 45°)$. Consequently, we have $67 \times 8 \times 3 \times 3 = 4,824$ facial images in total for our experiments. The data in our experiment are split into five folds, four for training and one for testing. Thus the size of the training and testing sets is 3,888 and 936 images respectively.

5.2. Experiment Setting

We evaluate our method under three cases on the two public datasets mentioned above : 1)comparison with two well-established features; 2) comparison with three well-established methods; 3) comparison with several state-of-the-art methods [49, 50, 19, 14, 18, 8]. The detailed settings are introduced as follows:

Comparison experiments of two well-established features: In order to evaluate the spatial-coherent features proposed in this paper, we compare it with two well-established features: LBP [51] and Sparse SIFT (SSIFT) [52]. Specifically, the method in [51] is used to extract LBP features from different number of key regions (68 for the facial image with pose 0°, 30° and 45°, 38 for the facial image with pose 60° and 90°) of facial images. For each key region, the LBP histogram is computed using 8 sampling points on a circle of radius 1. Then we concatenate all the LBP histograms into a single feature to represent a facial image. For the SIFT features, we use the method provided by [52]. We first extract 128-dimensional SIFT features around the key points and obtain a high dimensional feature vector. Then, Principal Component Analysis (PCA) is used to reduce the feature vector to 500-dimension to represent the facial image. Both LBP and SSIFT features are extracted per-pose, and then fed into the TDP classifiers.

Comparison experiments with three well-established methods: In order to verify the performance of our cml-TDP model, we compare it with three other state-of-the-art methods: the supervised

Latent Dirichlet Allocation (sLDA) [53] model, TDP [44] model and multi-SVM [13]. sLDA and TDP are trained based on the facial images with expressions regardless of poses. Similarly, the multi-SVM model consists of 5 SVMs, each trained separately with expression images under the specific pose. The parameters $\mathcal{C} \in [-1, 10]$ and $\sigma \in [-1, 1.5]$ of each SVM are tuned by grid search with a step of 0.1. During recognition, the pose of an input facial image is first estimated, and then the corresponding model is used to recognize the face into one expression category.

Comparison experiments with five state-of-the-art methods: We finally evaluate our method with respect to the classification accuracy and compare it with the previously proposed methods in [49, 50, 19, 14, 18]. More specifically, Hu et al. used normalized ground truth landmarks to build a facial expression system robust to non-frontal-view face in [49]. In [50], they investigated another multi-view FER method, with a goal to improve performance by taking into account the influence of viewing angles of facial images. Zheng et al. developed a theory of non-frontal FER, which is based on minimizing an estimated closed-form Bayes error [19]. In [14], the authors investigated the effects of pose on FER using variations of LBPs (Local Uniform Binary Patterns (LBP^{u^2}), multi-scale LBP (LBP^{m^s}), Local Gabor Binary Patterns ($LGBP$), and $LGBP/LBP^{m^s}$) at different resolutions and different grid sampling sizes. In [18], the author proposed a novel group sparse reduced-rank regression (GSRRR) model to describe the relationship between the multi-view facial feature vectors and the corresponding expression class label vectors, and obtained the best results so far.

In the sampling procedure, α in Eq.3 and γ in Eq.9 are empirically set as 0.1 and 1.5 respectively. The training and recognition experiments are done on a workstation having 12 CPUs with 3.5Ghz and 64GB memory, running Matlab R2015a on the Linux Operating System.

5.3. Parameter Selection

In this section, we shall evaluate the effects of three important paraments: 1) parameter *scale*, which control the density of local EIRs; 2) total number of codewords; 3) number of training iteration loop. The parameters are tuned by grid search using the validation set. We first set a fixed number of codewords and iteration loop and then evaluate the effect of different density of local EIRs. Afterwards, we evaluate the effect of number of codewords with the former chosen parameter *scale*. Similar, we evaluate the effect of number of training iteration loop with the former chosen parameter *scale* and codewords size. The selection experiments are conduct on BU-3DFE. According to section 5.1, the number of training examples is 7,680, and 1,920 images are used as the validation set to select parameters.

