Fusing Robust Face Region Descriptors via Multiple Metric Learning for Face Recognition in the Wild

Zhen Cui[1,3], Wen Li[2], Dong Xu[2], Shiguang Shan[1], Xilin Chen[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China
[2]School of Computer Engineering, Nanyang Technological University, Singapore
[3]School of Computer Science and Technology, Huaqiao University, Xiamen, China
zhen.cui@vipl.ict.ac.cn;WLI1@e.ntu.edu.sg;dongxu@ntu.edu.sg;{sgshan,xlchen}@ict.ac.cn

Abstract

In many real-world face recognition scenarios, face images can hardly be aligned accurately due to complex appearance variations or low-quality images. To address this issue, we propose a new approach to extract robust face region descriptors. Specifically, we divide each image (resp. video) into several spatial blocks (resp. spatial-temporal volumes) and then represent each block (resp. volume) by sum-pooling the nonnegative sparse codes of position-free patches sampled within the block (resp. volume). Whitened Principal Component Analysis (WPCA) is further utilized to reduce the feature dimension, which leads to our Spatial Face Region Descriptor (SFRD) (resp. Spatial-Temporal Face Region Descriptor, STFRD) for images (resp. videos). Moreover, we develop a new distance metric learning method for face verification called Pairwise-constrained Multiple Metric Learning (PMML) to effectively integrate the face region descriptors of all blocks (resp. volumes) from an image (resp. a video). Our work achieves the state-of-the-art performances on two real-world datasets LFW and YouTube Faces (YTF) according to the restricted protocol.

1. Introduction

Many face recognition systems have demonstrated promising results under well-controlled conditions with cooperative users. However, face recognition "in the wild" is still a challenging problem due to dramatic intra-class variations caused by pose, lighting and expression. Moreover, the faces in surveillance or internet videos (*e.g.* YouTube videos) are commonly with low-resolution and may be even blurred, which brings additional challenges for face recognition systems. According to the reported results, most face recognition algorithms degrade heavily on two real-world datasets: Labeled Faces in the Wild (LFW) [16] and YouTube Faces (YTF) [37]. Therefore, it is crucial to develop robust features and machine learning algorithms to improve the face recognition performances on these real-world datasets.

According to the features, face recognition methods can be roughly divided into two categories: global feature based approaches and local feature based approaches. Among global feature based approaches, "Eigenface" [31] and "Fisherface" [2] are two classical techniques using images [40] or even videos (*i.e.* a set of face images) [10,20, 34] as the input. However, the global features are not robust to local distortions due to expression, occlusion, etc. As a result, local feature based approaches were widely used in the past decades by developing hand-crafted local descriptors, *e.g.*, Gabor feature [11, 17], LBP feature [1] and their variants [18, 43]. Although some of these methods have shown promising performances on public face datasets collected in the controlled environment, these hand-crafted descriptors cannot work well for face recognition in the wild (See the results on the LFW dataset[1]). To improve the performance, researchers have tried different methods by integrating multiple types of local features [27,42] and borrowing an extra reference set [23, 41]. Moreover, it is also unclear how to effectively employ these features for the more challenging Video-based Face Recognition (VFR) task, in which the face images are usually of low-resolution and even low-quality.

Recently, the Bag-of-Feature (BoF) methods have attracted increasing attention and achieved excellent performance for object classification [24, 29]. However, faces are all similar in overall configuration and only different on subtle textons. If we directly apply the classical BoF model to face recognition, many facial details will be discarded, which might degrade the performance. More recently, a few

[1]http://vis-www.cs.umass.edu/lfw/

BoF-based methods, such as Random Projection Tree [4] and Local Quantized Patterns (LQP) [32], have attempted to apply BoF to face images, but it still lacks of a systematic study of BoF for face recognition with clear performance improvements.

In this work, we propose a new approach to extract robust features for real-world face recognition tasks. One major challenge is that in these scenarios the face images are only roughly aligned, thus considerable spatial misalignment exists in the cropped face images. To address this issue, in this work, we partition each image or one video keyframe into a set of blocks and only compare the features extracted from corresponding blocks. On the other hand, we represent each block by a set of position-free patches without enforcing spatial constraints for the patches within the block. This simple strategy in combination with the BoF framework leads to an effective feature, which is robust to face misalignment. Specifically, we first adopt the nonnegative sparse coding to quantize each patch according to a set of visual words in a pre-constructed visual vocabulary from k-means clustering. Then we extract Token-Frequency (TF) features from each image (*resp.* video) by sum-pooling the reconstruction coefficients over the patches within each block (*resp.* each spatial-temporal volume consisting of a set of blocks along the temporal dimension). Finally, we extract our Spatial Face Region Descriptor (SFRD) (*resp.* Spatial-Temporal Face Region Descriptor, STFRD) for images (*resp.* videos) by applying Whitened Principal Component analysis (WPCA) to reduce the dimension of TF features and suppress the noise in the leading eigenvectors.

