Joint Feature Learning for Face Recognition

Jiwen Lu, *Member, IEEE*, Venice Erin Liong, Gang Wang, *Member, IEEE*, and Pierre Moulin, *Fellow, IEEE*

***Abstract*— This paper presents a new joint feature learning (JFL) approach to automatically learn feature representation from raw pixels for face recognition. Unlike many existing face recognition systems, where conventional feature descriptors, such as local binary patterns and Gabor features, are used for face representation, we propose an unsupervised feature learning method to learn hierarchical feature representation. Since different face regions have different physical characteristics, we propose to use different feature dictionaries to represent them, and to learn multiple yet related feature projection matrices for these regions simultaneously. Hence position-specific discriminative information can be exploited for face representation. Having learned these feature projections for different face regions, we perform spatial pooling for face patches within each region to enhance the representative power of the learned features. Moreover, we stack our JFL model into a deep architecture to exploit hierarchical information for feature representation and further improve the recognition performance. Experimental results on five widely used face data sets show the effectiveness of our proposed approach.**

***Index Terms*— Face recognition, feature learning, joint learning, deep learning.**

I. Introduction

FACE recognition has been extensively investigated in computer vision and biometrics over the past two decades, and many face recognition methods have been proposed [13], [20], [22], [42], [58]. While many of these methods have achieved reasonably good performance in controlled environments, their performance is still far from satisfactory in unconstrained environments where diverse real-world imaging conditions such as varying poses, illuminations and expressions heavily affect the recognition performance. Hence, how to extract robust and discriminative features to enlarge the inter-personal margins and reduce the intra-personal variations simultaneously remains a central and challenging problem in face recognition.

Manuscript received August 15, 2014; revised January 19, 2015; accepted February 24, 2015. Date of publication March 3, 2015; date of current version May 15, 2015. This work was supported by the Human Cyber Security Systems Program within the Advanced Digital Sciences Center, Singapore, through the Agency for Science, Technology and Research, Singapore. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhenan Sun.

J. Lu and V. E. Liong are with the Advanced Digital Sciences Center, Singapore 138632 (e-mail: jiwen.lu@adsc.com.sg; venice.l@adsc.com.sg).

G. Wang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, and also with the Advanced Digital Sciences Center, Singapore 138632 (e-mail: wanggang@ntu.edu.sg).

P. Moulin is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA, and also with the Advanced Digital Sciences Center, Singapore 138632 (e-mail: moulin@ifp.uiuc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2015.2408431

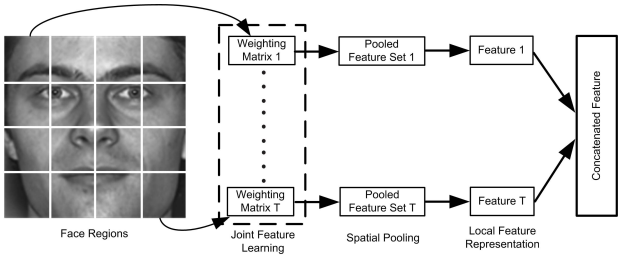


Fig. 1. The basic idea of our proposed approach. For each face image, we divide it into several non-overlapping regions and jointly learn feature weighting matrices. Then, the learned features in each region are pooled and represented as local histogram feature descriptors. Lastly, we combine and concatenate these local features into a longer feature vector for face representation. In this figure, we only show the basic learning module of our approach. How to stack this basic learning module into a deep architecture is detailed in Fig. 4.

A variety of facial feature representations have been proposed in recent years, and they can be mainly classified into two categories: holistic-feature based [5], [21], [23], [24], [59] and local-feature based [1], [40]. Typical holistic features include principal component analysis (PCA) [59] and linear discriminant analysis (LDA) [5], and representative local features include local binary patterns (LBP) [1] and Gabor wavelets [40]. While many local feature descriptors have achieved reasonably good performance in different face recognition applications [10], [37], [63], [69], [70], most of them are based on heuristics. Moreover, computing them is usually time-consuming.

In this paper, we propose an unsupervised feature learning approach for face recognition, as illustrated in Fig. 1. Unlike local face descriptors such as local binary patterns and Gabor features [1], [40], we propose a feature learning approach to learn data-adaptive features directly from raw pixel values for face representation and recognition, so that higher-order statistics information can be effectively characterized. While many feature learning methods have been proposed in recent years [34], [35], most of them learn a single feature projection matrix to transform all patches in images to achieve translation invariance, which is not very important for face recognition applications because there is usually a face alignment step in a practical face recognition system. For face recognition, different face regions usually have different structures and it is desirable to learn more region-specific features for face representation. A possible way to achieve this goal is to learn feature representations for different face regions individually. However, different face regions usually share some related information in feature representation and individual feature learning ignores this characteristic. To exploit shared information among different face regions, we jointly learn

features for different face regions, where both the relationship between different face regions and position-specific information are simultaneously exploited for representation. Having obtained features for each regions, we perform spatial pooling for different regions to increase their representative capability. Moreover, we stack our feature learning model into a deep architecture to exploit hierarchical and complementary information to further improve the recognition performance. Experimental results on the FERET, CAS-PEAL-R1, LFW, YTF and PaSC face datasets show that our approach achieves competitive or better performance than the state-of-the-art facial representation methods.

II. Related Work

In this section, we briefly review two related topics: 1) face representation, and 2) feature learning.

A. Face Representation

Face representation methods can be classified into two categories: holistic-based methods [4], [5], [59] and local-based methods [1], [40], [69], [70]. Holistic features lexicographically convert each face image into a high-dimensional feature vector and learn a feature subspace to preserve the statistical information of face images. Representative holistic features include PCA [59], LDA [5], [14], independent component analysis (ICA) [4], [15], and locality preserving projections (LPP) [25]. In contrast, local features first describe the structure of each local patch and then combine them into a concatenated feature vector. Typical local features include local binary patterns (LBP) [1], Gabor features [40] and their combinations such as local Gabor binary patterns (LGBP) [70], histogram of Gabor phase patterns [69], and Gabor volume based local binary patterns (GV-LBP) [37].

B. Feature Learning

A number of feature learning methods have been proposed in recent years [6], [26], [29], [32], [34], [52], and most of them have been successfully used in visual analysis applications such as pedestrian detection [50], action recognition [35], image classification [34], and visual tracking [18]. Representative feature learning methods include sparse auto-encoder [6], restricted Boltzmann machine [26], denoising auto-encoders [52], convolutional neural networks [29], independent subspace analysis [35], and reconstruction ICA (RICA) [34]. Recently, feature learning has also been used for face representation and several feature learning-based face recognition methods have been proposed. Most of them outperform the conventional local feature descriptors [9], [31], [38]. For example, Lei *et al.* [38] proposed a discriminant face descriptor (DFD) method by learning an image filter using the LDA criterion to obtain LBP-like features. Cao *et al.* [9] presented a learning-based (LE) feature representation method under the bag-of-word (BoW) framework. Hussain *et al.* [31] proposed a local quantized pattern (LQP) method by modifying the LBP method with a learned coding strategy. For face recognition, different face regions usually have different structures and it is desirable to learn more region-specific features for face representation.

III. Joint Feature Learning

In this section, we first present the basic joint feature learning module, and then detail the proposed stacked joint feature learning approach. Lastly, implementation details of the proposed method are presented.

A. Basic Joint Feature Learning

While many feature learning methods have achieved encouraging results on object and action recognition [6], [26], [29], [32], [34], [35], [52], most of them only learn a single feature projection matrix to transform all patches in images to achieve translation invariance. As mentioned in the introduction, this invariance property is not very important for face recognition. Instead, we should learn position-specific features to capture essential information from each local region because different face regions have different structures. As mentioned before, a possible way to achieve this goal is to learn features for different face regions individually. However, different face regions usually share some related information and individual learning will ignore this. Hence, we propose to jointly learn multiple different yet related features for different face regions to extract discriminative information for feature representation.

