Learning Discriminant Face Descriptor

Abstract—Local feature descriptor is an important module for face recognition and those like Gabor and local binary patterns (LBP) have proven effective face descriptors. Traditionally, the form of such local descriptors is predefined in a handcrafted way. In this paper, we propose a method to learn a discriminant face descriptor (DFD) in a data-driven way. The idea is to learn the most discriminant local features that minimize the difference of the features between images of the same person and maximize that between images from different people. In particular, we propose to enhance the discriminative ability of face representation in three aspects. First, the discriminant image filters are learned. Second, the optimal neighborhood sampling strategy is soft determined. Third, the dominant patterns are statistically constructed. Discriminative learning is incorporated to extract effective and robust features. We further apply the proposed method to the heterogeneous (cross-modality) face recognition problem and learn DFD in a coupled way (coupled DFD or C-DFD) to reduce the gap between features of heterogeneous face images to improve the performance of this challenging problem. Extensive experiments on FERET, CAS-PEAL-R1, LFW, and HFB face databases validate the effectiveness of the proposed DFD learning on both homogeneous and heterogeneous face recognition problems. The DFD improves POEM and LQP by about 4.5 percent on LFW database and the C-DFD enhances the heterogeneous face recognition performance of LBP by over 25 percent.

Index Terms—Face recognition, discriminant face descriptor, image filter learning, discriminant learning, heterogeneous face recognition

1 Introduction

Face recognition has attracted much attention due to the potential value for practical applications and its theoretical challenges. As a classical pattern recognition problem, it mainly involves two critical problems—feature representation and classifier construction. Most of the existing works focus on these two aspects to enhance the face recognition performance.

In many real applications, face recognition is a multiclass classification problem with uncertain class number. For example, in a face identification system, the number of face classes equals to the number of registered subjects. When a subject is added, the number of classes is changed. This is a characteristic of face recognition different from the general object recognition problems, where the number of classes is usually fixed. The face recognition algorithms are required to adapt to the variation of class numbers. Therefore, classification mechanisms successfully applied to general object recognition may not be applicable to face recognition.

Among various classification methods, the nearest neighbors (NN) classifier and its variants, i.e., nearest feature line [1] or nearest subspace [2], are the most popular methods in face recognition. In [3], Moghaddam et al. convert the face recognition into a two-class classification problem by constructing the intra and inter-face spaces. The intraspace is the difference between two images from the same person and the interspace is the difference between two images from different people. In this way, many two-class classifiers like Bayesian, SVM [4], Adaboost [5], and so on can be applied. Recently, Ma et al. [6] propose a sparse representation classifier (SRC), which formulates the probe image as a linear combination of gallery images. The combination coefficients corresponding to images from the same subject are set to be larger than others, and hence, the probe image can be recognized. For other works on classifier learning, please refer to [7].

Besides the classifier learning, feature representation is another important problem in face recognition. The face images in real world are affected by expressions, poses, occlusions, and illuminations; the difference of face images from the same person could be even larger than that from different ones. Therefore, how to extract robust and discriminant features that make the intraspace compact and enlarge the margin between different people is a critical and difficult problem in face recognition.

Up to now, many face representation approaches have been introduced, including subspace-based holistic features and local appearance features [8], [9]. Typical holistic features include the well-known principal component analysis (PCA) [10], linear discriminant analysis (LDA) [11], independent component analysis (ICA) [12], and so on. PCA provides an optimal linear transformation from the original image space to an orthogonal eigenspace with reduced dimensionality in sense of the least mean square reconstruction error. LDA seeks a linear transformation by maximizing the ratio of between-class variance and within-class variance. ICA is a generalization of PCA, which is

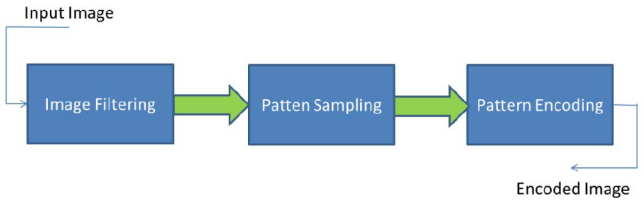

Fig. 1. Three-step way to extract LBP-like feature.

sensitive to the high-order correlation among the image pixels. Yan et al. [13] reinterpret the subspace learning from the view of graph embedding so that various methods, such as PCA, LDA, LPP [14], and so on, can all be interpreted under this framework. Recently, many advanced subspace analysis methods [15], [16] that address the problem of small sample size (SSS) are proposed to enhance the discriminative ability of the learned subspace.

Local appearance features, as opposed to holistic features like PCA and LDA, have certain advantages. They are more stable to local changes such as expression, occlusion, and inaccurate alignment. Gabor [17], [18] and local binary patterns (LBP) [19] are two representative features. Gabor wavelets capture the local structure corresponding to specific spatial frequency (scale), spatial locality, and selective orientation. It has been demonstrated to be discriminative and robust to illumination and expression changes. Local binary patterns that describe the neighboring changes around the central point are a simple yet effective way to represent faces. It is invariant to monotone transformation and is robust to illumination changes to some extent. The combination of Gabor and LBP further improves the face recognition performance. A lot of work has been proposed in this branch [20], [21], [22].

Recently, Kumar et al. [23] have proposed attribute and simile representations for face recognition. The attribute means the describable aspects of visual appearance (like gender, race, and age) and the binary classifiers are used to recognize them. For simile, the test image (or regions of the image) is compared to the images (or regions) in a reference set and the similarity between them is used as the face representation. The attribute and simile are finally combined to form a compact face description. Berg and Belhumeur [24] build a large and diverse collection of "Tom-versus-Pete" classifiers to extract discriminative attributes to represent face images. The attribution-based representation is shown to be effective and robust to face recognition in the real world.

In this work, we focus on LBP-like feature extraction and propose a novel discriminant face descriptor (DFD) that introduces the discriminant learning into feature extraction process.

1.1 Related Work

Generally, the LBP-like feature extraction can be decomposed into three steps (Fig. 1). First, an image filter is applied to reduce the noise affection and enhance the useful information. Second, certain pixel patterns on the filtered image are sampled and compared. Third, the encoded image is derived based on the pixel comparison results and encoding rules. In original LBP [19], the first filtering step is skipped and the LBP feature is extracted from the original image directly. The neighboring pixel values are compared with the central point and the LBP feature is encoded with a uniform pattern definition.

Many of the LBP variants [25] can be categorized to improve the original LBP at these three steps. In MBLBP [26], multiscale mean filters are applied at the first step, followed with the similar operation of LBP. In LGBP [20], HGPP [21], GV-LBP [22], a bank of Gabor filters with different scales and orientations are first applied and the local pattern is encoded from the Gabor magnitude/phase responses. Sobel-LBP [27] first extracts the gradient information from the original image and then LBP operator is applied to the gradient response images. In these methods, all the image filters are defined in a hand-crafted way.

There are also other variants focusing on the optimal neighborhood sampling and the encoder learning. Cao et al. [28] utilize unsupervised methods (random-projection tree and PCA tree) to learn the encoder and the PCA dimension reduction method is applied to get a compact face descriptor. Guo et al. [29] propose a supervised learning approach based on Fisher separation criterion to learn the encoder of LBP. The authors of [30] propose to construct a decision tree for each region to encode the pixel comparison result and in [31], a heuristic algorithm is used to find the optimal pixel comparison pairs for discriminative face representation. In local quantized patterns (LQP) [32], researchers adopt vector quantization to encode the local binary/ternary pattern values. TP-LBP and FP-LBP [33] adopt a specific sampling way to encode the relationship of patch difference. In object recognition, a number of encoding methods like sparse coding [34] and discriminant dictionary learning [35] are proposed to extract robust features.

As summarized, most LBP variants try to improve the ordinary LBP at one of the three steps. Most learning-based descriptors [28], [30], [31] focus on the improvement at the second or the third step. There is little work on image filter learning for feature extraction. The most related work is the "Volterrafaces" [36] in which various "Volterra" kernels are learned and a vote mechanism is adopted for face recognition.

1.2 Our Contribution

The proposed discriminant face descriptor improves the discriminative ability at all three steps of the LBP-like feature extraction. Fig. 2 illustrates the pipeline of the proposed method. First, it learns a discriminative image filter to enhance the effectiveness of the descriptor. Second, it adopts a soft way to determine the optimal neighborhood sampling strategy to best differentiate face images. Third, it learns dominant patterns in an unsupervised way to enhance the representative ability of the descriptor. By incorporating these improvements at three steps, a discriminant face descriptor is constructed. In the testing phase, after we encode the face image with the learned DFD, histogram features that describe the co-occurrence of encoded values are then extracted as the final face representation. Some preliminary results of this work have been published in [37], [38].