We also develop a new distance metric learning method called Pairwise-constrained Multiple Metric Learning (PMML) for face verification by integrating the SFRDs (*resp.* STFRDs) from all the blocks of an image (*resp.* all the volumes of a video). In contrast to the existing approaches which can only learn one distance metric for one type of feature, our method simultaneously learns multiple metrics for different descriptors, which better utilizes the correlations of these descriptors. We also introduce the first order continuous differentiable hinge loss to avoid unnecessary penalties on the pairs which satisfy the constraints. We conduct the experiments on two real-world face datasets, LFW [16] and YTF [37]. Extensive experiments demonstrate the effectiveness of our new descriptors, SFRD and STFRD, as well as our PMML metric learning method for face verification under unconstrained conditions.

2. Related Work

The BoF framework has been widely applied in object classification [24, 29]. It usually starts from low-level features and then encodes them into a set of visual words using hard or soft assignment. After that, a spatial pooling step is used to form the global representation. However, only few BoF methods were used for face recognition, because face images are different only on very subtle textons. Cao *et al.* [4] introduced the hard quantization by random-projection tree. After that, Cui *et al.* [9] developed a soft assignment method by using sparse coding. Similar methods were also proposed, such as block-based BoF [25] and LQP by hash table [32].

Most of the existing work on metric learning focuses on the Mahalanobis distance learning [12, 14, 21, 26, 36]. Specifically, the recently proposed methods such as Information Theoretic Metric Learning (ITML) [12], Logistic Discriminant Metric Learning (LDML) [14], Unsupervised metric learning [21], Pairwise Constrained Component Analysis (PCCA) [26] and KISS Metric Learning (KISSME) [21] were designed to deal with general pairwise constraints. However, they only optimize one metric without considering how to jointly learn multiple metrics.

Video-based face recognition underwent explosive developments in recent years [6, 7, 10, 15, 20, 34, 35, 37]. Such approaches usually consider face images in a video as an image set. Two major tasks are discussed in these methods: how to represent an image set and how to measure the similarity between two set representations. In the view of image set representation, these algorithms can be divided into subspace [6, 15, 20] and manifold [7, 10, 34, 35] based methods. Accordingly, principal angles [20] or the closest Euclidean distance [6, 15] are often used as the metric.

3. Overview

In this section, we take video based face verification as an example to briefly introduce our proposed method, and the image based face verification can be considered a special case by assuming each video only contains one frame.

For the input of our system, we first crop out face images by using Viola-Jones face detector [33] and then roughly align them by fixing the coordinates of automatically detected facial feature points [37][2]. As shown in Fig. 1, to address the misalignment problem, we partition each video into several spatial-temporal volumes and then represent each volume by using a set of position-free patches sampled from each keyframe within the volume without considering the spatial positions of these patches. For two videos, we only compare the corresponding volumes at the same spatial position, thus leading to the robustness to the small misalignments within the volumes.

To extract features from each volume, we employ the bag-of-feature model. Specifically, in the training stage we learn an overcomplete visual vocabulary by using k-means clustering. In the feature extraction stage, all the sampled

[2]In this work, we use the roughly aligned faces provided in the LFW and YTF datasets.

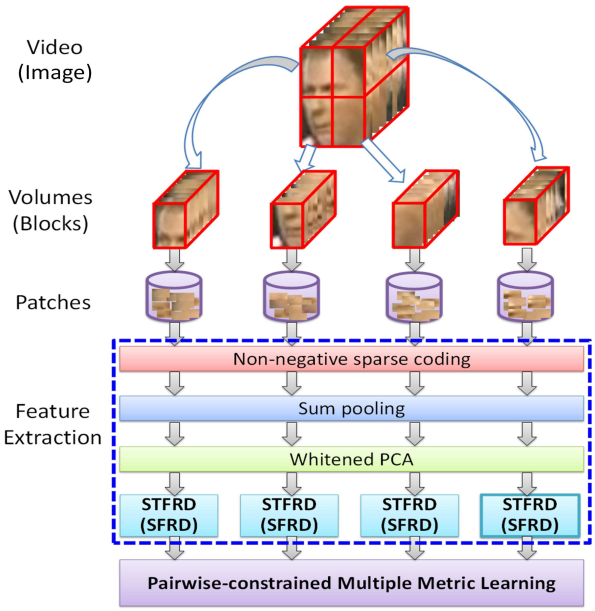


Figure 1: Illustration of the proposed approach for video based face verification. The image based face verification is a special case by treating each video as one frame.

patches are used as the input for the sparse coding method to obtain their reconstruction coefficients. We then apply sum-pooling over the reconstruction coefficients in each volume to extract the token-frequency feature. After that, whitened principal component analysis is utilized to further reduce the feature dimension, which leads to our spatial-temporal face region descriptor. For still images, spatial face region descriptors are extracted by a similar process without considering the temporal dimension. More details for extracting these descriptors can be found in Seciton 4.

Moreover, we further propose a pairwise-constrained multiple metric learning method to integrate the descriptors from all the volumes in each video. Different from the existing methods which can only learn one distance metric for one type of descriptor, we simultaneously learn multiple distance metrics, which can better exploit the correlations among the descriptors from different volumes. The detailed formulation and algorithm of PMML are introduced in Section 5.

Given a pair of samples, we decide whether they are from the same subject or not by calculating the distance between them using the learnt metrics. Moreover, by using multiple volume-partitioning modes, we can obtain multiple distances for each pair of samples, which can be directly used as the feature to train a classifier (*e.g.* SVM or Fisherface) for face recognition.

