Learning Compact Binary Face Descriptor for Face Recognition

Jiwen Lu, *Member, IEEE*, Venice Erin Liong, Xiuzhuang Zhou, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

Abstract—Binary feature descriptors such as local binary patterns (LBP) and its variations have been widely used in many face recognition systems due to their excellent robustness and strong discriminative power. However, most existing binary face descriptors are hand-crafted, which require strong prior knowledge to engineer them by hand. In this paper, we propose a compact binary face descriptor (CBFD) feature learning method for face representation and recognition. Given each face image, we first extract pixel difference vectors (PDVs) in local patches by computing the difference between each pixel and its neighboring pixels. Then, we learn a feature mapping to project these pixel difference vectors into low-dimensional binary vectors in an unsupervised manner, where 1) the variance of all binary codes in the training set is maximized, 2) the loss between the original real-valued codes and the learned binary codes is minimized, and 3) binary codes evenly distribute at each learned bin, so that the redundancy information in PDVs is removed and compact binary codes are obtained. Lastly, we cluster and pool these binary codes into a histogram feature as the final representation for each face image. Moreover, we propose a coupled CBFD (C-CBFD) method by reducing the modality gap of heterogeneous faces at the feature level to make our method applicable to heterogeneous face recognition. Extensive experimental results on five widely used face datasets show that our methods outperform state-of-the-art face descriptors.

Index Terms—Face recognition, heterogeneous face matching, feature learning, binary feature, compact feature, biometrics

1 Introduction

Face recognition is a longstanding computer vision problem and a variety of face recognition methods have been proposed over the past two decades in the literature [1], [4], [36], [38], [56], [71]. Generally, there are four stages in a conventional face recognition system: face detection, face alignment, face representation, and face matching. As a representative pattern recognition problem, face representation and face matching are the most two key stages in a face recognition system. For face representation, the objective is to extract discriminative features to make face images more separable. For face matching, the goal is to design effective classifiers to differentiate different face patterns.

Compared with face matching, face representation significantly affects the performance of a face recognition system because face images captured in real world environments are usually affected by many variations such as varying poses, expressions, illuminations, occlusions, resolutions, and backgrounds. These variations reduce the similarity of face samples from the same person and increase the similarity of face samples from different persons, which is one of the key challenges in face recognition. In recent years, a number of face representation methods have been proposed [1], [4], and they can be mainly classified into two categories: holistic features [4], [56] and local features [1], [36]. Representative holistic features include principal component analysis (PCA) [56] and linear discriminant analysis (LDA) [4], and typical local features are local binary pattern (LBP) [26] and Gabor wavelets [36]. While these face representation methods have achieved encouraging recognition performance in controlled environments, their performance is still far from satisfactory in unconstrained environments. Moreover, most of them are hand-crafted, which usually require strong priors to engineer them by hand. 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.

In this paper, we propose a compact binary face descriptor (CBFD) feature learning method for face representation. Inspired by the fact that binary codes are robust to local changes such as varying illuminations and expressions [26], we aim to learn compact binary codes directly from raw pixels for face representation. Specifically, we learn a feature mapping to project each local pixel difference vector (PDV) into a low-dimensional binary vector, where the variance of all binary codes in the training set is maximized so that the redundancy information in PDVs is removed. To make the learned binary codes compact, we expect that the loss between original PDVs and learned binary codes is minimized and the learned binary codes are evenly distributed at each bin. Then, we cluster and

- *J. Lu and V.E. Liong are with the Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign, 08-10, 1 Fusionopolis Way, Connexis North Tower, Singapore 138632, Singapore. E-mail: {jiwen.lu, venice.l}@adsc.com.sg.*
- *X. Zhou is with the College of Information Engineering, Capital Normal University, Beijing 100048, China. E-mail: zxz@xeehoo.com.*
- *J. Zhou is with the Department of Automation, Tsinghua University, Beijing, 100084, China. E-mail: jzhou@tsinghua.edu.cn.*

Manuscript received 26 July 2014; revised 20 Jan. 2015; accepted 18 Feb. 2015. Date of publication 0 . 0000; date of current version 0 . 0000. Recommended for acceptance by M. Tistarelli. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TPAMI.2015.2408359

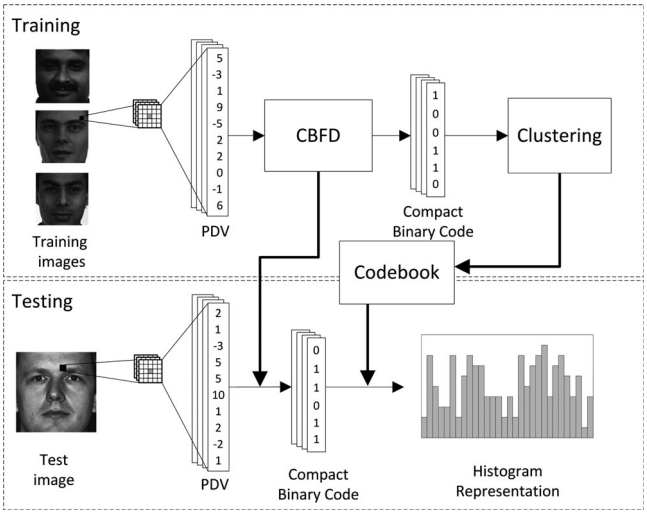


Fig. 1. The pipeline of our proposed feature learning-based face representation approach. For each training face image, we first extract PDVs and learn a feature mapping using CBFD to project each PDV into low-dimensional binary codes. Then, these binary codes are clustered to learn a codebook. For each test image, the PDVs are first extracted and encoded into binary codes using the learned feature mapping. Lastly, these binary codes are pooled as a histogram feature descriptor with the learned codebook.

pool these compact binary codes to obtain a histogram representation of each face image. Moreover, we propose a coupled CBFD (C-CBFD) method to reduce the modality gap at the feature level for heterogeneous face matching. Fig. 1 illustrates the pipeline of our proposed feature learning approach. Experimental results on five widely used face datasets show that our methods outperform state-of-the-art face representation methods.

The contributions of this work are summarized as follows:

1) We propose an unsupervised feature learning method to learn compact binary feature descriptor for face representation. With the learned feature filter, the redundancy information of the original raw pixels is removed in the obtained binary codes.
2) We develop a coupled learning method to learn compact binary face descriptor for heterogeneous face matching. With the learned coupled filters, a common discriminative binary feature space is obtained and the modality gap of heterogeneous faces is greatly reduced at the feature level.
3) We apply CBFD and C-CBFD to learn face features in a local manner so that position-specific information is exploited in the learned features.
4) We conduct extensive face recognition experiments on five widely used face datasets to demonstrate the efficacy of our proposed methods. Experimental results show that our methods are superior to most state-of-the-art face descriptors in both homogeneous and heterogeneous face recognition.

The rest of the paper is organized as follows. Section 2 briefly reviews some background. Section 3 and Section 4 detail the proposed CBFD and C-CBFD feature learning methods, respectively. Section 5 provides the experimental results, and Section 6 concludes the paper.

2 Background

In this section, we briefly review three related topics: 1) face representation, 2) feature learning, and 3) binary code learning.

2.1 Face Representation

There have been extensive work on face representation in the literature, and these methods can be mainly classified into two categories: holistic feature representation [4], [56] and local feature representation [26], [36]. 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 subspace-based face representation methods include PCA [56] and LDA [4]. Unlike holistic features, local features first describe the structure pattern of each local patch and then combine the statistics of all patches into a concatenated feature vector. Typical local features are LBP [26] and Gabor wavelets [36]. However, these local features are hand-crafted and usually require strong prior knowledge to design them by hand. Moreover, some of them are computationally expensive, which may limit their practical applications.

2.2 Feature Learning

There have been a number of feature learning methods proposed in recent years [5], [18], [21], [24], [27], [47]. Representative feature learning methods include sparse auto-encoders [5], denoising auto-encoders [47], restricted Boltzmann machine [18], convolutional neural networks [21], independent subspace analysis [24], and reconstruction independent component analysis [27]. Recently, there have also been some works on feature learning-based (LE) face representation, and some of them have achieved reasonably good performance in face recognition. For example, Lei et al. [32] proposed a discriminant face descriptor (DFD) method by learning an image filter using the LDA criterion to obtain LBP-like features. Cao et al. [8] presented a learning-based feature representation method by applying the bag-of-word (BoW) framework. Hussain et al. [23] proposed a local quantized pattern (LQP) method by modifying the LBP method with a learned coding strategy. Compared with hand-crafted feature descriptors, learning-based feature representation methods usually show better recognition performance because more data-adaptive information can be exploited in the learned features.

2.3 Binary Code Learning

Compared with real-valued feature descriptors, there are three advantages for binary codes: 1) they save memory, 2) they have faster computational speed, and 3) they are robust to local variations. Recently, there has been an increasing interest in binary code learning in computer vision [15], [55], [60], [62]. For example, Weiss et al. [62] proposed an efficient binary coding learning method by preserving the similarity of original features for image search. Norouzi et al. [44] learned binary codes by minimizing a triplet ranking loss for similar pairs. Wang et al. [60] presented a binary code learning method by maximizing the similarity of neighboring pairs and minimizing the

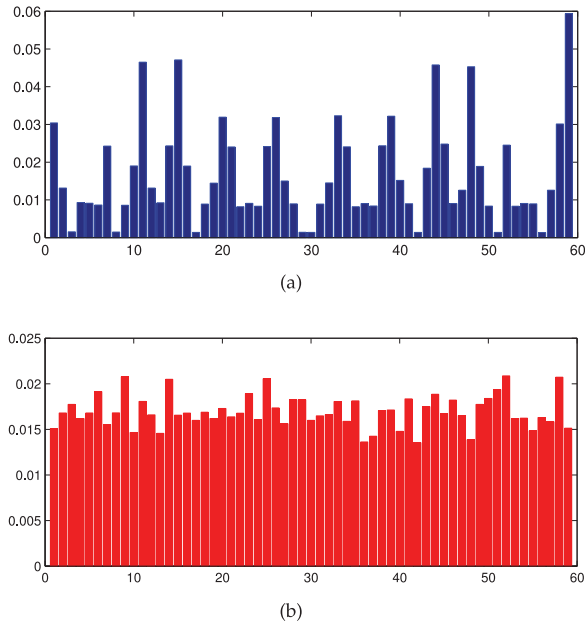


Fig. 2. The bin distributions of the (a) LBP and (b) our CBFD methods. We computed the bin distributions in the LBP histogram and our method in the FERET training set, which consists of 1,002 images from 429 subjects. For a fair comparison, both of them adopted the same number of bins for feature representation, which was set to 59 in this figure. We clearly see from this figure that the histogram distribution is uneven for LBP and is more uniform for our CBFD method.

similarity of non-neighboring pairs for image retrieval. Trzcinski and Lepetit [55] obtained binary descriptors from patches by learning several linear projections based on predefined filters during training. However, most existing binary code learning methods are developed for similarity search [15], [55] and visual tracking [34]. While binary features such as LBP and Haar-like descriptor have been used in face recognition and achieved encouraging performance, most of them are hand-crafted. In this work, we propose a feature learning approach to learn binary features directly from raw pixels for face representation.