Recent advances in multi-view learning have shown that parameters from different yet related learning objectives are usually assumed to lie in a low dimensional subspace [17], [33], so that different learning functions can have some overlapped bases, which are shared by different learning functions. Motivated by this finding, we assume that there are T learning objectives and the parameter of each objective can be represented as a linear combination of l ($< T$) latent bases. For each face image, we divide it into T non-overlapped regions and learn feature representation for each region jointly. Let $\mathbf{W}_t$ be the feature projection matrix for the tth region, $1 \leq t \leq T$. Then, the learned feature of the tth region is represented as $\mathbf{W}_t\mathbf{x}$, where $\mathbf{x}$ is a sample in this region, which consists of raw pixel values from a patch. We assume that there are l ($< T$) latent bases to form a dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_l] \in \mathbb{R}^{kn \times l}$ which is shared by different regions, and $\mathbf{W}_t$ is represented as $\mathbf{W}_t = \text{mat}(\mathbf{D}\alpha_t)$, where $\alpha_t \in \mathbb{R}^l$ is the representation coefficient vector for the tth region, and $\text{mat}(a)$ is the matrix form of a. We assume that α_t is sparse such that only a few of the latent bases in $\mathbf{D}$ are selected to represent $\mathbf{W}_t$. To achieve this, we formulate the following optimization problem:

$$\min_{\mathbf{D},\mathbf{A}} \mathbf{H}(\mathbf{D}, \mathbf{A}) = \sum_{t=1}^{T} F_t(\text{mat}(\mathbf{D}\alpha_t)) + \gamma_1 \left(\sum_{t=1}^{T} \|\text{vec}(\mathbf{W}_t^0) - \mathbf{D}\alpha_t\|_2^2 + \xi\|\alpha_t\|_1 \right)$$
$$\text{subject to: } \|\mathbf{d}_j\|^2 \leq 1, \quad 1 \leq j \leq l. \tag{1}$$

The first term in (1) is to learn a feature projection matrix $\text{mat}(\mathbf{D}\alpha_t)$ to extract sparse features for samples in

the tth region P_t. The second term in (1) favors $\mathbf{W}_t$ that is sparsely represented as a linear combination of a subset of $\mathbf{D}$. In (1), $\mathbf{A} = [\alpha_1, \alpha_2, \cdots, \alpha_T]$, γ_1 is a parameter to balance the importance of the two terms, ξ is a parameter to enforce the sparsity of α_t, vec($\mathbf{X}$) is the vectorization of the matrix $\mathbf{X}$, $\mathbf{W}_t^0$ is the initialized feature projection matrix for the tth region, which is learned individually for the tth region by using RICA [34] in (2). F_t is the objective function of the conventional RICA feature learning method for the tth region, which is defined as the following unconstrained optimization problem:

$$\min_{\mathbf{W}_t} F_t(\mathbf{W}_t) = \frac{\lambda}{m}\sum_{i=1}^{m}\|\mathbf{W}_t^T\mathbf{W}_t\mathbf{x}_{it} - \mathbf{x}_{it}\|_2^2 + \sum_{i=1}^{m}\sum_{j=1}^{k} h(\mathbf{W}_{tj}\mathbf{x}_{it}) \tag{2}$$

where $\{\mathbf{x}_{it}\}|_{i=1}^{m} \in \mathbb{R}^n$ is the unlabeled data set in the tth region, h is a nonlinear convex function, such as $h(\cdot) = \log(\cosh(\cdot))$ [32], $\mathbf{W}_t \in \mathbb{R}^{k\times n}$ is the feature weighting matrix, k is the number of features, and $\mathbf{W}_{tj}$ is the jth row feature in $\mathbf{W}_t$, $\mathbf{W}_t\mathbf{x}_{it}$ is the learned feature of $\mathbf{x}_{it}$, λ is a parameter to balance the importance of the two terms in (1).

Previous studies in face recognition [42], [54] have shown that face patches sampled from different positions can be considered as elements of a nonlinear manifold, especially when face patches are densely sampled. Moreover, Seo and Milanfar [54] have shown that spatial information of face patches is important for face feature representation because neighboring face patches are more similar to each other due to the local geometric structure constraint [54]. Motivated by this finding, we also expect neighboring regions to share similar feature representations in the learned space so that the spatial manifold constraint can be well preserved. To achieve this, we formulate the following optimization problem:

$$\min_{\mathbf{A}} \sum_{t=1}^{T}\sum_{t'=1}^{T}\|\alpha_t - \alpha_{t'}\|_2^2\mathbf{S}_{tt'} \tag{3}$$

where $\mathbf{S}_{tt'}$ is an affinity matrix to measure the spatial relation of the tth and t'th regions, which is defined as follows:

$$\mathbf{S}_{tt'} = \begin{cases} \exp(-d_{tt'}/\sigma), & \text{if } P_{t'} \in N_r(P_t) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $P_{t'}$ and P_t denote the t'th and tth regions in the face image, $N_r(P_t)$ denotes the r-nearest neighbors of P_t, r defines the neighborhood size, $d_{tt'}$ represents the distance between two neighboring regions, which is set as 1 for the horizontal and vertical neighboring regions, and $\sqrt{2}$ for the diagonal neighboring regions, respectively, and σ is set as 10 in our experiments. Fig. 2 illustrates one example to show how to determine the neighboring regions for one given face region.

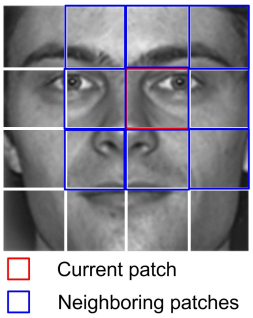


Fig. 2. Illustration of defining neighboring regions for one face region where r is selected as 8. The region with the red box is the current given region, and the other regions with blue boxes are the neighboring regions.

Combining (1) and (3), we formulate our joint feature learning model as the following optimization problem:

$$\begin{aligned}\min_{\mathbf{D},\mathbf{A}} \mathbf{H}(\mathbf{D},\mathbf{A}) = &\sum_{t=1}^{T} F_t(\text{mat}(\mathbf{D}\alpha_t)) \\ &+ \gamma_1\left(\sum_{t=1}^{T}\|\text{vec}(\mathbf{W}_t^0) - \mathbf{D}\alpha_t\|_2^2 + \xi\|\alpha_t\|_1\right) \\ &+ \gamma_2\sum_{t=1}^{T}\sum_{t'=1}^{T}\|\alpha_t - \alpha_{t'}\|_2^2\mathbf{S}_{tt'} \\ \text{subject to: } &\|d_j\|^2 \le 1, \quad 1 \le j \le l. \end{aligned} \tag{5}$$

where γ_2 is a parameter to balance the importance of (1) and (3).

We propose an alternating optimization method to iteratively optimize $\mathbf{D}$ and $\mathbf{A}$ in (5). For fixed $\mathbf{D}$, the cost function is additive over α_t, hence we solve n independent minimization problems:

$$\min_{\alpha_t} H(\alpha_t) = G(\alpha_t) + \xi\gamma_1\|\alpha_t\|_1, \quad 1 \le t \le n. \tag{6}$$

where

$$\begin{aligned} G(\alpha_t) = &\ F_t(\text{mat}(\mathbf{D}\alpha_t)) + \gamma_1\|\text{vec}(\mathbf{W}_t^0) - \mathbf{D}\alpha_t\|_2^2 \\ &-\gamma_2\sum_{t'=1}^{T} tr(\alpha_{t'}^T\alpha_t\mathbf{S}_{tt'}) - \gamma_2\sum_{t=1}^{T} tr(\alpha_t^T\alpha_{t'}\mathbf{S}_{tt'}) \end{aligned} \tag{7}$$

We use the feature sign search algorithm in [36] to solve for α_t in (6).

For fixed $\mathbf{A}$, we update the dictionary $\mathbf{D}$ by optimizing the following objective function

$$\begin{aligned}\min_{\mathbf{D}} H(\mathbf{D}) = &\sum_{t=1}^{T} F_t(\text{mat}(\mathbf{D}\alpha_t)) \\ &+ \gamma_1\sum_{t=1}^{T}\|\text{vec}(\mathbf{W}_t^0) - \mathbf{D}\alpha_t\|_2^2 \\ \text{subject to: } &\|\mathbf{d}_j\|^2 \le 1, \quad 1 \le j \le l. \end{aligned} \tag{8}$$

We use the conjugate gradient descent method in [34] to optimize $\mathbf{D}$ in (8). **Algorithm 1** summarizes the proposed joint feature learning method.