4. Face Region Descriptor

In this section, we introduce the details for extracting SFRDs and STFRDs from still images and videos respectively.

4.1. Spatial Face Region Descriptor

To enhance the facial textons and suppress Gaussian noises, we first apply a DoG filter on each face image. Then we uniformly partition each face image into K spatial blocks. For each block, a set of patches are sampled from a fixed grid of positions. We construct the intensity feature vector from each patch by sampling the pixels column by column, and each feature vector is further subtracted by its mean and divided by its L_2-norm for reducing the illumination changes and suppressing the scale effect in the subsequential sparse coding procedure. For better presentation, we still call these feature vectors as "*patches*" hereinafter unless necessary. Moreover, we denote each spatial block as S_k for $k = 1, \ldots, K$ and also denote a patch as $\mathbf{p} \in \mathbb{R}^D$ where D is the length of the intensity feature vector of the patch.

Nonnegative sparse coding: Since each patch only captures subtle facial textons, many patches are visually similar to each other. Directly using the intensity feature for face recognition may not be favorable, because even a tiny perturbation(e.g. expressions or noises) might induce a considerable change on the distances between two patches. To improve the robustness and enhance the discriminability, we employ the sparse coding method to map the patches from the low dimension feature space to a high dimension feature space.

In order to perform encoding, an overcomplete visual vocabulary (or dictionary) should be learnt in advance by using the clustering algorithm (*e.g.*, k-means). Let us denote the pre-learnt dictionary as $\mathbf{D} \in \mathbb{R}^{D \times M}$, where M is the number of visual words in the dictionary and $M \gg D$. Given a patch $\mathbf{p}$, to encode it into a high dimensional code, we employ the sparse coding method with nonnegative constraints, which can be formulated as follows:

$$\min_{\mathbf{c}} \|\mathbf{p} - \mathbf{D}\mathbf{c}\|^2 + \lambda \|\mathbf{c}\|_1, \text{s.t.} \quad \mathbf{c} \geq 0, \tag{1}$$

where $\mathbf{p}$ is a patch and $\mathbf{c}$ is the vector of reconstructing coefficients (*i.e.*, the code). The nonnegative constraint on the code $\mathbf{c}$ allows us to directly sum a set of codes without considering their negative values. To solve the above model, in our work we employ a greedy algorithm called Least Angle Regression (LAR) [13].

Sum pooling: We have represented each block as a set of high dimensional codes. To effectively describe the statistical properties of these position-free codes in each block, we extract the TF feature for each block by using the sum-pooling method. Specifically, the TF feature of the k-th block can be obtained as,

$$\mathbf{x}^k = \sum_{\mathbf{c}_i \in S_k} \mathbf{c}_i, \quad k = 1, 2, \cdots, K, \tag{2}$$

where $\mathbf{c}_i$ is the sparse code of the i-th patch in S_k.

Whitened PCA: The dimension of the pooled TF feature (*i.e.*, $\mathbf{x}^k$) can be very high since the dictionary is usually overcomplete (*i.e.*, M is much larger than D). To reduce the feature redundancy for efficient face recognition, we therefore seek a compact representation. A possible way is to apply Principle Component Analysis (PCA), which can remove some noises by discarding the eigenvectors corresponding to small eigenvalues. However, PCA is easily affected by high-frequent visual words [19]. Especially, for face images, the same visual word usually recurs many times in smooth facial areas, *e.g.* the cheek and the forehead. These high-frequent visual words contribute strong responses for the corresponding eigenvectors and eigenvalues when using PCA. A better way is to use Whitened PCA (*i.e.* WPCA) which suppresses the responses from larger eigenvalues. Formally, let us denote the covariance matrix of the TF features for the k-th block of all the training images as $\mathbf{C}_k \in \mathbb{R}^{M\times M}$ and represent the eigen-decomposition of the covariance matrix as $\mathbf{C}_k = \mathbf{P}_k\mathbf{\Lambda}_k\mathbf{P}_k^T$ where $\mathbf{\Lambda}_k \in \mathbb{R}^{M\times M}$ is a diagonal matrix of eigenvalues and $\mathbf{P}_k \in \mathbb{R}^{M\times M}$ is the matrix constructed from the corresponding eigenvectors. Then we obtain the SFRD $\mathbf{z}^k$ for the k-th block by using WPCA as follows:

$$\mathbf{z}^k = \frac{\tilde{\mathbf{\Lambda}}_k^{-\frac{1}{2}}\tilde{\mathbf{P}}_k^T\mathbf{x}^k}{\|\tilde{\mathbf{\Lambda}}_k^{-\frac{1}{2}}\tilde{\mathbf{P}}_k^T\mathbf{x}^k\|_2}, \tag{3}$$

where $\tilde{\mathbf{\Lambda}}_k \in \mathbb{R}^{m\times m}$ is a diagonal matrix of the $m < M$ largest eigenvalues and $\tilde{\mathbf{P}}_k \in \mathbb{R}^{M\times m}$ is the matrix of the corresponding eigenvectors. Thus we obtain a compact representation (*i.e.*, $\mathbf{z}^k$) for the k-th block of the face image.