3 Learning Compact Binary Face Descriptor

In this section, we first present the CBFD feature learning method and then introduce how to use CBFD for face representation.

3.1 CBFD Feature Learning

While binary features have been proven to very successful in face recognition [26], most existing binary face descriptors are all hand-crafted (e.g., LBP and its extensions) and they suffer from the following limitations:

1) It is generally impossible to sample large size neighborhoods for hand-crafted binary descriptors in feature encoding due to the high computational burden. However, a large sample size is more desirable because more discriminative information can be exploited in large neighborhoods.
2) It is difficult to manually design an optimal encoding method for hand-crafted binary descriptors. For example, the conventional LBP adopts a hand-crafted codebook for feature encoding, which is simple but not discriminative enough because the hand-crafted codebook cannot well exploit more contextual information.
3) Handcrafted binary codes such as those in LBP are usually unevenly distributed, as shown in Fig. 2a. Some codes appear less than others in many real-life face images, which means that some bins in the LBP histogram are less informative and compact. Therefore, these bins make LBP less discriminative.

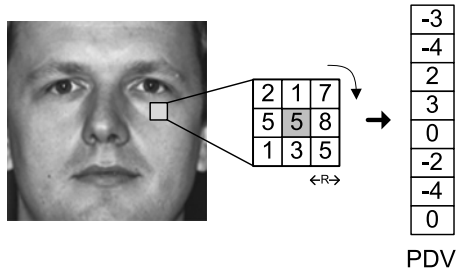


Fig. 3. One example to show how to extract a pixel difference vectors from the original face image. For any pixel in the image, we first identify its neighbors in a $(2R+1)\times(2R+1)$ space, where R is a parameter to define the neighborhood size and it is selected as 1 in this figure for easy illustration. Then, the difference between the center point and neighboring pixels is computed as the PDV.

To address these limitations, in this work, we propose a feature learning method to learn face descriptors directly from raw pixels. Unlike LBP which samples small-size neighboring pixels and computes binary codes with a fixed coding strategy, we sample large-size neighboring pixels and learn a feature filter to obtain binary codes automatically. Let $X=[x_1,x_2,\ldots,x_N]$ be the training set containing N samples, where $x_n\in\mathbb{R}^d$ is the nth pixel difference vector, and $1\le n\le N$. Unlike most previous feature learning methods [21], [27] which use the original raw pixel patch to learn the feature filters, we use PDVs for feature learning because PDV measures the difference between the center point and neighboring pixels within a patch so that it can better describe how pixel values change and implicitly encode important visual patterns such as edges and lines in face images. Moreover, PDV has been widely used in many local face feature descriptors, such as hand-crafted LBP [26] and learning-based DFD [32]. Fig. 3 illustrates how to extract one PDV from the original face image. For any pixel in the image, we first identify its neighbors in a $(2R+1)\times(2R+1)$ space, where R is a parameter to define the neighborhood size. Then, the difference between the center point and neighboring pixels is computed as the PDV. In our experiments, R is set as 3 so that each PDV is a 48-dimensional feature vector.

Our CBFD method aims to learn K hash functions to map and quantize each x_n into a binary vector $b_n=[b_{n1},\ldots,b_{nK}]\in\{0,1\}^{1\times K}$, which encodes more compact and discriminative information. Let $w_k\in\mathbb{R}^d$ be the projection vector for the kth function, the kth binary code b_{nk} of x_n can be computed as

$$b_{nk}=0.5\times\big(\mathrm{sgn}(w_k^T x_n)+1\big), \tag{1}$$

where $\mathrm{sgn}(v)$ equals to 1 if $v\ge 0$ and -1 otherwise.

To make b_n discriminative and compact, we enforce three important criterions to learn these binary codes:

1) The learned binary codes are compact. Since large-size neighboring pixels are sampled, there are some redundancy in the sampled vectors, making them compact can reduce these redundancy.
2) The learned binary codes well preserve the energy of the original samples vectors, so that less information is missed in the binary codes learning step.
3) The learned binary codes evenly distribute so that each bin in the histogram conveys more discriminative information, as shown in Fig. 2b.

To achieve these objectives, we formulate the following optimization objective function:

$$\begin{aligned}\min_{w_k} J(w_k) &= J_1(w_k) + \lambda_1 J_2(w_k) + \lambda_2 J_3(w_k) \\ &= -\sum_{n=1}^{N} \|b_{nk} - \mu_k\|^2 \\ &\quad + \lambda_1 \sum_{n=1}^{N} \|(b_{nk} - 0.5) - w_k^T x_n\|^2 \\ &\quad + \lambda_2 \|\sum_{n=1}^{N} (b_{nk} - 0.5)\|^2,\end{aligned} \tag{2}$$

where N is the number of PDVs extracted from the whole training set, μ_k is the mean of the kth binary code of all the PDVs in the training set, which is recomputed and updated in each iteration in our method, λ_1 and λ_2 are two parameters to balance the effects of different terms to make a good trade-off among these terms in the objective function.

The physical meanings of different terms in (2) are as follows:

1) The first term J_1 is to ensure that the variance of the learned binary codes are maximized so that we only need to select a few bins to represent the original PDVs in the learned binary codes.
2) The second term J_2 is to ensure that the quantization loss between the original feature and the encoded binary codes is minimized, which minimizes the information loss in the learning process.
3) The third term J_3 is to ensure that feature bins in the learned binary codes evenly distribute as much as possible, so that they are more compact and informative to enhance the discriminative power.

Let $W = [w_1, w_2, \dots w_K] \in \mathbb{R}^{d \times K}$ be the projection matrix. We map each sample x_n into a binary vector as follows:

$$b_n = 0.5 \times (\mathrm{sgn}(W^T x_n) + 1). \tag{3}$$

Then, (2) can be re-written as

$$\begin{aligned}\min_{W} J(W) &= J_1(W) + \lambda_1 J_2(W) + \lambda_2 J_3(W) \\ &= -\frac{1}{N} \times \mathrm{tr}((B-U)^T (B-U)) \\ &\quad + \lambda_1 \|(B - 0.5) - W^T X\|_F^2 \\ &\quad + \lambda_2 \|(B - 0.5) \times \mathbf{1}^{N\times 1}\|_F^2,\end{aligned} \tag{4}$$

where $B = 0.5 \times \big(\mathrm{sgn}(W^T X) + 1\big) \in \{0,1\}^{N \times K}$ is the binary code matrix and $U \in \mathbb{R}^{N \times K}$ is the mean matrix which are repeated column vector of the mean of all binary bits in the training set, respectively.

To our knowledge, (4) is an NP-hard problem due to the non-linear $\mathrm{sgn}(\cdot)$ function. To address this, we relax the $\mathrm{sgn}(\cdot)$ function as its signed magnitude [15], [60] and rewrite $J_1(W)$ as follows:

$$\begin{aligned}J_1(W) = &-\frac{1}{N} \times (\mathrm{tr}(W^T X X^T W)) \\ &- 2 \times \mathrm{tr}(W^T X M^T W) \\ &+ \mathrm{tr}(W^T M M^T W),\end{aligned} \tag{5}$$

where $M \in \mathbb{R}^{N \times d}$ is the mean matrix which are repeated column vector of the mean of all PDVs in the training set.

Similarly, $J_3(W)$ can be re-written as

$$\begin{aligned}J_3(W) &= \|(W^T X - 0.5) \times \mathbf{1}^{N\times 1}\|_2^2 \\ &= \|W^T X \times \mathbf{1}^{N\times 1}\|_2^2 \\ &\quad - N \times \mathrm{tr}(\mathbf{1}^{1\times K} \times W^T X \times \mathbf{1}^{N\times 1}) \\ &\quad + 0.5 \times \mathbf{1}^{1\times N} \times \mathbf{1}^{N\times 1} \times 0.5 \\ &= \mathrm{tr}(W^T X \mathbf{1}^{N\times 1} \mathbf{1}^{1\times N} X^T W) \\ &\quad - N \times \mathrm{tr}(\mathbf{1}^{1\times K} W^T X \mathbf{1}^{N\times 1}) \\ &\quad + H,\end{aligned} \tag{6}$$

where $H = 0.5 \times \mathbf{1}^{1\times N} \times \mathbf{1}^{N\times 1} \times 0.5$, which is a constant and is not influenced by W.

Combining (4)-(6), we have the following objective function for our CBFD model:

$$\begin{aligned}\min_{W} J(W) &= \mathrm{tr}(W^T Q W) + \lambda_1 \|(B - 0.5) - W^T X\|_2^2 \\ &\quad - \lambda_2 \times N \times \mathrm{tr}(\mathbf{1}^{1\times K} W^T X \mathbf{1}^{N\times 1}) \\ \text{subject to}: &\quad W^T W = I,\end{aligned} \tag{7}$$

where

$$\begin{aligned}Q \triangleq &-\frac{1}{N} \times (X X^T - 2 X M^T + M M^T) \\ &+ \lambda_2 X \mathbf{1}^{N\times 1} \mathbf{1}^{1\times N} X^T\end{aligned} \tag{8}$$

and the columns of W are constrained to be orthogonal.

While (7) is not convex for W and B simultaneously, it is convex to one of them when the other is fixed. Following the work in [15], we iteratively optimize W and B by using the following two-stage method.

