Neurocomputing 239 (2017) 143–152

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Locality-constrained linear coding based bi-layer model for multi-view facial expression recognition

Jianlong Wu [a,b], Zhouchen Lin [a,b,*], Wenming Zheng [c], Hongbin Zha [a,b]

[a] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing, 100871, P.R. China
[b] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, 200240, P.R. China
[c] Key Laboratory of Child Development and Learning Science (MOE), Research Center for Learning Science, Southeast University, Jiangsu, 210096, P.R. China

ARTICLE INFO

Article history:
Received 7 May 2016
Revised 2 October 2016
Accepted 3 February 2017
Available online 9 February 2017

Communicated by Prof. Zidong Wang

Keywords:
Multi-view facial expression recognition
Locality-constrained linear coding based bi-layer model
Bag-of-features

ABSTRACT

Multi-view facial expression recognition is a challenging and active research area in computer vision. In this paper, we propose a simple yet effective method, called the locality-constrained linear coding based bi-layer (LLCBL) model, to learn discriminative representation for multi-view facial expression recognition. To address the issue of large pose variations, locality-constrained linear coding is adopted to construct an overall bag-of-features model, which is then used to extract overall features as well as estimate poses in the first layer. In the second layer, we establish one specific view-dependent model for each view, respectively. After the pose information of the facial image is known, we use the corresponding view-dependent model in the second layer to further extract features. By combining all the features in these two layers, we obtain a unified representation of the image. To evaluate the proposed approach, we conduct extensive experiments on both BU-3DFE and Multi-PIE databases. Experimental results show that our approach outperforms the state-of-the-art methods.

1. Introduction

Facial expression recognition has become a hot research topic in computer vision because of its significant role in many applications, such as psychological research and human computer interaction (HCI). As defined by Ekman et al. [1], there are mainly six basic expression types, including angry, disgust, fear, happy, sad, and surprise. The major task of facial emotion recognition is to classify the given facial images into these six categories. During the past decades, a variety of methods [2,3] have been proposed in the literature towards this problem. For a comprehensive survey, one may refer to [4] and [5]. Most existing works focus on expression recognition from frontal or near-frontal face images. However, it is unrealistic to always acquire frontal facial images in real-world applications. Compared with frontal facial expression recognition, non-frontal facial expression recognition is more practical and challenging since we need to deal with the pose variations in the meantime.

Feature coding methods [6] have shown superior performance for image classification. Inspired by that, in this paper, we propose a novel locality-constrained linear coding based bi-layer (LLCBL) model, which is simple but effective for multi-view facial expression recognition. By sharing local bases, locality-constrained linear coding (LLC) [7] captures the correlations between similar descriptors and ensures that similar patches have similar codes, which are very helpful for facial expression recognition. Compared with sparse coding (SC) [8,9], LLC has analytical solution and hence the computation speed is much faster. Besides, for image feature encoding, locality is more essential than sparsity [10]. Therefore, we adopt the LLC based BoF method to represent the facial expression images.

For multi-view facial expression recognition, pose variation is the major problem we need to handle. Towards this issue, we construct an overall LLC based BoF model and a view-dependent model for each angle, respectively. The overall BoF model is used to extract overall features on all the facial images in the first layer. Then, we take advantage of the first layer features to estimate the pose, after which the corresponding view-specific BoF model is utilized to extract second layer pose-related features. By combining all of these two layers' features, we obtain a unified representation of the image. As the variances between different views are distinct, it is very easy to classify the pose. In the second layer, nearly all the train images and test images of view-dependent

* Corresponding author at: Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing 100871, P.R. China.
E-mail addresses: jlwu1992@pku.edu.cn (J. Wu), zlin@pku.edu.cn, zhouchenlin@gmail.com (Z. Lin), wenming_zheng@seu.edu.cn (W. Zheng), zha@cis.pku.edu.cn (H. Zha).

http://dx.doi.org/10.1016/j.neucom.2017.02.012
0925-2312/

model are under the same view, the representation of the test images is much more precise and it is easier to recognize the facial expression. While the overall BoF model preserves the correlated features between different angles, the view-dependent BoF model in the second layer can capture the characteristics of facial images under specific view as well as eliminating the influence of pose variations. In view of these above advantages, we construct LLCBL model for multi-view facial expression recognition. While facial landmark points are still very difficult to detect precisely in actual scene, another advantage of our method is that we extract local features on the dense and overlapped regions. Experimental results show that our LLCBL method can achieve a competitive multi-view facial expression recognition accuracy with a linear SVM [11] classifier.

The contributions of this paper are summarized as follows:

1. We apply the LLC based BoF method to recognize multi-view facial expression. The extracted features enjoy local smooth sparsity and can capture the correlations between similar descriptors very well.
2. Unlike existing works that just use classifiers to estimate the pose directly, our method can extract the first layer features, which is not only used for pose estimation but also part of the final facial image representation.
3. We propose a simple but effective LLCBL method for multi-view facial expression recognition. After the feature extraction and pose estimation in the first layer, we construct a specific model for each view in the second layer to extract view-dependent features, which can eliminate the impact of pose variations. By concatenating bi-layer features, we acquire a discriminative representation. In contrast to existing methods, the LLCBL method can preserve both the relationship between different views and characteristics of each single view.

The rest of the paper is organized as follows. We first introduce the framework of BoF model in Section 2. Then we present the LLC method and our bi-layer model in Section 3. We conduct multi-view facial expression recognition experiments on various databases in Section 4. Finally, Section 5 concludes our paper.

2. Related work

In this section, we introduce the related work on multi-view facial expression recognition and the BoF framework.

2.1. Multi-view facial expression recognition

Multi-view facial expression recognition has attracted many researchers' interest. According to which kinds of geometric information are based on to extract features, the existing approaches can mainly be classified into two different categories: facial landmark points based methods and facial regions based methods. As for the facial landmark points based methods, Hu et al. [12] used 2D displacements of facial landmark points around the eyes, eye-brow and mouth as facial features, and Rudovic et al. [13–15] mapped the 2D locations of landmark points of facial expressions in non-frontal poses to the corresponding locations in the frontal pose. Zhang et al. [16] utilized active appearance models (AAM) [17] to detect the locations of facial landmarks, after which they learnt multi-modal to combine texture and landmark modality. Instead of using geometric location features, Hu et al. [18] applied three different local descriptors (Local Binary Pattern (LBP) [19], Histograms of Oriented Gradients (HoG) [20], and Scale Invariant Feature Transform (SIFT) [21]) on key facial points to characterize facial expressions, and Zheng et al. [22,23] used sparse SIFT features extracted on the given 83 landmark points to represent the facial images.

However, all these facial landmarks based methods need to acquire the accurate locations of facial landmarks at the very beginning. According to the survey [24] of facial points detection, many landmarks detection methods have been proposed in the past several years, but it is still a challenging work to robustly detect accurate facial key points when the pose variations are very large. For these above landmarks based methods, inaccurate landmarks will influence the following feature extraction and representation procedures.

