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. Motivated by the fact that different face regions have different structures and the difference of feature representations of neighboring regions is smaller than that of non-neighboring regions, we propose learning multiple yet related feature projection matrices for different face regions simultaneously, so that position-specific discriminative information is 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 joint feature learning 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 datasets 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.

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 $l$ ($l < T$) latent bases. For each face image, we divide it into $T$ non-overlapped regions and learn feature representation for each region jointly. Let $W_t$ be the feature projection matrix for the $t$th region, $1 \leq t \leq T$. Then, the learned feature of the $t$th region is represented as $W_t x$, where $x$ is a sample in this region, which consists of raw pixel values from a patch. We assume that there are $l$ ($l < T$) latent bases to form a dictionary $D = [d_1, d_2, \ldots, d_l] \in \mathbb{R}^{kn \times l}$ which is shared by different regions, and $W_t$ is represented as $W_t = \text{mat}(D\alpha_t)$, where $\alpha_t \in \mathbb{R}^l$ is the representation coefficient vector for the $t$th region, and $\text{mat}(\alpha)$ is the matrix form of $\alpha$. We assume that $\alpha_t$ is sparse such that only a few of the latent bases in $D$ are selected to represent $W_t$. To achieve this, we formulate the following optimization
The first term in (1) is to learn a feature projection matrix \( \mathbf{D} \alpha_t \) to extract sparse features for samples in the \( t \)-th 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), \( \mathbf{A} = [\alpha_1, \alpha_2, \cdots, \alpha_T] \), \( \gamma_1 \) is a parameter to balance the importance of the two terms, \( \xi \) is a parameter to enforce the sparsity of \( \alpha_t \), \( \text{vec}(\mathbf{X}) \) is the vectorization of the matrix \( \mathbf{X} \). \( \mathbf{W}_t^0 \) is the initialized feature projection matrix for the \( t \)-th region, which is learned individually for the \( t \)-th region by using RICA [34] in (2). \( F_t \) is the objective function of the conventional RICA feature learning method for the \( t \)-th region, which is defined as the following unconstrained optimization problem:

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

(2)

where \( \{\mathbf{x}_{it}\}_{i=1}^{m} \in \mathbb{R}^n \) is the unlabeled data set in the \( t \)-th region, \( h \) is a nonlinear convex function, such as \( h(c) = \log(\cosh(c)) \) [32]. \( \mathbf{W}_t \in \mathbb{R}^{k \times n} \) is the feature weighting matrix, \( k \) is the number of features, and \( \mathbf{W}_{ij} \) is the \( j \)-th row feature in \( \mathbf{W}_t \), \( \mathbf{W}_t \mathbf{x}_{it} \) is the learned feature of \( \mathbf{x}_{it} \). \( \lambda \) 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:

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

(3)

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

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

(4)

where \( P_t \) and \( P_{t'} \) denote the \( t \)-th and \( t' \)-th 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 \( \sigma \) 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.

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

\[
\min_{\mathbf{D}, \mathbf{A}} H(\mathbf{D}, \mathbf{A}) = \sum_{t=1}^{T} F_t(\text{mat}(\mathbf{D} \alpha_t)) + \gamma_1 \left( \sum_{t=1}^{T} \| \text{vec}(\mathbf{W}_t^0) - \mathbf{D} \alpha_t \|_2^2 + \xi \| \alpha_t \|_1 \right) + \gamma_2 \sum_{t=1}^{T} \sum_{t'=1}^{T} \| \alpha_t - \alpha_{t'} \|_2^2 \mathbf{S}_{tt'}
\]

subject to: \( \| \mathbf{d}_j \|^2 \leq 1, \quad 1 \leq j \leq l. \)

(5)

where \( \gamma_2 \) 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 \( \alpha_t \), hence we solve \( n \) independent minimization problems:

\[
\min_{\alpha_t} H(\alpha_t) = G(\alpha_t) + \xi_1 \| \alpha_t \|_1, \quad 1 \leq t \leq n.
\]

(6)

where

\[
G(\alpha_t) = F_t(\text{mat}(\mathbf{D} \alpha_t)) + \gamma_1 \| \text{vec}(\mathbf{W}_t^0) - \mathbf{D} \alpha_t \|_2^2
\]

\[
- \gamma_2 \sum_{t'=1}^{T} \text{tr}(\alpha_t^T \alpha_{t'} \mathbf{S}_{tt'})
\]

\[
- \gamma_2 \sum_{t=1}^{T} \text{tr}(\alpha_t^T \alpha_{t'} \mathbf{S}_{tt'})
\]

(7)

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

For fixed \( \mathbf{A} \), we update the dictionary \( \mathbf{D} \) by optimizing
each region, we sample many small patches for feature learning. We perform pooling on the learned features, we apply an unsupervised clustering method to learn a set of codebooks (one for each 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.