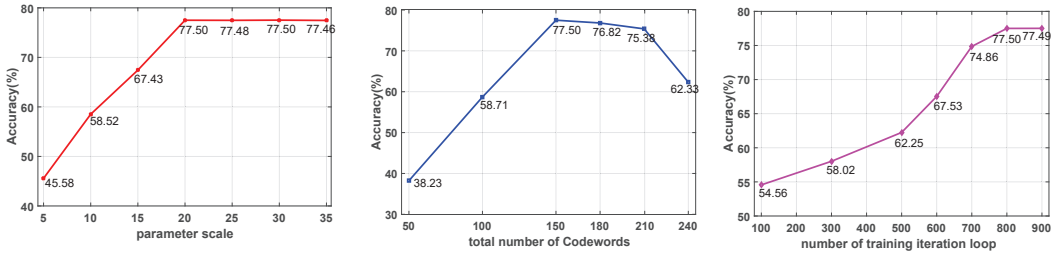
Local EIRs Sampling: In our experiment, we control the density of local EIRs by a parameter *scale*. To evaluate the effect of the density of local EIRs, we first adopt a fixed number of codewords 150 and training iteration loop 800. Then we consider parameter *scale* with 5, 10, 15, 20, 25, 30, 35. As shown in Figure 4(a), we can see that larger density of local EIRs achieve high performance. The reason may be that the higher the density of the local EIRs, the more detailed feature information we can get. However when the parameter *scale* reached 20, the increase of the accuracy tended to be flat. Besides, comparison with *scale* parameter 20, the training time is nearly doubled when it is 35 according to our experiment. Thus we set it as 20 for the trade-off between efficiency and accuracy.

Total Number of Codewords: We fix the parameter *scale* as 20 and training iteration loop as 800 and consider the number of codewords with 50, 100, 150, 180, 210, 240. Generally speaking, the capacity of the feature representation is influenced by the number of the codewords, i.e., when the codewords size is small, the feature representation is relatively simple. We do not use equal step size when select this parameter like the other parameters (parameter scale and number of training iteration loop). We first find a better result with a larger step size in order to save the training time, and then determine the final result with a smaller step size. Figure 4(b) illustrates the effects of the codebook size on the expression recognition accuracy. Notice that the highest average accuracy are achieved when the codebook size is 150. Thus, in our experiments, the size of the codewords is fixed as 150.

Number of Training Iteration Loop: We fix the parameter scale as 20 and number of codewords as 150 and consider the number of training iteration loop with 100, 300, 500, 600, 700, 800, 900. As shown in Figure 4(c), we can see that the average accuracy of the FER increased steady with the increase of the number of training iteration loop between 100 and 500. And then the results of cml-TDP changes sharply until it reach 800. After that, the recognition accuracy become insensible to the increase of the number of training iteration loop, thus we fix it as 800 in our experiments. The trend in this figure is possibly due to the fact that our model constructs the geometric constraints among features. When the number of training iteration loop is small, the dependency among the features cannot overcome the noise which is brought by the dependency among noise patches. As the iteration loop number increases, the usefulness of the dependency structure starts to play a more important role and the performance becomes much better.

5.4. Key Parts Segmentation

In this section, in order to provide intuitive understanding of the key parts segmentation in our cml-TDP, we visualize some exemplar results obtained in our experi-



(a) (b) (c)

FIGURE 4. Parameter selection: (a) Effect of parameter *scale*; (b) Effect of number of codewords; (c) Effect of number of training iteration loop.

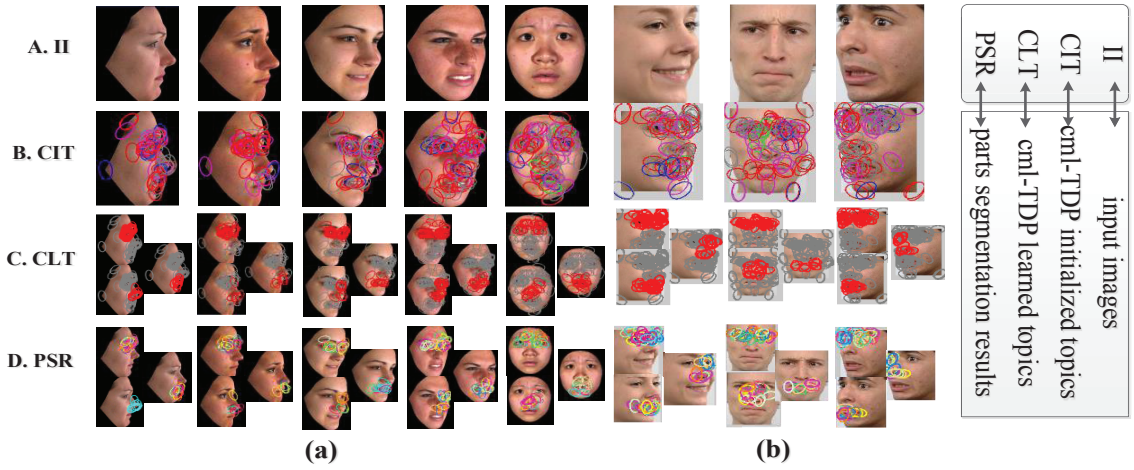


FIGURE 5. (This figure must be viewed in colors) Visualization of the segmentation results in cml-TDP: (a) key parts segmentation results on BU-3DFE, (b) key parts segmentation results on RAFD. II: input images. CIT: cml-TDP initialized topics. CLT: cml-TDP learned topics. PSR: parts segmentation results.