Update B with a fixed W: when W is fixed, (7) can be re-written as

$$\min_{B} J(B) = \|(B - 0.5) - W^T X\|_F^2. \tag{9}$$

The solution to (9) is $(B - 0.5) = W^T X$ if there is no constraint to B. Since B is a binary matrix, this solution is relaxed as

$$B = 0.5 \times \big(\mathrm{sgn}(W^T X) + 1\big). \tag{10}$$

Update W with a fixed B: when B is fixed, (7) can be re-written as

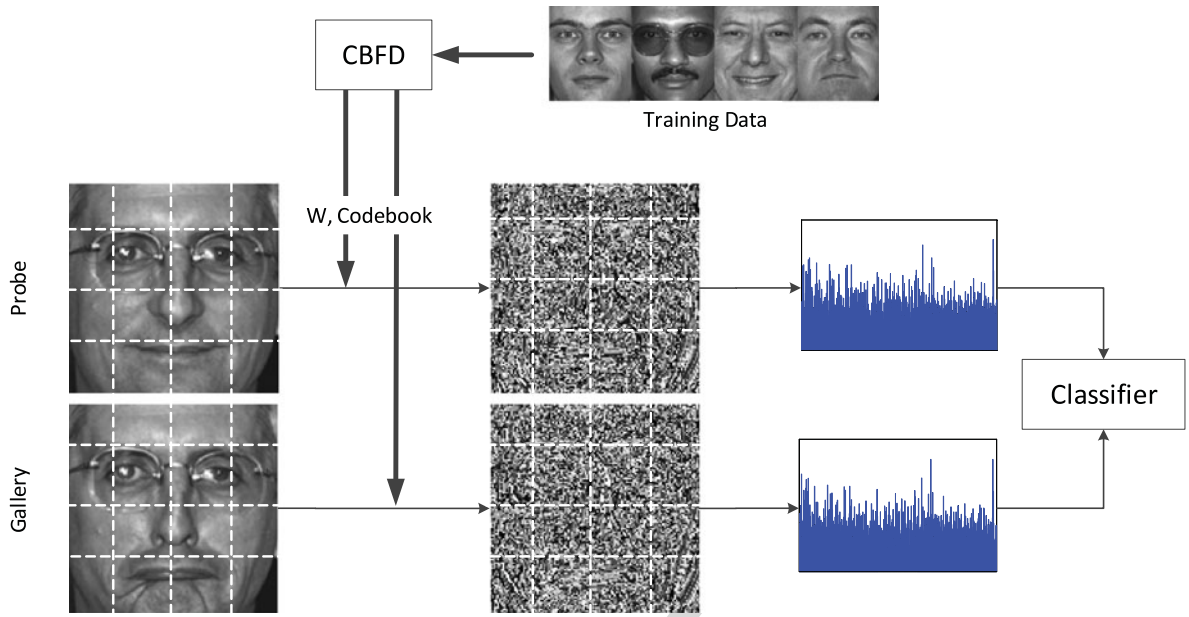


Fig. 4. The flow-chart of the CBFD-based face representation and recognition method. For each training face, we first divide it into several non-overlapped regions and learn the feature filter and dictionary for each region, individually. Then, we apply the learned filter and dictionary to extract histogram feature for each block and concatenate them into a longer feature vector for face representation. Finally, the nearest neighbor classifier is used to measure the sample similarity.

$$\begin{aligned}\min_{W} J(W) &= \mathrm{tr}(W^T Q W) + \lambda_1(\mathrm{tr}(W^T X X^T W)) \\ &\quad - 2 \times \mathrm{tr}((B - 0.5) \times X^T W)) \\ &\quad - \lambda_2 \times N \times tr(\mathbf{1}^{1\times K} W^T X \mathbf{1}^{N\times 1}) \\ \text{subject to} \quad & W^T W = I.\end{aligned} \tag{11}$$

We use the gradient descent method with the curvilinear search algorithm in [64] to solve W. *Algorithm 1* summarizes the detailed procedure of proposed CBFD method.

Algorithm 1. CBFD

Input: Training set $X = [x_1, x_2, \ldots, x_N]$, iteration number T, parameters λ_1 and λ_2, binary code length K, and convergence parameter ϵ.
Output: Feature projection matrix W.
Step 1 (Initialization):
Initialize W to be the top K eigenvectors of XX^T corresponding to the K largest eigenvalues.
Step 2 (Optimization):
For $t = 1, 2, \ldots, T$, repeat
2.1. Fix W and update B using (10).
2.2. Fix B and update W using (11).
2.3. If $|W^t - W^{t-1}| < \epsilon$ and $t > 2$, go to Step 3.
Step 3 (Output):
Output the matrix W.

3.2 CBFD-Based Face Representation

Having obtained the learned feature projection matrix W, we first project each PDV into a low-dimensional feature vector. Unlike many previous feature learning methods [27], [28] which usually perform feature pooling on the learned features directly, we apply an unsupervised clustering method to learn a codebook from the training set so that the learned codes are more data-adaptive. In our implementations, the conventional K-means method is applied to learn the codebook due to its simplicity. Then, each learned binary code feature is pooled as a bin and all PDVs within the same face image is represented as a histogram feature for face representation. Previous studies have shown different face regions have different structural information and it is desirable to learn position-specific features for face representation. Motivated by this finding, we divide each face image into many non-overlapped local regions and learn a CBFD feature descriptor for each local region. Lastly, features extracted from different regions are combined to form the final representation for the whole face image. Fig. 4 illustrates how to use the CBFD for face representation.

4 Learning Coupled Compact Binary Face Descriptor

Recently, many efforts have been made to heterogeneous face recognition [25], [35]. Heterogeneous faces mean that face images are captured in different environments or by different sensors, e.g., photo vs. near infrared, and photo vs. sketch. In this work, we also expect that our feature learning method is applicable for heterogeneous face recognition. Specifically, we propose a coupled CBFD feature learning method to minimize the modality gap at the feature level for heterogeneous face matching. Unlike CBFD, C-CBFD aims to seek K pairs of hash functions to obtain compact binary codes, and minimize the appearance difference of face samples from different modalities, simultaneously. To achieve this, we first extract coupled-PDVs for each face pair in different modalities, as shown in Fig. 5.

Let $X^1 = [x_1^1, x_2^1, \ldots, x_N^1]$ and $X^2 = [x_1^2, x_2^2, \ldots, x_N^2]$ be the PDVs extracted from two modalities of face image sets, where x_n^1 and x_n^2 are the nth PDV extracted from the first and the second modality at the same position, respectively, and $1 \le n \le N$. Our C-CBFD aims to seek K pairs of hash functions to map and quantize x_n^1 and x_n^2 into binary vectors $b_n^1 = [b_{n1}^1, \ldots, b_{nK}^1] \in \{0,1\}^{1\times K}$ and $b_n^2 = [b_{n1}^2, \ldots, b_{nK}^2] \in$

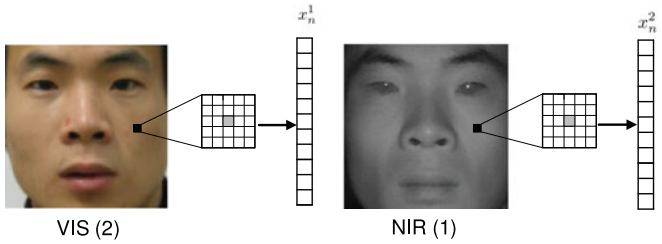


Fig. 5. One example to show how to extract coupled-PDVs from a face pair captured in two different modalities. There are two face images captured by the web camera and near-infrared camera, respectively, and they are aligned at the pixel level so that each pixel at the same position in these two images is aligned. Given any position, we extract two PDVs x_n^1 and x_n^2 to form a coupled-PDV for feature learning.

$\{0,1\}^{1\times K}$, and the kth binary codes b_{nk}^1 and b_{nk}^2 of x_n^1 and x_n^2 are computed as

$$b_{nk}^1 = 0.5 \times \Big(\mathrm{sgn}((w_k^1)^T x_n^1) + 1\Big) \tag{12}$$

$$b_{nk}^2 = 0.5 \times \Big(\mathrm{sgn}((w_k^2)^T x_n^2) + 1\Big), \tag{13}$$

where $w_k^1 \in \mathbb{R}^d$ and $w_k^2 \in \mathbb{R}^d$ are the kth function for the first and the second modality, respectively.

To make b_{nk}^1 and b_{nk}^2 compact and reduce their appearance difference, we formulate the following optimization objective function:

$$\begin{aligned}
&\min_{w_k^1, w_k^2} J(w_k^1, w_k^2) = J_1(w_k^1, w_k^2) + \lambda_1 J_2(w_k^1, w_k^2) \\
&\quad + \lambda_2 J_3(w_k^1, w_k^2) + \lambda_3 J_4(w_k^1, w_k^2) \\
&= -\left(\sum_{n=1}^N \|b_{nk}^1 - \mu_k^1\|^2 + \sum_{n=1}^N \|b_{nk}^2 - \mu_k^2\|^2\right) \\
&\quad + \lambda_1 \left(\sum_{n=1}^N \|(b_{nk}^1 - 0.5) - (w_k^1)^T x_n^1\|^2 \right. \\
&\quad \left. + \sum_{n=1}^N \|(b_{nk}^2 - 0.5) - (w_k^2)^T x_n^2\|^2\right) \\
&\quad + \lambda_2 (\|\sum_{n=1}^N (b_{nk}^1 - 0.5)\|^2 + \|\sum_{n=1}^N (b_{nk}^2 - 0.5)\|^2) \\
&\quad - \lambda_3 \sum_{n=1}^N corr(b_n^1, b_n^2),
\end{aligned} \tag{14}$$

where $corr(b_n^1, b_n^2)$ computes the correlation of two binary vectors b_n^1 and b_n^2.

The objectives of the first three terms in (14) are the same as those of CBFD, which enforce that the variance of the learned binary codes in each modality are maximized, the quantization loss between the original feature and the encoded binary codes in each modality is minimized, and feature bins of the learned binary codes in each modality evenly distribute as much as possible, respectively. The last term in (14) is to ensure that the difference between the binary codes of the corresponding PDVs is minimized, so that the feature modality difference can be reduced in the learned binary codes.