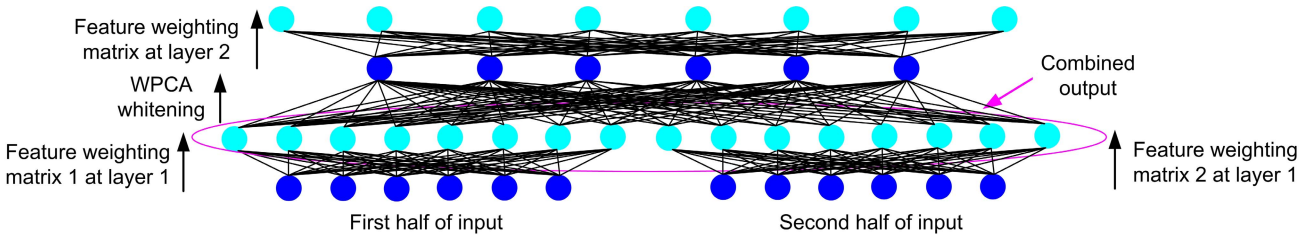


Fig. 3. Basic learning model of the stacked feature learning network. The network is built by "copying" the learned network and "pasting" it to the different places of the input data within the same region. The outputs of the first layer are utilized as inputs to the second layer in the network, and the filtering is performed with overlapping in our experiments. In this figure, the blue and green circles denote the input and output nodes in each network, and their sizes are set as 6 and 8, respectively, such that over-completed features can be learned by (2). We combine the outputs of the first layer and perform dimensionality reduction by WPCA to obtain a low-dimensional feature as the input to the second layer of the network. For clarity, the filtering step is shown here non-overlapping, but in our experiments it is performed with overlapping.

Algorithm 1 JFL

Input: Training set: $\{\mathbf{x}_{it}\}|_{i=1}^{m}$, $1 \leq t \leq T$, dictionary size l, iteration number R, and convergence error ϵ.
Output: Dictionary $\mathbf{D}$ and coefficient matrix $\mathbf{A}$.
Step 1 (Initialization):
1.1. Learn $\mathbf{W}_t^0$ by applying RICA on image patches in the tth region, $\forall t$.
1.2. Initialize $\mathbf{W}^0 = [\text{vec}(\mathbf{W}_1^0); \cdots; \text{vec}(\mathbf{W}_T^0)]^T$.
1.3. Compute covariance matrix $S_W = (\mathbf{W}^0)^T\mathbf{W}^0$.
1.4. Perform SVD on $S_W = \mathbf{USV}^T$.
1.5. Initialize $\mathbf{D}$ to be the first l columns of $\mathbf{U}$.
1.6. Initialize $\mathbf{A}$ with a nonnegative random matrix.
Step 2 (Local optimization):
For $r = 1, 2, \cdots, R$, repeat
2.1. Update $\mathbf{A}$ by solving (6).
2.2. Update $\mathbf{D}$ by solving (8).
2.3. If $r > 2$ and $|\mathbf{D}^r - \mathbf{D}^{r-1}| < \epsilon$, go to Step 3.
Step 3 (Output the dictionary and coefficient matrix):
Output $\mathbf{D} = \mathbf{D}^r$ and $\mathbf{A} = \mathbf{A}^r$.

B. Spatial Pooling

Having obtained the learned dictionary $\mathbf{D}$ and $\mathbf{A}$, we obtain the feature extraction matrix $\mathbf{W}_t = \text{mat}(\mathbf{D}\alpha_t)$ for the tth region. We first project each patch in the tth patch by using $\mathbf{W}_t$ and encode it into discrete codes.[1] Unlike previous feature learning methods [34], [35] which directly perform pooling on the learned features, we apply an unsupervised clustering method to learn a set of codebooks (one per region) from the training set such that the learned codes are more data-adaptive. In our implementation, K-means is used due to its simplicity. Hence, there are T codebooks to be learned from the training set. For each region, we extract a histogram feature by using the corresponding codebook and concatenate the histogram features from all regions into a longer feature vector for face representation.

C. Stacked Joint Feature Learning

Previous studies have shown that higher-level feature representations can be obtained if basic learning models are stacked [29], [34], [35], [50]. To make better use of joint feature learning to extract more hierarchical features for face representation, we develop a stacked feature learning architecture which progressively makes use of our basic feature learning module as sub-units for unsupervised feature learning. Fig. 3 shows the basic stacked feature learning model of our approach.

The basic idea is as follows. We first train the joint feature learning model on small face patches and then use the learned feature weighting matrices to filter sampled small patches within large patches.[2] Within each large image patch, the outputs of small patches are concatenated and taken as the input of the next layer. We employ weighted PCA (WPCA) [38] to map the combined responses of the first layer to a low-dimensional feature space to reduce the redundancy. Similarly to most deep learning methods [29], [34], [35], [50], the stacked joint feature learning network is trained greedily layer-wise. Specifically, we train the first layer until convergence before training the second layer. Having learned feature extraction matrices for the first and second layers, we project samples from each layer with the learned feature projections and learn dominant patterns to generate the codebook for each region and layer, respectively. In our experiments, we combine features extracted from both layers and combine them together for face recognition. In the experiment section, we show that this combination works better than using feature extracted from a single individual layer.

Fig. 4 shows how to use the stacked convolutional network to jointly learn feature presentation. For each face image, we divide it into several non-overlapping regions and learn local features for each region. In this figure, only 4 regions are used to illustrate the basic idea, but in our experiments, each face image is divided into more non-overlapping regions. We take the first region as an example to show how to extract features. We first sample a number of small face image patches in the first region and flatten them into feature vectors (denoted as $\mathbf{f}_{11}, \cdots, \mathbf{f}_{1m}$) as the input to the first layer of the network. Having been filtered by the first layer of the basic feature

[1] In this work, each face image is first divined into several regions. For each region, we sample many small patches for feature learning.

[2] In our implementations, the whole face is first divided into many regions. Then, for each region, we sample two sizes of patches. For small patches, they are sampled within each region for feature learning. For the second layer, the output of all small patches are concatenated and mapped into a low-dimensional feature first, and then filtered in the second layer.

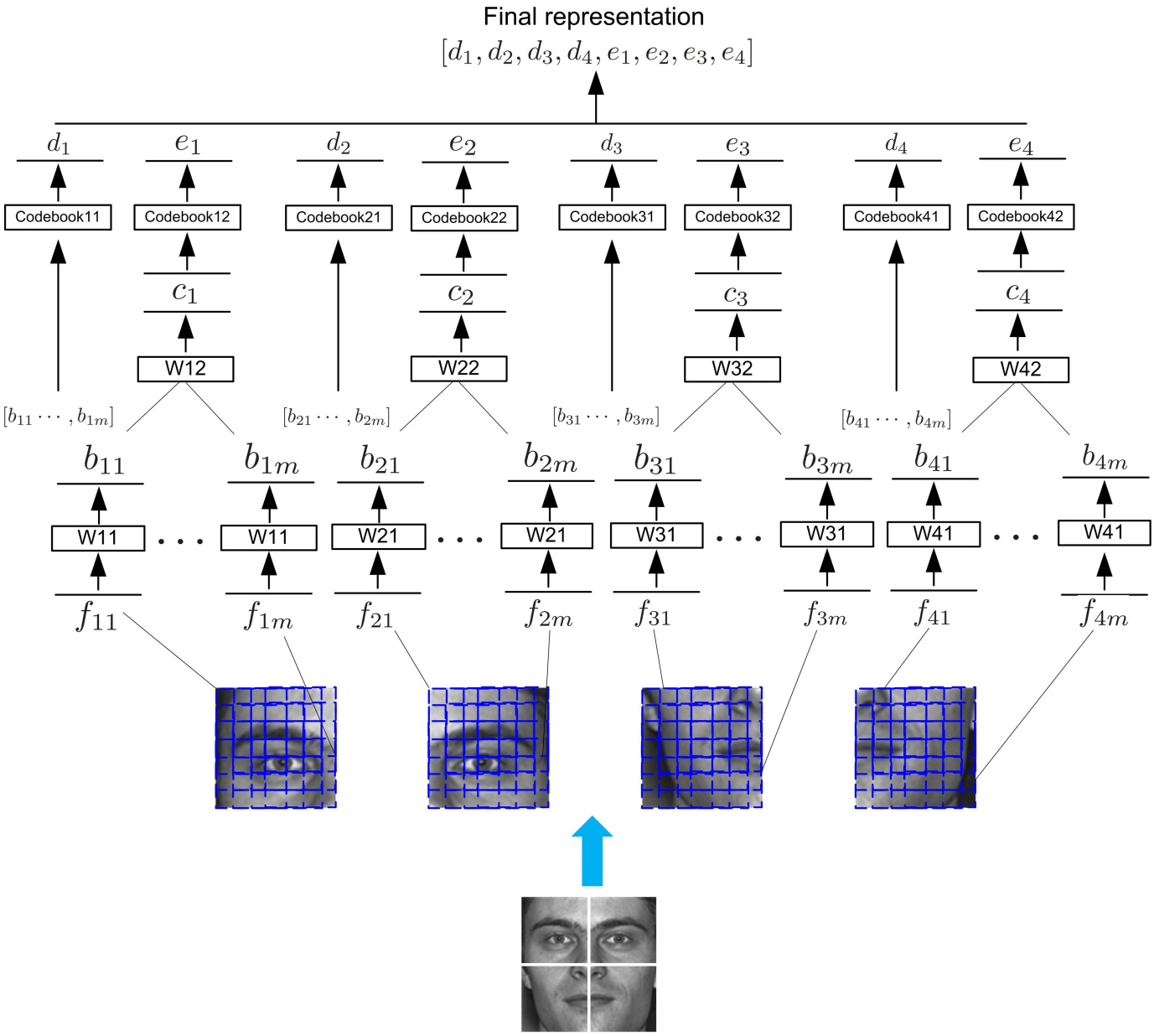


Fig. 4. Stacked joint feature learning for face representation.