To handle the above issue, researchers proposed some approaches to extract features on image regions. These regions based methods can be divided into two groups: overall model based methods and pose estimation based methods. The first group of methods generally consist of three parts: SIFT features extraction on overlapped dense patches, feature encoding or dimension reduction, and overall classifier learning. In the feature encoding process, Zheng et al. [25] calculated regional covariance matrices (RCM) first, and then proposed Bayes discriminant analysis via Gaussian mixture model (BDA/GMM) to reduce dimension. Tang et al. [26] learned super vectors based on ergodic hidden Markov models (EHMM). Tariq et al. [27–29] used different encoding methods of the bag-of-features (BoF) [30] model to represent facial images. However, as the differences between facial images under different views are very large, the existing overall models cannot capture the specific characteristics of each view. Towards this issue, the second group of methods propose to estimate the pose first, and then construct a view-dependent model for each view. Wu et al. [31,32] projected multi-view features into a common space for classification. Moore and Bowden [33,34] and Zheng et al. [23] divided each facial image into a set of grid regions and then extracted LBP [19] features on each subregion. After feature extraction and pose estimation, Moore and Bowden [33,34] used a view-specific support vector machine (SVM) [11] classifier for each view to recognize facial expressions. Zheng et al. [23] constructed a group sparse reduced-rank regression (GSRRR) model to encoding the features. Although view-dependent models can eliminate the influence of pose variations, they do not utilize the relationship among different views, which is also very important for recognition. Besides, the low level features they extracted and the final image representation are not discriminative enough to classify facial expressions. Besides these 2D facial image based methods, a few researchers tried to acquire pose invariant image based on 3D model. For example, Vieriu [35] first acquired 3D point cloud of the face, and then used two head landmarks and pose information to project the multi-view facial image into near frontal facial image. Then they extracted various features from the projected image and adopted random forest [36] for classification. This kind of methods can also handle pose variations problem, but the computation cost is very expensive.

2.2. Bag-of-features model

BoF is one of the most popular and effective image classification frameworks in the recent literature, which has achieved the state-of-the-art performance in many image classification tasks [6]. As shown in Fig. 1, a traditional BoF framework generally consists of four basic steps. These are, respectively:

1. Local features extraction: In this step, each image is divided into many landmark points or dense overlapped patches, and local features such as SIFT [21], HoG [20], and LBP [19] are extracted from each block or key point to represent the image.
2. Codebook generation: By learning from local features of the first step, this procedure generates a codebook. Sparse coding [8] and locality preservation [7] approaches are often used for learning the codebook. In order to improve the computa-

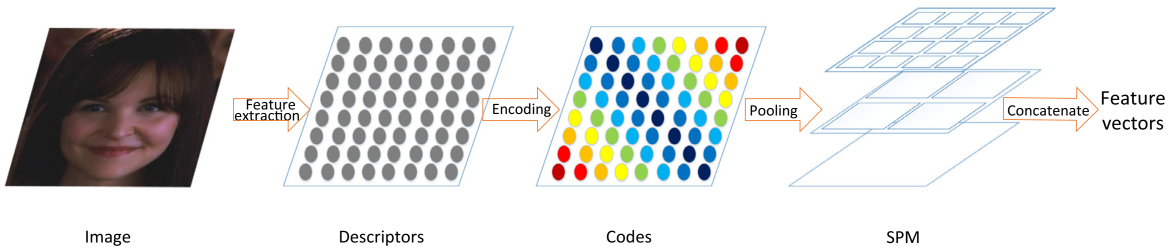


Fig. 1. A general framework of the traditional BoF model.

tional efficiency, clustering methods such as K-means [37] is often used to compute the cluster centres based on a subset of descriptors which are randomly selected from all these feature vectors. These cluster centres are called codewords, which compose the codebook.

3. Descriptors encoding: Each descriptor is encoded into a code vector with codewords in the codebook. By utilizing different encoding methods such as SC [8], saliency coding [38], and LLC [7], we can acquire code vectors with different properties. For a literature survey on encoding methods, one can refer to [39] and [6].
4. Spatial feature pooling: In this step, spatial pyramid matching(SPM) [40] first partitions the image into increasingly fine subregions, and then pooling process integrates all responses on each codeword in a subregion into one value. As SPM [40] can preserve the rough image geometric characteristics, it is commonly used in image categorization tasks, such as Li et al. [41,42] explored graph connections between images based on SPM. Besides, Li et al. [43] also proposed cellet-encoded SPM to improve the performance.

In pooling process, max-pooling and average pooling are two main pooling methods. By pooling code vectors in each spatial block across different spatial scales, we obtain the local description of every block. The final representation of the image is obtained by combining descriptions of all blocks.

3. Bi-layer model for multi-view facial expression recognition

In this section, we first describe the details of locality-constrained linear coding in Section 3.1. Then we illustrate the framework of our proposed bi-layer model in Section 3.2.

3.1. Locality-constrained linear coding

Let $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{D\times N}$ be a set of D-dimensional local features extracted from an image. $B = [b_1, b_2, \ldots, b_M] \in \mathbb{R}^{D\times M}$ denotes the codebook with M codewords. Encoded by different coding algorithms, local features X is converted to N coding vectors $C = [c_1, c_2, \ldots, c_N] \in \mathbb{R}^{M\times N}$. The Locality-constrained linear coding method [7] encodes each local descriptor x_i by solving the following problem:

$$\begin{aligned} &\arg\min_C \sum_{i=1}^{N} (\|x_i - Bc_i\|^2 + \lambda\|d_i \odot c_i\|^2), \\ &s.t.\quad \mathbf{1}^T c_i = 1, \forall i, \end{aligned} \tag{1}$$

where $\mathbf{1} \in \mathbb{R}^{M\times 1}$ is a column vector of all ones, $\odot$ denotes the element-wise multiplication and $d_i = \exp(\frac{\mathrm{dist}(x_i, B)}{\sigma}) \in \mathbb{R}^M$ is the locality adaptor. Specifically, $\mathrm{dist}(x_i, B) = [\|x_i - b_1\|_2, \ldots, \|x_i - b_M\|_2]^T$. σ is used for adjusting the weight decay speed.

Compared with SC [8], LLC adopts locality constraint, which is more crucial than sparsity in encoding [10]. Another advantage of LLC is that the problem defined in Eq. 1 has a closed-form solution:

$$\begin{aligned} &\hat{c}_i = ((B^T - \mathbf{1}x_i^T)(B^T - \mathbf{1}x_i^T)^T + \lambda \mathrm{diag}^2(d_i))^{-1}\mathbf{1}, \\ &c_i = \hat{c}_i / (\mathbf{1}^T \hat{c}_i). \end{aligned} \tag{2}$$

The above idea is also shown in [7].

During the process of solving Eq. 1, local bases for each descriptor are selected to form a local coordinate system. Instead of solving Eq. 1 directly, we can simply use the $K (K < D < M)$ nearest neighbors of x_i in the codebook as the local base $\tilde{B}$ to reconstruct the descriptor x_i:

$$\begin{aligned} &\min_C \sum_{i=1}^{N} \|x_i - \tilde{B}c_i\|^2, \\ &s.t.\quad \mathbf{1}^T c_i = 1, \forall i. \end{aligned} \tag{3}$$

Since $\tilde{B}$ is the K closest codewords of x and K is usually much smaller than the total number of codewords, approximated LLC achieves both locality and sparsity. The computation complexity can also be significantly reduced by this approximation.