Let $W^1 = [w_1^1, w_2^1, \dots w_K^1] \in \mathbb{R}^{d\times K}$ and $W^2 = [w_1^2, w_2^2, \dots w_K^2] \in \mathbb{R}^{d\times K}$ be the projection matrices of these two modalities. We map x_n^1 and x_n^2 into binary vectors as follows:

$$b_n^1 = 0.5 \times (\mathrm{sgn}((W^1)^T x_n^1) + 1) \tag{15}$$

$$b_n^2 = 0.5 \times (\mathrm{sgn}((W^2)^T x_n^2) + 1). \tag{16}$$

Then, (14) can be re-written as

$$\begin{aligned}
&\min_{W^1, W^2} J(W^1, W^2) = J_1(W^1, W^2) + \lambda_1 J_2(W^1, W^2) \\
&\quad + \lambda_2 J_3(W^1, W^2) + \lambda_3 J_4(W^1, W^2) \\
&= -\frac{1}{N} \times \mathrm{tr}((B^1 - U^1)^T (B^1 - U^1)) \\
&\quad - \frac{1}{N} \times \mathrm{tr}((B^2 - U^2)^T (B^2 - U^2)) \\
&\quad + \lambda_1 \|(B^1 - 0.5) - (W^1)^T X^1\|_F^2 \\
&\quad + \lambda_1 \|(B^2 - 0.5) - (W^2)^T X^2\|_F^2) \\
&\quad + \lambda_2 \|(B^1 - 0.5) \times \mathbf{1}^{N\times 1}\|_2^2 \\
&\quad + \lambda_2 \|(B^2 - 0.5) \times \mathbf{1}^{N\times 1}\|_2^2 \\
&\quad - \lambda_3 \frac{(W^1)^T C_{12} W^2}{\sqrt{(W^1)^T C_{11} W^1} \sqrt{(W^2)^T C_{22} W^2}},
\end{aligned} \tag{17}$$

where B^1 and B^2 are the binary codes matrices, U^1 and U^2 are mean matrices of the binary codes, C_{11} and C_{22} are the variance matrices of X^1 and X^2, respectively, and C_{12} is the covariance of X^1 and X^2.

Similar to CBFD, we also relax the $\mathrm{sgn}(\cdot)$ function as its signed magnitude. Let $W = \begin{bmatrix} W^1 \\ W^2 \end{bmatrix}$, (17) can be re-written as follows:

$$\begin{aligned}
\min_W J(W) &= J_1(W) + \lambda_1 J_2(W) \\
&\quad + \lambda_2 J_3(W) + \lambda_3 J_4(W) \\
&= -\frac{1}{N} tr(W^T \bar{X} \bar{X}^T W) \\
&\quad + \lambda_1 \|B - W^T X\|_F^2 \\
&\quad + \lambda_2 \|(W^T X - 0.5)\mathbf{1}^{2N\times 1}\|\|_F^2 \\
&\quad - \lambda_3 \mathrm{tr}(W^T \Phi W) \\
\text{subject to} \quad & W^T W = I,
\end{aligned} \tag{18}$$

where the columns of W are constrained to be orthogonal, and

$$\bar{X} = \begin{bmatrix} X^1 - M^1 & 0 \\ 0 & X^2 - M^2 \end{bmatrix} \tag{19}$$

$$X = \begin{bmatrix} X^1 & 0 \\ 0 & X^2 \end{bmatrix} \tag{20}$$

$$B = [B^1 - 0.5 \quad B^2 - 0.5] \tag{21}$$

$$\Phi = \begin{bmatrix} 0 & C_{12} \\ C_{21} & 0 \end{bmatrix} \tag{22}$$

$$\Psi = \begin{bmatrix} C_{11} & 0 \\ 0 & C_{22} \end{bmatrix} \tag{23}$$

Fig. 6. Several aligned and cropped face examples from the FERET dataset.

and M^1 and M^2 are the mean matrices of the corresponding PDVs.

We combine the first term $J_1(W)$ and the fourth term $J_4(W)$ in (18) into $J_1^{'}(W)$ and rewrite it as follows:

$$J_1^{'}(W) = -\mathrm{tr}\left(W^T\left(\frac{1}{N}\bar{X}\bar{X}^T - \lambda_3\Phi + \lambda_3\Psi\right)W\right). \quad (24)$$

Similar to CBFD, the objective in (18) can be simplified as

$$\begin{aligned} J(W) &= J_1^{'}(W) + \lambda_1\|B - W^TX\|_2^2 \\ &\quad - \lambda_2 \times 2N \times \mathrm{tr}(\mathbf{1}^{1\times 2K}W^TX\mathbf{1}^{2N\times 1}) \quad (25) \\ \text{subject to}: &\quad W^TW = I. \end{aligned}$$

Similar to the *Algorithm 1* used in CBFD, the projection matrix W of C-CBFD can be also solved by a gradient descent method. W is initialized by getting the top K eigenvectors corresponding to the K largest eigenvalues of $\left(\frac{1}{N}\bar{X}\bar{X}^T + \Phi + \Psi\right)$.

Having obtained the coupled filters, we first apply them on the heterogeneous face pairs to learn the codebook for different modalities, respectively. Similar to CBFD, we also learn local C-CBFD to better exploit the structure of face. Hence, we extract a histogram feature for each local region. Finally, these histogram features are concatenated into a long feature vector to measure the similarity of different faces.

5 Experiments

We evaluate our CBFD and C-CBFD methods on the widely used FERET [46], CAS-PEAL-R1 [13], LFW [22], PaSC [6], and CASIA NIR-VIS 2.0 [33] face datasets. Specifically, the FERET and CAS-PEAL datasets are employed to show the effectiveness of our CBFD method for face identification, the LFW and PaSC datasets are used to show the effectiveness of our CBFD method for face verification, and the CASIA NIR-VIS 2.0 dataset is used to show the effectiveness of our C-CBFD method for heterogeneous face recognition. The following describes the details of the experiments and results.[1]

5.1 Evaluation on FERET

The FERET dataset consists of 13,539 face images of 1,565 subjects who are diverse across age, gender, and ethnicity. We followed the standard FERET evaluation protocol [46], where six sets including the *training*, *fa*, *fb*, *fc*, *dup1*, and *dup2* were constructed for evaluation. All face images in these six sets are aligned and cropped into 128×128 pixels according to the provided eye coordinates. Fig. 6 shows some cropped example images from the FERET dataset. We performed feature learning on the *training* set, and applied the learned hash functions on the other five subsets for feature extraction. Finally, we took *fa* as the gallery set and the other four as probe sets. We set R as 3 and extracted a 48-dimensional PDV for each pixel and mapped it into K-bit binary codes by using the learned hashing functions. In our experiments, K was empirically set as 15. We determined the parameters λ_1 and λ_2 as 0.001 and 0.0001 using the cross-validation strategy on the FERET *training* set. The codebook size was set to 500, and the local region was fixed to 8×8, so that each face image was represented as a 32,000-dimensional feature vector after using CBFD ($32{,}000 = 500 \times 8 \times 8$). Lastly, we applied whitened PCA (WPCA) to reduce the feature dimension into 1,000 and used the nearest neighbor classifier with the cosine similarity for face matching. Similar to [23], [59], WPCA was only conducted on the *training* set.

TABLE 1
Rank-One Recognition Rates (Percent) Comparison with State-of-the-Art Feature Descriptors with the Standard FERET Evaluation Protocol

Method	fb	fc	dup1	dup2
LBP [26]	93.0	51.0	61.0	50.0
LBP+WPCA [26]	98.5	84.0	79.4	70.0
LGBP [70]	94.0	97.0	68.0	53.0
LGBP+WPCA [70]	98.1	99.0	83.8	85.0
LVP [41]	97.0	70.0	66.0	50.0
LGT [30]	97.0	90.0	71.0	67.0
HGGP [69]	97.6	98.9	77.7	76.1
HOG [42]	90.0	74.0	54.0	46.6
DT-LBP [39]	99.0	**100.0**	84.0	80.0
LDP [68]	94.0	83.0	62.0	53.0
GV-LBP-TOP [31]	98.4	99.0	82.0	81.6
DLBP [40]	99.0	99.0	86.0	85.0
GV-LBP [31]	98.1	98.5	80.9	81.2
LQP+WPCA [23]	99.8	94.3	85.5	78.6
POEM [59]	97.0	95.0	77.6	76.2
POEM+WPCA [59]	99.6	99.5	88.8	85.0
s-POEM+WPCA [58]	99.4	**100.0**	91.7	90.2
DFD [32]	99.2	98.5	85.0	82.9
DFD+WPCA [32]	99.4	**100.0**	91.8	92.3
CBFD	98.2	**100.0**	86.1	85.5
CBFD+WPCA	**99.8**	**100.0**	**93.5**	**93.2**

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

5.1.1 Comparison with the State-of-the-Art Face Descriptors

Table 1 tabulates the rank-one recognition rate of our CBFD and state-of-the-art feature descriptors on FERET with the standard evaluation protocol. We see that our proposed CBFD achieves higher recognition rates than the state-of-the-art feature descriptors such as HGGP, GV-LBP-TOP, GV-LBP, POEM and DFD. This is because our CBFD learns discriminative feature descriptors, which can encode more discriminative information than these hand-crafted descriptors. Compared with the recently proposed learning-based

1. The code is available at: https://sites.google.com/site/elujiwen/

TABLE 2
Rank-One Recognition Rates (Percent) of Our CBFD and Three Other Facial Feature Descriptors When PCA and WPCA Are Applied with the Standard FERET Evaluation Protocol

Method	fb	fc	dup1	dup2
LBP + PCA	94.6	94.8	70.9	67.1
LBP + WPCA	97.4	96.9	71.5	58.9
LQP + PCA	98.0	99.5	81.4	78.2
LQP + WPCA	99.7	**100.0**	91.4	88.5
DFD + PCA	97.9	96.4	76.2	68.8
DFD + WPCA	99.6	99.0	88.9	85.6
CBFD + PCA	98.2	**100.0**	86.1	85.5
CBFD + WPCA	**99.8**	**100.0**	**93.5**	**93.2**

TABLE 3
Rank-One Recognition Rates (Percent) of Our CBFD Method and Other Alternative Baselines with the Standard FERET Evaluation Protocol

Method	fb	fc	dup1	dup2
CBFD1 (J1)	99.7	100.0	93.1	91.4
CBFD2 (J2)	99.5	100.0	92.5	90.2
CBFD3 (J3)	99.7	100.0	93.4	91.5
CBFD4 (J1+J2)	99.7	100.0	93.4	92.7
CBFD5 (J1+J3)	99.7	100.0	93.1	92.7
CBFD6 (J2+J3)	99.7	100.0	93.0	91.8
CBFD (J1+J2+J3)	**99.8**	**100.0**	**93.5**	**93.2**

feature descriptor such as DFD, our CBFD is a binary feature learning method which can demonstrate stronger robustness to local variations. Hence, higher recognition rates are obtained in our CBFD method. Moreover, our CBFD achieves the best recognition performance when WPCA is applied.