weighting matrix ($\mathbf{W}_{11}$), they are mapped to $\mathbf{b}_{11}, \cdots, \mathbf{b}_{1m}$. We combine them together and apply WPCA to reduce the feature dimension of the concatenated feature vector. Then, we use the second layer of feature weighting matrix ($\mathbf{W}_{12}$) to project it to a feature vector $\mathbf{c}_1$. Similarly, we can obtain the outputs $\mathbf{b}_{21}, \cdots, \mathbf{b}_{2m}$, $\mathbf{b}_{31}, \cdots, \mathbf{b}_{3m}$, $\mathbf{b}_{41}, \cdots, \mathbf{b}_{4m}$ at the first layer, and $\mathbf{c}_2$, $\mathbf{c}_3$ and $\mathbf{c}_4$ at the second layer for the other three patches, respectively. Having extracted features at the first and second layers, we learn the corresponding codebooks and encode them as histogram features, where $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$ and $\mathbf{d}_4$ are histogram features for the fist layer, and $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ and $\mathbf{e}_4$ are histogram features for the second layer, respectively. Lastly, we concatenate these histogram features extracted from different layers and different regions together as the final feature representation of the whole face image.

D. Implementation Details

In our implementation of the proposed approach, each face image was divided into 8×8 non-overlapping regions and hence there are 64 regions in total to jointly learn the sparse features. The dictionary size l is set to 20 which indicates that only 20 latent bases are shared by these 64 regions in feature learning. The input sizes of the first and second layers of our approach are 7×7 and 9×9, and they were densely sampled from the image with an spacing of 1 and 1 pixels, respectively. Hence, there are nine 7×7 small patches within each 9×9 patch. For each 7×7 small patch, we learn a 49×100 projection matrix to map it into a 100 over-complete feature vector. The combined output of all the small patches in each 9×9 large patch are concatenated to a 900-dimensional feature vector, which is further reduced to

a 100-dimensional feature vector by WPCA. Then, we learn a 100×200 projection matrix to map it into a 200 over-complete feature vector. We apply WPCA to project these over-complete features extracted from the 7×7 and 9×9 patches into 15 and 15 feature dimension, respectively, such that the redundancy information can be further removed. Subsequently, we use the K-means method to learn codebooks to pool features extracted in the first and second layers. In our experiments, K is empirically set as 300 and 300 for the first and second layers, respectively. Lastly, we concatenate histogram features extracted from both layers together and use WPCA to map the concatenated feature vector into a low-dimensional feature space as the final feature representation of the whole face. In our experiments, all face images from different databases are cropped and scaled to size 128×128 pixels. The parameters γ_1, γ_2 and ξ of our approach are empirically tuned as 0.2, 0.2 and 0.005, respectively, by cross-validation on the FERET dataset. We find that our algorithm is not sensitive to these parameters. Having obtained the feature representation of each face image, we apply the cosine metric to measure their similarity for both our face identification and verification tasks.

E. Discussion

In this subsection, we highlight the difference between our joint feature learning model and several recently proposed methods.

1) Deep Convolutional Neural Networks [56], [57]: Deep convolutional neural networks (DCNN) have been used for feature learning in face recognition, and some of them [56], [57] have achieved very proposing performance. Generally, these methods require a large number of labeled data for training because there are extensive number of parameters to estimate. For some practical applications such as cross-modality face recognition, it is very hard to collect such large number of labeled data. In contrast, our feature learning method is a unsupervised approach. Hence, our feature learning applies to scenarios where labeled data are hard to collect.

2) Simultaneous Feature and Dictionary Learning (SFDL) [43]: In our recent work, we introduced SFDL to jointly learn discriminative features and dictionaries for image set based face recognition. The basic idea of SFDL is that some discriminative information for dictionary learning may be ignored in the feature learning stage if feature learning and dictionary learning are performed individually, so that jointly learning them in one framework can alleviate this shortcoming. Unlike SFDL, our JFL learns position-specific features for different face regions are learned while SFDL learns the feature projection matrix for the whole holistic face. Hence, the relationship between different face regions is exploited in JFL, which was ignored in SFDL.

IV. Experiments

We evaluated our proposed JFL method on five widely used face datasets including the FERET [51], CAS-PEAL-R1 [19], LFW [30], YTF [64] and PaSC [7]. The FERET and CAS-PEAL-R1 datasets are used to show the effectiveness of our approach for face identification in controlled environments, and the LFW, YTF and PaSC datasets are selected to demonstrate the efficacy of our approach for face verification in unconstrained environments.

Fig. 5. Several aligned and cropped face examples from the FERET dataset, showing two face images from the same person in each column.

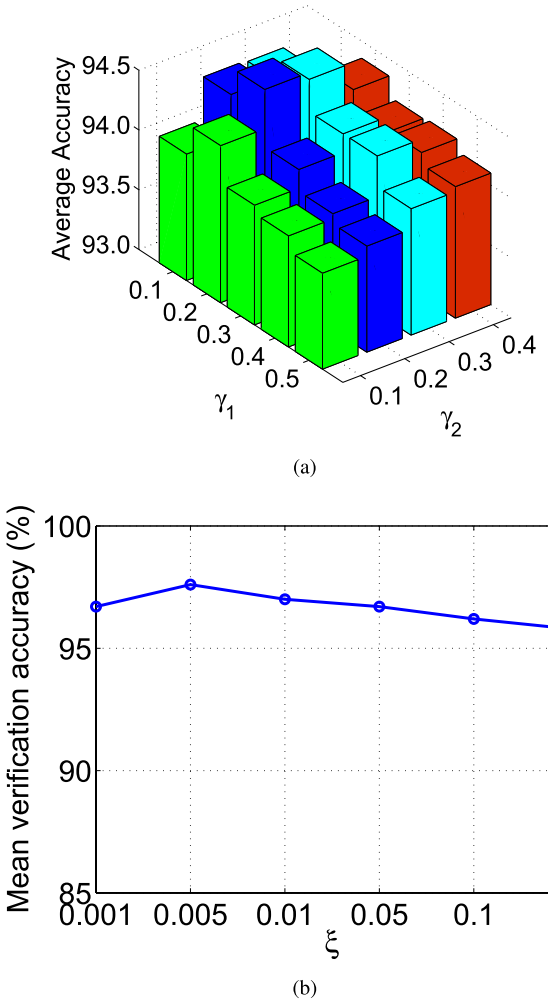


Fig. 6. Rank-one recognition rate of JFL on the training set of FERET using 10-fold cross-validation versus different values of (a) γ_1 and γ_2, and (b) ξ.

A. Evaluation on the FERET Dataset

The FERET database consists of 13539 facial images corresponding to 1565 subjects. In our experiments, we followed the standard FERET evaluation protocol [51], where six sets including training, fa, fb, fc, dup1, and dup2 were constructed for face recognition experiments. Fig. 5 shows some cropped example images from the FERET dataset.