The main contributions of this work are summarized below:

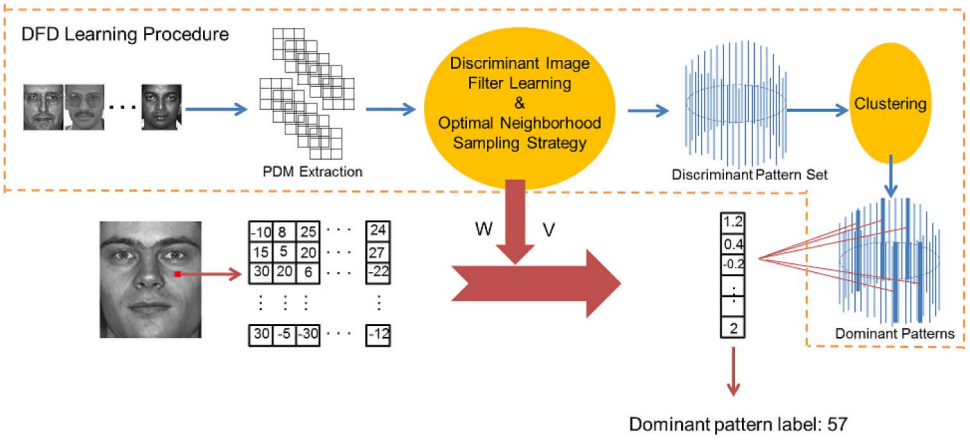

Fig. 2. The pipeline of discriminant face descriptor learning. In the learning phase, after extracting the pixel difference matrix (PDM), the discriminant learning is adopted to learn the discriminant image filters and the optimal neighborhood sampling strategy. The PDM is then projected and regrouped to form the discriminant pattern vector. The dominant patterns are finally obtained by using unsupervised clustering method. In the face labeling phase, for each pixel in face image, the PDM is first extracted and the discriminant pattern vector is then obtained by projecting the PDM using the learned discriminant image filters and the neighborhood sampling strategy. The pixel is finally labeled to the ID of dominant pattern, which is the one most similar to the discriminant pattern vector.

1. A discriminant image filter learning method is proposed. With the learned image filters, more useful face information helpful for face recognition is explored.
2. The optimal neighborhood sampling strategy in LBP-like feature extraction is learned. Different from the previous work, we determine the best sampling method in a soft way, in which a soft sampling matrix (SSM) is learned to differentiate the importance of each neighbor. This soft sampling strategy is more flexible to extract the discriminant face patterns.
3. By incorporating the discriminant image filter and the optimal soft sampling learning, a discriminant face descriptor is proposed with the formulation and solution. Moreover, local DFDs are learned for different parts of faces to improve the discriminative power and obtain more precise image description.
4. The coupled discriminant face descriptors (C-DFD) is proposed to address the heterogeneous face data. The coupled discriminant filters and the optimal soft sampling strategy are learned iteratively to obtain the common discriminant face representation.

2 Discriminant Face Descriptor Learning

The motivation of DFD learning is directly related to achieving high face recognition accuracy, that is, to reduce the intradifference and enlarge the interdifference of face images, so that they can be correctly classified. To achieve this goal, we incorporate the discriminant learning into an LBP-like feature extraction process. Specifically, the discriminant image filters learning and the optimal soft neighborhood sampling are proposed to enhance the essential face patterns and suppress the external variations. In the following, we introduce the formulation of discriminant image filters learning, optimal soft neighborhood sampling strategy, and its optimization to learn an effective DFD from face images. Some abbreviations used in this paper are summarized in Table 1.

2.1 Discriminant Image Filters Learning

Given a face image I, its filtered image is denoted as $f(I)$. In this work, we apply LBP-like operator on filtered image, where the neighboring pixels are compared with the center. For position p, the pixels in neighboring region R^p are grouped as $df(I)^p = [f(I)^{p_1} - f(I)^{p_0}, f(I)^{p_2} - f(I)^{p_0}, \ldots, f(I)^{p_d} - f(I)^{p_0}]$, where $f(I)^{p_0}$ is the pixel value of filtered image at position p and $f(I)^{p_i}$ denotes the pixel value of filtered image at position p_i. $\{p_1, p_2, \ldots, p_d\} \in R^p$ are the neighbors of position p and d is the number of neighbors. The vector $df(I)^p$ is named pixel difference vector (PDV) in the following. Intuitively, the purpose of discriminant image filter learning is to find a filter f so that after the image filtering, the PDVs of images from the same person are similar and the differences of PDVs from different people are enlarged. Following Fisher criterion [5], it can be formulated to maximize the ratio of between-class scatter S'_b to the within-class scatter S'_w. Let $df(I)^p_{ij}$ be the pth PDV of the jth sample from class i; the between-class scatter S'_b and within-class scatters S'_w can be defined as

$$S'_w = \sum_{i=1}^{L} \sum_{j=1}^{C_i} \sum_{p=1}^{N} \left(df(I)^p_{ij} - df(m)^p_i\right)\left(df(I)^p_{ij} - df(m)^p_i\right)^T,$$
$$S'_b = \sum_{i=1}^{L} \sum_{p=1}^{N} C_i \left(df(m)^p_i - df(m)^p\right)\left(df(m)^p_i - df(m)^p\right)^T, \quad (1)$$

TABLE 1
Summary of Some Abbreviations Used in This Paper

Abbrev.	Full Name
PDV	Pixel Difference Vector
PDM	Pixel Difference Matrix
SSM	Soft Sampling Matrix
DPV	Discriminant Pattern Vector
DFD	Discriminant Face Descriptor
C-DFD	Coupled Discriminant Face Descriptor

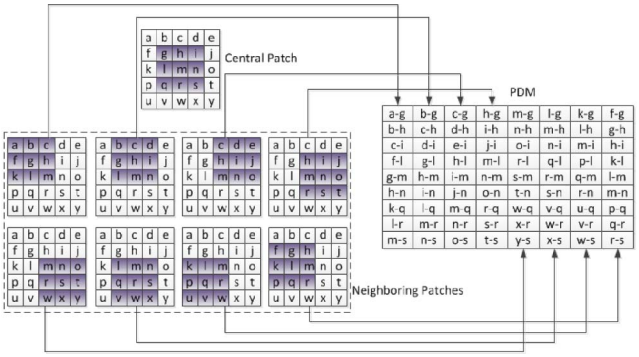

Fig. 3. An example of pixel difference matrix extraction for discriminant face descriptor learning. The image filter size is 3×3 and the neighboring radius is 1. For every central patch, eight neighboring patches are compared, respectively, and then grouped to form the pixel difference matrix.

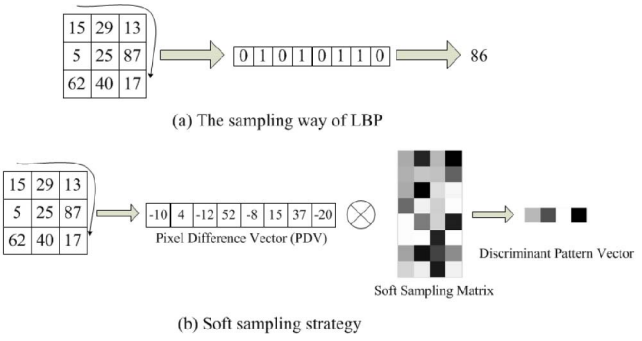

Fig. 4. The difference of the proposed soft neighborhood sampling strategy and the ordinary LBP sampling method. In ordinary LBP, the compared string is binarized and converted into a decimal value (i.e., LBP code). In the proposed method, the extracted pixel difference vector is multiplied with an optimal learned soft sampling matrix to obtain the discriminant pattern vector, which will be further encoded with dominant patterns.

where L is the number of face classes and C_i is the number of samples from the ith class. $df(m)_i^p$ is the mean vector of pth PDVs on filtered images from the ith class and $df(m)^p$ is the total mean vector of pth PDVs over the sample set.