To fairly compare our CBFD with previous facial feature descriptors, we compared it with three state-of-the-art feature representation methods such as LBP, LQP and DFD under the same experimental settings on FERET. For these three compared methods, we implemented them and conducted face recognition experiments with the same cropped and aligned face images which were used in CBFD. Moreover, the same nearest neighbor classifier with the cosine similarity was used for identification. Table 2 shows the rank-one recognition rates of our CBFD and the other compared feature descriptors on FERET with the standard evaluation protocol.[2] In this experiment, we applied PCA and WPCA to project each descriptor into a 1,000-dimensional feature vector for recognition, respectively, and both PCA and WPCA were conducted on the *training* set. We see that our CBFD consistently outperforms these compared feature descriptors on all subsets of the FERET dataset. Moreover, WPCA further improves the recognition rates for all face representation methods, which is consistent to previous studies [32], [59].

5.1.2 Influence of Different Learning Strategies

We investigated the contributions of different terms in our CBFD model. We defined the following six alternative baselines to study the importance of different terms in our feature learning model:

1) CBFD-1: learning W only from J_1.
2) CBFD-2: learning W only from J_2.
3) CBFD-3: learning W only from J_3.
4) CBFD-4: learning W from J_1 and J_2.
5) CBFD-5: learning W from J_1 and J_3
6) CBFD-6: learning W from J_2 and J_3.

CBFD-1 has a closed-form solution, which can be computed by the eigen-decomposition of J_1. CBFD-2, CBFD-3 and CBFD-6 perform an iterative optimization which is similar to CBFD where J_2, J_3 and J_2+J_3 are used in the objective function of CBFD, respectively. CBFD-4 and CBFD-5 perform *Algorithm 1* by setting the parameters λ_2 and λ_1 to 0 and 0, respectively. For all these six learning model models, WPCA is applied to project each face sample into a 1,000-dimensional feature vector. Table 3 shows the rank-one recognition rates of CBFD and the other six alternative variations on the FERET dataset. We see that all three terms our CBFD model contain discriminative information in our feature descriptor, and J_2 contributes more than J_3 to exploit discriminative information. Moreover, the highest recognition rate can be obtained when all the three terms are used together.

5.1.3 Comparison with the Real-Valued Coded Learning

To better show the advantage of the binary code learning in our CBFD, we also learn features by the relaxed model without any binary processing. Specifically, we developed the following two real-valued codes learning methods to show the importance of the binary codes learning in our method:

1) CRFD-1. Learning real-valued codes only from the J_1 in (2) by discarding the sgn$(\cdot)$ function.
2) CRFD-2. Learning real-valued codes only from the J_1 and J_3 in (2) by discarding the sgn$(\cdot)$ function.

For other steps in CRFD-1 and CRFD-2, we followed the same procedure in our CBFD to learn the feature representations. Table 4 shows the recognition rates on the FERET of different real-valued codes learning and our CBFD method. We see that our CBFD outperforms the compared real-valued codes learning methods. This is because binary codes are more robust to local variations in face images than real-valued codes.

TABLE 4
Rank-One Recognition Rates (Percent) of Our CBFD Method and Other Real-Valued Codes Feature Learning Methods with the Standard FERET Evaluation Protocol

Method	fb	fc	dup1	dup2
CRFD-1 (J1)	**99.8**	**100.0**	90.9	88.5
CRFD-2 (J1 + J3)	99.7	**100.0**	91.1	88.5
CBFD	**99.8**	**100.0**	93.5	**93.2**

2. The results reported in this table are different from those obtained by previous methods. That is because different settings such as different sizes of image regions, different pre-processing methods, and different classifiers were used, respectively.

TABLE 5
Rank-One Recognition Rates (Percent) of Our CBFD Method and Four Existing Binary Codes Learning Methods with the Standard FERET Evaluation Protocol

Method	fb	fc	dup1	dup2
PCA-ITQ [15]	99.7	99.3	89.4	89.9
AQBC [14]	99.7	99.1	88.8	86.9
SH [62]	99.4	90.4	71.9	64.7
SPH [17]	**99.8**	99.8	86.9	87.3
CBFD	**99.8**	**100.0**	**93.5**	**93.2**

TABLE 6
Computational Time (ms) Comparison of Different Face Feature Representation Methods

Method	Feature dimension	Time
LBP [26]	3,776	22.9
SIFT[37]	8,192	63.7
DFD [32]	50,176	1,511.2
CBFD	32,000	227.3

5.1.4 Comparison with the Existing Binary Codes Learning Methods

While most previous binary codes learning methods were developed for visual search [14], [15], [17], [55], [60], [62], they are also applicable to face recognition even if there is no such study before. In this section, we applied four widely used binary codes learning methods including PCA-iterative quantization (PCA-ITQ) [15], angular quantization-based binary coding (AQBC) [14], spectral hashing (SH) [62], and Spherical hashing (SPH) [17] to learn face feature representation. The standard implementations of all these four binary codes learning methods were provided by the original authors. We applied these methods to learn binary codes for face representation in our CBFD model. Specifically, we applied these binary codes learning methods to learn binary codes by replacing the objective function of CBFD in (2). All other steps in CBFD were kept the same so that these binary codes learning methods can be fairly compared with our CBFD. We also employed WPCA to project each face sample into a 1000-dimensional feature vector. Table 5 tabulates the rank-one recognition rates of our CBFD and state-of-the-art binary codes learning methods on FERET with the standard evaluation protocol. We clearly see that our CBFD outperforms the other existing binary codes learning methods in our face recognition task. Compared with the other existing binary codes learning methods, our CBFD is elaborately designed for face feature representation to extract compact and discriminative features for face representation, so that better recognition rate is obtained.

5.1.5 Parameter Analysis

Fig. 7a shows the average recognition rate of CBFD versus different values of λ_1 and λ_2 on FERET. We see that CBFD achieves the best recognition performance when λ_1 and λ_2 were set to 0.001 and 0.0001, respectively.

Fig. 7b shows the objective function value of CBFD versus varying number of iterations on FERET. We see that CBFD converges in 30 iterations.

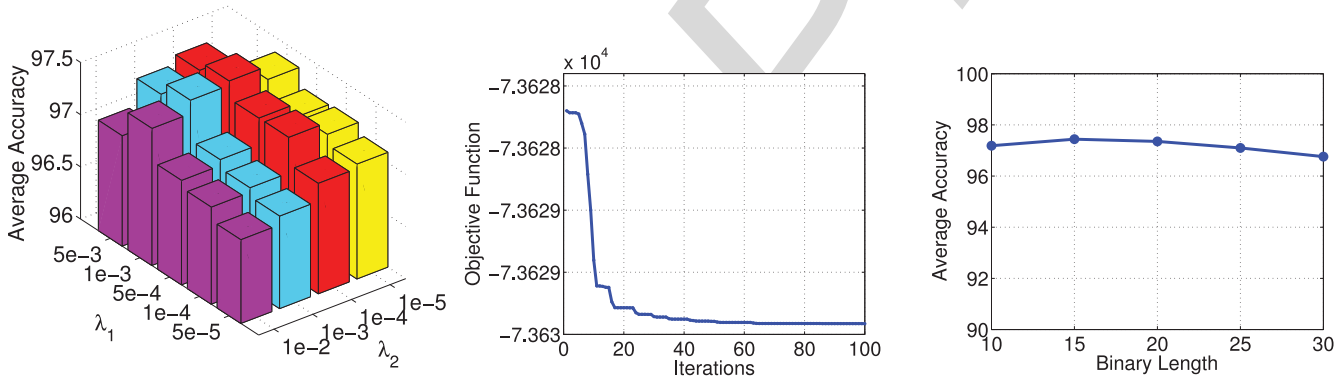
Fig. 7c shows the average rank-one recognition rate of CBFD versus different binary codes length on FERET. We observe that the best recognition rate of CBFD is obtained when the binary code length is set between 15 to 20.

5.1.6 Computational Time

Lastly, we compared the computational time of different feature extraction methods. Our hardware configuration comprises of a 3.4-GHz CPU and a 24 GB RAM. Table 6 shows the feature dimension and average feature extraction time of different methods. It can be seen that both DFD and CBFD improve the recognition performance with higher feature dimensions than LBP and SIFT. Moreover, we observe that our CBFD is more efficient DFD in terms of the feature extraction time. The reason is that our CBFD extracts one PDV for each pixel and DFD extracts a set of PDVs, so the number of filtering in our CBFD is less and the feature extraction of CBFD is faster.

5.2 Evaluation on CAS-PEAL-R1

The CAS-PEAL-R1 dataset contains 9,060 face images from 1,040 subjects with varying variations of pose, expression, accessory, and lighting (PEAL). In our experiments, we



(a) (b) (c)

Fig. 7. (a) Rank-one recognition rate of CBFD on FERET versus different values of λ_1 and λ_2. (b) Objective function value of CBFD versus different number of iterations on FERET. (c) Rank-one recognition rate of CBFD on FERET versus different values of the binary codes length.

Fig. 8. Several aligned and cropped face examples from the CAS-PEAL-R1 dataset.