TABLE I
RANK-ONE RECOGNITION RATES (%) COMPARISON WITH THE STATE-OF-THE-ART FACIAL DESCRIPTORS TESTED WITH THE STANDARD FERET EVALUATION PROTOCOL

Method	fb	fc	dup1	dup2	year
LBP [1]	93.0	51.0	61.0	50.0	2004
LGBP [70]	94.0	97.0	68.0	53.0	2005
HGGP [69]	97.6	98.9	77.7	76.1	2007
HOG [48]	90.0	74.0	54.0	46.6	2008
LLGP [66]	99.0	99.0	80.0	78.0	2009
LDP [68]	94.0	83.0	62.0	53.0	2010
GV-LBP-TOP [37]	98.1	98.5	80.9	81.2	2011
PDO [62]	99.7	**100.0**	91.7	90.6	2011
LQP [31]	99.8	94.3	85.5	78.6	2012
EPLS [53]	97.2	98.5	85.3	85.5	2012
POEM [63]	97.0	95.0	77.6	76.2	2012
s-POEM [61]	99.4	**100.0**	91.7	90.2	2013
DFD [38]	99.4	**100.0**	91.8	92.3	2014
JFL	**99.9**	**100.0**	**93.7**	**93.6**	

*The results of other methods are from the original papers.

Fig. 7. Several aligned and cropped face examples from the CAS-PEAL-R1 dataset, showing two face images from the same person in each column.

We first performed feature learning on the generic training set, and then applied the learned features on the other five sets for feature extraction. Finally, we applied WPCA to reduce the feature dimension to 1100 and used the cosine metric to compute similarity. We took fa as the gallery set and used the other four sets as the probe sets.

1) Parameter Selection: Parameters γ_1, γ_2 and ξ are selected by cross-validation on the FERET training set. There are 1002 images in the set, and we used 10-fold cross-validation. Fig. 6(a) and (b) show the average recognition rate of JFL versus different values of (γ_1, γ_2), and ξ on the training set of FERET, respectively. We see that JFL achieves the best recognition performance when γ_1, γ_2 and ξ are set to 0.2, 0.2, and 0.005, respectively.

2) Comparison With the State-of-the-Art Feature Descriptors: Table I shows the rank-one recognition rate of our proposed approach, compared with the state-of-the-art facial descriptors. Our approach outperforms the existing best results with gains in accuracy of 0.1%, 1.4% and 2.4% on the fb, dup1 and dup2 sets, and also achieves the best recognition rate on the fc set.

B. Evaluation on the CAS-PEAL-R1 Dataset

The CAS-PEAL-R1 database contains 9060 face images from 1040 subjects with varying pose, expression, accessory, and lighting (PEAL). In our experiments, we followed the standard evaluation protocol [19], where five sets including training, gallery, expression, lighting and accessory are constructed for face recognition experiments. Fig. 7 shows some aligned and cropped example images from the dataset. We first performed feature learning on the training set, and then applied the learned features on the other four sets for feature extraction. Finally, we applied WPCA to reduce the feature dimension into 1039 and used the cosine metric to compute similarity. Table II shows the rank-one recognition rate of our approach on the CAS-PEAL-R1 dataset, compared with the state-of-the-art facial descriptors. As can be seen, our approach improves the previous best results by 0.1%, 0.3% and 3.5% on the expression, accessory and lighting probe sets, respectively.

TABLE II
RANK-ONE RECOGNITION RATES (%) COMPARISON WITH THE STATE-OF-THE-ART FACIAL DESCRIPTORS TESTED WITH THE STANDARD CAS-PEAL-R1 EVALUATION PROTOCOL

Method	Expression	Accessory	Lighting	year
LBP [1]	97.0	89.0	29.0	2004
LGBP [70]	95.0	87.0	51.0	2005
LVP [47]	96.0	86.0	29.0	2006
HGGP [69]	96.0	92.0	62.0	2007
LLGP [66]	96.0	90.0	52.0	2009
DT-LBP [44]	98.0	92.0	41.0	2009
DLBP [45]	99.0	92.0	41.0	2011
DFD [38]	99.6	96.9	63.9	2014
JFL	**99.7**	**97.2**	**67.4**	

*The results of other methods are from the original papers.

Fig. 8. Several aligned and cropped face examples from the deep funneled LFW dataset, where two face images in the first three columns are from the same person and those in the last three columns are from different persons.

C. Evaluation on the LFW Dataset

The LFW dataset [30] contains more than 13000 face images of 5749 subjects captured from the web with variations in expression, pose, age, illumination, resolution, background, and so on. We followed the standard evaluation protocol on the "View 2" dataset [30] which includes 3000 matched pairs and 3000 mismatched pairs. The dataset is divided into 10 folds, and each fold consists of 300 matched (positive) pairs and 300 mismatched (negative) pairs. There are six evaluation protocols nn this dataset [28], and we evaluated our JFL with two different settings in our experiments: 1) *unsupervised* and 2) *image-restricted with label-free outside data*.

1) Unsupervised Setting: Here we used the deep funneled images for feature learning. Fig. 8 shows several cropped and aligned face images from the deep funneled version of the LFW dataset. We first performed feature learning on the training set, and applied the learned features on both the training and testing sets for feature extraction. Then, we applied WPCA to reduce the feature dimension into 700 for

TABLE III

AUC (%) COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON LFW WITH THE UNSUPERVISED SETTING

Method	AUC
LBP [60]	75.47
SIFT [60]	54.07
LARK [54]	78.30
LHS [55]	81.07
PAF [67]	94.05
MRF-MLBP [2]	89.94
DFD [38]	83.07
JFL	**91.03**

*The results of other methods are from the original papers.

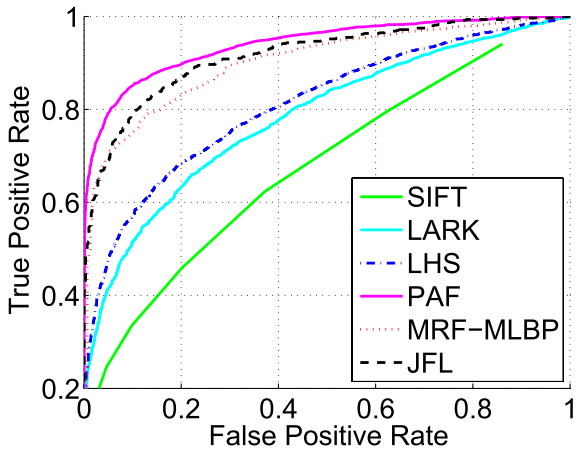


Fig. 9. ROC curves of different methods on LFW with the unsupervised setting.

Fig. 10. Several cropped and aligned face examples from the LFW-a dataset, showing two face images from the same person in the first three columns and two face images from different persons in the last three columns.

each aligned face image. Finally, we used the cosine metric to compute similarity. Table III and Fig. 9 show the area under curve (AUC) and the ROC curves of our JFL and other existing methods under the unsupervised setting on LFW. We see that our JFL method outperforms the other methods with the unsupervised setting.

2) Image-Restricted Setting With Label-Free Outside Data: Here we used the LFW-a[3] version of the LFW dataset for face verification. Fig 10 shows several example face images. We performed feature learning on the training set, and applied the learned features on both the training and testing sets for feature extraction. WPCA was applied to reduce the feature dimension into 700 for each aligned face image. We applied the discriminative deep metric learning (DDML) [27] to learn a

[3]Available: http://www.openu.ac.il/home/hassner/data/lfwa/.

TABLE IV

COMPARISONS OF THE MEAN VERIFICATION RATE AND STANDARD ERROR (%) WITH THE STATE-OF-THE-ART RESULTS ON LFW UNDER THE IMAGE RESTRICTED SETTING WITH LABEL-FREE OUTSIDE DATA

Method	Accuracy
CSML+SVM, aligned [49]	88.00 ± 0.37
PAF [67]	87.77 ± 0.51
SFRD+PMML [12]	89.35 ± 0.50
Sub-SML [8]	89.73 ± 0.38
VMRS [3]	91.10 ± 0.59
DDML [27]	90.68 ± 1.41
JFL	**87.12 ± 1.70**
JFL (combine)	**92.93 ± 1.26**

*The results of other methods are from the original papers.

