Similarity Metric Learning for Face Recognition

Qiong Cao, Yiming Ying
Department of Computer Science
University of Exeter, UK
{qc218,y.ying}@exeter.ac.uk

Peng Li
Department of Engineering Mathematics
University of Bristol, UK
lileopold@gmail.com

Abstract

Recently, there is a considerable amount of efforts devoted to the problem of unconstrained face verification, where the task is to predict whether pairs of images are from the same person or not. This problem is challenging and difficult due to the large variations in face images. In this paper, we develop a novel regularization framework to learn similarity metrics for unconstrained face verification. We formulate its objective function by incorporating the robustness to the large intra-personal variations and the discriminative power of novel similarity metrics. In addition, our formulation is a convex optimization problem which guarantees the existence of its global solution. Experiments show that our proposed method achieves the state-of-the-art results on the challenging Labeled Faces in the Wild (LFW) database [10].

1. Introduction

Face recognition has attracted increasing attentions due to its applications in biometrics and surveillance. Recently, considerable research efforts are devoted to the unconstrained face verification problem [8, 17, 18, 20, 23, 24], the task of which is to predict whether two face images represent the same person or not. The face images are taken under unconstrained conditions and show significant variations in complex background, lighting, pose, and expression (see e.g. Figure 1). In addition, the evaluation procedure for face verification typically assumes that the person identities in the training and test sets are exclusive, requiring the prediction of never-seen-before faces. Both factors make face verification very challenging.

Similarity metric learning aims to learn an appropriate distance or similarity measure to compare pairs of examples. This provides a natural solution for the verification task. Metric learning [5,7,22,25,26] usually focuses on the (squared) Mahalanobis distance defined, for any $x, t \in \mathbb{R}^d$, by $d_M(x,t) = (x-t)^T M(x-t)$, where M is a positive semi-definite (p.s.d.) matrix. It was observed in [8, 27] that directly applying metric learning methods only yields a modest performance for face verification. This may be partly because most of such methods deal with the specific tasks of improving kNN classification, which may be not necessarily suitable for face verification. Similarity learning aims to learn the bilinear similarity function [3, 19] defined by $s_M(x,t) = x^T M t$ or the cosine similarity $CS_M(x,t) = x^T M t / \left(\sqrt{x^T M x}\sqrt{t^T M t}\right)$ [14], which has successful applications in image searching and face verification.

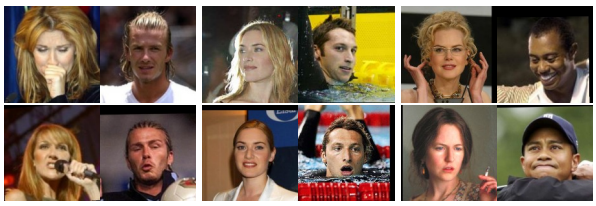

Figure 1: Example images from the Labeled Faces in the Wild (LFW) database exhibit large intra-personal variations: each column is a pair of images from the same person.

In this paper, we build on previous studies [7, 8, 11, 14, 22, 25, 27] to show the great potential of similarity metric learning methods to boost the verification performance using low-level feature descriptors such as Scale-Invariant Feature Transform (SIFT) [8] and Local Binary Pattern (LBP) [16]. To this end, we develop a novel regularization framework to learn similarity metrics for unconstrained face verification, which is referred to as similarity metric learning over the intra-personal subspace. We formulate its objective function by considering both the robustness to the large intra-personal variations and the discriminative power, a property that most metric learning methods do not hold. In addition, our formulation is a convex optimization problem, and hence a global solution can be efficiently found by existing algorithms. This is, for instance, not the case for the current similarity metric learning model [14].

We report experimental results on the Labeled Faces in the Wild (LFW) [10] dataset, a standard testbed for un-

constrained face verification. The face images collected directly from the website Yahoo! News contain significant intra-personal differences that may be encountered in our daily life. Our proposed method achieves $\mathbf{89.73}\%$ in the restricted setting, which outperforms the current best result 88.13% in [18]. Shifting to the unrestricted setting, our method achieves 90.75%, which is competitive with the current state-of-the-art result 90.90% in [4].

The paper is organized as follows. Section 2 presents the proposed model and Section 3 discusses the related work. Experimental results are reported in Section 4. Section 5 concludes the paper.