Under linear assumption, suppose the image filter vector to be w, and the value of filtered image at position p can be represented as $f(I)^p = w^T I^p$, where I^p denotes the image patch vector centered at position p. Similarly, the PDV $df(I)^p$ can be represented as $df(I)^p = w^T dI^p$. Substituting it into (1), we get

$$\begin{aligned} S'_w &= \sum_{i=1}^{L}\sum_{j=1}^{C_i}\sum_{p=1}^{N} w^T\left(dI_{ij}^p - dm_i^p\right)\left(dI_{ij}^p - dm_i^p\right)^T w, \\ S'_b &= \sum_{i=1}^{L}\sum_{p=1}^{N} C_i w^T\left(dm_i^p - dm^p\right)\left(dm_i^p - dm^p\right)^T w, \end{aligned} \tag{2}$$

where dI_{ij}^p is pixel difference matrix extracted from the jth image of class i at position p, dm_i^p is the mean PDM for the ith class, and dm^p is total mean PDM at position p. Fig. 3 shows an example of how to extract PDM from each pixel.

2.2 Optimal Neighborhood Sampling Strategy

In ordinary LBP, the neighboring pixels are compared with the center and the resulted binary values are then converted into a decimal value. The neighboring pixels in LBP are treated equally. However, different neighboring pixels could be of different contribution to face description. Careful selection of neighboring pixels may help to improve the face recognition performance. In [31], researchers adopted a heuristical way to find the best pixel sampling pairs in local regions. In this work, we propose a soft way to determine the optimal neighborhood sampling strategy. Different from previous methods, we try to learn a weight matrix corresponding to each PDV, named soft sampling matrix. Each column of SSM is a soft sampling vector, which is a weight assignment for the elements in PDV. The number of column is set empirically to extract sufficient complementary information. Fig. 4 illustrates how the soft neighborhood sampling strategy works with SSM. By multiplying the PDV with its corresponding SSM, the neighboring pixels in PDV are assigned to different weights that reflect the contribution difference. In this way, the pixels helpful to face recognition are enhanced and irrelevant ones are suppressed. The derived vector can be further encoded with dominant patterns. Suppose the SSM as $v = [v_1, v_2, \ldots, v_d]^T$, where d is number of neighboring pixels in local region, after weight combination, the resulted vectors from the same person are supposed to be consistent and those from different people are different. By appropriate formulation, this problem can also be formulated as maximizing the ratio of between-class scatter $S_b^{''}$ and within-class scatter $S_w^{''}$, which are computed as follows:

$$\begin{aligned} S_w^{''} &= \sum_{i=1}^{L}\sum_{j=1}^{C_i}\sum_{p=1}^{N}\left(df(I)_{ij}^p - df(m)_i^p\right)vv^T\left(df(I)_{ij}^p - df(m)_i^p\right)^T, \\ S_b^{''} &= \sum_{i=1}^{L}\sum_{p=1}^{N} C_i\left(df(m)_i^p - df(m)^p\right)vv^T\left(df(m)_i^p - df(m)^p\right)^T. \end{aligned} \tag{3}$$

2.3 Optimization

By combining (2) and (3), the between-class scatter S_b and within-class scatter S_w can be reformulated as

$$\begin{aligned} S_w &= \sum_{i=1}^{L}\sum_{j=1}^{C_i}\sum_{p=1}^{N} w^T\left(dI_{ij}^p - dm_i^p\right)vv^T\left(dI_{ij}^p - dm_i^p\right)^T w, \\ S_b &= \sum_{i=1}^{L}\sum_{p=1}^{N} C_i w^T\left(dm_i^p - dm^p\right)vv^T\left(dm_i^p - dm^p\right)^T w. \end{aligned} \tag{4}$$

Following Fisher criterion, the objective of DFD learning is to find image filter vectors w and soft sampling matrix v, so that the ratio of between-class scatter matrix to the within-class scatter can be maximized. It is easy to find that this formulation is similar to the two-dimensional linear discriminant analysis (2D-LDA) [39], where the PDM is the basic matrix to compute the between and within-class scatter and the left (discriminant image filter) and right projections (soft sampling matrix) are required to be computed. Like 2D-LDA, we solve the above optimization problem in an iterative way. At each iteration, one of the variables w, v is fixed and the optimal solution for another one is derived by solving the generalized eigenvalue problem. As indicated in [39], one loop of iteration is

enough to achieve good performance while reducing the computational cost. The whole algorithm of DFD learning is illustrated in Algorithm 1.

Algorithm 1. Discriminant face descriptor learning algorithm.

Input: A set of sampled patch difference matrices (PDMs) $\{dI_{ij}^p, i = 1, \ldots, L, j = 1, \ldots, C_i, p = 1, \ldots, N\}$, where $dI_{ij} \in R^{d_1 \times d_2}$, d_1 is dimension of image patch and d_2 is the number of neighbors for each PDM. The reduced dimension d_1', d_2' and the number of iteration T are set in advance.

Output: Image filter projections $w \in R^{d_1 \times d_1'}$ and soft sampling matrix $v \in R^{d_2 \times d_2'}$, where d_1' and d_2' are the reduced dimension.

1: **Initialize:** $w = I$ and $v = I$, where I is the identity matrix.

2: **for** $t = 1, \ldots, T$ **do**:

3: a) Compute within and between-class scatter matrices:

$$S_w^1 = \sum_{i=1}^{L} \sum_{j=1}^{C_i} \sum_{p=1}^{N} (dI_{ij}^p - dm_i^p) v v^T (dI_{ij}^p - dm_i^p)^T;$$

$$S_b^1 = \sum_{i=1}^{L} \sum_{p=1}^{N} C_i (dm_i^p - dm^p) v v^T (dm_i^p - dm^p)^T;$$

4: b) Solve the generalized eigenvalue problem and obtain the eigenvectors w_0 with d_1' largest eigenvalues. $S_b^1 w = \lambda S_w^1 w$

5: c) $w \leftarrow w_0$

6: d) Compute within and between-class scatter matrices

$$S_w^2 = \sum_{i=1}^{L} \sum_{j=1}^{C_i} \sum_{p=1}^{N} (dI_{ij}^p - dm_i^p)^T w w^T (dI_{ij}^p - dm_i^p);$$

$$S_b^2 = \sum_{i=1}^{L} \sum_{p=1}^{N} C_i (dm_i^p - dm^p)^T w w^T (dm_i^p - dm^p);$$

7: e) Solve the generalized eigenvalue problem and obtain the eigenvectors v_0 with d_2' largest eigenvalues. $S_b^2 v = \lambda S_w^2 v$

8: f) $v \leftarrow v_0$

9: **end for**

10: **Return:** w and v

2.4 Dominant Patterns Learning

With the learned image filters and soft sampling matrix, the PDM can be projected onto a discriminant subspace. Suppose we finally preserve d_1' image filter vectors and d_2' soft sampling vectors, after left and right projections, the PDM is projected onto a $d_1' \times d_2'$ matrix. This matrix is then transformed into a vector of $d_1' \times d_2'$ dimension, which is named discriminant pattern vector in the following. Here, we simply use the unsupervised clustering method (K-means) to learn the dominant patterns. Recent complex methods like random-tree [28] and the dominant pattern determination method used in [29] were also tried. Because of the discriminant learning in DPV extraction, there is no significant difference in performance among various dominant learning methods and the K-means is adopted due to its simplicity.

2.5 DFD-Based Face Representation

The structures and useful information for local face regions are different. To describe the face image precisely, we propose to learn a local DFD for each local face region, respectively. In feature extraction phase, a face image is first divided into multiple regions. The pixels in each region are encoded according to the locally learned DFD (including discriminant image filters, optimal soft sampling matrix, and the dominant patterns), respectively. For each pixel, the extracted PDM is projected with the learned image filters and soft sampling matrix. The resultant submatrix is then transformed to form a DPV. Finally, the pixel is labeled with the ID of dominant pattern which is the most similar with the extracted DPV.

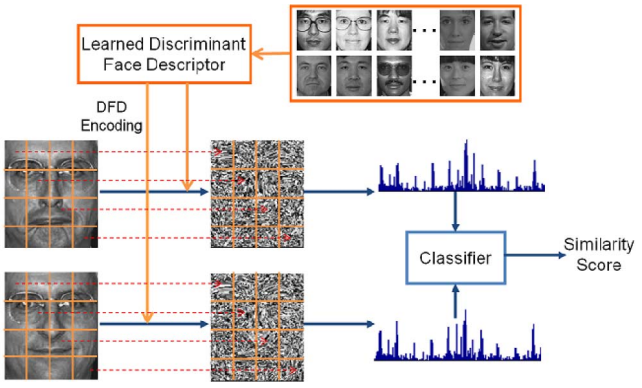

Fig. 5. The process of DFD-based face representation and recognition.