.

ment on BU-3DFE, and RAFD databases.

Figure 5 gives several examples about the key parts segmentation. Specifically, Figure 5(a) shows the segmented key pats associated with different persons and different poses in BU-3DFE. Figure 5(b) shows some of the examples of the segmentation results on RAFD.

In Figure 5, row A provides the raw images of different persons with different poses; row B shows the EIRs detected and their initial latent topic assignments; row C gives the topic assignments after TDP learning; and row D gives the final key parts segmentation results. From the images in row B, we can see that the EIRs are mainly distributed in the regions where the texture has significant changes. A closer look at row C clearly shows that most of the EIRs on the key facial subregions are colored in red, indicating that ml-TDP can successfully perform key parts segmentation on multi-pose images.

5.5. Effect of Key Parts Segmentation and Pose Estimation Error on FER

In this section, experiments are conducted on the BU-3DFE dataset to analyze how the FER accuracy is affected by erroneous key parts segmentation and pose estimation.

Effect of Key Parts Segmentation Noise on FER: In this experiment, the key parts segmentation results of the facial images in the validation set are deliberately corrupted by translating the segmentation subregions with different levels of noise in φ, with $\varphi = 0, 2, 5, 13, 22$ pixels. The mean FER accuracies of noise data with different intervals are reported in Table 1. In each column in Table 1, the results are achieved on the data translated by the same level of noise, and averaged over all expressions for each pose. The last row of Table 1 gives the accuracy averaged over all poses and expressions for each interval of noise.

Clearly, the highest accuracy averaged over all poses and expressions are achieved by the clean data. The mean FER accuracy waves when φ is smaller than 5. Generally, our model is quite tolerable regarding key parts segmentation errors.

Effect of Pose Estimation Noise on FER: In this experiment, pose estimation results of the facial images in the validation set are deliberately corrupted by different levels of noises: $\Delta = 0\%, 2\%, 5\%, 10\%, 20\%$, where $\Delta = i\%$ indicates that $i\%$ pose estimation

TABLE 1. Expression recognition accuracies with different levels of noise ($\varphi = 0$, 2, 5, 13, 22 pixels) through translating the segmented key parts on facial images in BU-3DFE, reported in %. The highest accuracy is highlighted in bold.

Poses	$\varphi = 0$	$\varphi = 2$	$\varphi = 5$	$\varphi = 13$	$\varphi = 22$
0°	80.21	80.10	77.09	72.66	70.31
45°	79.95	77.34	81.25	76.04	69.53
90°	85.68	84.38	85.87	80.47	77.34
135°	73.70	74.48	74.74	67.19	64.32
180°	66.93	64.32	65.10	59.64	54.69
mean	**77.29**	76.12	76.81	71.20	67.24

TABLE 2. FER accuracy with different levels of noise (Δ= 0%, 2%, 5%, 10%, and 20%) added to poses estimation results on facial images in BU-3DFE, reported in %. The highest accuracy is highlighted in bold.

Poses	$\Delta = 0$	$\Delta = 2$	$\Delta = 5$	$\Delta = 10$	$\Delta = 20$
0°	80.21	80.21	79.69	79.43	75.78
45°	79.95	79.43	80.47	78.39	74.22
90°	85.68	86.46	85.42	82.81	78.39
135°	73.70	74.22	72.66	72.40	69.01
180°	66.93	64.84	63.02	60.68	56.77
mean	**77.29**	77.03	76.25	74.74	70.83

results of each pose are randomly changed. The FER accuracies of noise-corrupted data with different levels of noises are reported in Tabel 2. In each column in Tabel 2, the results are achieved on the data corrupted with the same level of noises, and average over all expressions with each pose. The last row of Table 2 shows the accuracy averaged over all poses and expressions.

Clearly, FER accuracy in our model is more sensitive to pose estimation noise than that to key parts segmentation. Regardless of the pose, the highest accuracy are always achieved by the noiseless data. This is mainly because pose is explicitly introduced in our hierarchical theme model.

5.6. Facial Expression Recognition Results

Latent Themes: Figure 6 illustrates the latent theme model learned for each expression of three poses in RAFD. A small panel in the figure shows the feature distribution over the 60 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 FER in the multi-pose conditions.