Notations: For any $X, Y \in \mathbb{R}^{d\times n}$, $\langle X, Y\rangle = \mathbf{Tr}(X^\top Y)$ where $\mathbf{Tr}(\cdot)$ denotes the trace of a matrix. The space of symmetric d times d matrices is denoted by $\mathbb{S}^d$, and the set of positive semi-definite matrices is denoted by $\mathbb{S}^d_+$. The standard Euclidean norm on vectors is denoted by $\|\cdot\|$ and the Frobenius norm on matrices by $\|\cdot\|_F$. In the following sections, a face image is represented by a feature vector in $\mathbb{R}^p$. The notations $\mathcal{S}$ and $\mathcal{D}$, respectively, denote the index set of similar pairs (from the same person) and that of dissimilar pairs (from different persons), *i.e.* $(i, j) \in \mathcal{S}$ means a similar image-pair (x_i, x_j).

2. Similarity Metric Learning Over the Intra-Personal Subspace

In this section, we develop a new method of learning a similarity metric for face verification, which will be described step by step as follows.

2.1. Formulation of the Learning Problem

To obtain a good similarity function measuring the similarity between face images, we formulate the learning objective by considering both the *robustness* to the large intra-personal variations and the *discrimination* for separating similar image-pairs from dissimilar image-pairs.

Robustness. One challenging issue in face verification is to retain the robustness of the similarity metric to the noise and the large intra-personal variations in face images.

To remove the noise, one commonly used method is to apply the principal component analysis (PCA). PCA computes the d eigenvectors with the largest eigenvalues of the covariance matrix defined by $C = \sum_{i=1}^n (x_i - \mathbf{m})(x_i - \mathbf{m})^T \in \mathbb{R}^{p\times p}$, where $\mathbf{m}$ is the mean of the data. The PCA-reduced images are usually referred to as Eigenfaces (e.g. [2]).

To reduce the effect of large intra-personal variations, we follow the idea in [9, 13, 21] by further mapping d-dimensional Eigenfaces to the intra-personal subspace. Specifically, let the intra-personal covariance matrix be defined by

$$C_{\mathcal{S}} = \sum_{(i,j)\in\mathcal{S}} (x_i - x_j)(x_i - x_j)^T, \tag{1}$$

and $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ and $V = (v_1, \cdots, v_k)$ be the top-leading k eigenvalues and eigenvectors of C_S. The mapping of the Eigenfaces to the k-dimensional intra-personal subspace ($k \le d$) is defined by the whitening process:

$$\tilde{x} = \text{diag}(\lambda_1^{-1/2}, \ldots, \lambda_k^{-1/2})V^T x. \tag{2}$$

Note that the features are weighted by the inverse of the eigenvalues, which penalizes the eigenvectors with large eigenvalues and therefore reduces the variance of the features, *i.e.* the intra-personal variations.

Throughout this paper, we only consider the special case where the dimension of the intra-personal subspace equals the dimension of PCA, i.e. $k = d$. In this case, if $C_{\mathcal{S}}$ is invertible and denote

$$L_{\mathcal{S}} = V\text{diag}(\lambda_1^{1/2}, \ldots, \lambda_d^{1/2}), \tag{3}$$

then $C_{\mathcal{S}} = L_{\mathcal{S}} L_{\mathcal{S}}^T$ and equation (2) becomes $\tilde{x} = L_{\mathcal{S}}^{-1} x$.

Discrimination. After the images are mapped to the intra-personal subspace, we now consider the discrimination using a similarity metric function, a property that discriminates similar image-pairs from dissimilar image-pairs. To this end, one option is to use the cosine similarity function CS_M which was observed to outperform the distance measurement d_M in face verification [14]. However, it is not a convex function with respect to M. Recent studies [3, 19] observed that the similarity function s_M has a promising performance on image similarity search. Motivated by these observations, we combine the similarity function s_M and the distance d_M and propose a *generalized similarity metric* $f_{(M,G)}$ to measure the similarity of an image pair $(\tilde{x}_i, \tilde{x}_j)$:

$$f_{(M,G)}(\tilde{x}_i, \tilde{x}_j) = s_G(\tilde{x}_i, \tilde{x}_j) - d_M(\tilde{x}_i, \tilde{x}_j). \tag{4}$$

Apparently, $f_{(M,G)}$ is linear and convex with respect to variable (M, G).