\(^2\)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.

---

**Algorithm 1: JFL**

**Input:** Training set: \(\{x_t\}_{t=1}^{n_t}, 1 \leq t \leq T\), dictionary size \(l\), iteration number \(R\), and convergence error \(\epsilon\).

**Output:** Dictionary \(D\) and coefficient matrix \(A\).

**Step 1 (Initialization):**

1.1. Learn \(W^1_{\gamma}\) by applying RICA on image patches in the \(\gamma\)th region, \(\forall \gamma\).
1.2. Initialize \(W^0 = [\text{vec}(W^1_{\gamma})]_t^T\).
1.3. Compute covariance matrix \(S_W = (W^0)^T W^0\).
1.4. Perform SVD on \(S_W = U \Sigma V^T\).
1.5. Initialize \(D\) to be the first \(l\) columns of \(U\).
1.6. Initialize \(A\) with a nonnegative random matrix.

**Step 2 (Local optimization):**

For \(r = 1, 2, \ldots, R\), repeat
2.1. Update \(A\) by solving (6).
2.2. Update \(D\) by solving (8).
2.3. If \(r > 2\) and \(|D^r - D^{r-1}| < \epsilon\), go to Step 3.

**Step 3 (Output the dictionary and coefficient matrix):**

Output \(D^*\) and \(A^*\).

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\(^3\). 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.

\(^3\)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 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 \( f_{11}, \cdots, f_{1m} \)) as the input to the first layer of the network. Having been filtered by the first layer of the basic feature weighting matrix \( (W_{11}) \), they are mapped to \( b_{11}, \cdots, b_{1m} \). 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 \( (W_{12}) \) to project it to a feature vector \( c_1 \). Similarly, we can obtain the outputs \( b_{21}, b_{22}, \cdots, b_{2m}, b_{31}, b_{32}, \cdots, b_{3m}, b_{41}, \cdots, b_{4m} \) at the first layer, and \( c_2, c_3, \) and \( c_4 \) 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 \( d_1, d_2, d_3 \) and \( d_4 \) are histogram features for the first layer, and \( e_1, e_2, e_3, \) and \( e_4 \) 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 \( 8 \times 8 \) 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 \( 7 \times 7 \) and \( 9 \times 9 \), and they were densely sampled from the image with an spacing of 1 and 1 pixels, respectively. Hence, there are nine \( 7 \times 7 \) small patches within each \( 9 \times 9 \) patch. For each \( 7 \times 7 \) small patch, we learn a \( 49 \times 100 \) projection matrix to map it into a 100 over-complete feature vector. The combined output of all the small patches in each \( 9 \times 9 \) 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 \( 100 \times 200 \) projection matrix to map it into a 200 over-complete feature vector. We apply WPCA to project these over-complete features extracted from the \( 7 \times 7 \) and \( 9 \times 9 \) 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 \( 128 \times 128 \) pixels. The parameters \( \gamma_1, \gamma_2 \) and \( \xi \) 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.

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.

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.

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. 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 $\gamma_1$, $\gamma_2$ and $\xi$ 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 ($\gamma_1$, $\gamma_2$), and $\xi$ on the training set of FERET, respectively. We see that JFL achieves the best recognition performance when $\gamma_1$, $\gamma_2$ and $\xi$ 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
TABLE I
RANK-ONE RECOGNITION RATES (%) COMPARISON WITH THE STATE-OF-THE-ART FACIAL DESCRIPTORS TESTED WITH THE STANDARD FERET EVALUATION PROTOCOL.