1) Experiments on the BU-3DFE Database: In this section, we report the recognition accuracy on BU-3DFE in four cases: first, detailed recognition accuracy on BU-3DFE. Second, comparison with two well-established features to verify the validity of our features. Third, comparison with three well-established methods to verify the validity of our model. At last, comparison with five previously proposed methods in the literature [49, 50, 19, 14, 18].

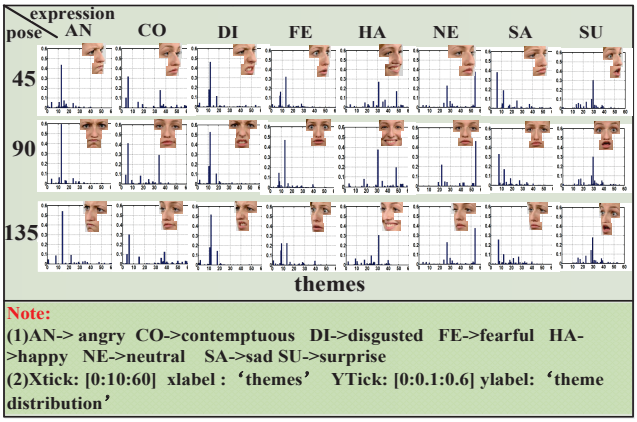


FIGURE 6. Theme distributions. Each row represents one pose and each column represents one expression. The panel shows the mean distribution of the 60 latent expression themes on different poses and expressions. AN: angry; CO: contemptuous; DI: disgusted; FE: fearful; HA: happy; NE: neutral; SA: sad; SU: surprise. XTick: [0:10:60]; xlabel: 'themes'. YTick: [0:01:0.6]; ylabel: 'theme distribution'.

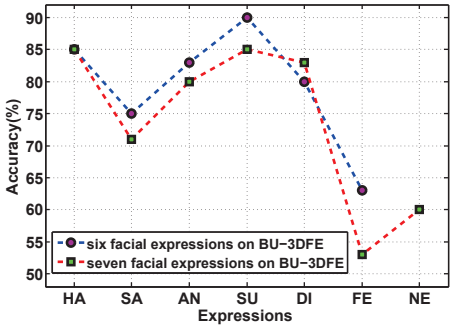


FIGURE 7. Overall performance on BU-3DFE for six expressions and seven expressions, respectively.

A. Recognition accuracy on BU-3DFE: Figure 7 shows the overall performance on BU-3DFE for six expressions and seven expressions respectively. A closer look at the figure reveals that, among the seven expressions, happiness and surprise are easier to be recognized with accuracy over 85%. This is most likely due to the fact that the muscle deformations of both expressions are relatively large compared with others. Moreover, fear is the most difficult to be recognized, followed by neutral.

Figure 8(a) and 8(b) provides the confusion matrix of recognizing each facial expression using our method. One could interpret that a contributing factor to the poor performance of fear is its confusion with happiness. This coincides with the findings of Moore and Bowden in [Moore and Bowden 2011], where the authors point out that the confusion is due to the expressions of fear and happiness having similar muscle deformation around the mouth. In addition, the fear expression also has some confusion with disgust. This is most likely due to the fact that fear is a relatively subtle

TABLE 3. Compared with two well-established features to verify the validity of our features on BU-3DFE, reported in %. The highest one is highlighted in bold.

Features	Classifiers	Poses	Expressions		Rec. Rates(%)
			number	levels	
LBP	cml-TDP	5	7	1,2,3,4	65.3
sift+83 landmarks(SSIFT)	cml-TDP	5	7	1,2,3,4	71.7
FVGC (ours)	cml-TDP	5	7	1,2,3,4	**73.9**

TABLE 4. Recognition accuracy comparison across five poses among our model, sLDA, TDP and multi-SVM on BU-3DFE, reported in %. The highest one for each pose is highlighted in bold.

Methods	Features	Poses					Expressions		Rec. Rates (%)
		0°	30°	45°	60°	90°	number	levels	
sLDA	FVGC (sift+points)	61.7	59.4	63.1	59.6	49.7	7	1,2,3,4	58.7
TDP	FVGC (sift+points)	65.8	62.3	64.9	61.5	57.5	7	1,2,3,4	62.4
multi-SVM	FVGC (sift+points)	69.4	64.7	70.2	63.5	60.7	7	1,2,3,4	65.7
cml-TDP(ours)	FVGC (sift+points)	**78.4**	**73.8**	**81.3**	**70.6**	**65.4**	7	1,2,3,4	**73.9**

expression and thus is hard to disambiguate between other expressions. Another two expressions likely to be confused are sadness and anger. This confusion may attribute to the similar low muscle deformations. Besides, the neutral is easily confused with all of the other expressions. This may due to the neutral has the least amount of facial movement and thus are difficult to distinguish.