Let $\mathcal{P} = \mathcal{S} \cup \mathcal{D}$ denotes the index set of all pairwise constraints. If image $\tilde{x}_i$ is similar to $\tilde{x}_j$ (*i.e.* images from the same individual), define its associated binary output $y_{ij} = 1$ and -1 otherwise. To better discriminate similar image-pairs from dissimilar image-pairs, we should learn M and G from the available data such that $f_{(M,G)}(\tilde{x}_i, \tilde{x}_j)$ reports a large score for $y_{ij} = 1$ and a small score otherwise. Based on this rationale, we derive the formulation of the empirical discrimination using the hinge loss:

$$\mathcal{E}_{\text{emp}}(M, G) = \sum_{(i,j)\in\mathcal{P}} (1 - y_{ij} f_{(M,G)}(\tilde{x}_i, \tilde{x}_j))_+. \tag{5}$$

Minimizing the above empirical error with respect to M and G will encourage the discrimination of similar image-pairs from dissimilar ones. However, directly minimizing the functional $\mathcal{E}_{\text{emp}}$ does not guarantee a robust similar metric $f_{(M,G)}$ to large intra-personal variations and also will lead to overfitting. Below, we propose a novel regularization framework which learns a robust and discriminative similarity metric.

Proposed Regularization Framework. Based on the above discussions, our target now is to learn matrices M and G such that $f_{(M,G)}$ not only retains the robustness to the large intra-personal variations but also preserves a good discriminative information. To this end, we propose a new method referred to as *similarity metric learning over the intra-personal subspace* which is given by

$$\min_{M,G\in\mathbb{S}^d} \mathcal{E}_{\text{emp}}(M,G) + \frac{\gamma}{2}(\|M-I\|_F^2 + \|G-I\|_F^2). \quad (6)$$

By introducing the slacking variables, the above formulation is identical to:

$$\begin{aligned} \min_{M,G\in\mathbb{S}^d} \quad & \sum_{t\in\mathcal{P}} \xi_t + \frac{\gamma}{2}(\|M-I\|_F^2 + \|G-I\|_F^2), \\ \text{s.t.} \quad & y_{ij}[f_{(M,G)}(\tilde{x}_i,\tilde{x}_j)] \geq 1-\xi_{ij}, \\ & \xi_t \geq 0, \ \ \forall t=(i,j)\in\mathcal{P}. \end{aligned} \quad (7)$$

The regularization term $\|M-I\|_F^2+\|G-I\|_F^2$ in our formulation (7) prevents image vectors $\tilde{x}$ in the intra-personal subspace from being distorted too much, and hence retains the most robustness of the intra-personal subspace. Minimizing the empirical term $\sum_{(i,j)\in\mathcal{P}}\xi_{ij}$ promotes the discriminative power of $f_{M,G}$ for discriminating similar image-pairs from dissimilar ones. The positive parameter γ is trade-offing the effects of the two terms in the objective function of (7). We emphasize here that we did not constrain M or G to be positive semi-definite in the above formulation. Later on, formulation (7) is referred to as ***Sub-SML*** for similarity metric learning over the intra-personal subspace.

2.2. Dual Formulation and Algorithm

We now turn our attention to the computational algorithm of (7). It is easy to see that Sub-SML is a convex optimization problem which guarantees a global solution. For notational simplicity, for any $t=(i,j)\in\mathcal{P}$, let $\widetilde{X}_t = (\tilde{x}_i-\tilde{x}_j)(\tilde{x}_i-\tilde{x}_j)^T$ and $\widetilde{\widetilde{X}}_t = \tilde{x}_i\tilde{x}_j^T$. We can establish the dual problem of Sub-SML as follows.

Theorem 1. *The dual formulation of Sub-SML (*i.e. *formulation (7)) can be written as*

$$\begin{aligned} \max_{0\leq\alpha\leq 1} & \sum_{t\in\mathcal{P}}\alpha_t + \sum_{t=(i,j)\in\mathcal{P}}\alpha_t y_t(\|\tilde{x}_i-\tilde{x}_j\|^2 - \tilde{x}_i^T\tilde{x}_j) \\ & -\tfrac{1}{2\gamma}(\big\|\sum_{t\in\mathcal{P}} y_t\alpha_t\widetilde{X}_t\big\|_F^2 + \big\|\sum_{t\in\mathcal{P}} y_t\alpha_t\widetilde{\widetilde{X}}_t\big\|_F^2). \end{aligned} \quad (8)$$

Moreover, if the optimal solution of (8) is denoted by α^* then the optimal solution (M^*,G^*) of (7) is given by $M^* = I-\frac{1}{\gamma}\sum_{t\in\mathcal{P}} y_t\alpha_t^*\widetilde{X}_t$ and $G^* = I+\frac{1}{\gamma}\sum_{t\in\mathcal{P}} y_t\alpha_t^*\widetilde{\widetilde{X}}_t$.