After the face image labeling, histogram-based features in each region that describe the co-occurrence of patterns are extracted and concatenated. Directly matching metrics like L1/L2 distance and histogram intersection can be adopted to measure the dissimilarity of different face images. Dimensionality reduction technique like PCA can also be applied to further improve the matching efficiency. The process of DFD-based face representation is illustrated in Fig. 5.

3 Coupled Discriminant Face Descriptor for Heterogeneous Images

Recently, more and more attention has been paid to heterogeneous face image matching problem. Heterogeneous faces are defined as faces which are captured in different environments or by different devices, for example, visual (VIS) versus near infrared (NIR), VIS versus Sketch, and so on, which are common in many real applications like law enforcement and video surveillance. Previous works mainly focus on transforming the heterogeneous face images into the same modality for matching or developing an advanced classifier that is robust to the modality gap of extracted features.

In this work, we try to reduce the modality gap at the feature level to simplify the heterogeneous face recognition problem while traditional face descriptors could fail to reduce the appearance gap. From the three-step view (Fig. 1), Zhang et al. [40] have proposed a coupled encoding method at the third step to reduce the difference of heterogeneous features. Analogized from the DFD, we propose to learn a coupled discriminant face descriptor by incorporating discriminative learning into the three steps of feature extraction. Specifically, we adopt a coupled image filter pair to model the difference of the images from different modalities. After the coupled image filtering, the responses of heterogeneous images from the same

person are as similar as possible, and therefore, the appearance gap of different modalities is reduced.

As mentioned above, the objective of DFD learning is to extract discriminative features that are robust to image variations. Similar to DFD learning for homogeneous face images, the purpose of C-DFD is to reduce the difference of PDVs for the heterogeneous images from the same person and meanwhile enlarge that from different subjects. Let I^V and I^M be the face images with two modalities (e.g., VIS and NIR modalities) and their filtered images are denoted as $f(I^V)$ and $f(I^M)$, respectively. Suppose $df(I^V)_{ij}^p$ and $df(I^M)_{ij}^p$ are the pth heterogeneous PDVs of jth sample pair from the ith class. Following Fisher criterion, the objective of coupled image filters learning can be formulated to maximize the ratio of between-class scatter and within-class scatter, which can be formulated as

$$\begin{aligned} S_w &= S_w^{VV} + S_w^{MM} + S_w^{VM} + S_w^{MV}, \\ S_b &= S_b^{VV} + S_b^{MM} + S_b^{VM} + S_b^{MV}, \end{aligned} \tag{5}$$

where S_b^{VM}, S_w^{VM} are the between- and within-class matrices between modality V and M, which are defined as

$$\begin{aligned} S_w^{VM} = &\sum_{i=1}^{L}\sum_{j=1}^{C_i}\sum_{p=1}^{N}\big(df(I^V)_{ij}^p - df(m^M)_i^p\big)(df(I^V)_{ij}^p \\ &- df(m^M)_i^p)^T, \\ S_b^{VM} = &\sum_{i=1}^{L}\sum_{p=1}^{N} C_i(df(m^V)_i^p - df(m^M)^p)(df(m^V)_i^p \\ &- df(m^M)^p)^T, \end{aligned} \tag{6}$$

where $df(I^V)_{ij}^p$, $df(I^M)_{ij}^p$, $df(m^V)_i^p$, $df(m^M)_i^p$, $df(m^V)^p$, $df(m^M)^p$ are defined similarly as in Section 2.1 and the superscript ^{V} or ^{M} is the modality indicator. By introducing the optimal soft sampling matrix learning as in Section 2.2, the S_b^{VM} and S_w^{VM} can be reformulated as

$$\begin{aligned} S_w^{VM} = &\sum_{i=1}^{L}\sum_{j=1}^{C_i}\sum_{p=1}^{N}\big(df(I^V)_{ij}^p - df(m^M)_i\big)^p \\ &vv^T\big(df(I^V)_{ij}^p - df(m^M)_i^p\big)^T, \\ S_b^{VM} = &\sum_{i=1}^{L}\sum_{p=1}^{N} C_i\big(df(m^V)_i^p - df(m^M)^p\big) \\ &vv^T\big(df(m^V)_i^p - df(m^M)^p\big)^T. \end{aligned} \tag{7}$$

Differently from Section 2.1, in this part, we learn a couple of discriminant image filters to better deal with the heterogeneous face appearance variation. Under linear assumption, the filtered images $f(I^V)$ and $f(I^M)$ at position p can be formulated as $f(I^V)^p = (w^V)^T I^{Vp}$ and $f(I^M)^p = (w^M)^T I^{Mp}$, respectively, where I^{Vp} and I^{Mp} are original image patch vectors centered at position p for heterogeneous image pair and w^V and w^M are coupled image filter vectors. As in Section 2.1, we obtain the optimal solution of w^M, w^V, v in an iterative way. First, the soft sampling matrix v is fixed, and the (5) can be formulated in the form of $S_w = w^T\sum_{i,j\in\{V,M\}} A_w^{ij} w$ and $S_b = w^T\sum_{i,j\in\{V,M\}} A_b^{ij} w$ (see Appendix A, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.112), where $w = [w^V; w^M]$. The solution w to the problem of maximizing S_b/S_w can be obtained by solving the generalized eigenvalue problem $\sum_{i,j\in\{V,M\}} A_b^{ij} w = \lambda\sum_{i,j\in\{V,M\}} A_w^{ij} w$ with the eigenvectors corresponding to the leading eigenvalues. The coupled discriminative filters w^V and w^M can then be obtained by splitting the w appropriately.

The next step is to learn the optimal v by fixing the w^V, w^M. As we know, the trace of S_w^{VM} and S_b^{VM} is equivalent to the $\tilde{S}_w^{VM}$ and $\tilde{S}_b^{VM}$, where $\tilde{S}_w^{VM}, \tilde{S}_b^{VM}$ are defined as

$$\begin{aligned} \tilde{S}_w^{VM} = &\sum_{i=1}^{L}\sum_{j=1}^{C_i}\sum_{p=1}^{N} v^T\big(df(I^V)_{ij}^p - df(m^M)_i^p\big)^T \\ &(df(I^V)_{ij}^p - df(m^M)_i^p)v = v^T\hat{S}_w^{VM}v, \\ \tilde{S}_b^{VM} = &\sum_{i=1}^{L}\sum_{p=1}^{N} C_i v^T\big(df(m^V)_i^p - df(m^M)^p\big)^T \\ &\big(df(m^V)_i^p - df(m^M)^p\big)v = v^T\hat{S}_b^{VM}v, \end{aligned} \tag{8}$$

where

$$\begin{aligned} \hat{S}_w^{VM} = &\sum_{i=1}^{L}\sum_{j=1}^{C_i}\sum_{p=1}^{N}\big(df(I^V)_{ij}^p - df(m^M)_i^p\big)^T\big(df(I^V)_{ij}^p \\ &- df(m^M)_i^p\big), \\ \hat{S}_b^{VM} = &\sum_{i=1}^{L}\sum_{p=1}^{N} C_i\big(df(m^V)_i^p - df(m^M)^p\big)^T(df(m^V)_i^p \\ &- df(m^M)^p). \end{aligned} \tag{9}$$

By defining $\hat{S}_w = \hat{S}_w^{VV} + \hat{S}_w^{VM} + \hat{S}_w^{MV} + \hat{S}_w^{MM}$, $\hat{S}_b = S_b^{VV} + S_b^{VM} + S_b^{MV} + S_b^{MM}$, we can get the optimal v by solving the generalized eigenvalue problem $\hat{S}_b v = \lambda\hat{S}_w v$ with its leading eigenvalues.

After we obtain the coupled image filters and the soft sampling matrix, the dominant patterns can then be determined by K-means clustering method as introduced in Section 2.5. Similarly to DFD learning, we learn local C-DFDs in practice to model the face image precisely. The face labeling phase is similar to what is described in Section 2.4 by replacing the discriminant image filters with coupled discriminant image filters. The histogram-based features are finally extracted and compared to measure the dissimilarity of different images.