B. Comparison with two well-established features: To verify the validity of our feature, we compare it with two well-established features: LBP [51] and SSIFT [52]. As mentioned above (Section 5.2), we obtain the LBP by dividing the key parts into small regions from which LBP histograms are extracted and concatenated to representing the image. SSIFT are obtained by extracting 128-dimensional SIFT features around key points and concatenate these to form a high dimensional feature vector. Then PCA is used to reduce the dimensions of the feature vector to 500-dimension to represent the facial image. Table 3 gives an overview of the results obtained on BU-3DFE dataset by comparing our features with the other features. The average recognition rates are shown in the last column, which reveal that our method achieves clear performance gain, especially when compare it with LBP (from 65.3% of LBP to 73.9% of our model).

C. Comparison with three well-established methods: To verify the validity of our model, we compare with three state-of-the-art methods, namely, sLDA [53], TDP [44] and multi-SVM [13]. As mentioned above, sLDA and TDP are trained based on the facial images with expressions regardless of poses. the multi-SVM model consists of 5 SVMs, each trained separately with expression images under the specific pose. Table 4 gives an overview of the results obtained on BU-3DFE dataset by comparing our model with the other models based on the FER accuracy. The average results are shown in the last column, which clearly reveal that our cml-TDP model has a very significant improvement than the others. The significant accuracy gain over sLDA and TDP shows the advantage of explicitly introducing poses in our model. In addition, our unified model avoids the sperate training and parameter tuning in multi-SVM, and thus highly scalable to the large number of poses seen in multi-character images.

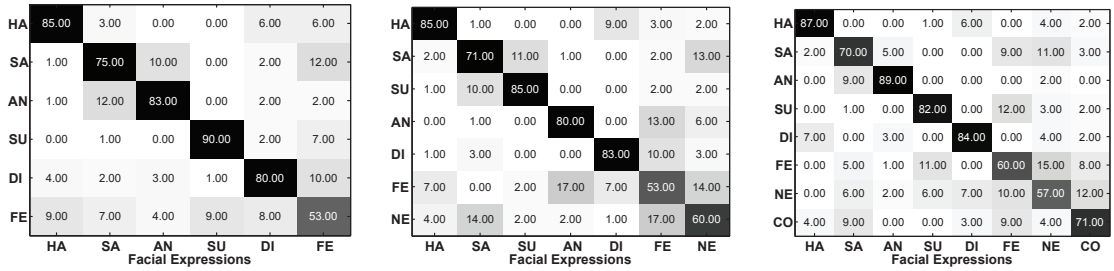
D. Comparison with five previously proposed methods: We further evaluate cml-TDP by comparing its performance with multi-pose FER results recently reported in the literature [49, 50, 19, 14, 18]. Details regarding each reported results, e.g., features, classifiers, the number of poses, and the number and level of expressions, are summarized in Table 5. cml-TDP is evaluated under the same experiment setting, and the FER accuracy is reported in the last column. The results clearly show that cml-TDP outperforms the existing methods with a 0.4% to 16.2% improvement on FER accuracy. This may attribute to the intermediate features learned and the geometric constraints among training in our model. In [49] and [19], the authors only used the geometric features for FER. However these maybe unable to cope well with the complex expression changes. Thus, in [18], the author extracted dense SIFT features from images on 83 key points, then using a PCA to reduce the dimensions of the SIFT feature vectors. Afterwards, they combined the features with 83 key points as the final features for FER and obtained the best results so far with an average recognition rate of 78.9%. However, as we known accurate landmark detection was still challenging. Our approach made up for this deficiency by automatic getting the local elliptical interest regions. Then we can extract SIFT features and the corresponding geometric information of them, which also made us get better results.

2) Experiments on RAFD Database: In order to further evaluate the performance of our model, we conduct the experiment on another database, RAFD. We report the recognition accuracy on RAFD in two cases: first, report the performance on RAFD in details. And then, comparison with three other methods: the sLDA [53], TDP [44], and multi-SVM [13].

A. Recognition accuracy on RAFD: The overall performances on RAFD over each pose and expression are shown in Figure 9(a) and Figure 9(b). A closer look at Figure 9(a) reveals that, like the results in BU-3DFE, among the eight expressions, happy, surprise,

TABLE 5. Compared with five previously proposed methods for a comprehensive comparison on BU-3DFE, reported in %. The highest one is highlighted in bold.