Proof. We use the Lagrangian multiplier theorem to prove the desired result. By introducing Lagrangian multipliers $\alpha,\beta\geq 0$, define the Lagrangian function related to (7) by $\mathcal{L}(\alpha,\beta;M,G,\xi) = \sum_{t\in\mathcal{P}}\xi_t + \frac{\gamma}{2}(\|M-I\|_F^2 + \|G-I\|_F^2) - \sum_{t=(i,j)\in\mathcal{P}}\alpha_t\big(y_{ij}[s_G(\tilde{x}_i,\tilde{x}_j) - d_M(\tilde{x}_i,\tilde{x}_j)] - 1 + \xi_t\big) - \sum_{t\in\mathcal{P}}\beta_t\xi_t$. Then, taking the derivatives of $\mathcal{L}$ with respect to the primal variables M, G and ξ implies that $M = I-\frac{1}{\gamma}\sum_{t\in\mathcal{P}} y_t\alpha_t\widetilde{X}_t$, $G = I+\frac{1}{\gamma}\sum_{t\in\mathcal{P}} y_t\alpha_t\widetilde{\widetilde{X}}_t$, and $\alpha_t+\beta_t=1$. Substituting these equalities back to $\mathcal{L}$, we get the desired result. This completes the proof of the theorem. □

Formulation (8) is a standard quadratic programming (QP) problem, which can be solved by the standard MATLAB subroutine `quadprog.m`. However, these QP solvers employed the interior-point methods which use the second-order information (Hessian matrix) of the objective function. In the dual problem (8), the number of variables equals the number of image-pairs which is usually very large. Hence, the interior methods quickly become infeasible when the number of image-pairs increases. Instead, we use the accelerated first-order (gradient-based) algorithm proposed in [1,15] which is suitable for large-sized datasets. This method is guaranteed to converge to the global solution with rate $\mathcal{O}(1/k^2)$ where k is the iteration number.

3. Related Work and Discussion

There is a large amount of work on learning similarity metrics. Below we review metric learning models [11, 22, 25, 27] which are closely related to our proposed method Sub-SML, and show the inherent relationship among these models.

Xing et al. [25] proposed to maximize the sum of distances between dissimilar pairs, while maintaining an upper bound on the sum of squared distances between similar pairs. Specifically, the following formulation was proposed:

$$\begin{aligned} \max_{M\in\mathbb{S}_+^d} \quad & \textstyle\sum_{(i,j)\in\mathcal{D}}\sqrt{d_M(x_i,x_j)} \\ \text{s.t.} \quad & \textstyle\sum_{(i,j)\in\mathcal{S}} d_M(x_i,x_j)\leq 1. \end{aligned} \quad (9)$$

Weinberger et al. [22] developed the method called LMNN to learn a Mahalanobis distance metric in kNN classification settings. It aims to explore a large margin nearest neighbor classifier by exploiting nearest neighbor samples as side information in the training set. Specifically, given

a similar set $\mathcal{S} = \{(i,j) : x_i \text{ similar to } x_j\}$ and a triplet set $\mathcal{T} = \{(i,j,k) : x_i \text{ similar to } x_j, x_j \text{ dissimilar to } x_k\}$. LMNN can be rewritten as the following:

$$\begin{aligned} \min_{M,\xi} \quad & \sum_{\tau=(i,j,k)\in\mathcal{T}} \xi_{ijk} + \gamma \sum_{(i,j)\in\mathcal{S}} d_M(x_i,x_j) \\ & d_M(x_j,x_k) - d_M(x_i,x_j) \geq 1 - \xi_{ijk}, \\ & M \in \mathbb{S}_+^d,\ \ \xi_{ijk} \geq 0,\ \ \forall\,(i,j,k) \in \mathcal{T}. \end{aligned} \tag{10}$$

The recent proposal by Ying and Li [27] is very similar to the method in [25]. Specifically, the authors proposed the following method:

$$\begin{aligned} \max_{M\in\mathbb{S}_+^d} \quad & \min_{(i,j)\in\mathcal{D}} d_M(x_i,x_j) \\ \text{s.t.} \quad & \sum_{(i,j)\in\mathcal{S}} d_M(x_i,x_j) \leq 1. \end{aligned} \tag{11}$$

This method was further shown to be an eigenvalue optimization problem which, hence, is referred to as DML-eig.