Fig. 9. Several aligned and cropped face examples with different similarity transformations from the deep funneled LFW dataset. (a) 150 $\times$130 with contour, (b) 150 $\times$ 130 without contour, and (c) 128 $\times$ 128 without contour.

followed the standard evaluation protocol [13], where five sets including *training*, *gallery*, *expression*, *lighting* and *accessory* were constructed for face recognition evaluation. The generic *training* set contains 1,200 images of 300 persons, four images per person. The *gallery* set contains 1,040 frontal images of 1,040 subjects, one image per subject. The *expression*, *lighting* and *accessory* sets are used as probe sets, which contains 1,570, 2,243, and 2,285 face images, respectively. All face images in these four sets are aligned and cropped into 128×128 pixels according to the provided eye coordinates. Fig. 8 shows some aligned and cropped example images from the CAS-PEAL-R1 dataset. We first performed feature learning on the *gallery* set, and then applied the learned features on the other three sets for feature extraction. Parameters of our model are the same as those used on the FERET dataset. Finally, we applied WPCA to reduce the feature dimension into 1,039 and used the cosine metric to compute the similarity. Table 7 tabulates the rank-one recognition rate of different feature descriptors on the CAS-PEAL-R1 dataset. Compared with the state-of-the-art facial descriptors, our CBFD with WPCA improves the previous best DFD+WPCA results by 0.1, 0.3 and 3.5 percent on the *expression*, *accessory* and *lighting* probe sets, respectively.

5.3 Evaluation on LFW

The LFW dataset [22] contains more than 13,000 face images of 5,749 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 [22] which includes 3,000 matched pairs and 3,000 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 paradigms on this dataset [20]. In our experiments, we evaluated our proposed CBFD with three different settings: 1) *unsupervised*, 2) *image-restricted with label-free outside data*, and 3) *image-unrestricted with label-free outside data*.

TABLE 7
Rank-One Recognition Rates (Percent) Comparison with the State-of-the-Art Facial Descriptors Tested with the Standard CAS-PEAL-R1 Evaluation Protocol

Method	Expression	Accessory	Lighting
LGBP [70]	95.0	87.0	51.0
LVP [41]	96.0	86.0	29.0
HGGP [69]	96.0	92.0	62.0
LLGP [66]	96.0	90.0	52.0
DT-LBP [39]	98.0	92.0	41.0
DLBP [40]	99.0	92.0	41.0
DFD [32]	99.3	94.4	59.0
DFD [32]+WPCA	99.6	96.9	63.9
CBFD	99.4	94.8	59.5
CBFD+WPCA	**99.7**	**97.2**	**67.4**

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

5.3.1 Unsupervised Setting

In the unsupervised setting evaluation, we used the deep funneled images for feature learning, where the provided eye coordinates were used for face cropping. For each face image, we used the similarity transformation with three different parameters to obtain three different aligned face images. Specifically, we aligned each image into (a) 150×130 with contour, (b) 150×130 without contour, and (c) 128×128 without contour. Let d_1 be the distance between the eyes to the upper boundary, d_2 the distance between the left eye to the left boundary, and d_3 the distance between two eye centers, (d_1, d_2, d_3) were set to (50, 40, 50), (50, 36, 58), and (42, 36, 56) for the aligned version (a), (b) and (c), respectively. Fig 9 shows several cropped and aligned face images with different similarity transformations. We see that the first aligned version encodes both facial contour and facial components, the second version encodes only facial components, and the last one extracts complementary facial component information.

We first performed feature learning on the training set, and applied the learned features on both the training and testing sets for feature extraction. Parameters of our model are the same as those used on the FERET dataset. Then, we applied WPCA to reduce the feature dimension into 700 for each aligned face image. Finally, we used the cosine metric to compute the similarity. Table 8 and Fig. 10 show the AUC and ROC curve of our CBFD and other existing methods under the unsupervised setting on LFW, where CBFD (a), CBFD (b), CBFD (c), and CBFD mean (a, b, c) indicate the result obtained on the aligned images with the similarity transformation (a), (b), and (c), and the average of these three learned descriptors, respectively. We see that our method achieves better performance than most existing state-of-the-art methods with the unsupervised setting.

TABLE 8
AUC (Percent) Comparisons with the State-of-the-Art Methods on LFW with the Unsupervised Setting

Method	AUC
LBP [57]	75.47
SIFT [57]	54.07
LARK [48]	78.30
LHS [51]	81.07
PAF [67]	94.05
MRF-MLBP [2]	89.94
CBFD (a)	82.32
CBFD+WPCA (a)	88.75
CBFD+WPCA (b)	88.89
CBFD+WPCA (c)	88.65
CBFD+WPCA (mean: a, b, c)	**90.91**

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

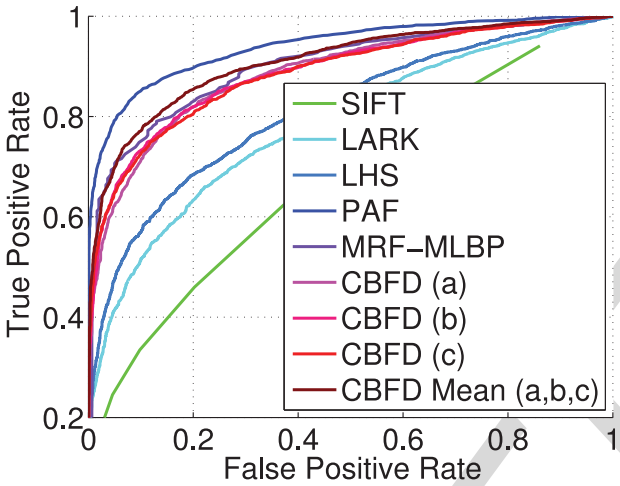


Fig. 10. ROC curves of different methods on LFW with the unsupervised setting. CBFD (a), CBFD (b), and CBFD (c) are face representations obtained by extracting CBFD in 150×130 with contour, 150×130 without contour, and 128×128 without contour face images, respectively. CBFD mean (a, b, c) is the combination of these three representations.

Similar to the comparison on FERET, we also compared CBFD with LBP, LQP and DFD under the same experimental settings. For these three compared feature representation methods, we implemented them and conducted face verification experiments with the same cropped and aligned face images which were used in CBFD. Moreover, the cosine similarity is used for verification. Table 9 shows the AUC of our CBFD and the other feature descriptors under the unsupervised setting on LFW. We see that our CBFD consistently outperforms these compared feature descriptors. Moreover, WPCA further improves the verification rates for all face representation methods.

5.3.2 *Image-Restricted Setting with Label-free Outside Data*

In the image-restricted setting with label-free outside data evaluation, we used the LFW-a[3] version for feature learning, where the provided eye coordinates were used for face alignment. Similar to the deep funneled version, we also

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

TABLE 9
AUC (Percent) Comparisons with the Existing Face Feature Descriptors on LFW with the Unsupervised Setting

Method	AUC
LBP	80.9
LBP + WPCA	84.2
LQP	81.3
LQP + WPCA	87.0
DFD	80.3
DFD + WPCA	83.7
CBFD	82.3
CBFD + WPCA	**88.8**

Fig. 11. Several cropped and aligned face examples with different similarity transformations from the LFW-a dataset. (a) 150×130 with contour, (b) 150×130 without contour, and (c) 128×128 without contour.

align each face images into three versions with different similarity transformations. Fig 11 shows several example face images with different similarity transformations, where the parameters of (d_1, d_2, d_3) were set the same as those in the unsupervised setting. We performed feature learning on the training set, and applied the learned features on both the training and testing sets for feature extraction. WPCA was applied to reduce the feature dimension into 700 for each aligned face image. We applied the discriminative deep metric learning (DDML) [19] to learn a distance metric to compute the similarity of each face pair. To further improve the verification performance, we combined our CBFD with five existing feature descriptors: 1) HDLBP [10], 2) LBP [19], 3) Sparse SIFT [19], 4) Dense SIFT [19] and 5) HOG [12]. We used these five descriptors provided by the original authors in our experiments, respectively.

Having obtained these eight feature descriptors, we used the square root of each feature for face verification and employed SVM to fuse the score to get the final verification result. Table 10 and Fig. 12 show the mean verification rates with standard errors and ROC curves of our CBFD and other existing methods, where CBFD+WPCA (a), CBFD +WPCA (b), CBFD+WPCA (c), CBFD+WPCA (mean), CBFD+WPCA (svm) and CBFD+WPCA (combine) indicate the results obtained on the aligned images with the similarity transformation (a), (b), (c), the average, the score level fusion with SVM, and the score level fusion with SVM on eight feature descriptors, respectively. We see that our method achieves competitive performance with existing

TABLE 10
Comparisons of the Mean Verification Rate and Standard Error (Percent) with the State-of-the-Art Results on LFW under the Image Restricted Setting with Label-Free Outside Data

Method	Accuracy
CSML+SVM, aligned+WPCA [43]	88.00 ± 0.37
PAF [67]	87.77 ± 0.51
SFRD+PMML+WPCA [11]	89.35 ± 0.50
Sub-SML [7]	89.73 ± 0.38
VMRS+WPCA [3]	91.10 ± 0.59
DDML+WPCA [19]	90.68 ± 1.41
CBFD+WPCA(a)	87.33 ± 2.42
CBFD+WPCA(b)	87.57 ± 1.43
CBFD+WPCA(c)	87.23 ± 1.68
CBFD+WPCA(mean: a, b, c)	89.05 ± 1.51
CBFD+WPCA(svm: a, b, c)	$\mathbf{89.07 \pm 1.51}$
CBFD+WPCA(combine)	$\mathbf{92.62 \pm 1.08}$

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

TABLE 11
Comparisons of the Mean Verification Rate and Standard Error (Percent) with the State-of-the-Art Results on LFW under the Image Unrestricted Setting with Label-Free Outside Data

Method	Accuracy
Combined Joint Bayesian [9]	90.90 ± 1.48
Sub-SML [7]	90.75 ± 0.64
ConvNet - RBM [53]	91.75 ± 0.48
VMRS+WPCA [3]	92.05 ± 0.45
Fisher vector faces+WPCA [52]	93.03 ± 1.05
High-dim LBP [10]	93.18 ± 1.07
CBFD+WPCA(a)	87.87 ± 1.86
CBFD+WPCA(b)	88.90 ± 1.81
CBFD+WPCA(c)	88.35 ± 1.61
CBFD+WPCA(mean: a, b, c)	90.90 ± 1.40
CBFD+WPCA(svm: a, b, c)	$\mathbf{90.75 \pm 1.10}$
CBFD+WPCA(combine)	$\mathbf{93.80 \pm 1.31}$

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

state-of-the-art methods. Moreover, our method achieves the best recognition rate (92.62 percent) when the other five feature descriptors are combined, while the current best is 91.10 percent with this setting.