Method	Classifier	Features	Poses	Expressions		Rec. Rates(%)
				number	levels	
Hu et al. 2008a [49]	pose-wise svm	41 landmarks	5	6	1,2,3,4	66.7
Hu et al. 2008b [50]	single knn	SIFT+LPP	5	6	2,3,4	73.8
Zheng et al. 2009 [19]	knn	83 landmark points	5	6	1,2,3,4	78.3
Moore and Bowden 2011 [14]	pose-wise svm	LBP^{u2}	5	6	1,2,3,4	58.4
Moore and Bowden 2011 [14]	pose-wise svm	LBP^{ms}	5	6	1,2,3,4	65.0
Moore and Bowden 2011 [14]	pose-wise svm	LGBP	5	6	1,2,3,4	68.0
Moore and Bowden 2011 [14]	pose-wise svm	$LGBP/LBP^{ms}$	5	6	1,2,3,4	71.1
Zheng 2014 [18]	svm	LBP^{u}	5	6	1,2,3,4	66.0
Zheng 2014 [18]	svm	Sparse SIFT	5	6	1,2,3,4	78.9
Zheng 2014 [18]	svm	83 landmark points	5	6	1,2,3,4	71.4
cml-TDP(ours)	transformed dirichlet processes	FVGC	5	6	1,2,3,4	**79.33**

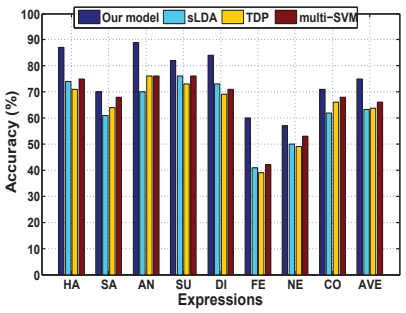


(a) (b) (c)

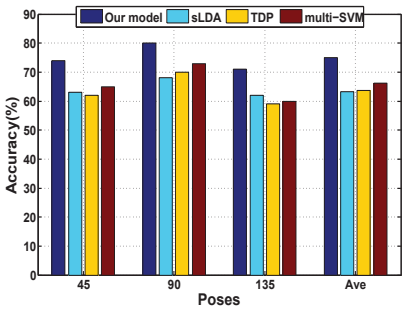
FIGURE 8. (a) The average confusion matrix of six facial expressions on BU-3DFE. The average recognition rate is 79.33%; (b) The average confusion matrix of seven facial expressions on BU-3DFE. The average recognition rate is 73.86%; (c) The average confusion matrix of eight facial expressions on RAFD. The average recognition rate is 75.00%.

disgust and angry are easier to be recognized, fear and neutral are more difficult to be recognized. To inspect this phenomenon, we check the facial images, and found that comparison with other expressions, fear and neutral have relatively few facial movements, and thus are difficult to recognize. The confusion matrix in Figure 8(c) provides details of the performance of our model on each facial expression in RAFD. A closer look at this figure reveals that neutral and fear are the most confused expressions (15%), which may be due to these two expressions have similar muscle deformation around the nose. In addition, fear also has high confusion with surprise (11%). This is most likely due to the fact that both of these two expressions have wide-opened eyes and high eyebrows. Whats more, neutral is likely confused with most of the expressions except happiness and anger. This may be due to the fact that the facial movement of neutral is the least, and it is hard to distinguish. These confusions mentioned above, in turn, lead to the lower recognition accuracy of fear and neutral expressions.

B. Comparison with three well-established methods: In this experiment, we compare with three state-of-the-art methods, namely, sLDA [53], TDP [44] and multi-SVM [13]. sLDA and TDP are trained based on the facial images with expressions regardless of poses. the multi-SVM model consists of 3 SVMs, each trained separately with expression images under the specific pose. The average recognition accuracies of each method are reported in Figure 9(b). Clearly, our method gets the highest recognition accuracy of 75.00%



(a)



(b)

FIGURE 9. Performance comparison among our model, sLDA, TDP and multi-SVM. (a) Recognition accuracy for each expression. (b) Recognition accuracy for each pose.

on RAFD.

From the comparison of these three databases in our experiments, we note that our approach yields better performance compared with the other methods. In our experiments, the performance across different poses and emotions are, however, comparable to one another. This observation strongly supports that our approach is robust to the varying poses and emotions of the facial images. This may attribute to the intermediate features learning as well as the geometric constraints among training in our cml-TDP model.

6. CONCLUSIONS

In this paper, we propose a novel graphical model, cml-TDP, cascaded for key parts segmentation and recognition in multi-pose FER. Pose is explicitly introduced in cml-TDP so that separate training and parameter tuning for each pose is not required. By sharing the pool of features over expressions and poses, we provide a scalable solution for multi-pose FER. In addition, the geometric constraints among different facial parts are implicitly encoded in our model and integrated with the local features to improve FER accuracy. Experiments on two benchmark facial expression databases show the superior performance of our system.