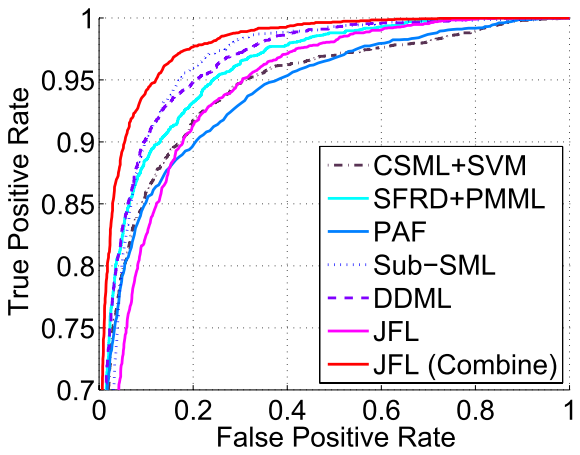


Fig. 11. ROC curves of different methods on LFW with the image-restricted setting with label-free outside data.

distance metric network to compute the similarity of each face pair. To further improve the verification performance, we combined our JFL with five feature descriptors: 1) HDLBP [11], 2) LBP [27], 3) Sparse SIFT [27], 4) Dense SIFT [27] and 5) HOG [16].[4] Having obtained eight feature descriptors, we used the square root of each element in each feature vector for face verification and employed SVM to fuse the scores to get the final verification result. Table IV and Fig. 11 show the mean verification rate with standard error and the ROC curve of our JFL and other existing methods on this dataset. Our JFL achieves competitive performance with existing state-of-the-art methods. Moreover, JFL achieves the best recognition rate (92.93%) when other five feature descriptors are combined, while the current best is only 91.10% with this setting.

D. Evaluation on the YTF Dataset

The YTF [64] dataset contains 3425 videos of 1596 subjects which were downloaded from YouTube. Fig. 12 shows several example face images. The average length of each video clip is about 180 frames. There are large variations

[4]We used these five feature representations provided by the original authors, respectively.

Fig. 12. Several cropped and aligned face examples from the YTF dataset, showing two face images from the same person in the first three columns and two face images from different persons in the last three columns.

TABLE V
VERIFICATION PERFORMANCE (MEAN ACCURACY ± STANDARD ERROR %) COMPARISON WITH THE STATE-OF-THE-ART RESULTS ON THE YTF DATASET

Method	Accuracy
LBP+MBGS [64]	76.4 ± 1.8
APEM (fusion) [39]	79.1 ± 1.5
STFRD+PMML [12]	79.5 ± 2.0
MBGS+SVM- [65]	78.9 ± 1.9
VSOF+OSS (Adaboost) [46]	79.7 ± 1.8
DDML(LBP) [27]	81.3 ± 1.6
DDML(combined) [27]	82.3 ± 1.5
JFL	**81.9 ± 0.9**
JFL (combined)	**83.9 ± 0.9**

*The results of other methods are from the original papers.

TABLE VI
VERIFICATION RATE (%) AT THE 1.0% FAR OF DIFFERENT METHODS ON THE PaSC DATASET

Method	Verification rate
LRPCA [7]	10.0
LBP [1]	25.1
SIFT [41]	23.2
DFD [38]	30.6
JFL	**32.6**

in pose, illumination, and expression in each video. In our experiments, we followed the standard evaluation protocol and tested our approach for unconstrained face verification by using 5000 video pairs which were randomly selected in [64], where half of them were from the same subject and the remaining half were from different subjects. There pairs were equally divided into 10 folds with each fold has 250 intra-personal pairs and 250 inter-personal pairs. Different from the LFW dataset, the image restricted training setting has been widely used in the YTF dataset and we also followed the same setting with 10-fold cross validation in our experiments [64]. Specifically, we cropped each image frame to size of 128 × 128 according to the provided aligned data[5] to learn the feature representation for each frame. Then, we applied WPCA to reduce the feature dimension of each image frame to 500 dimensions. Considering that all the faces are aligned by fixing the detected facial key points [64], the features extracted from all the frames within one video clip were averaged to output a mean feature vector for further processing. Lastly, we used discriminative deep metric learning (DDML) [27] for face verification. Table V tabulates the verification performance of our approach, compared with the state-of-the-art results with the restricted setting. Fig. 13 shows the ROC curves of different feature descriptors. As can be seen, our proposed approach achieves the state-of-the-art performance on the YTF dataset.

[5]Available at: http://www.cs.tau.ac.il/~/ytfaces/.

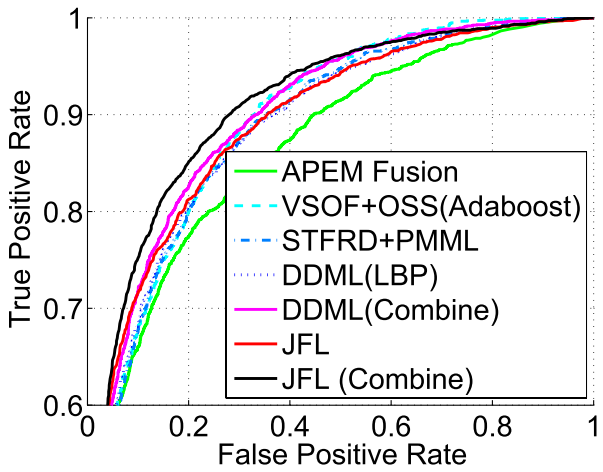


Fig. 13. ROC curves of different feature descriptors on the YTF dataset.

Fig. 14. Several cropped face examples from the PaSC dataset, showing two face images from the same person in the same column.

E. Evaluation on the PaSC Dataset

The PaSC dataset consists of 9376 still images of 293 people, collected at different locations, poses and distances from the camera. There is one query set and one target set, and each has 4688 images. Each image is aligned and cropped to 128 × 128 pixels according to the provided eye coordinates. Fig. 14 shows some cropped example images. We performed feature learning on the target sets and then used WPCA to project each face image into a 500-dimensional feature vector as the final face representation. We used the standard evaluation protocol in [7] where all images in the query set are compared with those in the target set so that a similarity matrix is computed to generate the ROC curve.

Besides the LRPCA baseline result provided in [7], we also compared our method with two conventional feature descriptors and one learning-based feature descriptor. For the conventional local feature descriptors, the LBP and SIFT were compared. For the learning-based feature, the recently proposed DFD method [38] was compared. Specifically, we first divided each image into 8 × 8 non-overlapping blocks, where the size of each block is 16 × 16. Then, we

TABLE VII
RECOGNITION RATES (%) OF THE JOINT AND INDIVIDUAL FEATURE LEARNING METHODS ON DIFFERENT FACE DATASETS

Method	FERET				CAS-PEAL-R1			LFW		YTF	PaSC
	fa	fb	dup1	dup2	Expression	Accessory	Lighting	Unsupervised	Restricted		
IFL	99.4	99.8	92.1	92.5	99.3	96.5	64.3	89.6	91.4	82.2	30.6
JFL	**99.9**	**100.0**	**93.7**	**93.6**	**99.7**	**97.2**	**67.4**	**91.0**	**92.9**	**83.9**	**32.6**

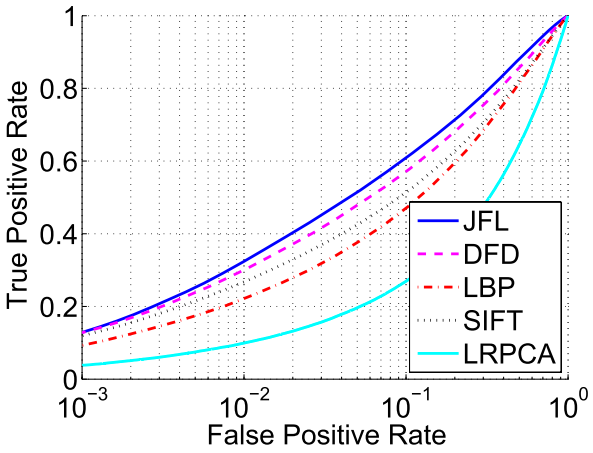


Fig. 15. ROC curves of different feature descriptors on the PaSC dataset with the unsupervised setting.

extracted a 59-dimensional LBP feature and 128-dimensional SIFT feature for each block and concatenated them to form 3776-dimensional and 8192-dimensional feature vectors, respectively. Finally, we employed WPCA to reduce each of them into a 500-dimensional feature vector as the final representation. For DFD, we followed the same setting in [38]. We also divided each face image to 8×8 non-overlapping blocks and learned local features for each block. Finally, the learned features from all blocks were concatenated and the WPCA was applied to reduce it into a 500-dimensional feature vector. Table VI tabulates the verification rate at the 1.0% FAR and Fig. 15 shows the ROC curve of these methods, respectively. As can seen, JFL significantly outperforms the other four compared methods, as the improvement of the verification rate is at least 2.0%.

F. Analysis

1) Comparison With Individual Feature Learning: We compared our joint feature learning (JFL) approach with individual feature learning (IFL). Individual feature learning means hierarchical features are learned from the different face regions separately. Table VII shows the recognition results of individual and joint feature learning methods on different face datasets. As can be seen, our joint feature learning approach achieves better recognition accuracy than the individual feature learning method in all the five face datasets. This is because joint feature learning can exploit more shared information across different face regions and exploit more discriminative information.