5.3.3 Image-Unrestricted Setting with Label-Free Outside Data

For the image-unrestricted setting with label-free outside data evaluation, we used the same LFW-a dataset in our experiments. Since the label information of each training sample is known in this setting, we trained a joint Bayesian metric learning [9] for face verification. We also combined the above five other descriptors with our CBFD descriptors to further improve the verification performance. Table 11 and Fig. 13 show the mean verification rate with standard errors and ROC curves of different methods with the image-unrestricted setting with label-free outside data. We see that our method achieves comparable performance with existing state-of-the-art methods. Moreover, our method achieves the best recognition rate (93.80 percent) when other five feature descriptors are combined, while the current best is 93.18 percent with this setting.

5.4 Evaluation on PaSC

The PaSC dataset consists of 9,376 still images of 293 people, where face images were collected in different locations, poses and distances from the camera. There is one query set and one target set, and each has 4,688 images. Each image is aligned and cropped into 128×128 pixels according to the provided eye coordinates. Fig. 14 shows some cropped example images from PaSC. 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. Parameters of our model are the same as those used on the FERET dataset. We used the standard evaluation protocol in [6] 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 [6], we also compared our method with the conventional LBP and SIFT hand-crafted feature descriptors. 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 a 3,776-dimensional and 8,192-dimensional feature vector, respec-

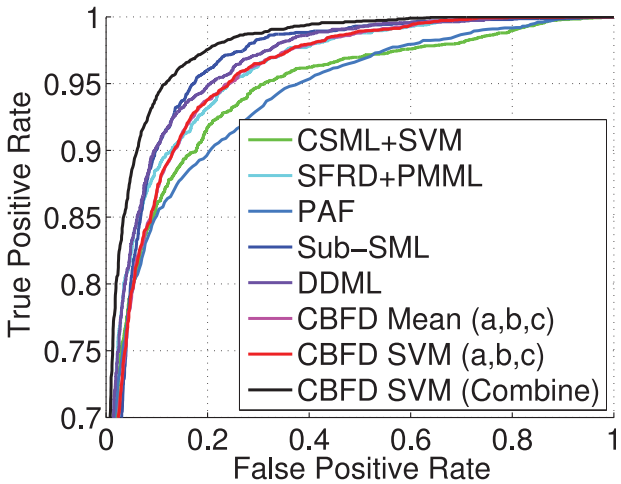


Fig. 12. ROC curves of different methods on LFW with the image-restricted setting with label-free outside data.

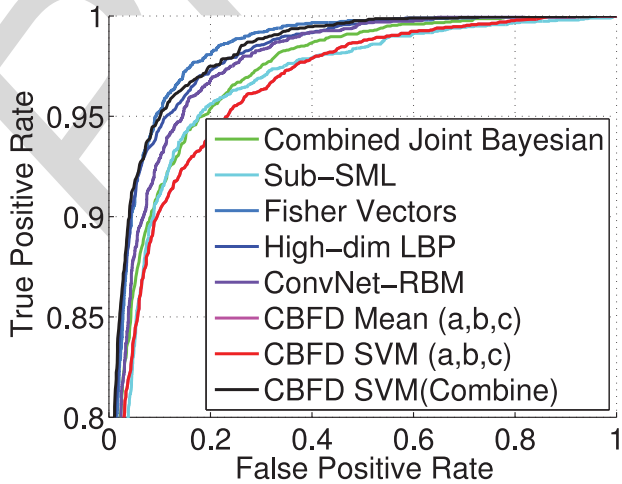


Fig. 13. ROC curves of different methods on the LFW dataset with the image-unrestricted setting with label-free outside data.

Fig. 14. Several cropped face examples from the PaSC dataset.

TABLE 12
Verification Rate (Percent) at the 1.0 Percent FAR of Different Methods on the PaSC Dataset

Method	Verification rate
LRPCA [6]	10.0
LBP [26]	25.1
SIFT [37]	23.2
CBFD	**32.7**

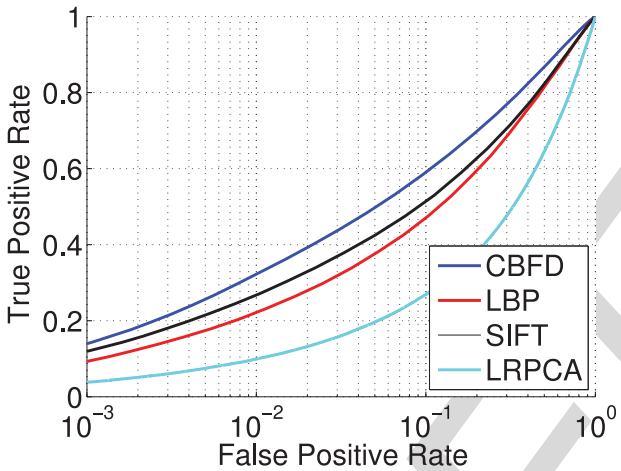


Fig. 15. ROC curves of different feature descriptors on the PaSC dataset with the unsupervised setting.

tively. Finally, we employed WPCA to reduce each of them into a 500-dimensional feature vector as the final representation. Table 12 tabulates the verification rate at the 1.0 percent FAR of these methods and Fig. 15 shows the ROC curves, respectively. As can seen, our proposed CBFD significantly outperforms the other three compared methods, where the minimal improvement of verification rate is 7.6 percent.

5.5 Evaluation on CASIA NIR-VIS 2.0

Lastly, we evaluated our C-CBFD method on heterogeneous face matching to further demonstrate the effectiveness of our feature learning method. The CASIS NIR-VIS 2.0 dataset [33] was used for heterogeneous face matching evaluation. This dataset contains 275 subjects, and each subject has 1-22 VIS and 5-50 NIR face images, respectively. Each face image was aligned and cropped into 128×128 according to the provided eyes' positions. Fig. 16 shows some aligned and cropped example images from the CASIA NIR-VIS 2.0 dataset. We followed the evaluation protocol of this dataset where the VIS images are utilized as the gallery set and the NIR images as the probe set.

Fig. 16. Several aligned and cropped face examples from the CASIA VIS-NIR 2.0 dataset, where face images in the first and second rows are the NIR and VIS images, respectively.

TABLE 13
Performance Comparison on the CASIA VIS-NIR 2.0 Dataset, where VR1 and VR2 Denote the Mean Verification Rate When the FAR is Set to 0.1 and 1.0 Percent, Respectively

Method	Rank 1	VR1	VR2
CCA [16]	28.5 ± 3.4	10.8	30.7
PLS [49]	17.7 ± 1.9	2.3	9.5
CDFE [35]	27.9 ± 2.9	6.9	23.3
MvDA [25]	41.6 ± 4.1	19.2	42.8
LCFS [61]	35.4 ± 2.8	16.7	35.7
GMLDA [50]	23.7 ± 1.4	5.1	16.6
GMMFA [50]	24.8 ± 1.1	7.6	19.5
LBP [26]	35.4 ± 2.7	4.2	31.8
LTP [54]	35.1 ± 2.2	8.2	34.7
TP-LBP [65]	36.2 ± 1.6	3.7	12.9
FP-LBP [65]	23.2 ± 1.0	1.7	9.0
LPQ [45]	47.5 ± 0.9	4.0	17.2
SIFT [37]	49.1 ± 2.3	14.3	40.8
C-CBFD	$\mathbf{56.6 \pm 2.4}$	**20.4**	**44.3**
C-CBFD + LDA	$\mathbf{81.8 \pm 2.3}$	**47.3**	**75.3**

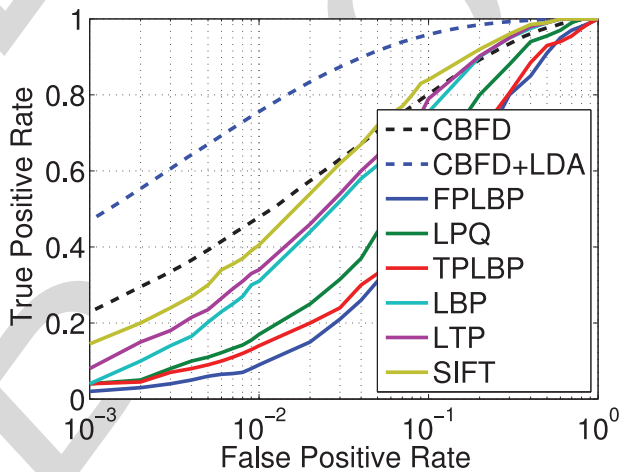


Fig. 17. ROC curves of different methods on the CASIA NIR-VIS 2.0 dataset.

We used DOG to pre-process each face image to remove the immunization effects. Having obtained the coupled feature projection matrices, each image is projected into a 400-dimensional feature vector by WPCA, and the nearest neighbor classifier with the cosine similarity is used for face matching. λ_1 and λ_2 are set to 0.001 and 0.0001, respectively, similar to the FERET experiment. While the parameter λ_3 was empirically set to 0.01 by cross validation in the training set. Table 13 tabulates the rank-one recognition rates of state-of-the-art methods and Fig. 17 shows the ROC

curves of different feature descriptors for heterogeneous face recognition. As can be seen, our proposed C-CBFD outperforms all the other compared methods. Moreover, the performance is significantly improved when LDA is used for classification. Compared with the existing descriptor-based methods, our C-CBFD can represent face images from two modalities into a common feature space, which is more effective to heterogeneous face matching because the modality appearance difference is heavily reduced. Compared with the existing model-based methods which learn a latent common subspace to reduce the appearance of different modalities, our C-CBFD learn modality-invariant feature descriptors at the feature level, which are more effective to reduce the gap and obtain higher recognition rate.

5.6 Discussion

The above experimental results suggest the following five key observations:

1) With WPCA and cosine similarity, our CBFD and C-CBFD achieve the best performance than state-of-the-art feature descriptors in homogeneous and heterogeneous face recognition, respectively. This is because our CBFD and C-CBFD automatically learn feature representation from raw data, which are more data-adaptive than existing hand-crafted descriptors. Moreover, both CBFD and C-CBFD are binary feature descriptors, which demonstrate stronger robustness to local variations. Hence, higher recognition rates are obtained.
2) Each of the three criterions in our CBFD model is effective to extract discriminative information in our feature descriptor. Hence, the best recognition performance is obtained when all these three terms are used together for feature learning.
3) Facial images aligned with different similarity transforms provide complementary information for feature extraction, so that better recognition performance is obtained when they are combined together.
4) Since the whitening procedure reduce the correlation of the extracted features and make all features have the same variance, the importance of different features can be balanced and the dominance of a few features in the original space can be avoided. Hence, higher recognition rate can be obtained, which is consistent to most previous findings [23], [43], [59].
5) Our CBFD is more efficient because encoding high-dimensional features to compact binary codes is a simple operation that yields low memory usage and fast computation speed during quantization [15], [23], [63].