4 Experiments

We compare our DFD with some of state-of-the-art descriptors. For homogeneous face recognition, the FERET [41], CAS-PEAL-R1 [42], and LFW [43] databases are used to evaluate the performance of different methods. For heterogeneous face image matching, we compare the performance of different methods on a publicly available HFB (VIS versus NIR) [44] and a self-collected heterogeneous face database, named HFB-S.

4.1 FERET

The FERET database is one of the largest publicly available databases. The training set contains 1,002 images. In test phase, there are one gallery set with 1,196 images from 1,196 subjects and four probe sets (fb, fc, dup I, and dup II)

Fig. 6. Cropped face examples from the FERET database.

including expression, illumination, and aging variations. All the images are rotated, scaled, and cropped into 150×130 size according to the provided eye coordinates. Some cropped example images are shown in Fig. 6.

4.1.1 Parameter Clarification

Fig. 7 shows the neighboring pixels considered in this work. Note that the neighbor selection is not very critical in our method as long as sufficient neighbors are considered because the optimal soft sampling learning will select the most discriminative neighbors adaptively.

In the following experiments, the images are equally divided into 7×7 nonoverlapping regions. We learn in total $7 \times 7 = 49$ local DFDs for the whole image. Suppose two feature vectors extracted from image i and j to be $H^i = [h_1^i, h_2^i, \ldots, h_N^i]$ and $H^j = [h_1^j, h_2^j, \ldots, h_N^j]$, where h_k^i, h_k^j, $k = 1, 2, \ldots, N$, are histogram features extracted from the kth region. The histogram intersection metric (10) is used to measure the similarity of H^i and H^j. The effect of three parameters, including the size of image filter S ($d_1 = S \times S$), the size of neighborhood region R (shown in Fig. 7, $d_2 = 16$) and the number of dominant patterns K, is examined on the FERET fb probe set. The reduced dimension d_1' and d_2' are set to 5 and 4 empirically in the following experiments:

$$d(H_i, H_j) = \sum_{k=1}^{N} \sum_{m} min\big(h_k^i(m), h_k^j(m)\big), \tag{10}$$

where $h_k^i(m)$ is the mth bin value of histogram h_k^i.

We first set the size of image filter S and the neighborhood region R to 5 and vary the number of dominant patterns K from $\{16, 32, 64, 128, 256, 512, 1{,}024, 2{,}048\}$. Fig. 8 shows the recognition rates with respect to different numbers of dominant patterns. A larger value of K achieves better face recognition performance. However, larger number of dominant patterns also leads to higher dimension of extracted features, which increases the computational cost in feature matching. Considering the tradeoff between the face recognition accuracy and the computational efficiency, we finally set the number of dominant patterns K to 1,024 in the following experiments.

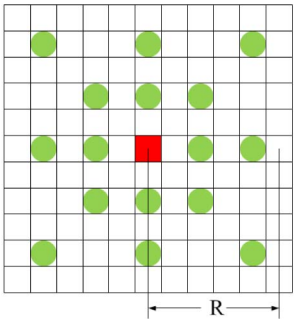

Fig. 7. The neighborhood region and the neighboring pixels considered in this work. The green points (neighboring pixels) are compared with the red one (central pixel) to extract the PDM.

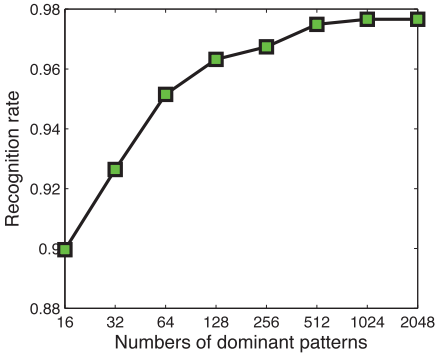

Fig. 8. Recognition rates with respect to different numbers of dominant patterns.

By setting K to 1,024, we further examine the face recognition performance by varying the values of S and R. The values of S and R vary from $\{3, 5, 7, 9, 11\}$ and $\{3, 5, 7, 9\}$, respectively. Table 2 lists the recognition results with different configurations of S and R on the FERET fb probe set. The results show the values of S and R have an effect on the face recognition performance, but not significantly. In the following experiments, to reduce the complexity of the proposed method, we always set S and R to be the same value.

4.1.2 Recognition Results and Discussions

We compare DFD with popular descriptors like LBP, LGBP, LLGP, LQP, and so on. For DFD, three scales of image filters, S = 3, 5, 7 are tested. The DFD is learned from FERET training set. All the methods are tested following the four standard testing protocols (fb, fc, dup I, dup II). There are expression and lighting variations in fb and fc probe sets, respectively. Dup I and dup II probe sets are used to test face recognition performance across aging.

The work in [19], [20], [22] shows that different regions of face images make different contribution to face recognition. As adopted in these methods, we adopt a weighted histogram intersection metric to measure the dissimilarity of two images. Features extracted from the parts like eyes, nose should be assigned with larger weights to emphasize the importance of these regions. Given two feature vectors $H^i = [h_1^i, h_2^i, \ldots, h_N^i]$ and $H^j = [h_1^j, h_2^j, \ldots, h_N^j]$, where h_k^i, h_k^j, $k = 1, 2, \ldots, N$, are histogram features extracted from the

TABLE 2
Face Recognition Rates (Percent) with Different Scales of Image Filters and Neighborhood Radius on FERET fb Probe Set

	R=3	R=5	R=7	R=9
S=3	96.9	97.8	97.2	97.2
S=5	97.4	97.7	98.0	97.7
S=7	97.9	98.1	98.2	98.0
S=9	98.1	98.2	98.2	97.7
S=11	98.3	98.0	98.1	97.7

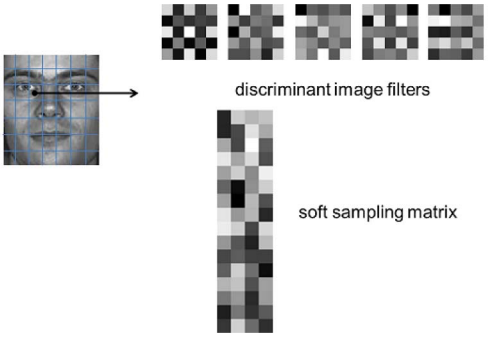

Fig. 9. Illustration of learned discriminant image filters and soft sampling matrix from local patch. Five image filters with the size of 5×5 are learned and the dimension of soft sampling matrix is $d_2 \times d_2' = 16 \times 4$ corresponding to 16 neighbors shown in Fig. 7.

kth region. The weighted histogram intersection metric is defined as

$$d(H_i, H_j) = \sum_{k}^{N} \sum_{m} w_k \cdot min\big(h_k^i(m), h_k^j(m)\big), \qquad (11)$$

where w_k is the weight for the kth region, which is learned following Fisher criterion as in [18] from the training set.

We also apply whitened PCA (WPCA), followed with cosine metric to measure the dissimilarity of different face images. As adopted in [45], [32], WPCA is conducted on the gallery set only.

Fig. 9 illustrates an example of discriminant image filters and the soft sampling matrix learned from local patch. Five discriminant and complementary image filters and the optimal SSM are learned to explore the discriminant and sufficient information for face recognition.

Table 3 lists the face recognition performance of proposed DFD, compared with state-of-the-art descriptors. The results indicate that

1. A number of LBP variants improve the face recognition performance of ordinary LBP. The combination of Gabor and LBP response is an effective way to enhance performance.
2. The learning-based methods, like DT-LBP, DLBP, and the proposed DFD, achieve higher recognition rates than the combination of Gabor and LBP, especially on dup I and II probe sets, where DT-LBP, DLBP, and DFD outperform LGBP by over 10 percent. It indicates that the learning-based descriptor is able to extract more discriminative and proper information for face recognition than the hand-crafted ones.
3. With weighted histogram intersection metric, DFD achieves recognition rates comparable to DT-LBP and DLBP, and outperforms recently proposed POEM and LQP methods. The DFD achieves significantly higher recognition rates than POEM and LQP on dup I and II probe sets, indicating that DFD is more robust to aging variation than POEM and LQP.
4. Regarding the scale of DFD, the recognition performance of three DFDs ($S = 3, 5, 7$) on fb and fc probe set are similar. On dup I and II sets, the DFD ($S = 5$) achieves higher face recognition accuracy than other two DFDs ($S = 3$ and $S = 7$).
5. With WPCA and cosine metric, the proposed DFD ($S = 5$) achieves the best face recognition results on fc, dup I and dup II probe sets and the third highest on fb probe set. Especially on dup II probe set, which has the largest time lapse, it improves the performance of POEM and LQP by 7 and 13 percent, respectively, validating the proposed DFD is able to extract discriminative and stable face representation and has demonstrated its potential to enhance the state-of-the-art face recognition performance.