2) Effects of Features Extracted From Different Layers: In our approach, we combined features extracted from layers 1 and 2 for face feature representation. A natural question is: what is the individual contribution of each single layer of feature? To answer this question, we used the features extracted from each single layer for face recognition. Table VIII shows the recognition accuracy with features extracted from different layers on different subsets of the FERET dataset. As can be seen, features extracted in both layers contain discriminative information for face recognition. Specifically, the combined feature outperforms single feature extracted from layers 1 and 2 with the gains in accuracy of 0.9% and 1.4%, 1.0% and 2.5%, 1.4% and 1.9%, and 1.9% and 1.6% on the fb, fc, dup1 and dup2 sets, respectively.

TABLE VIII
RECOGNITION RATES (%) OF THE FEATURES EXTRACTED FROM DIFFERENT LAYERS ON FERET

Method	FERET			
	fa	fb	dup1	dup2
Layer 1	99.0	99.0	92.3	91.7
Layer 2	98.5	97.5	91.8	92.0
Layer 1 + Layer 2	**99.9**	**100.0**	**93.7**	**93.6**

3) Effects of the Spatial Constraint: Besides learning position-specific features, we also exploit the spatial constraint on the learned features in our approach. To examine the effect of the spatial constraint in our model, we set the parameter γ_2 to 0 and develop a new baseline method called JFL0, which means that our joint feature learning model is employed without exploiting the constraint on different regions. Table IX shows the recognition accuracy of JFL and JFL0. As can be seen, features extracted with the spatial constraint achieve higher recognition performance than those without the constraint, which means the spatial constraint plays an important role in our feature learning approach.

4) Feature Visualization: Generally, it is challenging to visualize and analyze higher level feature representations. We followed [35] and visualized the filters learned in early stage of our JFL model. Fig. 16 shows 100 filters learned in the early stage of our JFL method from FERET, where the size of each filter is 7×7. As can seen, our proposed JFL can learn features that detect edges from different directions.

G. Computational Time

In the training stage, our approach took 4 hours to learn the feature weighting matrices on 1002 face images from the FERET dataset. Having learned these parameters, feature extraction is very efficient because only filtering operations are required for feature extraction. For practical applications, training can be performed offline and testing is required in real

TABLE IX
RECOGNITION RATES (%) OF OUR JOINT FEATURE LEARNING METHOD WITH AND WITHOUT THE SPATIAL CONSTRAINT ON DIFFERENT FACE DATASETS

Method	FERET				CAS-PEAL-R1			LFW		YTF	PaSC
	fa	fb	dup1	dup2	Expression	Accessory	Lighting	Unsupervised	Restricted		
JFL0	99.1	99.4	91.8	92.2	99.1	96.1	64.1	88.6	91.2	81.8	29.8
JFL	**99.9**	**100.0**	**93.7**	**93.6**	**99.7**	**97.2**	**67.4**	**91.0**	**92.9**	**83.9**	**32.6**

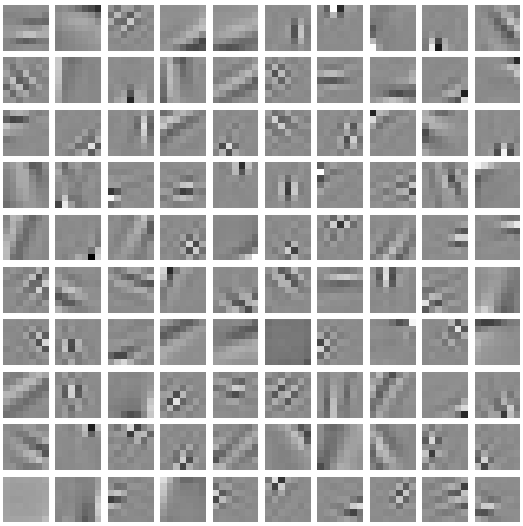


Fig. 16. 100 filters learned in the early stage of our JFL method from FERET, where the size of each filter is 7×7.

TABLE X
FEATURE EXTRACTION TIME (SECOND/IMAGE) OF DIFFERENT FEATURE DESCRIPTORS

Method	Time
LBP	0.27
TPLBP	0.31
FPLBP	0.33
LARK	0.37
DFD	1.51
JFL	**0.28**

time. We compared the feature extraction computational time of our approach and other local feature methods. Our hardware configuration includes a 2.8-GHz CPU and a 10GB RAM. Table X shows the feature extraction time of different feature representation methods. The feature extraction time of the proposed approach is comparable to that of other local feature descriptors.

H. Discussion

We make the following two observations from these experimental results:

1) JFL consistently outperform state-of-the-art feature descriptors in all datasets. This is because JFL learns feature representations from raw data, which are more data-adaptive than existing feature descriptors.
2) JFL outperforms DFD even though DFD is a supervised feature learning approach. This is because JFL exploits the relationship between different face regions in feature learning so that more position-specific feature can be exploited. Unlike DFD, JFL extracts more hierarchical information by the stacked model in feature representation.

V. CONCLUSION

We have proposed a new unsupervised feature learning approach to learn hierarchical feature representations for face recognition. By jointly learning multiple and related sparse features for different face regions, we extract more position-specific discriminative information for face representation. Experimental results on both controlled and unconstrained face datasets have shown the efficacy of our approach.

There are three interesting directions for future work:

1) Extension of JFL to learn feature representations for homogeneous face recognition, such as photo vs. sketch and 2D vs 3D face matching.
2) Extension of JFL into a supervised version to improve the discriminative power of the learned features for recognition.
3) Use of alternative deep learning architectures such as deep neural networks to improve the feature learning performance.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proc. 8th ECCV*, 2004, pp. 469–481.
[2] S. R. Arashloo and J. Kittler, "Efficient processing of MRFs for unconstrained-pose face recognition," in *Proc. IEEE 6th Int. Conf. BTAS*, Sep./Oct. 2013, pp. 1–8.
[3] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz, "Fast high dimensional vector multiplication face recognition," in *Proc. IEEE ICCV*, Dec. 2013, pp. 1960–1967.
[4] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Netw.*, vol. 13, no. 6, pp. 1450–1464, Nov. 2002.
[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[6] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. NIPS*, 2007, pp. 153–160.
[7] J. R. Beveridge *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *Proc. IEEE 6th Int. Conf. BTAS*, Sep./Oct. 2013, pp. 1–8.
[8] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *Proc. IEEE ICCV*, Dec. 2013, pp. 2408–2415.
[9] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proc. IEEE CVPR*, Jun. 2010, pp. 2707–2714.
[10] Z. Chai, Z. Sun, H. Méndez-Vázquez, R. He, and T. Tan, "Gabor ordinal measures for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 1, pp. 14–26, Jan. 2014.

[11] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3025–3032.

[12] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3554–3561.

[13] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1864–1870, Sep. 2012.

[14] W. Deng, J. Hu, X. Zhou, and J. Guo, "Equidistant prototypes embedding for single sample based face recognition with generic learning and incremental learning," *Pattern Recognit.*, vol. 47, no. 12, pp. 3738–3749, 2014.

[15] W. Deng, Y. Liu, J. Hu, and J. Guo, "The small sample size problem of ICA: A comparative study and analysis," *Pattern Recognit.*, vol. 45, no. 12, pp. 4438–4450, 2012.

[16] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using histograms of oriented gradients," *Pattern Recognit. Lett.*, vol. 32, no. 12, pp. 1598–1603, 2011.

[17] A. Evgeniou and M. Pontil, "Multi-task feature learning," in *Proc. NIPS*, vol. 19. 2007, pp. 41–48.

[18] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *IEEE Trans. Neural Netw.*, vol. 21, no. 10, pp. 1610–1623, Oct. 2010.

[19] W. Gao *et al.*, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.

[20] J. Gui, W. Jia, L. Zhu, S.-L. Wang, and D.-S. Huang, "Locality preserving discriminant projections for face and palmprint recognition," *Neurocomputing*, vol. 73, nos. 13–15, pp. 2696–2707, 2010.

[21] J. Gui, Z. Sun, J. Cheng, S. Ji, and X. Wu, "How to estimate the regularization parameter for spectral regression discriminant analysis and its kernel version?" *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 211–223, Feb. 2014.

[22] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, and S. Ji, "Discriminant sparse neighborhood preserving embedding for face recognition," *Pattern Recognit.*, vol. 45, no. 8, pp. 2884–2893, 2012.