3.2. Framework of our bi-layer model

Based on LLC, we propose a bi-layer model, LLCBL, for multi-view facial expression recognition. The framework and pipeline of the proposed method are illustrated in Figs. 2 and3, respectively. The detailed process is illustrated as follows. During the training process, in the first layer, we first extract the dense SIFT features of all the images. Let $X = [X_1, X_2, \ldots, X_S]$ be the local descriptors set of all the images, where S is the total number of different angles and X_i $(i = 1, 2, \ldots, S)$ denotes the local features subset of the i-th angle facial expression images. Based on the features set X of all local descriptors, an overall dictionary B is learnt, which is then utilized to encoding all the descriptors with the LLC method. The final image representation of the first layer $F^1 = [f_1, f_2, \ldots, f_M]$ is obtained by max-pooling over the LLC codes in each spatial block across different spatial scales. In the second layer, we train the dictionary B_i of specific i-th view with the corresponding features subset X_i, respectively. For each view, the descriptors in subset X_i $(i = 1, 2, \ldots, S)$ is encoded by the corresponding dictionary B_i with LLC, respectively. Then spatial pyramid pooling is applied to get the final representation of the second layer F^2. The main difference of these two layers lies in the dictionary learning process. By combining all the features of these two layers, we obtain a unified representation of the image $F = [F^1, F^2]$. For the test facial expression images, we first extract the first layer features F^1 with the overall dictionary B with the same process as that in the training process. With the pose information of train samples, the first layer representation F^1 of test image is then used to estimate the angle of test image by a linear SVM [11] classifier. As the variances between different views are distinct, it is very easy to correctly classify the pose. After the view of the test image is acquired, the second layer features F^2 are calculated based on the corresponding dictionary B_i. We concatenate features $F = [F^1, F^2]$ as the final

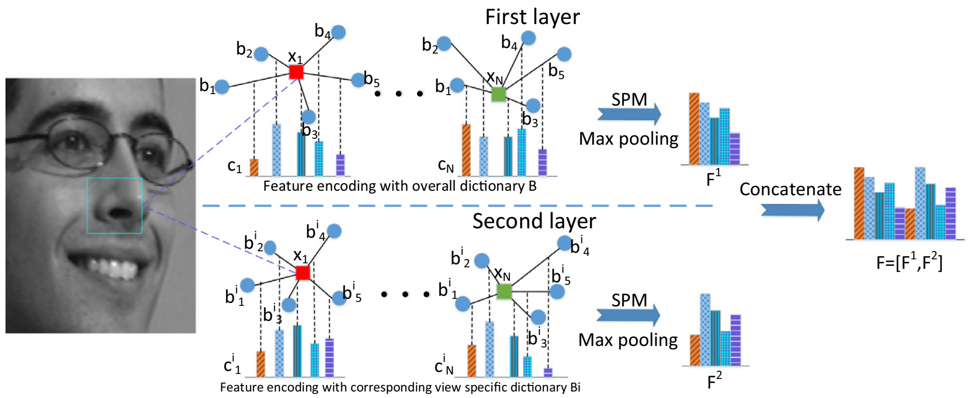


Fig. 2. Framework of the proposed LLCBL model. For the input expression images with pose variations, we first extract local features in each block. Then in the first layer, we use the overall dictionary to encode local features. In the second layer, we use the corresponding view specific dictionary to encode local features. After encoding, SPM based max pooling is adopted to pool the coding vectors on each codeword. By concatenating the features of two layers, we get the final representation of the image. For test images, the pooled features of first layer F^1 are used to estimate the view.

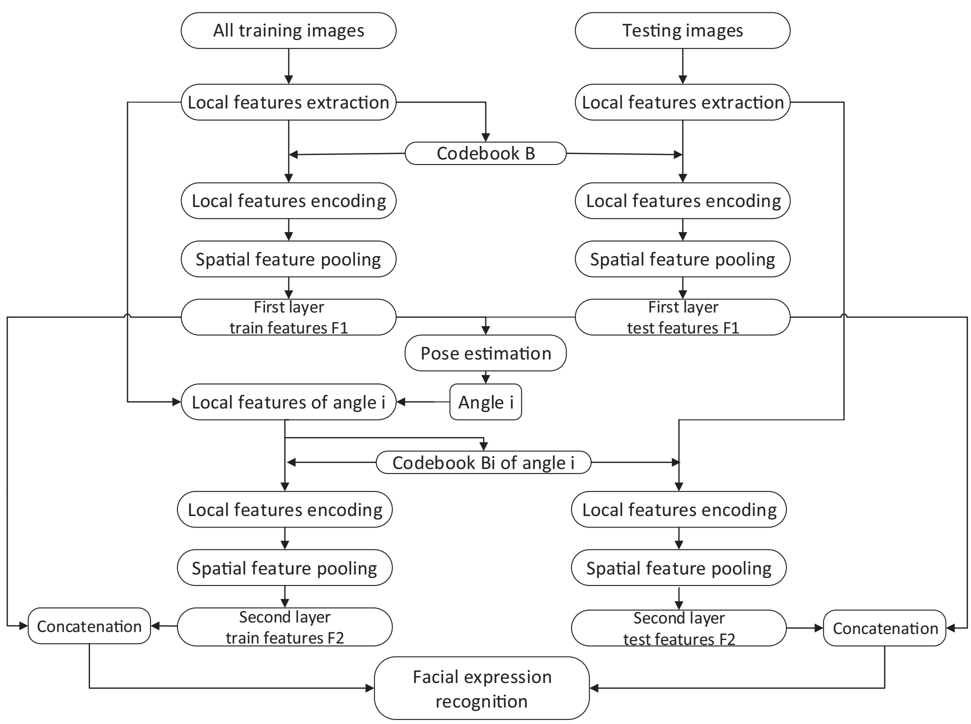


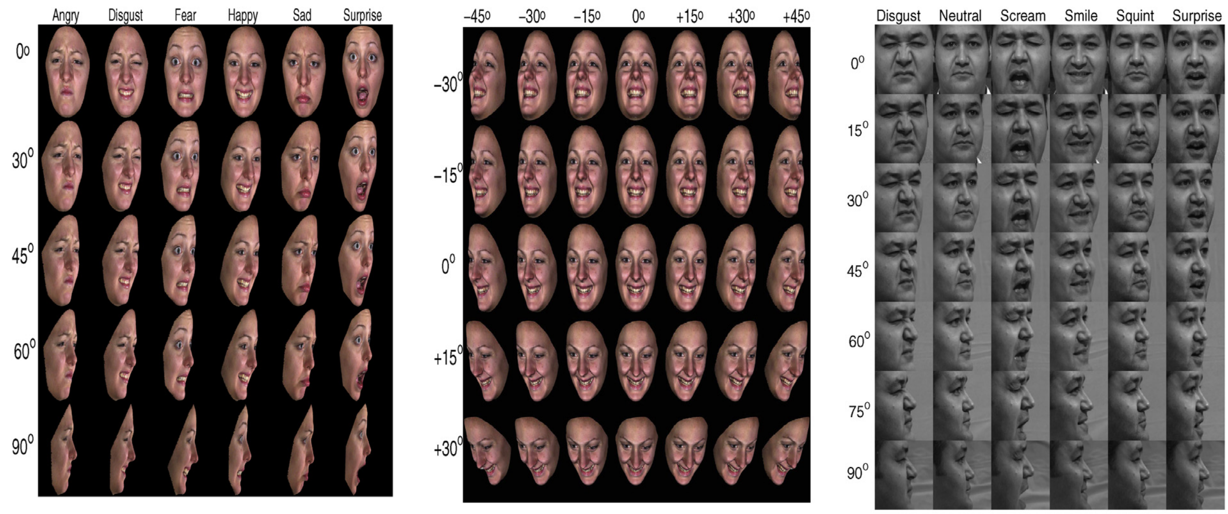
Fig. 3. Pipeline of the proposed LLCBL model. An overall LLC based BoF model is constructed in the first layer, and several view-dependent BoF models are established in the second layer.

representation of test image. Finally, a linear SVM [11] classifier is applied to final representation F to recognize the facial expression of the test image.