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

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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Expression</th>
<th>Accessory</th>
<th>Lighting</th>
<th>year</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP [1]</td>
<td>97.0</td>
<td>89.0</td>
<td>29.0</td>
<td>2004</td>
</tr>
<tr>
<td>LGBP [70]</td>
<td>95.0</td>
<td>87.0</td>
<td>51.0</td>
<td>2005</td>
</tr>
<tr>
<td>LVP [47]</td>
<td>96.0</td>
<td>86.0</td>
<td>29.0</td>
<td>2006</td>
</tr>
<tr>
<td>HGDP [69]</td>
<td>96.0</td>
<td>92.0</td>
<td>62.0</td>
<td>2007</td>
</tr>
<tr>
<td>LLGP [66]</td>
<td>98.0</td>
<td>92.0</td>
<td>41.0</td>
<td>2009</td>
</tr>
<tr>
<td>DT-LBP [44]</td>
<td>99.0</td>
<td>92.0</td>
<td>41.0</td>
<td>2011</td>
</tr>
<tr>
<td>DFD [38]</td>
<td>99.6</td>
<td>96.9</td>
<td>63.9</td>
<td>2014</td>
</tr>
<tr>
<td>JFL</td>
<td>99.7</td>
<td>97.2</td>
<td>67.4</td>
<td>2014</td>
</tr>
</tbody>
</table>

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

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 on 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 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.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP [60]</td>
<td>75.4</td>
</tr>
<tr>
<td>SIFT [60]</td>
<td>54.07</td>
</tr>
<tr>
<td>LARK [54]</td>
<td>78.30</td>
</tr>
<tr>
<td>LHS [55]</td>
<td>81.07</td>
</tr>
<tr>
<td>PAF [67]</td>
<td>94.05</td>
</tr>
<tr>
<td>MRF-MLBP [2]</td>
<td>89.94</td>
</tr>
<tr>
<td>DFD [38]</td>
<td>83.07</td>
</tr>
<tr>
<td>JFL</td>
<td>91.03</td>
</tr>
</tbody>
</table>

*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.

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

In our experiments, we followed the standard evaluations in pose, illumination, and expression in each video. Each video clip is about 180 frames. There are large variations in pose, illumination, and expression in each video. To further reduce the feature dimension into 700 for each aligned face image. We applied the discriminative deep metric learning (DDML) [27] to learn a 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.

**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.

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

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

### 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 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 $128 \times 128$ according to the

**Table V**

Verification performance (mean accuracy ± standard error %) comparison with the state-of-the-art results on the YTF dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy ± Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP+MBGS [64]</td>
<td>76.4 ± 1.8</td>
</tr>
<tr>
<td>APEM (fusion) [39]</td>
<td>79.1 ± 1.5</td>
</tr>
<tr>
<td>STFRD+PMML [12]</td>
<td>79.5 ± 2.0</td>
</tr>
<tr>
<td>MBGS+SVM [65]</td>
<td>78.9 ± 1.9</td>
</tr>
<tr>
<td>VSOF+OSS (Adaboost) [46]</td>
<td>79.7 ± 1.8</td>
</tr>
<tr>
<td>DDML(LBP) [27]</td>
<td>81.3 ± 1.6</td>
</tr>
<tr>
<td>DDML(combined) [27]</td>
<td>82.3 ± 1.5</td>
</tr>
<tr>
<td>JFL</td>
<td>81.9 ± 0.9</td>
</tr>
<tr>
<td>JFL (combined)</td>
<td>83.9 ± 0.9</td>
</tr>
</tbody>
</table>

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

### Table V

<table>
<thead>
<tr>
<th>Method</th>
<th>Verification rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRPCA [7]</td>
<td>10.0</td>
</tr>
<tr>
<td>LBP [1]</td>
<td>25.1</td>
</tr>
<tr>
<td>SIFT [41]</td>
<td>23.2</td>
</tr>
<tr>
<td>DFD [38]</td>
<td>30.6</td>
</tr>
<tr>
<td>JFL</td>
<td>32.6</td>
</tr>
</tbody>
</table>

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 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 be seen, JFL significantly outperforms the other four compared methods, as the improvement of the verification rate is at least 2.0%.