4.2. Spatial-Temporal Face Region Descriptor

For the videos, there are many similar patches along the temporal dimension. Therefore, instead of using pixelwise sampling, we adopt the sparse sampling along the spatial dimension with a fixed step because it is sufficient to capture the texton information. We then perform the nonnegative sparse coding as similarly used in SFRD. In the pooling step, the sum pooling is applied to all the patches within the entire spatial-temporal volume along both the spatial and temporal dimensions. Such a spatial-temporal pooling strategy essentially characterizes the statistics of a certain region of the face in the video. Formally we extract the TF feature from each volume as:

$$\mathbf{x}^k = \frac{1}{V}\sum_{v=1}^{V}\sum_{\mathbf{c}_i\in S_k^v}\mathbf{c}_i, \quad k = 1, 2, \cdots, K, \tag{4}$$

where V is the number of keyframes of the video, S_k^v is the v-th keyframe in the k-th volume, and $\mathbf{c}_i$ represents the i-th code in S_k^v. Finally, we perform WPCA on these TF features to extract the STFRDs for the video.

5. Pairwise-constrained Multiple Metric Learning

With above descriptors extracted for each block, we need to further consider distance metric to compare them, as well as the method fusing these block-wise descriptors. For this purpose, we propose a new multiple metric learning method to jointly learn a discriminative distance. Formally, let us denote each image/video as $\mathbf{z} = \{\mathbf{z}^1, \ldots, \mathbf{z}^K\}$ where $\mathbf{z}^k|_{k=1}^K$ is the face region descriptor and K is the total number of blocks/volumns for each image/video. Given a pair of training samples, we define the distance as:

$$d_{\mathbf{W}_1,\ldots,\mathbf{W}_K}(\mathbf{z}_i, \mathbf{z}_j) = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{z}_i^k - \mathbf{z}_j^k)^T\mathbf{W}_k(\mathbf{z}_i^k - \mathbf{z}_j^k), \tag{5}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are respectively the i-th and j-th samples (*i.e.*, two sets of face region descriptors from two images/videos), $\mathbf{z}_i^k$ and $\mathbf{z}_j^k$ are respectively the k-th face region descriptors for the i-th and j-th samples, and $\mathbf{W}_k$ is the Mahalanobis matrix. Intuitively, given two descriptors $\mathbf{z}_i$ and $\mathbf{z}_j$ of two images/videos, their learnt distance $d(\mathbf{z}_i, \mathbf{z}_j)$ should be smaller than a certain threshold ρ when the two images/videos are from the same person, and it must be larger than the threshold ρ when the two images/videos are from different subjects. Therefore, our task is to jointly optimize the unknown $\mathbf{W}_k$ for $k = 1, \ldots, K$ using the provided intra-class pairs and inter-class pairs.

An alternative way is to separately learn a distance metrics $\mathbf{W}_k$ for each face region descriptor, and then fuse the K distances as the final distance. However, considering that these K face region descriptors are from the same face image, jointly learning these distance metrics is more favorable by implicitly taking global information into account. Another possible way is to concatenate the K descriptors together into a single feature vector for each sample, however, which results in a very large distance metric ($Km\times Km$). In contrast, our proposed method reduces the number of parameters, and hence improves the generalization capability.

5.1. Formulation

In distance metric learning methods, a regularizer is usually imposed on the Mahalanobis matrix to prevent overfitting due to the small training set and high model complexity. In this work, we adopt the LogDet divergence [12], *i.e.*,

$$H(\mathbf{W}_k, \mathbf{W}_0) = \mathrm{tr}(\mathbf{W}_k\mathbf{W}_0^{-1}) - \log|\mathbf{W}_k\mathbf{W}_0^{-1}| - m, \tag{6}$$

where $\mathrm{tr}(\cdot)$ is the trace norm, $|\cdot|$ is the matrix determinate and m is the dimension of the descriptors. On one hand, it controls the complexity of Mahalanobis matrix $\mathbf{W}_k$ by forcing it to be close to a given matrix $\mathbf{W}_0$, which is usually defined as the identity matrix. On the other hand, the LogDet divergence implicitly pushes $\mathbf{W}_k$ to be a symmetric positive definite matrix during the optimization [22].

With the regularizer in (6), we formulate the multiple metric learning problem as follows:

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_K} \quad \frac{1}{K}\sum_{k=1}^{K} H(\mathbf{W}_k, \mathbf{W}_0), \tag{7}$$

$$\text{s.t.} \quad \frac{\delta_{ij}}{K}\sum_{k=1}^{K} d_{\mathbf{W}_k}(\mathbf{z}_i^k, \mathbf{z}_j^k) \le \delta_{ij}\rho - \tau, \tag{8}$$

where $d_{\mathbf{W}_k} = (\mathbf{z}_i^k - \mathbf{z}_j^k)^T\mathbf{W}_k(\mathbf{z}_i^k - \mathbf{z}_j^k)$ is the distance between the two face region descriptors, ρ is the threshold for distance comparison, $\tau > 0$ is the margin, $\delta_{ij} = 1$ if the two samples belong to the same subject, and $\delta_{ij} = -1$ otherwise. In other words, the distance of the two samples should be smaller than $\rho - \tau$ if they are from the same subject, and larger than $\rho + \tau$ otherwise.