3.3. LLCBL properties analysis

In the proposed LLCBL method, the first layer is the overall LLC based BoF model, and the second layer is view-specific LLC based BoF model. LLCBL concatenates the features of these two layers to get the final discriminative representation. Our bi-layer model possesses several attractive properties. Firstly, this model can well preserve those nice properties of LLC. As the extracted dense SIFT features are invariant to scale and rotation transformations, the final image representation can also well handle these transformations. During the encoding process, under the locality regularization, codewords of LLC can well reconstruct the local features and similar descriptors are encoded with similar codes. Besides, as the LLC method has close-form solution, the encoding speed is very fast. Secondly, view-dependent models in the second layer can represent the facial images more precisely while the angle of facial expression images has a large variation. For multi-view facial expression recognition, the most challenging aspect is the view transformation. Towards this issue, we construct view-dependent model for each specific view in the second layer of LLCBL, and it will eliminate the influence of various views. Last but not least, the final concatenated features F of LLCBL contain both the overall and view-dependent characteristics. It can represent the original facial image with richer information, which benefits classification. As both global and local features are important for image representation, we concatenate features of both two layers to acquire better representation.

Compared with single layer LLC for multi-view expression recognition, LLCBL takes view variations into consideration and constructs view-dependent models. While single layer LLC



(a) First set of BU-3DFE (b) Second set of BU-3DFE (c) Multi-PIE

Fig. 4. Example images from the BU-3DFE and Multi-PIE datasets. (a) Examples of the 2D facial images of one subject in the first set of the BU-3DFE database with respect to the six facial expressions and five facial views. (b) Examples of the 2D facial images of one expression for one subject in the second set of the BU-3DFE database with respect to five tilt angles and seven pan angles. (c) All 42 facial expression images of one subject in the Multi-PIE database with respect to six facial expressions and seven facial views.

only contains overall features, LLCBL contains both overall and view-dependent features, which help LLCBL to represent expression images better.

4. Experimental results and analysis

We do extensive experiments on two widely used databases that are BU-3DFE database [44] which is synthetic and Multi-PIE database [45] which is close to real-world scenes to validate our proposed method. Some example images of these databases are shown in Fig. 4. As we introduced in the Section 1, facial landmark points are still very challenging to detect precisely in practice, especially for those facial expression images with large variations. While the detected key points are not as exact as the ground truth, the recognition accuracy may decrease significantly. Therefore, we only compare the performance of our proposed LLCBL model with that of other region based state-of-the-art methods in this paper, including the pose estimation based LBP method [34] and GSRRR model [23], overall model based BDA/GMM [25], EHMM [26], generic sparse coding (GSC) method [27], supervised soft vector quantization (SSVQ) method [28] and supervised super-vector encoding (SSE) method [29]. We also apply the original LLC method [7] to multi-view facial expression recognition and compare the performance.

Throughout the experiments, we only use one single descriptor SIFT [21] and a set of 128-dimension SIFT features is extracted from patches densely located with step size of 3 pixels in both horizontal and vertical directions under only one scale 16×16. K-means [37] is used to generate the codebook with 1024 centres. The number of neighbors is set to 5 during feature coding with LLC, after which we use the spatial pyramid matching (SPM) [40] with levels of $[1 \times 1, 2 \times 2, 4 \times 4, 8 \times 8]$ and adopt max-pooling method to pool the feature codes in each spatial block. We set all above optimal parameters according to [6], which give a detailed analysis about parameters selection. In the classification process, we use the liblinear SVM [11] classifier to estimate the pose and classify the facial expressions. To evaluate the performance, we do 5-fold cross validation and then average the results. All the subjects are randomly divided into 5 portions. In each of 5 folds, images from one portion (20% subjects) are used as test samples and images from the remaining four portions (80% subjects) are used as train samples. Under this circumstance, there are no overlapped subjects between train subjects and test subjects, and images of each subject are regarded as test samples exactly once.

4.1. Experimental results on BU-3DFE database

We first test the performance on the synthetic database. The BU-3DFE database [44] has been widely used for multi-view facial expression recognition. It contains 3D models of 100 subjects (56 female and 44 male), each of whom has six universal facial expressions (anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU)) under four different levels of intensities. The OpenGL software is utilized to render these 3D models and generate 2D facial expression images under different angles for classification. Among all of the existing multi-view facial expression recognition works, there are mainly two widely used 2D facial image sets of BU-3DFE database. We conduct experiments on both two image sets.

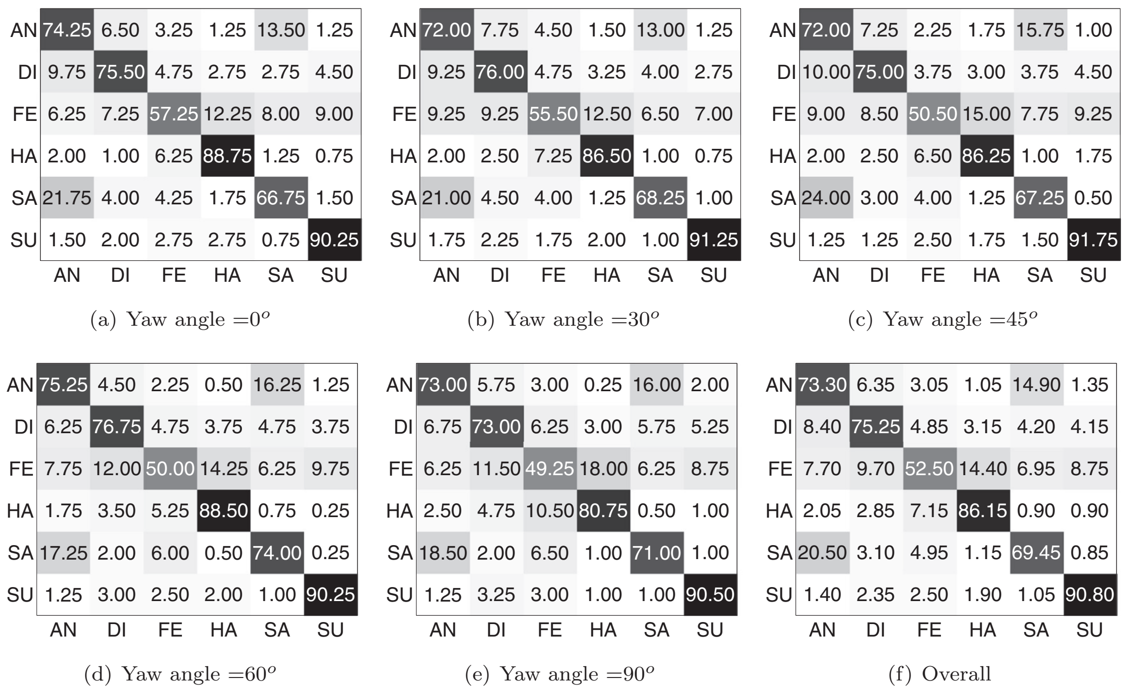
For the first image set of BU-3DFE database, it consists of $5 \times 4 \times 6 \times 100 = 12000$ 2D facial expression images under five different yaw angles (0°, 30°, 45°, 60°, 90°) and four levels of intensities (levels 1,2,3, and 4). Some sample images of one subject in various yaw angles are shown in Fig. 4a. [23,34] also adopt this image set for experiments. We evaluate the performance of our algorithm with these images under all five yaw views and four levels of intensities.

In Table 1, we compare the performance of our bi-layer model with that of other state-of-the-art methods under the same experiments setting on the first set of BU-3DFE database. We need to mention that this dataset contains all 4 levels intensities expression images. On this dataset, the single layer LLC can achieve the accuracy 72.1%, which is a little higher than the best performance 71.1% of other region based methods. Our LLCBL can further improve the recognition accuracy to 74.6%.