6 Conclusion

We have proposed a new feature learning method called compact binary face descriptor for face representation and recognition. To make CBFD applicable to heterogeneous face recognition, we have further proposed a coupled CBFD method to reduce the modality gap for heterogeneous face matching. Experiments on five benchmark face databases verify that our methods achieve better or competitive recognition performance than the state-of-the-art feature descriptors.

There are two interesting directions for future work:

1) Our CBFD and C-CBFD are general feature learning methods. It is interesting to apply them to other computer vision applications such as object recognition and visual tracking to further demonstrate their effectiveness.
2) In this work, we only learned features from one single layer. Recent advances in deep learning have shown that learning hierarchal features in a deep networks can usually achieve better performance because more abstract information can be usually exploited [18], [29]. Hence, it is interesting to exploit more hierarchical features under the deep learning framework to further improve the recognition performance.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61225008 and Grant 61373090, the National Basic Research Program of China under Grant 2014CB349304, the Ministry of Education of China under Grant 20120002110033, the Tsinghua University Initiative Scientific Research Program, and a research grant for the Human Centric Cyber Systems (HCCS) Program at the Advanced Digital Sciences Center (ADSC) from the Agency for Science, Technology and Research (A*STAR) of Singapore.

References

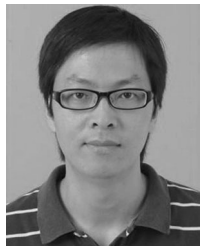
[1] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[2] S. R. Arashloo and J. Kittler, "Efficient processing of MRFs for unconstrained-pose face recognition," in *Proc. 6th Int. Conf. Biometrics: Theory, Appl. Syst.*, 2013, pp. 1–8.
[3] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz, "Fast high dimensional vector multiplication face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1960–1967.
[4] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[5] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 153–160.
[6] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Glynn, and S. Cheng, "The challenge of face recognition from digital point-and-shoot cameras," in *Proc. 6th Int. Conf. Biometrics: Theory, Appl. Syst.*, 2013, pp. 1–8.
[7] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2408–2415.
[8] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 2707–2714.
[9] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 566–579.

[10] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3025–3032.
[11] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3554–3561.
[12] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using histograms of oriented gradients," *Pattern Recog. Lett.*, vol. 32, no. 12, pp. 1598–1603, 2011.
[13] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man Cybern., Part A: Syst. Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.
[14] Y. Gong, S. Kumar, V. Verma, and S. Lazebnik, "Angular quantization-based binary codes for fast similarity search," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1205–1213.
[15] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 817–824.
[16] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[17] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, and S.-E. Yoon, "Spherical hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2957–2964.
[18] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[19] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1875–1882.
[20] G. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," Dept. Comput. Sci., Univ. Massachussets Amherst, Amherst, MA, USA, Tech. Rep. UM-CS-2014-003, 2014.
[21] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2518–2525.
[22] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, Amherst, MA, USA, Tech. Rep. 07-49, 2007.
[23] S. U. Hussain, T. Napoléon, F. Jurie, "Face recognition using local quantized patterns," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–12.
[24] A. Hyvärinen, J. Hurri, and P. O. Hoyer, "Independent component analysis," *Natural Image Statist.*, vol. 39, pp. 151–175, 2009.
[25] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 808–821.
[26] J. Kittler, A. Hilton, M. Hamouz, and J. Illingworth, "3d assisted face recognition: A survey of 3d imaging, modelling and recognition approachest," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 469–481.
[27] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "Ica with reconstruction cost for efficient overcomplete feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1017–1025.
[28] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 3361–3368.
[29] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1096–1104.
[30] Z. Lei, S. Z. Li, R. Chu, and X. Zhu, "Face recognition with local Gabor textons," in *Proc. Int. Conf. Adv. Biometrics*, 2007, pp. 49–57.
[31] Z. Lei, S. Liao, M. Pietikainen, and S. Z. Li, "Face recognition by exploring information jointly in space, scale and orientation," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 247–256, Jan. 2011.
[32] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 289–302, Feb. 2014.
[33] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshop*, 2013, pp. 348–353.
[34] X. Li, C. Shen, A. R. Dick, and A. van den Hengel, "Learning compact binary codes for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 2419–2426.
[35] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 13–26.
[36] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.
[37] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[38] J. Lu, Y.-P. Tan, and G. Wang, "Discriminative multimanifold analysis for face recognition from a single training sample per person," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 39–51, Jan. 2013.
[39] D. Maturana, D. Mery, and A. Soto, "Face recognition with decision tree-based local binary patterns," in *Proc. 10th Asian Conf. Comput. Vis.*, 2010, pp. 618–629.
[40] D. Maturana, D. Mery, and A. Soto, "Learning discriminative local binary patterns for face recognition," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recog. Workshops*, 2011, pp. 470–475.
[41] X. Meng, S. Shan, X. Chen, and W. Gao, "Local visual primitives (LVP) for face modelling and recognition," in *Proc. Int. Conf. Pattern Recog.*, 2006, pp. 536–539.
[42] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 93–104, 2008.
[43] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 709–720.
[44] M. Norouzi, D. Fleet, and R. Salakhutdinov, "Hamming distance metric learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1070–1078.
[45] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proc. 3rd Int. Conf. Image Signal Process.*, 2008, pp. 236–243.
[46] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[47] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 833–840.
[48] H. J. Seo and P. Milanfar, "Face verification using the lark representation," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 4, pp. 1275–1286, Dec. 2011.
[49] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 593–600.
[50] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2160–2167.
[51] G. Sharma, S. ul Hussain, and F. Jurie, "Local higher-order statistics (LHS) for texture categorization and facial analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 1–12.
[52] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 1–12.
[53] Y. Sun, X. Wang, X. Tang, "Hybrid deep learning for face verification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1489–1496.
[54] X. Tan and B. Triggs, "Fusing Gabor and LBP feature sets for kernel-based face recognition," in *Proc. 3rd Int. Conf. Anal. Model. Faces Gestures*, 2007, pp. 235–249.
[55] T. Trzcinski and V. Lepetit, "Efficient discriminative projections for compact binary descriptors," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 228–242.
[56] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
[57] R. Verschae, J. Ruiz-del Solar, M. Correa, "Face recognition in unconstrained environments: A comparative study," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2008, pp. 1–12.
[58] N.-S. Vu, "Exploring patterns of gradient orientations and magnitudes for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 2, pp. 295–304, Feb. 2013.
[59] N.-S. Vu and A. Caplier, "Enhanced patterns of oriented edge magnitudes for face recognition and image matching," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1352–1365, Mar. 2012.

[60] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3424–3431.
[61] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2407–2414.
[62] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1753–1760.
[63] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1753–1760.
[64] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Programm.*, vol. 142, pp. 397–434, 2013.
[65] L. Wolf, T. Hassner, Y. Taigman, "Descriptor based methods in the wild," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2008, pp. 1–14.
[66] S. Xie, S. Shan, X. Chen, X. Meng, and W. Gao, "Learned local gabor patterns for face representation and recognition," *Signal Process.*, vol. 89, no. 12, pp. 2333–2344, 2009.
[67] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3539–3545.
[68] B. Zhang, Y. Gao, S. Zhao, and J. Liu, "Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 533–544, Feb. 2010.
[69] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of Gabor phase patterns (HGPP): A novel object representation approach for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 57–68, Jan. 2007.
[70] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 786–791.
[71] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

Jiwen Lu received the BEng degree in mechanical engineering and the MEng degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, and the PhD degree in electrical engineering from the Nanyang Technological University, Singapore, respectively. He is currently a research scientist at the Advanced Digital Sciences Center (ADSC), Singapore. His research interests include computer vision, pattern recognition, and machine learning. He has authored/coauthored more than 100 scientific papers in these areas, where more than 30 papers are in the *IEEE Transactions Journals* (TPAMI/TIP/TIFS/TCSVT) and the top-tier computer vision conferences (ICCV/CVPR/ECCV). He serves as a area chair for 2015 IEEE International Conference on Multimedia and Expo (ICME 2015), 2015 IAPR/IEEE International Conference on Biometrics (ICB 2015), and Special Session Chair for 2015 IEEE Conference on Visual Communications and Image Processing (VCIP 2015). He received the First-Prize National Scholarship and the National Outstanding Student Award from the Ministry of Education of China in 2002 and 2003, the Best Student Paper Award from PREMIA of Singapore in 2012, and the Top 10 percent Best Paper Award from MMSP2014, respectively. Recently, he gives tutorials at some conferences such as CVPR2015, FG2015, ACCV2014, ICME2014, and IJCB2014. He is a member of the IEEE.

Venice Erin Liong received the BS degree from the University of the Philippines Diliman, Quezon City, Philippines, in 2010, and the MS degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon City, South Korea, in 2013. She is currently a research engineer at the Advanced Digital Sciences Center (ADSC), Singapore. Her research interests include computer vision, pattern recognition, and machine learning.



Xiuzhuang Zhou received BSci degree from the Department of Atmosphere Physics, Chengdu University of Information Technology, Chengdu, China, in 1996. He received the MEng degree and PhD degree from the School of Computer Science, Beijing Institute of Technology, Beijing, China, in 2005 and 2011, respectively. He is currently an assistant professor in the College of Information Engineering, Capital Normal University, Beijing, China. His research interests include computer vision, pattern recognition, and machine learning. He has authored several scientific papers in peer-reviewed journals and conferences including some top venues such as the *IEEE Transactions on Image Processing*, *IEEE Transactions on Information Forensics and Security*, CVPR and *ACM MM*. He is a member of the IEEE.

Jie Zhou received the BS and MS degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the PhD degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 30 papers have been published in top journals and conferences such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, and CVPR. He is an associate editor for the *International Journal of Robotics and Automation* and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a senior member of the IEEE.