However, there may not exist a feasible solution to the above problem in some cases. To address this problem, we introduce a slack variable ξ_{ij} for each constraint, and reformulate the objective as:

$$\min_{\mathbf{W}_k,\xi_{ij}} \quad \frac{1}{K}\sum_{k=1}^{K} H(\mathbf{W}_k, \mathbf{W}_0) + \frac{\gamma}{n}\sum_{i,j}\ell(\xi_{ij}, \delta_{ij}\rho - \tau), \tag{9}$$

$$\text{s.t.} \quad \frac{\delta_{ij}}{K}\sum_{k=1}^{K} d_{\mathbf{W}_k}(\mathbf{z}_i^k, \mathbf{z}_j^k) \le \xi_{ij}, \tag{10}$$

where n is the number of pairs of training samples, γ is a tradeoff parameter, and $\ell(\cdot,\cdot)$ is the first order continuous differentiable hinge loss function defined as:

$$\ell(x, x_0) = \begin{cases} 0 & x \le x_0 \\ (x - x_0)^2 & x > x_0. \end{cases} \tag{11}$$

Compared with the original LogDet divergence loss in [12], this hinge loss not only avoids unnecessary penalties on those pairs that already satisfy the constraint, but also accelerates the algorithm when only considering the violated constraints in each cyclic projection (See Section 5.2).

5.2. Optimization

To solve the problem in (9), we adopt the cyclic projection method [3, 5], in which we iteratively update $\mathbf{W}_k$ and ξ_{ij} by projecting the current solution into the feasible domain of one violated constraint. The convergence of cyclic projection method can be found in [3,5]. The updating rules for our PMML are shown in the following proposition:

Proposition 1. *Given the solution at the t-th iteration $\mathbf{W}_k^t$ for $k = 1, \ldots, K$, if there exists a pair of descriptors ($\mathbf{z}_i$, $\mathbf{z}_j$) for which the constraint in (10) is not satisfied, then we update $\mathbf{W}_k$'s and the corresponding ξ_{ij} as follows:*

$$\begin{cases} \mathbf{W}_k^{t+1} = \mathbf{W}_k^t + \mu(\mathbf{W}_k^t(\mathbf{z}_i^k - \mathbf{z}_j^k)(\mathbf{z}_i^k - \mathbf{z}_j^k)^T\mathbf{W}_k^t), & (12) \\ \xi_{ij}^{t+1} = \xi_{ij}^t - \dfrac{n}{2\gamma}\alpha, & (13) \end{cases}$$

where $\mu = \delta_{ij}\alpha/(1 - \delta_{ij}\alpha d_{\mathbf{W}_k^t}(\mathbf{z}_i^k, \mathbf{z}_j^k))$ and α can be solved by

$$\frac{\delta_{ij}}{K}\sum_{k=1}^{K}\frac{d_{\mathbf{W}_k^t}(\mathbf{z}_i^k, \mathbf{z}_j^k)}{1 - \delta_{ij}\alpha d_{\mathbf{W}_k^t}(\mathbf{z}_i^k, \mathbf{z}_j^k)} - (\xi_{ij}^t - \frac{n}{2\gamma}\alpha) = 0. \tag{14}$$

Proof. Based on the cyclic projection method [3,5], we can project our current solution $\mathbf{W}_k^t$ into the feasible domain of the violated constraint to obtain $\mathbf{W}_k^{t+1}$ and simultaneously update ξ_{ij} by solving the following equations [22]:

$$\begin{cases} \nabla H(\mathbf{W}_k^{t+1}) = \nabla H(\mathbf{W}_k^t) + \alpha\delta_{ij}(\mathbf{z}_i^k - \mathbf{z}_j^k)(\mathbf{z}_i^k - \mathbf{z}_j^k)^T, & (15) \\ \dfrac{\gamma}{n}\nabla\ell(\xi_{ij}^{t+1}) = \dfrac{\gamma}{n}\nabla\ell(\xi_{ij}^t) - \alpha, & (16) \\ \dfrac{\delta_{ij}}{K}\displaystyle\sum_{k=1}^{K}(\mathbf{z}_i^k - \mathbf{z}_j^k)^T\mathbf{W}_k^{t+1}(\mathbf{z}_i^k - \mathbf{z}_j^k) = \xi_{ij}^{t+1}. & (17) \end{cases}$$

Then, we can derive (12) and (13) from (15) and (16), respectively. Substituting (12) and (13) into (17), we obtain the the equation related to α as in (14). □

We list the algorithm in Algorithm 1. The main time cost is to update $\mathbf{W}_k^{t+1}$ in the step 6, which is $O(Km^2)$ for each constraint. Therefore, the total time cost is $O(LKm^2)$ where L is the total number of the updating in Step 6 executed by the algorithm. In practice, L is not very large since we only need to deal with the pairs for which the constrains are not satisfied.