TABLE 3
Recognition Rates (Percent) of Proposed Method with State-of-the-Art Methods on FERET Database

Methods	fb	fc	dup I	dup II
LBP [19]*	97.0	79.0	66.0	64.0
LGBP [20]*	98.0	97.0	74.0	71.0
LVP [46]*	97.0	70.0	66.0	50.0
LGT [18]*	97.0	90.0	71.0	67.0
HGPP [21]*	97.5	99.5	79.5	77.8
LLGP [47]*	99.0	99.0	80.0	78.0
DT-LBP [30]*	99.0	**100.0**	84.0	80.0
DLBP [31]*	99.0	99.0	86.0	85.0
POEM [45]*	97.6	95.0	77.6	76.2
LQP [32]*	99.2	69.6	65.8	48.3
DFD(S=3)	99.0	99.0	80.8	80.8
DFD(S=5)	99.2	98.5	85.0	82.9
DFD(S=7)	99.0	95.9	80.9	81.2
POEM+WPCA [45]*	99.6	99.5	88.8	85.0
LQP+WPCA [32]*	**99.8**	94.3	85.5	78.6
DFD(S=3)+WPCA	99.3	99.0	88.8	87.6
DFD(S=5)+WPCA	99.4	**100.0**	**91.8**	**92.3**
DFD(S=7)+WPCA	99.3	96.4	87.7	86.3

Note that the results are from the original paper.

4.1.3 Impact Analysis of Discriminant Filters and Soft Sampling Strategy

In this part, we investigate the effectiveness of discriminant filters and soft sampling strategy individually. First, we only learn the discriminant filters without soft sampling strategy, denoted as DFD^l. In this case, the projection v is set to be $v = [1, \ldots, 1]^T$ so that the neighborhood samplings are treated equally. Second, we apply the soft sampling strategy without discriminant filters learning, denoted as DFD^r. The projection w in this case is set to be $w = [0, \ldots, 0, 1, 0, \ldots, 0]^T$, in which the central elements is set to 1 and other elements 0, so that only the central pixel value in the patch is preserved. The original LBP[1] is also compared as the baseline. The scale/radius size of all the descriptors is set to 5. Table 4 illustrates the comparison results. We can see that both the learned discriminant filters and the soft sampling strategy help to improve the face recognition performance, compared with the original LBP. The combination of them further enhances the face recognition accuracy, indicating that the proposed DFD is effective for face recognition. Comparing DFD^l with LGBP listed in Table 3, we can see that DFD^l slightly outperforms LGBP. Note that LGBP applies 40 Gabor filters, while DFD

1. The LBP matlab code is downloaded from http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab.

TABLE 4
The Effectiveness Comparison (Percent) of Discriminant Filters and Soft Sampling Strategy on FERET Database

Methods	fb	fc	dup I	dup II
LBP	97.1	91.2	67.3	69.2
DFD^l	98.4	97.9	77.2	76.1
DFD^r	99.2	94.3	82.6	78.6
DFD	99.2	98.5	85.0	82.9

TABLE 5
The Comparison Results (Percent) of Global and Local DFD on FERET Database

Methods	fb	fc	dup I	dup II
Global DFD	98.5	96.9	78.4	79.1
Local DFD	99.2	98.5	85.0	82.9

TABLE 6
Computational Cost Comparison of Different Face Representations

Methods	Feature dimension	Feature Extraction Time (ms)
LBP	3776	9.68
LGBP	655360	647.7
HGPP	1474560	1280.1
DFD	50176	179.0

only preserves four linear filters. It validates that the leaning-based linear filters have advantage over Gabor filters to explore effective information for face recognition.

4.1.4 Global DFD versus Local DFD

To describe the face image precisely, in this paper, we divided the face image into different parts and a group of local DFDs is learned from each part, respectively, which is then used to encode the face image pixels correspondingly. To examine the spatial dependency of DFD, we also implement the global DFD, which is learned from the whole face and applied to encode all the image pixels. Table 5 lists the face recognition comparison results of global DFD and local DFD. Because different face regions contain different face structures, local DFD, which describes the face structures locally and precisely, achieves better face recognition performance than global DFD. It validates that the proposed locally learning way is useful to achieve high face recognition accuracy.

4.1.5 Computational Cost

In our experimental configuration, the face image is divided into 49 nonoverlapping regions, each of which corresponds to a 1,024 dimensional feature. Therefore, the feature dimension of DFD is $1{,}024 \times 49 = 50{,}176$. With WPCA, the dimension of extracted feature is reduced to be 1,100 for more efficient matching. We compare the feature dimension and computational costs of DFD extraction with LBP, LGBP, and HGPP representations (shown in Table 6). LBP is implemented by the original authors and other descriptors are implemented by us. All the computational cost is computed on a PC with 3.20 GHZ i5 CPU and 4 G RAM using matlab implementation. It can be seen that LGBP, HGPP, and DFD significantly enlarge the feature size of LBP descriptor with rich improvements on face recognition performance. Compared with LGBP and HGPP, DFD is of lower feature dimension, but with higher face recognition accuracy. It is promising and has great potential to be adopted in real applications.

Fig. 10. Cropped face examples from the CAS-PEAL-R1 database.

TABLE 7
Recognition Rates (Percent) of Proposed Method with State-of-the-Art Methods on CAS-PEAL-R1 Database

Methods	Expression	Accessory	Lighting
LGBP [20]*	95.0	87.0	51.0
LVP [46]*	96.0	86.0	33.0
HGPP [21]*	96.8	92.5	62.9
LLGP [47]*	98.0	92.0	55.0
DT-LBP [30]*	98.0	92.0	41.0
DLBP [31]*	99.0	92.0	41.0
DFD(S=3)	99.3	94.4	59.0
DFD(S=5)	99.0	93.7	47.1
DFD(S=7)	98.2	89.9	38.3
DFD(S=3)+WPCA	99.0	**96.9**	**63.9**
DFD(S=5)+WPCA	**99.6**	**96.9**	58.9
DFD(S=7)+WPCA	98.9	94.9	50.0

**Note that the results are from the original paper.*

4.2 CAS-PEAL-R1

The CAS-PEAL-R1 database is a large-scale Chinese face database for face recognition algorithm training and evaluation. This database provides large-scale face images with different sources of variations, including pose, expression, accessory, and lighting. In this experiment, we follow the standard testing protocols. The gallery set includes 1,040 images from 1,040 people. For probe sets, we use the expression, lighting, accessory subsets, which contain 1,570, 2,243, and 2,285 images, respectively. All the images are cropped to 150×130 size according to the provided eye coordinates. Fig. 10 shows cropped face examples from CAS-PEAL-R1 database.

To examine the generalization performance of the learned DFD, we apply DFD learned from FERET training set on CAS-PEAL-R1 face database directly to test its performance. We compare the performance of proposed DFD with LGBP, LLGP, DT-LBP, DLBP, and so on. Three probe sets including expression, lighting, accessory variations are used to evaluate different methods. The comparison results are listed in Table 7. From the results, we can see that:

1. The performance of DFD ($S = 3$) is better than other two DFDs ($S = 5$ and $S = 7$). It seems that the DFD with smaller scale has advantage in this database.
2. Comparing the results of DFD with previously reported results on this database, it shows that DFD achieves higher face recognition results on expression and accessory probe sets. It indicates that the DFD learned from FERET training set has good