### Table VI

**Verification rate (%) at the 1.0% FAR of different methods on the PaSC dataset.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Verification rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRPCA [7]</td>
<td>10.0</td>
</tr>
<tr>
<td>LBP [1]</td>
<td>25.1</td>
</tr>
<tr>
<td>SIFT [41]</td>
<td>23.2</td>
</tr>
<tr>
<td>DFD [38]</td>
<td>30.6</td>
</tr>
<tr>
<td>JFL</td>
<td>32.6</td>
</tr>
</tbody>
</table>

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

---

\(^5\)Available at: [http://www.cs.tau.ac.il/~wolf/ytfaces/](http://www.cs.tau.ac.il/~wolf/ytfaces/).
TABLE VII
RECOGNITION RATES (%) OF THE JOINT AND INDIVIDUAL FEATURE LEARNING METHODS ON DIFFERENT FACE DATASETS.

<table>
<thead>
<tr>
<th>Method</th>
<th>FERET</th>
<th>CAS-PEAL-R1</th>
<th>LFW</th>
<th>YTF</th>
<th>PaSC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ia</td>
<td>fb</td>
<td>dup1</td>
<td>dup2</td>
<td></td>
</tr>
<tr>
<td>IFL</td>
<td>99.4%</td>
<td>99.8%</td>
<td>92.1%</td>
<td>92.5%</td>
<td></td>
</tr>
<tr>
<td>JFL</td>
<td>99.9%</td>
<td>100.0%</td>
<td>93.7%</td>
<td>93.6%</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>FERET</th>
<th>CAS-PEAL-R1</th>
<th>LFW</th>
<th>YTF</th>
<th>PaSC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ia</td>
<td>fb</td>
<td>dup1</td>
<td>dup2</td>
<td></td>
</tr>
<tr>
<td>JFL0</td>
<td>99.1%</td>
<td>99.4%</td>
<td>91.8%</td>
<td>92.2%</td>
<td></td>
</tr>
<tr>
<td>JFL</td>
<td>99.9%</td>
<td>100.0%</td>
<td>93.7%</td>
<td>93.6%</td>
<td></td>
</tr>
</tbody>
</table>

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

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

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.

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 $\gamma_2$ 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 be 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 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 outperforms 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.

### TABLE X

<table>
<thead>
<tr>
<th>Method</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP</td>
<td>0.27</td>
</tr>
<tr>
<td>TPLBP</td>
<td>0.31</td>
</tr>
<tr>
<td>FPLBP</td>
<td>0.33</td>
</tr>
<tr>
<td>LARK</td>
<td>0.37</td>
</tr>
<tr>
<td>DFD</td>
<td>1.51</td>
</tr>
<tr>
<td>JFL</td>
<td>0.28</td>
</tr>
</tbody>
</table>

### REFERENCES


Gang Wang received the B.S. degree from Harbin Institute of Technology in electrical engineering in 2005 and PhD degree in the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign in 2010. He is currently an Assistant Professor in the School of Electrical and Electronic Engineering, Nanyang Technological University, and a research scientist at the Advanced Digital Sciences Center, Singapore. His research interests include computer vision and machine learning. Particularly, he is focusing on object recognition, scene analysis, large scale machine learning, and deep learning. He is a member of IEEE.

Pierre Moulin received his doctoral degree from Washington University in St. Louis in 1990, after which he joined at Bell Communications Research in Morristown, New Jersey, as a Research Scientist. In 1996, he joined the University of Illinois at Urbana-Champaign, where he is currently Professor in the Department of Electrical and Computer Engineering, Research Professor at the Beckman Institute and the Coordinated Science Laboratory, and affiliate professor in 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.

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 IEEE. He currently serves on the editorial boards of Foundations and Trends in Signal Processing. He was co-founding Editor-in-Chief of the IEEE Transactions on Information Forensics and Security (2005-2008), member of the IEEE Signal Processing Society Board of Governors (2005-2007), and has served IEEE in various other capacities.

He received a 1997 Career award from the National Science Foundation and an IEEE Signal Processing Society 1997 Senior Best Paper award. He is also co-author (with Juan Liu) of a paper that received an IEEE Signal Processing Society 2002 Young Author Best Paper award. In 2003 he became IEEE Fellow and Beckman Associate of UIUC’s Center for Advanced Study. In 2007-2009 he was Sony Faculty scholar at UIUC. He was plenary speaker for ICASSP 2006, ICIP 2011, and several other conferences. He is Distinguished Lecturer of the IEEE Signal Processing Society for 2012-2013.