Algorithm 1 Pairwise-constrained Multiple Metric Learning

Input: Training pairs $\{(\mathbf{z}_i^k, \mathbf{z}_j^k), \delta_{ij}\}$, and ρ, τ, γ, $\mathbf{W}_0$

1: Initialize $t = 1$, $\mathbf{W}_k^1 = \mathbf{W}_0$, $\eta_{ij} = 0$, $\xi_{ij}^t = \delta_{ij}\rho - \tau$.
2: **repeat**
3: Pick a pair of samples ($\mathbf{z}_i$, $\mathbf{z}_j$) and compute the distances $d_{\mathbf{W}_k^t}(\mathbf{z}_i^k, \mathbf{z}_j^k)$ for $k = 1, \ldots, K$.
4: **if** $\frac{\delta_{ij}}{K}\sum_{k=1}^{K} d_{\mathbf{W}_k^t}(\mathbf{z}_i^k, \mathbf{z}_j^k) > \delta_{ij}\rho - \tau$ **then**
5: Solve α in (14), and set $\alpha \leftarrow \min(\alpha, \eta_{ij})$ and $\eta_{ij} \leftarrow \eta_{ij} - \alpha$
6: Update $\mathbf{W}_k^{t+1}$ by using (12) and set $\mathbf{W}_k = \mathbf{W}_k^{t+1}$ for $k = 1, \ldots, K$.
7: Update ξ_{ij}^{t+1} by using (13).
8: $t = t + 1$.
9: **end if**
10: **until** The objective converges.

Output: Mahalanobis matrices $\mathbf{W}_1, \ldots, \mathbf{W}_K$.

6. Experiments

We evaluate our method for the unconstrained face verification task by using two real-world face datasets: LFW [16] and YTF [37] datasets.

6.1. Experimental Setup

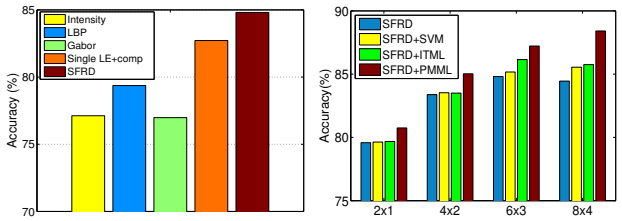
LFW is an image dataset for unconstrained face verification. It contains more than 13,000 face images collected from the web with large variations in pose, age, expression, illumination, etc. We consider the restricted protocol and follow the standard setting in [16], which splits the dataset into ten subsets with each subset containing 300 intra-class pairs and 300 inter-class pairs. The performances are measured by using 10-fold cross validation.

YTF is a video dataset which contains 3,425 videos of 1,595 different subjects downloaded from YouTube. The average length of each video clip is about 180 frames. It is collected for unconstrained VFR and also contains large variations in pose, age, expression, illumination, etc. We also consider the restricted mode and strictly follow the standard setting in [37]. We use the 5,000 video pairs randomly selected in [37] for unconstrained face verification, in which one half of the pairs of videos are from the same subject, while the remaining half of the pairs are from different subjects. These pairs are also officially divided into 10 splits [37] with each split containing 250 intra-class pairs and 250 inter-class pairs.

We directly crop the face images according to the provided data and then resize them into 110×60 pixels for LFW and 40×24 pixels for YTF. The DoG filter is set to $\sigma_1 = 0$ (*i.e.* no filter) and $\sigma_2 = 2$. The sampling template size is set to 9×9 pixels. The parameter λ in sparse coding is set to 0.1. For the dictionary size, we set $M = 512$ as the default value for a good tradeoff between the performance and the efficiency (see Fig. 4(a)). The dimension of WPCA is set to $m = 60$. For our PMML, the parameters τ and γ are decided by using the cross validation within the range of $\{1.0, 1.2, 1.4, 1.6\}$ and the range of $\{0.1, 1, 50, 100\}$, respectively. To fuse multiple partition modes, the SVM classifier is used with RBF kernel by setting the bandwidth parameter as the mean distance and the regularization parameter $C = 1$.

6.2. Face Region Descriptor and PMML

Our approach consists of two key components: face region descriptors and the learning algorithm PMML. So we evaluate our approach from the two aspects by taking LFW dataset as an example. For descriptors, we compare our spatial face region descriptor (SFRD) with four types of features: intensity feature, LBP and Gabor in [27] as well as "Single LE + comp" in [4], in which the first three features are extracted from the whole face images and Single LE is extracted from facial components (*e.g.* eyes, nose, etc.). For our SFRD, we set the number of blocks for each image as $K = 8 \times 4$. Following [4, 27], the performances are measured by using 10-fold cross validation with Euclidean distance and we report the accuracies in Fig. 2(a). The results of intensity, LBP and Gabor are from [27] and the result of



(a) Comparisons with other features using the Euclidean distance.

(b) Comparisons with other metrics using our SFRD feature.

Figure 2: Comparisons between our work and other features and metrics on the LFW dataset.

"Single LE + comp" is from [4]. We observe that our SFRD is the best and the improvement over the second best (*i.e.* "Single LE + comp") is 2.09% (84.81% *v.s.* 82.72%) while "Single LE + comp" requires more accurate face alignment technique for cropping the facial components. Moreover, we also observe that the learnt descriptors (*i.e.* our SFRD and "Single LE + comp") are better than the hand-crafted descriptors (*i.e.* intensity feature, LBP and Gabor), which indicates that it is beneficial to learn the descriptors for unconstrained face recognition.