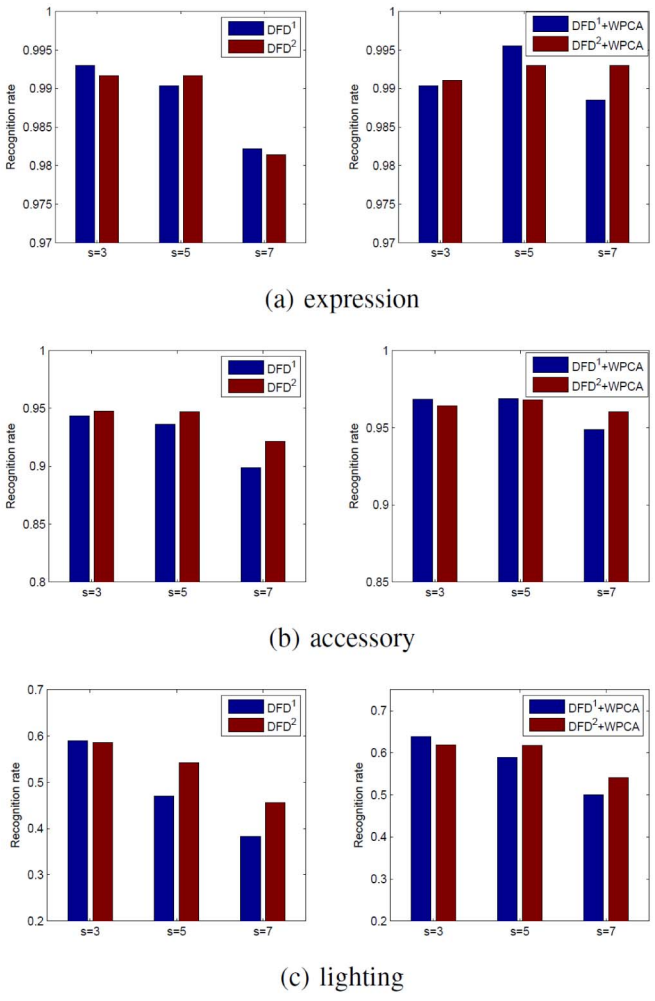

Fig. 11. Face recognition comparison between DFD^1 trained on FERET training set and DFD^2 trained on CAS-PEAL training set.

generalization ability and is robust to variations of expression and accessory.

3. In lighting probe set, DFD ($\mathrm{S} = 3$) is significantly better than other two DFDs ($\mathrm{S} = 5$ and $\mathrm{S} = 7$). Most of the DFD results on this probe sets are worse than HGPP and LLGP, which utilizes Gabor filters to extract the illumination robust representation. It is worth noting that the proposed DFD is a data-driven method. The lack of face samples with lighting variations on FERET training set may result in the unsatisfactory performance on lighting probe set. Nonetheless, the performance of DFD on lighting probe set is still much better than other learning-based methods like DT-LBP and D-LBP, indicating the superiority of DFD over previous learning-based descriptors.
4. With WPCA, the proposed DFD further improves the face recognition performance. DFD (S = 3)+ WPCA achieves higher performance on all three probe sets than the previous reported results. It improves the previous best results by 4.4 and 1 percent on accessory and lighting probe sets, validating that the proposed DFD has good generalization ability and is able to extract robust information to image variations.

Fig. 12. Cropped face examples from LFW database.

To examine the effect of training set on the face recognition performance of DFD, we further learn the DFD on the training set of CAS-PEAL-R1 database and evaluate its performance on expression, accessory, and lighting probe sets. The CAS-PEAL-R1 training set contains 1,200 images from 300 subjects. There are 207 images with lighting variations.

Fig. 11 shows the comparison results of DFD learned from FERET training set (denoted as DFD^1) and DFD learned from CAS-PEAL-R1 training set (denoted as DFD^2). It shows that the DFD^1 and DFD^2 have similar performance on expression and accessory probe sets, validating that DFD does have good generalization. On lighting probe set, DFD^2 achieves higher recognition accuracy than DFD^1. Since DFD is a data-driven method, more samples with lighting variation on CAS-PEAL-R1 training set are helpful to improve the robustness of DFD to lighting. We can also find that the generalization of DFD with smaller scales (e.g., $\mathrm{S} = 3$) is better than those with larger scales. This is because the smaller scale DFD is of lower model complexity (i.e., fewer variables in image filters to be learned) and the FERET training set is large enough to learn a robust descriptor. In contrast, larger scale DFD needs more training data to improve the generalization ability of the learned DFD. With WPCA, the differences of DFD^1 and DFD^2 are reduced and the generalization performance can be further improved.

4.3 LFW

Labeled Faces in the Wild (LFW) is a database collected from the web for studying the problem of unconstrained face recognition. There are 13,233 images from 5,749 different people, with large pose, occlusion, expression variations. In testing phase, researchers are suggested to report performance as 10-fold cross validation using splits that are randomly generated and provided by the organizers. In this experiment, we use the aligned images (LFW-a) [48] and crop the images with the size of 150×130 from the original images. The cropped examples are shown in Fig. 12.

To better evaluate the effectiveness of DFD in real applications, we examine DFD on LFW database. The DFD is learned from the FERET training set. We test on the "View 2" set of LFW, which consists of 10 folds of 300 positive and 300 negative image pairs randomly selected from the original image set. In this experiment, all descriptors are compared in an unsupervised way (i.e., no class label information is involved in classifier/metric learning). The mean recognition accuracy with its standard error is reported. Strictly speaking, the proposed DFD does not follow the LFW protocols because it uses the external face database (FERET) for training. Note that the image

TABLE 8
Mean Recognition Accuracy (Percent) for Different Descriptors on LFW Database

Descriptor	Accuracy
LBP [49]*	69.45±0.5
SIFT [49]*	64.10±0.6
Hybrid descriptor [33]*	78.47±0.5
LARK [50]*	72.23±0.5
POEM [45]*	75.22±0.7
LQP [32]*	75.30±0.8
DFD(S=3)	78.35±0.5
DFD(S=5)	**80.02**±0.5
DFD(S=7)	79.47±0.5
POEM+WPCA [45]*	82.71±0.6
LQP+WPCA [32]*	**86.20**±0.5
DFD(S=3)+WPCA	82.90±0.5
DFD(S=5)+WPCA	84.02±0.4
DFD(S=7)+WPCA	83.13±0.5

Note that the results are from the original paper.

quality in FERET database is very different from that in LFW. This experiment can be considered as an examination of DFD's generalization ability (learned from constrained images) to unconstrained scenarios.

The accuracy defined in [43] is adopted to compare the DFD with other descriptors previously reported on LFW database, including LBP, SIFT, hybrid descriptor, LARK, POEM, and LQP. Table 8 lists the recognition accuracy of different methods and Fig. 13 shows the ROC curves of some descriptors. Note that we only plot the ROC curves which are available on the LFW website. It is clear to see that without whitened PCA, the proposed DFD achieves the best face verification accuracy compared to state-of-the-art descriptors on LFW like LARK, POEM, and LQP. It improves LARK by 7.8 percent, POEM by 4.8 percent, LQP by 4.7 percent, and hybrid descriptor (combination of LBP, TPLBP, FPLBP, etc.) by 1.5 percent, indicating that DFD is a good face descriptor for face recognition. With whitened PCA, DFD also outperforms POEM and achieves comparable results with LQP. Different from previous methods, DFD is a learning-based descriptor rather than manually designed one. Although the face appearance of FERET is very different from that on LFW, the DFD learned from FERET can still work well in the unconstrained case. This is promising, indicating that the generalization ability of DFD is good and it is feasible to deploy DFD in real application.

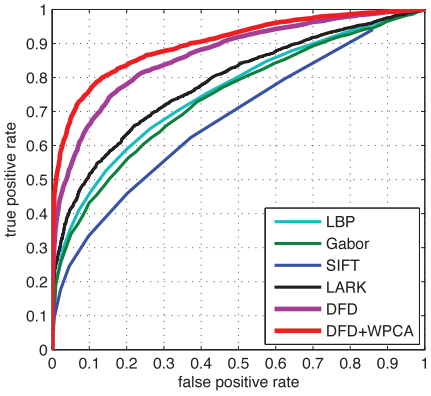

Fig. 13. ROC curves over View 2 on LFW database. The results of LBP, Gabor, SIFT, and LARK are cited from the website (http://vis-www.cs.umass.edu/lfw/results.html) directly.

Fig. 14. Cropped examples from HFB database. The first row is VIS images and the second row is its NIR ones from the same subject.

Considering the results of DFD on FERET, CAS-PEAL-R1, and LFW databases, the DFD with $S = 5$ achieves the best and most stable face recognition performance and is a good choice for face representation in practice.

4.4 HFB

The HFB database was collected by CBSR for heterogeneous biometric research. There are totally 5,097 images, including 2,095 VIS and 3,002 NIR ones from 202 people in the database. In this experiment, we use the former 100 people with their VIS and NIR images as training set. The left images from 102 people form the testing set. There is no overlap of images or subjects between training and testing sets. In testing phase, the gallery set consists of VIS images and the NIR images are used as the probe ones. All the images are cropped into 128×128 size according to the automatically detected eye coordinates. Cropped example images are shown in Fig. 14. The DoG-based preprocessing method [51] is applied to VIS and NIR images to reduce the effect of illumination.