Table 2 shows the recognition rates of each expression under every yaw angle and Fig. 5 shows the confusion matrix among these six expressions. From Table 2, we can see that the recognition rates vary form view to view, and the optimal view for

Table 1
Performance comparison with state-of-the-art methods on the first image set of BU-3DFE database.

Method	Poses		Expressions		Feature	Overall(%)
	number	pan	number	levels		
Moore and Bowden [34]	5	$(0^o, +90^o)$	6	1,2,3,4	LBP^{u2}	58.4
Moore and Bowden [34]	5	$(0^o, +90^o)$	6	1,2,3,4	LBP^{ms}	65.0
Moore and Bowden [34]	5	$(0^o, +90^o)$	6	1,2,3,4	LGBP	68.0
Moore and Bowden [34]	5	$(0^o, +90^o)$	6	1,2,3,4	$LGBP/LBP^{ms}$	71.1
GSRRR [23]	5	$(0^o, +90^o)$	6	1,2,3,4	LBP^{u2}	66.0
Single layer LLC	5	$(0^o, +90^o)$	6	1,2,3,4	Dense SIFT	72.1
LLCBL method (Ours)	5	$(0^o, +90^o)$	6	1,2,3,4	Dense SIFT	**74.6**



(a) Yaw angle $=0^o$ (b) Yaw angle $=30^o$ (c) Yaw angle $=45^o$
(d) Yaw angle $=60^o$ (e) Yaw angle $=90^o$ (f) Overall

Fig. 5. Confusion matrices for facial expressions over different facial views on the first image set of BU-3DFE database. (a)–(e) The confusion matrices corresponding to five facial views. (f) The overall recognition confusion matrix.

Table 2
Recognition accuracies under different expressions and facial views on the first image set of BU-3DFE database.

Expressions	Results (%)					
	0^0	30^0	45^0	60^0	90^0	Average
Angry	74.25	72.00	72.00	**75.25**	73.00	73.30
Disgust	75.50	76.00	75.00	**76.75**	73.00	75.25
Fear	**57.25**	55.50	50.50	50.00	49.25	52.50
Happy	**88.75**	86.50	86.25	88.50	80.75	86.15
Sad	66.75	68.25	67.25	**74.00**	71.00	69.45
Surprise	90.25	91.25	**91.75**	90.25	90.50	90.80
Average	75.46	74.82	73.79	**75.79**	72.92	74.58

recognition is 60^o with accuracy 75.79% while the worst view is 90^o with accuracy 72.92%. For different expressions, the recognition accuracies are also different. Surprise and happy expressions are much easier to recognize while fear is very difficult to classify, whose average recognition accuracy is less than 55%. In the meantime, we can see from Fig. 5 that angry and sad expressions are more likely to be misclassified, which result in the low accuracies of these two expressions. The recognition rates of these two expressions are less than the average recognition result. All the misclassification rates of fear expression to other expressions are relatively high, which make the accuracy of fear expression lowest among all these six expressions.

For the second image set of BU-3DFE database, it consists of $7 \times 5 \times 6 \times 100 = 21000$ 2D facial expression images under seven different pan angles (0°, ±15°, ±30°, ±45°), five different tilt yaw angles (0°, ±15°, ±30°) and with the strongest expression intensity (level 4). Some sample images of one expression in various views are shown in Fig. 4a. There are mainly three fundamental differences between these two image sets of BU-3DFE database. First of all, The second image set contains facial expressions under not only different pan views but also different tilt views. The largest variation pan angle of this image set is only ± 45°, while the pan angle of the first image set vary from 0° to 90°. Besides, the second image set only contains images with the strongest expression intensity, but images in the first image set are under four different intensities. Many works, such as [25–29], adopt the second image set for experiments and achieve state-of-the-art recognition results. We evaluate the performance of our algorithm with all these expression images under all seven yaw views and five tilt views.

As shown in Table 3, the performance of single layer LLC is better than other state-of-the-art methods, which demonstrates that LLC is better than SC [8], SSVQ[28], and SSE [29] during encoding. Our proposed LLCBL method can achieve 80.2% while the highest recognition accuracy among all existing methods is only 76.6%, which is much lower than that of our method. In the meantime, the recognition accuracy of LLCBL surpasses that of LLC with 2.5%.

Table 3
Performance comparison with state-of-the-art methods on the second subset of BU-3DFE database with the strongest expression intensity.

Method	Poses			Expressions		Feature	Overall (%)
	number	pan	tilt	number	level		
Zheng et al.[25](DBA/GMM)	35	$(-45^o, +45^o)$	$(-30^o, +30^o)$	6	4	Dense SIFT	68.2
Tang et al.[26] (EHMM)	35	$(-45^o, +45^o)$	$(-30^o, +30^o)$	6	4	Dense SIFT	75.3
Tariq et al.[27] (GSC)	35	$(-45^o, +45^o)$	$(-30^o, +30^o)$	6	4	Dense SIFT	76.1
Tariq et al.[28] (SSVQ)	35	$(-45^o, +45^o)$	$(-30^o, +30^o)$	6	4	Dense SIFT	76.1
Tariq et al.[29] (SSE)	35	$(-45^o, +45^o)$	$(-30^o, +30^o)$	6	4	Dense SIFT	76.6
Single layer LLC	35	$(-45^o, +45^o)$	$(-30^o, +30^o)$	6	4	Dense SIFT	77.7
LLCBL method (Ours)	35	$(-45^o, +45^o)$	$(-30^o, +30^o)$	6	4	Dense SIFT	**80.2**

Table 4
Performance comparison with state-of-the-art methods on the Multi-PIE database.

Method	Poses		Expressions number	Feature	Overall (%)
	number	pan			
Moore and Bowden [34]	7	$(0^o, +90^o)$	6	LBP^{ms}	73.3
Moore and Bowden [34]	7	$(0^o, +90^o)$	6	LGBP	80.4
GSRRR [23]	7	$(0^o, +90^o)$	6	LBP^{u2}	81.7
Single layer LLC	7	$(0^o, +90^o)$	6	Dense SIFT	84.4
LLCBL method (Ours)	7	$(0^o, +90^o)$	6	Dense SIFT	**86.3**

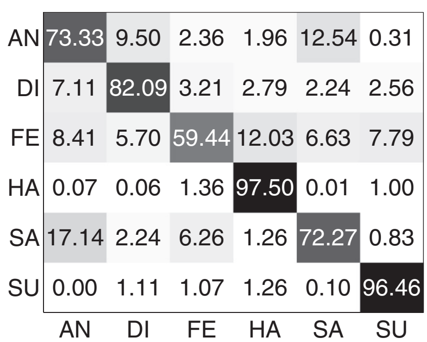


Fig. 6. The overall recognition confusion matrix for facial expressions on the second image set of BU-3DFE database.

As there are so many different views, we only present the overall recognition confusion matrix for facial expressions over different facial views in Fig. 6. The recognition results are roughly consistent with that shown in Fig. 5, except for the recognition improvement of disgust, happy and surprise three expressions.