In the future, we plan to optimize our algorithm to improve the FER accuracy and reduce the error propagation between segmentation level and recognition level. The framework is general and can be easily applied to other classification tasks as well, such as object recognition, image classification, and audio event recognition, which we leave as future work. More specifically, the first level (key parts segmentation) in our model can be used for object recognition, and segmentation if we can get more accurate masks. Besides, the latent topic distribution learned in our model is served as the feature for FER, which can also be seen as attributes for an image, and used in fine-grained recognition tasks.

ACKNOWLEDGEMENTS

This work is supported in part by the National Nature Science Foundation of China under Grants 61272211 and 61672267, the Open Project Program of the National Laboratory of Pattern Recognition under Grant 201700022, the General Financial Grant from the China Postdoctoral Science Foundation 2015M570413, the Graduate Student Scientific Research Innovation Projects in Jiangsu Province under Grant KYCX17_1811, the Innovation Project of Undergraduate Students in Jiangsu University under Grant 16A235, and the Nature Science Foundation of Jiangsu Province under Grant BK20140571.

REFERENCES

[1] Sariyanidi, E., Gunes, H., and Cavallaro, A. (2017) Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing*, **26**, 1965–1978.

[2] Chu, W. S., la Torre, F. D., and Cohn, J. (2017) Selective transfer machine for personalized facial expression analysis. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, **39**, 529–545.

[3] Jinwei, G., Xiaodong, Y., Shalini, D. M., and Jan, K. (2017) Dynamic facial analysis: From bayesian filtering to recurrent neural networks. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[4] Shu, X., Cai, Y., Yang, L., Zhang, L., and Tang, J. (2017) Computational face reader based on facial attribute estimation. *Neurocomputing*, **236**, 153–163.

[5] Corneanu, C. A., Simón, M. O., Cohn, J. F., and Guerrero, S. E. (2016) Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, **38**, 1548–1568.

[6] Benitez Quiroz, C. F., Srinivasan, R., and Martinez, A. M. (2016) Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[7] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009) A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, **31**, 39–58.

[8] Eleftheriadis, S., Rudovic, O., and Pantic, M. (2015) Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing*, **24**, 189–204.

[9] Liu, M., Shan, S., Wang, R., and Chen, X. (2014) Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1749–1756. IEEE.

[10] Rudovic, O., Pantic, M., and Patras, I. (2013) Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, **35**, 1357–1369.

[11] Sikka, K., Wu, T., Susskind, J., and Bartlett, M. (2012) Exploring bag of words architectures in the facial expression domain. *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 250–259. Springer.

[12] Tariq, U., Yang, J., and Huang, T. S. (2012) Multi-view facial expression recognition analysis with generic sparse coding feature. *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 578–588. Springer.

[13] Hesse, N., Gehrig, T., Gao, H., and Ekenel, H. K. (2012) Multi-view facial expression recognition using local appearance features. *Proceedings of 2012*

21st International Conference on Pattern Recognition (ICPR), pp. 3533–3536. IEEE.

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

[15] Sangineto, E., Zen, G., Ricci, E., and Sebe, N. (2014) We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 357–366. ACM.

[16] Chen, J., Ariki, Y., and Takiguchi, T. (2013) Robust facial expressions recognition using 3d average face and ameliorated adaboost. *Proceedings of the 21st ACM international conference on Multimedia*, pp. 661–664. ACM.

[17] Zhang, L. and Tjondronegoro, D. (2011) Facial expression recognition using facial movement features. *IEEE Transactions on Affective Computing*, **2**, 219–229.

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

[19] Zheng, W., Tang, H., Lin, Z., and Huang, T. S. (2009) A novel approach to expression recognition from non-frontal face images. *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pp. 1901–1908. IEEE.

[20] Tariq, U., Yang, J., and Huang, T. S. (2013) Maximum margin gmm learning for facial expression recognition. *Proceedings of 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6. IEEE.

[21] Zheng, W., Tang, H., Lin, Z., and Huang, T. S. (2010) Emotion recognition from arbitrary view facial images. *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 490–503. Springer.

[22] Tang, H., Hasegawa-Johnson, M., and Huang, T. (2010) Non-frontal view facial expression recognition based on ergodic hidden markov model supervectors. *Proceedings of 2010 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1202–1207. IEEE.

[23] Mao, Q., Rao, Q., Yu, Y., and Dong, M. (2017) Hierarchical bayesian theme models for multi-pose facial expression recognition. *IEEE Transactions on Multimedia*, **16**, 861–873.

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

[25] Goodfellow, I. J., Courville, A., and Bengio, Y. (2013) Scaling up spike-and-slab models for unsupervised feature learning. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, **35**, 1902–1914.