[23] J. Gui, D. Tao, Z. Sun, Y. Luo, X. You, and Y. Y. Tang, "Group sparse multiview patch alignment framework with view consistency for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3126–3137, Jul. 2014.

[24] J. Gui, S.-L. Wang, and Y.-K. Lei, "Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data," *Artif. Intell. Med.*, vol. 50, no. 3, pp. 181–191, 2010.

[25] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[26] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[27] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. IEEE CVPR*, Jun. 2014, pp. 1875–1882.

[28] G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. UM-CS-2014-003, 2014.

[29] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Proc. IEEE CVPR*, Jun. 2012, pp. 2518–2525.

[30] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[31] S. U. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *Proc. BMVC*, 2012, pp. 1–12.

[32] A. Hyvärinen, J. Hurri, and P. O. Hoyer, "Independent component analysis," *Natural Image Statist.*, vol. 39, pp. 151–175, Jan. 2009.

[33] A. Kumar and H. Daume, III, "Learning task grouping and overlap in multi-task learning," in *Proc. 29th ICML*, 2012, pp. 1–8.

[34] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. NIPS*, 2011, pp. 1017–1025.

[35] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE CVPR*, Jun. 2011, pp. 3361–3368.

[36] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. NIPS*, 2006, pp. 801–808.

[37] Z. Lei, S. Liao, M. Pietikainen, and S. Z. Li, "Face recognition by exploring information jointly in space, scale and orientation," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 247–256, Jan. 2011.

[38] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 289–302, Feb. 2014.

[39] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3499–3506.

[40] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[41] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[42] J. Lu, Y.-P. Tan, and G. Wang, "Discriminative multimanifold analysis for face recognition from a single training sample per person," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 39–51, Jan. 2013.

[43] J. Lu, G. Wang, W. Deng, and P. Moulin, "Simultaneous feature and dictionary learning for image set based face recognition," in *Proc. 13th ECCV*, 2014, pp. 265–280.

[44] D. Maturana, D. Mery, and A. Soto, "Face recognition with decision tree-based local binary patterns," in *Proc. 10th ACCV*, 2010, pp. 618–629.

[45] D. Maturana, D. Mery, and A. Soto, "Learning discriminative local binary patterns for face recognition," in *Proc. IEEE Int. Conf. FG*, Mar. 2011, pp. 470–475.

[46] H. Méndez-Vázquez, Y. Martinez-Diaz, and Z. Chai, "Volume structured ordinal features with background similarity measure for video face recognition," in *Proc. ICB*, Jun. 2013, pp. 1–6.

[47] X. Meng, S. Shan, X. Chen, and W. Gao, "Local visual primitives (LVP) for face modelling and recognition," in *Proc. 18th ICPR*, 2006, pp. 536–539.

[48] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 93–104, 2008.

[49] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. 10th ACCV*, 2010, pp. 709–720.

[50] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE CVPR*, Jun. 2012, pp. 3258–3265.

[51] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[52] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. 28th ICML*, 2011, pp. 833–840.

[53] W. R. Schwartz, H. Guo, J. Choi, and L. S. Davis, "Face identification using large feature sets," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2245–2255, Apr. 2012.

[54] H. J. Seo and P. Milanfar, "Face verification using the LARK representation," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 4, pp. 1275–1286, Dec. 2011.

[55] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE CVPR*, Jun. 2012, pp. 2160–2167.

[56] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014, pp. 1988–1996.

[57] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE CVPR*, Jun. 2014, pp. 1891–1898.

[58] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognit.*, vol. 39, no. 9, pp. 1725–1745, 2006.

[59] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[60] R. Verschae, J. Ruiz-Del-Solar, and M. Correa, "Face recognition in unconstrained environments: A comparative study," in *Proc. ECCVW*, 2008, pp. 1–12.

[61] N.-S. Vu, "Exploring patterns of gradient orientations and magnitudes for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 2, pp. 295–304, Feb. 2013.

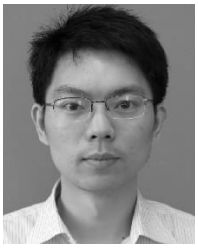
[62] N.-S. Vu and A. Caplier, "Mining patterns of orientations and magnitudes for face recognition," in *Proc. IJCB*, Oct. 2011, pp. 1–8.
[63] N.-S. Vu and A. Caplier, "Enhanced patterns of oriented edge magnitudes for face recognition and image matching," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1352–1365, Mar. 2012.
[64] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. IEEE CVPR*, Jun. 2011, pp. 529–534.
[65] L. Wolf and N. Levy, "The SVM-minus similarity score for video face recognition," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3523–3530.
[66] S. Xie, S. Shan, X. Chen, X. Meng, and W. Gao, "Learned local Gabor patterns for face representation and recognition," *Signal Process.*, vol. 89, no. 12, pp. 2333–2344, 2009.
[67] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3539–3545.
[68] B. Zhang, Y. Gao, S. Zhao, and J. Liu, "Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 533–544, Feb. 2010.
[69] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of Gabor phase patterns (HGPP): A novel object representation approach for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 57–68, Jan. 2007.
[70] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," in *Proc. 10th IEEE ICCV*, Oct. 2005, pp. 786–791.

Jiwen Lu (S'10–M'11) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore. He is currently a Research Scientist with the Advanced Digital Sciences Center, Singapore. His research interests include computer vision, pattern recognition, and machine learning.

He has authored or coauthored over 100 scientific papers in his research areas, where more than 30 papers are in the IEEE TRANSACTIONS journals (IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE/IEEE TRANSACTIONS ON IMAGE PROCESSING/ IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY/IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY) and the top-tier computer vision conferences (ICCV/CVPR/ECCV). He serves as an Area Chair of the 2015 IEEE International Conference on Multimedia and Expo (ICME), and the 2015 IAPR/IEEE International Conference on Biometrics, and the Special Session Chair of the 2015 IEEE Conference on Visual Communications and Image Processing.

Dr. Lu was a recipient of the First Prize National Scholarship and the National Outstanding Student Award from the Ministry of Education of China in 2002 and 2003, the Best Student Paper Award from PREMIA, Singapore, in 2012, and the Top 10% Best Paper Award from MMSP in 2014. Recently, he gives tutorials at some conferences, such as CVPR 2015, FG 2015, ACCV 2014, ICME 2014, and IJCB 2014.

Venice Erin Liong received the B.S. degree from the University of the Philippines Diliman, Quezon City, Philippines, in 2010, and the M.S. degree from the Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 2013. She is currently a Research Engineer with the Advanced Digital Sciences Center, Singapore. Her research interests include computer vision, pattern recognition, and machine learning.



Gang Wang (M'11) received the B.S. degree in electrical engineering from the Harbin Institute of Technology, in 2005, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, in 2010. He is currently an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, and a Research Scientist with the Advanced Digital Sciences Center, Singapore. In particular, he is focusing on object recognition, scene analysis, large-scale machine learning, and deep learning. His research interests include computer vision and machine learning.

Pierre Moulin (F'03) received the Ph.D. degree from Washington University in St. Louis, in 1990, after which he joined Bell Communications Research, Morristown, NJ, as a Research Scientist. In 1996, he joined the University of Illinois at Urbana-Champaign (UIUC), where he is currently a Professor with the Department of Electrical and Computer Engineering, a Research Professor with the Beckman Institute and the Coordinated Science Laboratory, and an Affiliate Professor with the Department of Statistics. His fields of professional interest include image and video processing, compression, statistical signal processing and modeling, media security, decision theory, and information theory.

He was a recipient of the Career Award from the National Science Foundation in 1997 and the IEEE Signal Processing Society Senior Best Paper Award in 1997. He has coauthored (with J. Liu) a paper that received the IEEE Signal Processing Society Young Author Best Paper Award in 2002. He was the Beckman Associate with the UIUC's Center for Advanced Study. From 2007 to 2009, he was a Sony Faculty Scholar with UIUC. He was a Plenary Speaker for ICASSP 2006, ICIP 2011, and several other conferences. He was the Distinguished Lecturer of the IEEE Signal Processing Society from 2012 to 2013.

Dr. Moulin has served on the Editorial Boards of the IEEE TRANSACTIONS ON INFORMATION THEORY, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the PROCEEDINGS OF THE IEEE. He serves on the Editorial Boards of *Foundations and Trends in Signal Processing*. He was the cofounding Editor-in-Chief of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (2005–2008), and a member of the IEEE Signal Processing Society Board of Governors (2005–2007). He has served for the IEEE in various other capacities.