4.2. Experimental results on Multi-PIE database

Now we test the performance of LLCBL on the real-world image database. The Multi-PIE database [45] contains 755,370 images from 337 different subjects. The facial images of this database are captured under 15 view points and 19 illumination conditions in up to four recording sessions. We adopt the same subset of Multi-PIE for expression recognition as in [23,34], and only select these 100 subjects presented in all four recording. For each sample of the selected subjects, six types of expressions (neutral, smile, surprise, squint, disgust, and scream) and seven different poses (0°, 15°, 30°, 45°, 60°, 75°, and 90° yaw angles) are considered in the experiments. Consequently, there are 4200 images in total. Before the experiments, we first crop and normalize each facial image, and then down-sample it into the same size of 120×160 pixels. Fig. 4a shows all 42 facial images of one subject.

We compare the average recognition accuracy of our bi-layer model with the ones achieved by [34] and [23] under the same experiments settings, and the results are shown in Table 4. While the recognition rates of [34] and [23] are 80.4% and 81.7%, respectively,

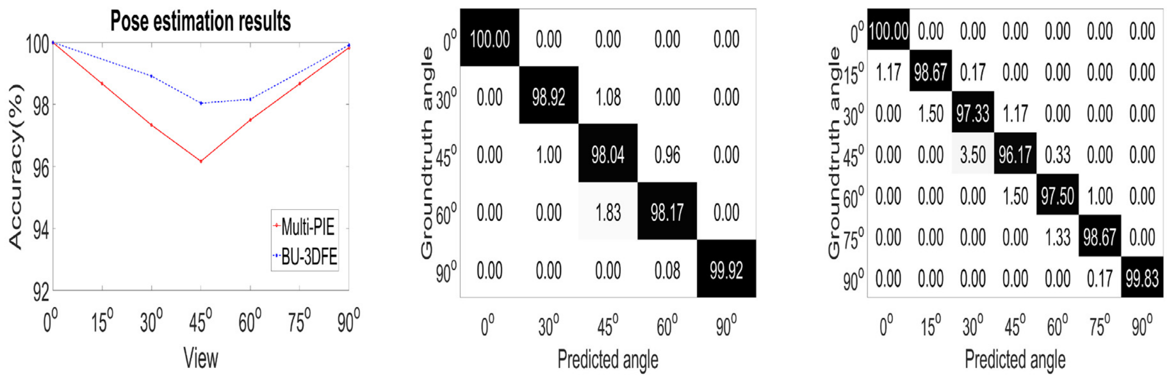
Table 5
Recognition accuracies of expressions under different facial views on the Multi-PIE database.

Expressions	Results (%)							
	0^o	15^o	30^o	45^o	60^o	75^o	90^o	Average
Disgust	74.0	77.0	78.0	78.0	80.0	79.0	**85.0**	78.7
Neutral	90.0	**93.0**	**93.0**	91.0	91.0	84.0	86.0	89.7
Scream	97.0	96.0	**98.0**	95.0	94.0	96.0	**98.0**	96.3
Smile	**86.0**	84.0	82.0	83.0	85.0	84.0	85.0	84.1
Squint	**82.0**	81.0	80.0	74.0	74.0	61.0	72.0	74.9
Surprise	96.0	94.0	96.0	**97.0**	93.0	90.0	94.0	94.3
Average	87.5	87.5	**87.8**	86.3	86.2	82.3	86.7	86.3

the recognition accuracy of our method can achieve 86.3%, which is much higher than that of other methods. Table 5 shows the recognition rates of each expression under different facial views. We can see that the optimal facial view for each expression is different, and the corresponding view of the highest average recognition accuracy among all expressions is 30°. Fig. 9 shows the confusion matrices of each yaw angle as well as the overall confusion matrix of the experiments. From both Table 5 and Fig. 9, we can see that, among these six expressions, scream and surprise expressions are easier to be recognized while their recognition accuracies are 96.3% and 94.3%, respectively. Squint and disgust expressions are more difficult to recognize while their recognition rates are less than 80%. Besides, it can be found in Fig. 9 that disgust and squint expressions are more likely to be misclassified, which is due to the fact that these two expressions are relatively similar to each other.

4.3. Influence of pose estimation

In the end of the first layer of Fig. 2, we simply use the liblinear SVM [11] classifier to estimate the pose. The accuracy and confusion matrix of pose estimation are presented in Fig. 7. For simplification, we just show the pose estimation results on the Multi-PIE and the first subset of BU-3DFE dataset. Since there are obvious differences between facial expressions of different views, the accuracy of pose estimation is very high. The average accuracies of pose estimation on the Multi-PIE and the first subset of BU-3DFE are 98.31% and 99.01%, respectively. According to Fig. 7, we can see that only a few images are misclassified and those misclassified samples are all classified into their adjacent views. Even though



(a) Pose estimation accuracy on two datasets (b) Pose estimation confusion matrix on the first BU-3DFE dataset (c) Pose estimation confusion matrix on the Multi-PIE dataset

Fig. 7. Pose estimation accuracy and confusion matrix on two datasets.

the ground truth views of them are slightly different, they also share the similar patterns. Therefore, it is reasonable to neglect the influence of pose estimation in our method. Through experiments, we also find that most of those expressions whose views are misclassified can be correctly recognized in the second layer of our framework. In this case, our LLCBL model can well eliminate the influence of view variations.

4.4. Experimental results analysis

According to the recognition results on these three datasets, we can easily find that single layer LLC can achieve comparable accuracy with the state-of-the-art methods as LLC can well preserve the local similarity during encoding. Compared with single layer LLC, our LLCBL can further improve the accuracy with nearly 2.3%. There are mainly two reasons accounting for this improvement. The first one is that we construct view-dependent models in the second layer of LLCBL to reduce the impact of view variations, and the second reason is that LLCBL contains both overall and view-dependent features, both of which are important for classification.

For the impact of expressions intensities, compared with the recognition result on the second image set of BU-3DFE database, the recognition accuracy of LLCBL method on the first BU-3DFE image set is much lower, which can be ascribed to the inconspicuous deformations of the low intensities expressions and large variation of the facial expression views. To further investigate the influence of expression intensity, we show the recognition accuracy of different expression intensities on the first subset of BU-3DFE dataset in Fig. 8. As introduced in the Section 4.1, the first subset of BU-3DFE dataset consists of facial expressions across four different intensities. According to Fig. 8, we can easily find that the recognition accuracy goes up with the increase of expression intensity. For expressions with high level intensity, the change of action units is very large. In this case, the difference between expressions of different class under high level intensity is much larger than that under low level intensity. Therefore, expressions with high level intensity are relatively easier to be correctly recognized than that with low level intensity.