For the distance metric learning method, we compare our PMML with three baselines: SFRD, SFRD+SVM and SFRD+ITML. SFRD uses the Euclidean distance based on the original spatial face region descriptor; SFRD+SVM uses the SVM classifier for multiple SFRDs; SFRD+ITML is the average of distance metrics which are individually learnt by using ITML [12] for each face region descriptor. Fig. 2(b) shows the results with different partition modes, *i.e.*, 2×1, 4×2, 6×3 and 8×4. Our SFRD+PMML is the best for all cases, which demonstrates that PMML can better integrate the descriptors from different facial regions. It is worth noting that *our method only using* 8×4 *blocks achieves the accuracy of* ***88.41%****, which outperforms the current state-of-the-art on LFW dataset by using the restricted protocol*[3]. Moreover, our method only uses one single type of feature without requiring any reference set.

6.3. Experimental Comparisons on LFW

We compare our proposed method with the state-of-the-art methods [4, 8, 14, 23, 27, 28, 30, 38, 42] under the restricted protocol. Some baseline methods exploited multiple features to improve the performances, *e.g.* "CSML+SVM" [27] fused three features: intensity, LBP and Gabor, "High-Throughput Brain-Inspired Features" [8] selected the robust features by searching a large set of features, "Multiple LE + comp" [4] fused multiple local descriptors extracted from different facial regions, "DML-eig combined" [42] combined four types of descriptors. Moreover,

[3]Currently, the best result in the official website of LFW is 88.13% as reported in [8].

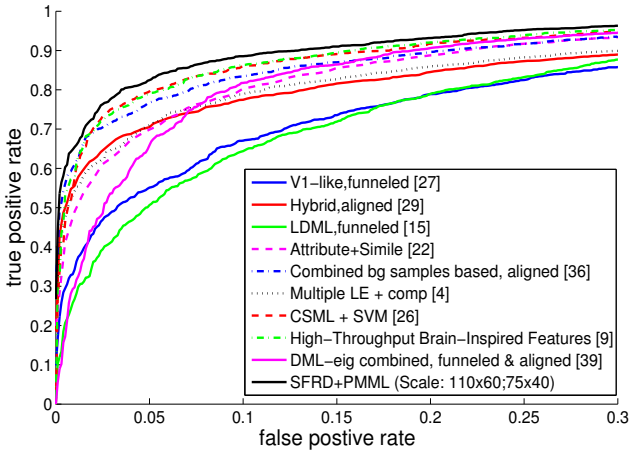


Figure 3: Comparisons of ROC curves between our work and the state-of-the-art methods on the LFW dataset by using the standard restricted protocol.

some work also learnt the metrics, *e.g.*, LDML [14], DML [42], and cosine metric [27]. For our SFRD+PMML, we use four different partitions, 2×1, 4×2, 6×3 and 8×4, and two different scales of face images, *i.e.*, 110×60 and 75×40, and finally concatenate the eight distances for each pair of samples as the feature for the SVM classifier.

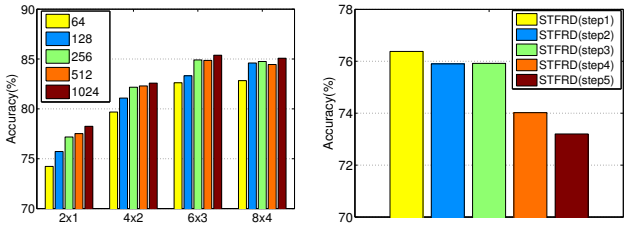
The ROC curves of different methods are shown in Fig. 3. The results of baselines are obtained from the official website of LFW. The improvement of our method over other state-of-the-art methods is about 2∼3 percentages when the false positive rate ranges from 0.1 to 0.2. Note that *the recognition accuracy of our method reaches* ***89.35%****, which is better than the current state-of-the-art on LFW dataset by* ***1.22%*** *using the restricted protocol*. The above results clearly demonstrate the effectiveness of the proposed SFRD descriptor and the pairwise-constrained multiple metric learning method.

6.4. Experimental Comparisons on YTF

For video-based face verification on the YTF dataset, we compare our proposed method with the several existing VFR methods, including MSM [39], DCC(pair) [20], MMD [35], AHISD [6], CHISD [6] and SANP [15]. We run experiments using the source codes provided by the authors and follow their parameter settings. Among them, an important parameter is the PCA dimension which is used in the image set representation. We search the PCA energy in the range of $\{80\%, 85\%, 90\%, 95\%\}$, and report the best results for their methods. For MMD, the threshold θ is searched in the range of $\{1.1, 1.4, 1.7, 2.0\}$. Since DCC [20] was not specifically designed for face verification , we modify it by constructing the within-class scatter matrix from intra-class pairs and the between-class scatter matrix form inter-class pairs. For our method, due to the small image size, we use three partitions, $K = 2 \times 1$, 4×2 and 6×3. The sparsely sampling step is set as 3 pixels, which can achieve the comparable performance with denser sampling steps as shown in Fig. 4(b). The other parameters are the same as that in the LFW dataset.

Table 1: The comparisons on the YouTube Faces dataset (Mean Accuracy ± Standard Deviation in %).

Method	Result
LBP(Min dist) [37]	65.70±1.70
MBGS+LBP [37]	76.40±1.80
MSM [39]	62.54±1.47
DCC(pair) [20]	70.84±1.57
MMD [35]	64.96±1.00
AHISD [6]	66.50±2.03
CHISD [6]	66.24±1.70
SANP [15]	63.74±1.69
STFRD	75.92±2.00
STFRD+PMML	**79.48±2.52**



(a) (b)

Figure 4: The performances of our SFRD and STFRD features using the Euclidean distance under different settings: (a) different codebook sizes on the LFW dataset. (b) different sampling steps on the YTF dataset.