The coupled DFD is learned from HFB training set. The neighboring pixel selection way is the same as in homogeneous face recognition (Fig. 7). We also test the performance of DFD, which is learned from the HFB training set by combining the VIS and NIR images together. Besides LBP and its variants (TPLBP, FPLBP [33]), LPQ [52], SIFT [53], and HOG [54],[2] which are popular descriptors in heterogeneous face recognition [55], [56], are also compared. In the following experiments, the NIR images are compared with the VIS images and the rank-1 face recognition rate, the face verification rate (VR) when the false accept rate (FAR) is 0.001 and 0.01, and the equal error rate (EER) are reported.

Fig. 15 illustrates the score matrices of LBP and C-DFD. One pair of VIS and NIR images is selected for each subject in testing set. The VIS and NIR images form the row and column of the score matrix, respectively. The diagonal is the image pair from the same subject. That is, for the point at the diagonal, the brighter the better. For other area, the darker the better. It shows that the discriminative ability of scores derived from C-DFD is much better than that of LBP, indicating that C-DFD is more effective to match the NIR and VIS images correctly.

Table 9 shows the face recognition performance of different descriptors on VIS versus NIR face matching problem and Fig. 16 illustrates the ROC curves. Not only the C-DFD, but also the DFD, outperform other methods in

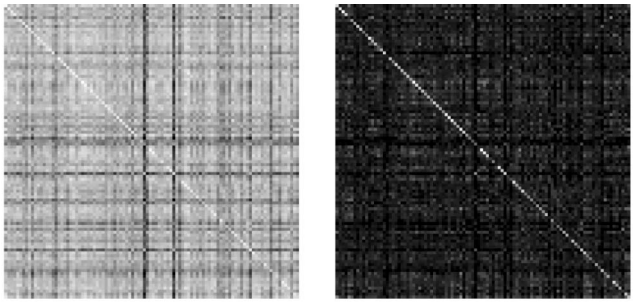

Fig. 15. Illustration of score matrices generated by (a) LBP and (b) C-DFD.

TABLE 9
Heterogeneous Face Recognition Performance (Percent) on HFB Database

Methods	Rank-1	VR @FAR=1%	VR @FAR=0.1%	EER
LBP	55.2	32.1	10.3	24.2
TPLBP	47.9	17.8	3.2	23.4
FPLBP	51.8	13.4	3.4	25.2
LPQ	65.4	26.9	13.3	20.4
SIFT	64.5	43.0	16.1	19.0
HOG	51.2	21.1	7.0	19.0
DFD(S=3)	91.5	83.3	58.9	6.4
DFD(S=5)	69.3	64.2	31.5	8.8
DFD(S=7)	58.8	45.9	21.1	14.4
C-DFD(S=3)	**92.2**	**85.6**	**65.5**	**5.5**
C-DFD(S=5)	85.2	79.1	53.6	6.9
C-DFD(S=7)	74.5	71.2	34.2	8.4

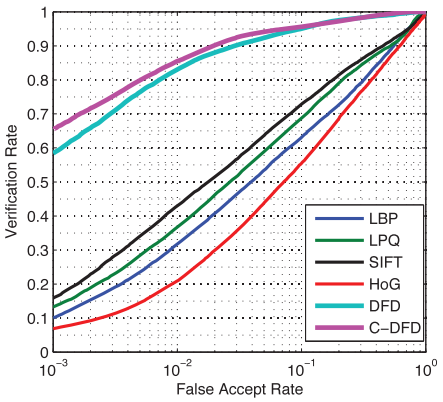

Fig. 16. ROC curves of LBP, LPQ, SIFT, HOG, DFD, and C-DFD on HFB database.

terms of all indices. C-DFD ($S = 3$) beats the best performance of traditional descriptor (SIFT) by 28 percent or so in terms of rank-1 recognition rate, improves the VR ($FAR = 1$ percent) from 43.0 to 85.6 percent, and reduces the EER from 19.0 to 5.5 percent. It significantly improves the LBP, TPLBP, FPLBP, LPQ, SIFT, and HOG descriptors, validating the superiority of the learning-based descriptor. The C-DFD, which models the heterogeneous appearance, achieves better recognition performance than DFD, indicating that C-DFD is effective and necessary for highly accurate heterogeneous face recognition.

4.5 HFB-S

The HFB-S database was collected by us as a supplemental database to the above HFB. There are in total

Fig. 17. Cropped examples from HFB-S database. The first row is VIS images and the second row is its NIR images from the same subject.

TABLE 10
Heterogeneous Face Recognition Performance (Percent) on HFB-S Database

Methods	Rank-1	VR @FAR=1%	VR @FAR=0.1%	EER
LBP	66.6	54.3	33.8	17.9
TPLBP	40.9	29.6	10.3	24.4
FPLBP	35.9	25.8	10.3	24.2
LPQ	75.4	55.8	36.5	17.3
SIFT	71.9	52.8	38.3	20.8
HOG	48.2	32.0	12.6	23.6
DFD(S=3)	89.5	73.2	57.9	10.9
DFD(S=5)	81.2	68.6	52.3	11.9
DFD(S=7)	73.2	61.2	44.0	14.2
C-DFD(S=3)	**90.2**	**78.6**	**64.5**	**9.5**
C-DFD(S=5)	89.2	76.8	62.6	9.7
C-DFD(S=7)	84.5	72.7	56.5	11.1

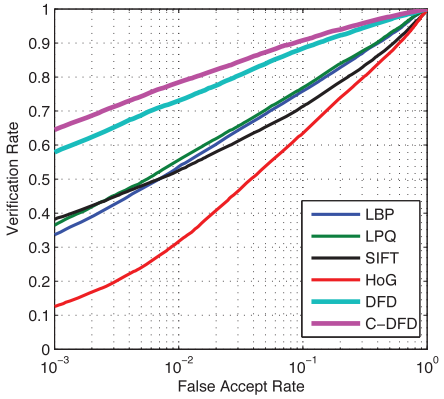

Fig. 18. ROC curves of LBP, LPQ, SIFT, HOG, DFD, and C-DFD on HFB-S database.

5,000 images, including 2,500 VIS images and 2,500 NIR images, from 500 subjects. There are five VIS images and five NIR images for each subject. In experiment, the VIS images are used as the gallery set and the NIR images form the probe set. All the images are cropped to the size of 128×128 according to the automatically detected eye coordinates (Fig. 17). There are more pose variation included in HFB-S database. The DoG-based preprocessing method [51] is applied.

We directly apply the C-DFD/DFD learned from HFB to examine its generalization performance. Table 10 lists the comparison performance with other popular descriptors and Fig. 18 illustrates the corresponding ROC results. It is promising that C-DFD has good generalization. C-DFD trained on HFB also significantly outperforms the state-of-the-art descriptors like LBP, LPQ, SIFT, and HOG (over 20 percent in terms of the rank-1 recognition rate and the verification rates) on HFB-S database, validating that the proposed C-DFD is effective and practical in real applications.

5 Conclusions

This paper proposes a learning-based discriminant face descriptor for face recognition. It enhances the face recognition performance by introducing the discriminative learning into three steps of LBP-like feature extraction. The discriminant image filters, the optimal soft sampling matrix and the dominant patterns are all learned from images. By applying DFD on face images, the appearance difference from different people is maximized and the difference from the same person is minimized. Regarding the heterogeneous (cross-modality) face recognition, we further extend the DFD and propose coupled DFD. Coupled image filters are learned to reduce the feature gap of heterogeneous face images. The DFD is examined on both constrained face databases (FERET and CAS-PEAL-R1) and unconstrained one (LFW). The results show that DFD outperforms state-of-the-art descriptors in most cases, validating the effectiveness of DFD. The C-DFD is compared with LBP, LPQ, SIFT, HOG, and DFD on large VIS and NIR face databases comprehensively, showing that C-DFD is reasonable and effective to address the heterogeneous face recognition problem. Extensive experimental results show that the proposed DFD has good generalization and is a competitive descriptor for face recognition under various circumstances. One of our future work is to investigate DFD in video-based face analysis.

Acknowledgments

This work was supported by the Chinese National Natural Science Foundation Project #61103156, National IoT R&D Project #2150510, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, European Union FP7 Project #257289 (TABULA RASA), and AuthenMetric R&D Funds. The authors would like to thank Dr. Guoying Zhao for her valuable comments on this paper.