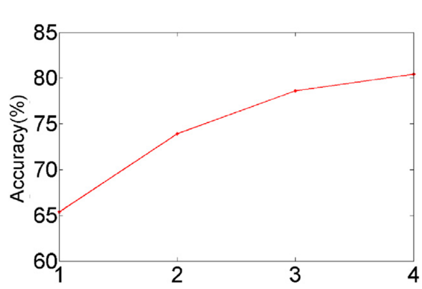


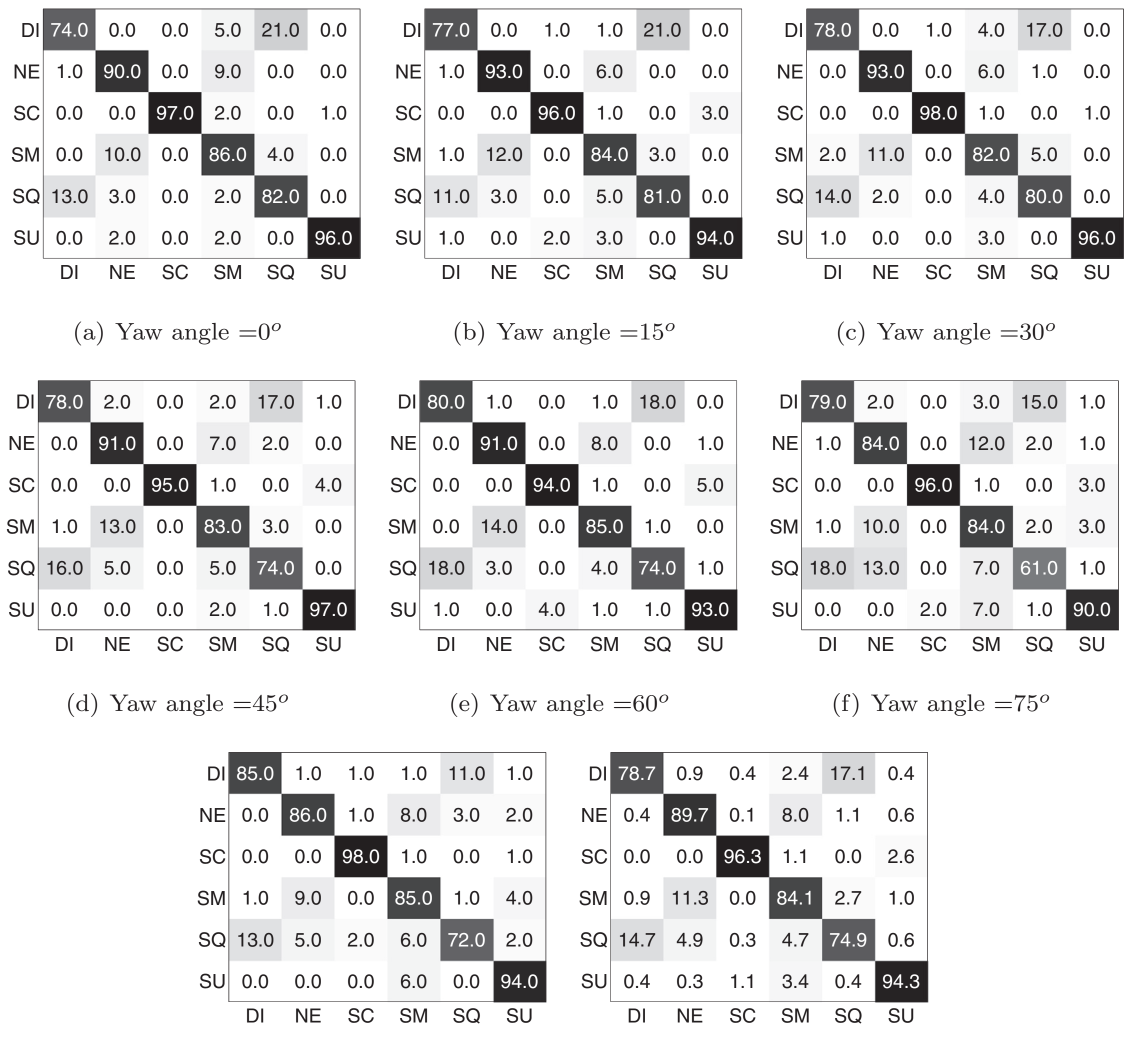
Fig. 8. Influence of expression intensity on the first image set of BU-3DFE database.

For the misclassification between expressions, it is related to the similar action units of facial expressions. Facial action unit system (FACS) [46] defined 44 action units, and each expression can be regarded as a combination of several action units. When two facial expressions share similar action units, they are easy to be misclassified. For example, as shown in Figs. 5 and 6, misclassification rate of angry and sad expressions is very high. This can be due to the fact forehead action units of angry and sad expressions are very similar to each other, which can be seen in Fig. 4a. when the action units are very different, misclassification rate is relatively low.

For different views, facial images with pan angle that is less than 60^o benefit the recognition. When the pan angle is larger than 60^o, the overall recognition accuracy decreases. According to the FACS [46], action units of different expressions mainly lie in the mouth, eyes and forehead. However, when the pan angle is larger than 60^o, these moving parts of face are not obvious any longer, which will reduce the recognition result. On the other hand, we notice that the pan angle corresponding to the highest recognition accuracy among all different angles is not 0^o but between 30^o and 60^o. For frontal images, they are nearly symmetric. Half of the frontal image can well represent the original image. In this way, the frontal facial expression images contain much redundant information. Compared with frontal facial images, facial images with small pan angle can also preserve the information of one face side and add some detailed side face information. Therefore, facial images with small pan angle might benefit the classification.

5. Conclusions

In this paper, we investigate the multi-view facial expression recognition problem and propose a simple yet effective bi-layer model based on the classical LLC method. Our bi-layer model, LLCBL, can extract discriminative features to represent the facial expression images. The extracted features contain both overall and view-dependent characteristics, which benefit the classification. As we extracted SIFT features from dense local patches, another major advantage of our method is that we do not need to detect the facial landmark points which is still a challenging work. To evaluate the performance of our proposed method, we conduct extensive



(a) Yaw angle $=0^o$ (b) Yaw angle $=15^o$ (c) Yaw angle $=30^o$ (d) Yaw angle $=45^o$ (e) Yaw angle $=60^o$ (f) Yaw angle $=75^o$ (g) Yaw angle $=90^o$ (h) Overall

Fig. 9. Confusion matrices for facial expressions over different facial views on the Multi-PIE database. (a)–(g) The confusion matrices corresponding to seven facial views. (h) The overall recognition confusion matrix.

experiments on both BU-3DFE database and Multi-PIE database. The results show that the recognition accuracy of our method is much higher than that of other state-of-the-art methods under the same experimental settings, especially on the Multi-PIE database. While the dimension of the final unified representation is very high and the structure is very sparse, we would like to design an approach to reduce the dimension of these features in the future. Other efficient classifiers such as kernel SVM can be applied to multi-view facial expression recognition to further improve the recognition accuracy.

Acknowledgment

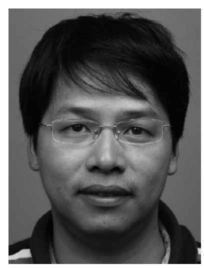
Zhouchen Lin is supported by National Basic Research Program of China (973 Program) (Grant no. 2015CB352502), National Natural Science Foundation (NSF) of China (Grant nos. 61625301 and 61231002), and Qualcomm. Wenming Zheng is supported by NSFC (Grant nos. 61231002 and 61572009), 973 Program (Grant no. 2015CB351704), and Natural Science Foundation of Jiangsu Province (Grant no. BK20130020). Hongbin Zha is supported by Beijing Municipal Natural Science Foundation (Grant no. 4152006).

References

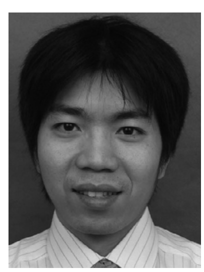
[1] P. Ekman, W.V. Friesen, Pictures of facial affect, Consulting Psychologists Press, 1975.
[2] K. Yu, Z. Wang, M. Hagenbuchner, D.D. Feng, Spectral embedding based facial expression recognition with multiple features, Neurocomputing 129 (2014) 136–145.
[3] L. Zhang, D. Tjondronegoro, V. Chandran, Random gabor based templates for facial expression recognition in images with facial occlusion, Neurocomputing 145 (2014) 451–464.
[4] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.
[5] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recogn. 36 (1) (2003) 259–275.
[6] Y. Huang, Z. Wu, L. Wang, T. Tan, Feature coding in image classification: a comprehensive study, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 493–506.
[7] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2010, pp. 3360–3367.
[8] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 1794–1801.
[9] F. Long, M.S. Bartlett, Video-based facial expression recognition using learned spatiotemporal pyramid sparse coding features, Neurocomputing 173 (2016) 2049–2054.
[10] K. Yu, T. Zhang, Y. Gong, Nonlinear learning using local coordinate coding, in: Proceedings of Advances in Neural Information Processing Systems, 2009, pp. 2223–2231.