[26] Pickup, L. C., Capel, D. P., Roberts, S. J., and Zisserman, A. (2009) Bayesian methods for image super-resolution. *The Computer Journal*, **52**, 101–113.

[27] Ekman, P. and Friesen, W. V. (1977) Facial action coding system. *Consulting Psychologists Press, Stanford University, Palo Alto.*

[28] Cao, L. and Li, F. F. (2007) Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. *Proceedings of 2007 IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8. IEEE.

[29] Zhang, F., Mao, Q., Dong, M., and Zhan, Y. (2016) Multi-pose facial expression recognition using transformed dirichlet process. *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 347–351. ACM.

[30] Zhang, K., Huang, Y., Du, Y., and Wang, L. (2017) Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, **26**, 4193–4203.

[31] Zhao, R., Gan, Q., Wang, S., and Ji, Q. (2016) Facial expression intensity estimation using ordinal information. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3466–3474. IEEE.

[32] Liu, P., Han, S., Meng, Z., and Tong, Y. (2014) Facial expression recognition via a boosted deep belief network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1805–1812. IEEE.

[33] Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., and Movellan, J. (2009) Toward practical smile detection. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, **31**, 2106–2111.

[34] Girisha, H., Sreepathi, B., and Karibasappa, K. (2014) Multi-view face recognition using local binary pattern. *International Journal of Computer Science and Information Technologies*, **5**, 2978–2981.

[35] Biswas, A. and Ghose, M. (2014) Expression invariant face recognition using dwt sift features. *International Journal of Computer Applications*, **92**.

[36] Rudovic, O., Patras, I., and Pantic, M. (2010) Coupled gaussian process regression for pose-invariant facial expression recognition. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 350–363. Springer.

[37] Gupta, S. K., Agrwal, S., Meena, Y. K., and Nain, N. (2011) A hybrid method of feature extraction for facial expression recognition. *Proceedings of 2011 Seventh International Conference on Signal-Image Technology and Internet-Based Systems*, pp. 422–425. IEEE.

[38] Lv, Y., Feng, Z., and Xu, C. (2014) Facial expression recognition via deep learning. *Smart Computing (SMARTCOMP), 2014 International Conference on*, pp. 303–308. IEEE.

[39] Khorrami, P., Paine, T., and Huang, T. (2015) Do deep neural networks learn facial action units when doing expression recognition? *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 19–27.

[40] Zhu, X. and Ramanan, D. (2012) Face detection, pose estimation, and landmark localization in the wild. *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2879–2886. IEEE.

[41] Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, **90**, 577–588.

[42] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006) Hierarchical dirichlet processes. *Journal of the american statistical association*, **101**, 1566–1581.

[43] Torralba, A., Willsky, A. S., Sudderth, E. B., and Freeman, W. T. (2005) Describing visual scenes using transformed dirichlet processes. *Proceedings of Advances in neural information processing systems (NIPS)*, pp. 1297–1304.

[44] Sudderth, E. B., Torralba, A., Freeman, W. T., and Willsky, A. S. (2008) Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, **77**, 291–330.

[45] Li, L. J. and Li, F. F. (2010) Optimol: automatic online picture collection via incremental model learning. *International journal of computer vision*, **88**, 147–168.

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

[47] Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. J. (2006) A 3d facial expression database for facial behavior research. *Proceedings of 7th international conference on automatic face and gesture recognition (FGR06)*, pp. 211–216. IEEE.

[48] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., and van Knippenberg, A. (2010) Presentation and validation of the radboud faces database. *Cognition and emotion*, **24**, 1377–1388.

[49] Hu, Y., Zeng, Z., Yin, L., Wei, X., Tu, J., and Huang, T. S. (2008) A study of non-frontal-view facial expressions recognition. *Proceedings of the 19th International Conference on Pattern Recognition (ICPR)*, pp. 1–4. IEEE.

[50] Hu, Y., Zeng, Z., Yin, L., Wei, X., Zhou, X., and Huang, T. S. (2008) Multi-view facial expression recognition. *Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1–6. IEEE.

[51] Ahonen, T., Hadid, A., and Pietikäinen, M. (2004) Face recognition with local binary patterns. *Proceedings of European conference on computer vision (ECCV)*, pp. 469–481. Springer.

[52] Tariq, U., Yang, J., and Huang, T. S. (2012) Multi-view facial expression recognition analysis with generic sparse coding feature. *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 578–588. Springer.

[53] Mcauliffe, J. D. and Blei, D. M. (2008) Supervised topic models. *Proceedings of Advances in neural information processing systems (NIPS)*, pp. 121–128.

[54] Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010) Multi-pie. *Image and Vision Computing*, **28**, 807–813.