The recognition accuracies and standard deviations of different methods are reported in Table 1. We also report the results of two methods from [37], LBP(Min dist) and MBGS+LBP, which are the current published state-of-the-art results on YTF [4]. Note that LBP(Min dist) and MBGS+LBP are generally better than other baselines, possibly because higher resolution face images were used in their methods. From this table, the performance of STFRD can reach 75.92%, which demonstrates the effectiveness of our proposed STFRD. By combining STFRD with PMML, the performance can be further improved to 79.48%, which outperforms the current state-of-the-art methods on the YTF dataset.

7. Conclusion

In this paper, we have proposed two robust face region descriptors SFRD and STFRD for image-based and video-based face recognition, respectively. To handle the misalignment problem, we partition each image (*resp.* video)

[4] More results can be found in: http://www.cs.tau.ac.il/ wolf/ytfaces/

into several spatial blocks (*resp.* spatial-temporal volumes), and then apply the BoF model and the sum pooling method in each block (*resp.* each volume) to extract the TF features. WPCA is finally adopted to generate robust face region descriptors. Furthermore, we develop a new pairwise-constrained multiple metric learning (PMML) method to integrate the face region descriptors from different regions. Our proposed method achieves the state-of-the-art performances on two public real-world datasets LFW and YTF.

Acknowledgement

The work is partially supported by National Basic Research Program of China (973 Program) under contract 2009CB320902; Natural Science Foundation of China under contracts nos. 61025010, 61173065, 61222211 and 61202297. This research is also partially supported by Multi-plAtform Game Innovation Centre (MAGIC) in Nanyang Technological University. MAGIC is funded by the Interactive Digital Media Programme Office (IDMPO) hosted by the Media Development Authority of Singapore.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *T-PAMI*, 28(12):2037–2041, 2006.

[2] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *T-PAMI*, 19(7):711–720, 1997.

[3] L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[4] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, 2010.

[5] Y. Censor and S. Zenios. *Parallel optimization: Theory, algorithms, and applications*. Oxford University Press, USA, 1998.

[6] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, 2010.

[7] Y. Chen, V. Patel, P. Phillips, and R. Chellappa. Dictionary-based face recognition from video. *ECCV*, 2012.

[8] D. Cox and N. Pinto. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *FG*, 2011.

[9] Z. Cui, S. Shan, X. Chen, and L. Zhang. Sparsely encoded local descriptor for face recognition. In *FG*, 2011.

[10] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for video-based face recognition. In *CVPR*, 2012.

[11] J. Daugman et al. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Optical Society of America, Journal, A: Optics and Image Science*, 2:1160–1169, 1985.

[12] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[13] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[14] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

[15] Y. Hu, A. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, 2011.

[16] G. Huang, M. Mattar, T. Berg, E. Learned-Miller, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[17] Y. Huang, D. Xu, and T. J. Cham. Face and human gait recognition using image-to-class distance. *T-CSVT*, 20(3):431–438, 2010.

[18] Y. Huang, D. Xu, and F. Nie. Patch distribution compatible semi-supervised dimension reduction for face and human gait recognition. *T-CSVT*, 22(3):479–488, 2012.

[19] H. Jégou, O. Chum, et al. Negative evidences and co-occurrences in image retrieval: the benefit of pca and whitening. In *ECCV*, 2012.

[20] T. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *T-PAMI*, 29(6):1005–1018, 2007.

[21] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[22] B. Kulis, M. Sustik, and I. Dhillon. Low-rank kernel learning with bregman matrix divergences. *JMLR*, 10:341–376, 2009.

[23] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[24] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[25] Z. Li, J. Imai, and M. Kaneko. Robust face recognition using block-based bag of words. In *ICPR*, 2010.

[26] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012.

[27] H. Nguyen and L. Bai. Cosine similarity metric learning for face verification. *ACCV*, 2011.

[28] N. Pinto, J. DiCarlo, and D. Cox. How far can you get with a modern face recognition test set using only simple features? In *CVPR*, 2009.

[29] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[30] Y. Taigman, L. Wolf, and T. Hassner. Multiple one-shots for utilizing class label information. In *BMVC*, 2009.

[31] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[32] S. ul Hussain, T. Napoléon, and F. Jurie. Face recognition using local quantized patterns. In *BMVC*, 2012.

[33] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[34] R. Wang and X. Chen. Manifold discriminant analysis. In *CVPR*, 2009.

[35] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, 2008.

[36] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

[37] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011.

[38] L. Wolf, T. Hassner, and Y. Taigman. Similarity scores based on background samples. *ACCV*, 2010.

[39] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *FG*, 1998.

[40] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang. Multilinear discriminant analysis for face recognition. *T-IP*, 16(1):212–220, 2007.

[41] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *CVPR*, 2011.

[42] Y. Ying and P. Li. Distance metric learning with eigenvalue optimization. *JMLR*, 13:1–26, 2012.

[43] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *T-PAMI*, 29(6):915–928, 2007.