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[12] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, T.S. Huang, A study of non-frontal-view facial expressions recognition, in: Proceedings of the IEEE International Conference on Pattern Recognition, 2008, pp. 1–4.
[13] O. Rudovic, I. Patras, M. Pantic, Coupled Gaussian process regression for pose-invariant facial expression recognition, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 350–363.
[14] O. Rudovic, M. Pantic, I. Patras, Coupled Gaussian processes for pose-invariant facial expression recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (6) (2013) 1357–1369.
[15] S. Eleftheriadis, O. Rudovic, M. Pantic, Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition, IEEE Trans. Image Process. 24 (1) (2015) 189–204.
[16] W. Zhang, Y. Zhang, L. Ma, J. Guan, S. Gong, Multimodal learning for facial expression recognition, Pattern Recognit. 48 (10) (2015) 3191–3202.
[17] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
[18] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, T. Huang, Multi-view facial expression recognition, in: Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition, 2008, pp. 1–6.
[19] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[20] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 1, 2005, pp. 886–893.
[21] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[22] W. Zheng, H. Tang, Z. Lin, T.S. Huang, A novel approach to expression recognition from non-frontal face images, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 1901–1908.
[23] W. Zheng, Multi-view facial expression recognition based on group sparse reduced-rank regression., IEEE Trans. Affect. Comput. 5 (1) (2014) 71–85.
[24] N. Wang, X. Gao, D. Tao, X. Li, Facial feature point detection: a comprehensive survey, arXiv preprint arXiv:1410.1037, 2014.
[25] W. Zheng, H. Tang, Z. Lin, T.S. Huang, Emotion recognition from arbitrary view facial images, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 490–503.
[26] H. Tang, M. Hasegawa-Johnson, T. Huang, Non-frontal view facial expression recognition based on ergodic hidden Markov model supervectors, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2010, pp. 1202–1207.
[27] U. Tariq, J. Yang, T.S. Huang, Multi-view facial expression recognition analysis with generic sparse coding feature, in: Proceedings of the European Conference on Computer Vision Workshops and Demonstrations, 2012, pp. 578–588.
[28] U. Tariq, J. Yang, T.S. Huang, Maximum margin GMM learning for facial expression recognition, in: Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition, 2013, pp. 1–6.
[29] U. Tariq, J. Yang, T.S. Huang, Supervised super-vector encoding for facial expression recognition, Pattern Recognit. Lett. (2014).
[30] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Proceedings of the European Conference Computer Vision Workshop on Statistical Learning in Computer Vision, 1, 2004, pp. 1–2.
[31] J. Wu, Z. Lin, H. Zha, Multiple models fusion for emotion recognition in the wild, in: Proceedings of the ACM International Conference on Multimodal Interaction, 2015, pp. 475–481.
[32] J. Wu, Z. Lin, H. Zha, Multi-view common space learning for emotion recognition in the wild, in: Proceedings of the ACM International Conference on Multimodal Interaction, 2016.
[33] S. Moore, R. Bowden, The effects of pose on facial expression recognition, in: Proceedings of the British Machine Vision Conference, 2009, pp. 1–11.
[34] S. Moore, R. Bowden, Local binary patterns for multi-view facial expression recognition, Comput. Vision. Image Underst. 115 (4) (2011) 541–558.
[35] R.-L. Vieriu, S. Tulyakov, S. Semeniuta, E. Sangineto, N. Sebe, Facial expression recognition under a wide range of head poses, in: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, 2015, pp. 1–7.
[36] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[37] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (2) (1982) 129–137.
[38] Y. Huang, K. Huang, Y. Yu, T. Tan, Salient coding for image classification, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2011, pp. 1753–1760.
[39] A.V. Ken Chatfield Victor Lempitsky, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, in: Proceedings of the British Machine Vision Conference, 2011, pp. 76.1–76.12.
[40] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2, 2006, pp. 2169–2178.
[41] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, Q. Tian, Discovering discriminative graphlets for aerial image categories recognition, IEEE Trans. Image Process. 22 (12) (2013) 5071–5084.
[42] L. Zhang, Y. Yang, Y. Gao, Y. Yu, C. Wang, X. Li, A probabilistic associative model for segmenting weakly supervised images, IEEE Trans. Image Process. 23 (9) (2014) 4150–4159.
[43] L. Zhang, Y. Gao, Y. Xia, Q. Dai, X. Li, A fine-grained image categorization system by cellet-encoded spatial pyramid modeling, IEEE Trans. Ind. Electron. 62 (1) (2015) 564–571.
[44] L. Yin, X. Wei, Y. Sun, J. Wang, M.J. Rosato, A 3D facial expression database for facial behavior research, in: Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition, 2006, pp. 211–216.
[45] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE, Image Vis. Comput. 28 (5) (2010) 807–813.
[46] P. Ekman, W.V. Friesen, Facial action Coding System, Palo Alto: Consulting Psychologists Press, 1977.

Jianlong Wu received the bachelors degree in electronics and information engineering from Huazhong University of Science and Technology in 2014. He is currently pursuing the Ph.D. degree with the School of Electronics Engineering and Computer Science, Peking University. His research interests include computer vision, pattern recognition and machine learning.



Zhouchen Lin received the Ph.D. degree in applied mathematics from Peking University, Beijing, China, in 2000. He is currently a Professor with the Key Laboratory of Machine Perception, school of Electronics Engineering and Computer Science, Peking University. He is a Chair Professor with Northeast Normal University, Changchun, China. In 2012, he was a Lead Researcher with the Visual Computing Group, Microsoft Research Asia. He was a Guest Professor with Shanghai Jiaotong University, Shanghai, China, Beijing Jiao Tong University, Beijing, and Southeast University, Nanjing, China. He was a Guest Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His current research interests include computer vision, image processing, computer graphics, machine learning, pattern recognition, and numerical computation and optimization. He is an area chair of CVPR 2014, ICCV 2015, NIPS 2015, AAAI 2016, CVPR 2016, IJCAI 2016, and AAAI 2017. He is an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence and the International Journal of Computer Vision. He is an IAPR Fellow.



Wenming Zheng received the B.S. degree in computer science from Fuzhou University, Fuzhou, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, China, in 2004. He has been with the Research Center for Learning Science, Southeast University, since 2004. Currently, he is a Professor with the Key Laboratory of Child Development and Learning Science of the Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing. His current research interests include affective computing, neural computation, pattern recognition, machine learning, and computer vision.

Hongbin Zha received the B.E. degree in electrical engineering from Hefei University of Technology, Hefei, China, in 1983 and the M.S. and Ph.D. degrees in electrical engineering from Kyushu University, Fukuoka, Japan, in 1987 and 1990, respectively. After working as a Research Associate with Kyushu Institute of Technology, he joined Kyushu University as an Associate Professor in 1991. In 1999, he was also a Visiting Professor with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Surrey, U.K. Since 2000, he has been with Peking University, Beijing, China as a Professor with the Key Laboratory of Machine Perception (MOE), and also with the School of Electronics Engineering and Computer Science. His research interests include computer vision, digital geometry processing, and robotics. He received the Franklin V. Taylor Award from the IEEE Systems, Man, and Cybernetics Society in 1999. He is a member of the IEEE Computer Society.