Kan et al. [11] proposed a side-information based linear discriminant analysis (SILD) approach for face verification. SILD is a modification of LDA which is given by: $\arg\max \mathbf{Tr}(WC_{\mathcal{D}}W^T)/\mathbf{Tr}(WC_{\mathcal{S}}W^T)$, where $C_{\mathcal{D}} = \sum_{(i,j)\in\mathcal{D}}(x_i - x_j)(x_i - x_j)^T$. Let $M = W^TW$ then SILD can be rewritten as

$$\max_{M\in\mathbb{S}_+^d} \frac{\mathbf{Tr}(C_{\mathcal{D}}M)}{\mathbf{Tr}(C_{\mathcal{S}}M)} = \max_{M\in\mathbb{S}_+^d} \left[\frac{\sum_{(i,j)\in\mathcal{D}} d_M(x_i,x_j)}{\sum_{(i,j)\in\mathcal{S}} d_M(x_i,x_j)}\right]. \tag{12}$$

A common term in the above three formulations is the summation of distances between similar image-pairs, *i.e.* $\sum_{(i,j)\in\mathcal{S}} d_M(x_i,x_j) = \mathbf{Tr}(C_{\mathcal{S}}M) = \mathbf{Tr}(L_{\mathcal{S}}^T M L_{\mathcal{S}})$ (recalling that $C_{\mathcal{S}} = L_{\mathcal{S}}L_{\mathcal{S}}^T$). Let $\widetilde{M} = L_{\mathcal{S}}^T M L_{\mathcal{S}}$, and then formulation (9) is equivalent to the following

$$\begin{aligned} \max_{\widetilde{M}\in\mathbb{S}_+^d} \quad & \sum_{(i,j)\in\mathcal{D}} \sqrt{(\tilde{x}_i - \tilde{x}_j)^T \widetilde{M} (\tilde{x}_i - \tilde{x}_j)} \\ \text{s.t.} \quad & \mathbf{Tr}(\widetilde{M}) \leq 1, \end{aligned} \tag{13}$$

where, according to the definition of $L_{\mathcal{S}}$ (equation (3)) in Section 2, $\tilde{x}_i = L_{\mathcal{S}}^{-1}x_i$ is the mapped vector of x_i in the intra-personal subspace. In analogy to the above argument, LMNN is equivalent to

$$\begin{aligned} \arg\min_{\widetilde{M},\xi} \quad & \sum_{\tau=(i,j,k)\in\mathcal{T}} \xi_{ijk} + \gamma\mathbf{Tr}(\widetilde{M}) \\ & d_{\widetilde{M}}(\tilde{x}_j,\tilde{x}_k) - d_{\widetilde{M}}(\tilde{x}_i,\tilde{x}_j) \geq 1 - \xi_{ijk} \\ & \xi_{ijk} \geq 0,\ \ \forall\,(i,j,k) \in \mathcal{T}, \widetilde{M} \in \mathbb{S}_+^d, \end{aligned} \tag{14}$$

and DML-eig can be rewritten as

$$\begin{aligned} \max_{\widetilde{M}\in\mathbb{S}_+^d} \quad & \min_{(i,j)\in\mathcal{D}} d_{\widetilde{M}}(\tilde{x}_i,\tilde{x}_j) \\ \text{s.t.} \quad & \mathbf{Tr}(\widetilde{M}) \leq 1. \end{aligned} \tag{15}$$

SILD is equivalent to

$$\max_{\widetilde{M}\in\mathbb{S}_+^d} \left[\frac{\sum_{(i,j)\in\mathcal{D}} d_{\widetilde{M}}(\tilde{x}_i,\tilde{x}_j)}{\mathbf{Tr}(\widetilde{M})}\right]. \tag{16}$$

We should mention that the image-vectors x_i and x_j in formulations (9), (10), (11) and (12) for face verification are PCA-reduced vectors (i.e. d-dimensional Eigenfaces). They all aim to maintain the average distance $\sum_{(i,j)\in\mathcal{S}} d_M(x_i,x_j)$ between similar images small. We can observe, from their equivalent formulations (13), (14), (15), and (16), that they can also be regarded as metric learning over the intra-personal subspace. In this sense, we can say that minimizing the average distance between similar images plays a similar role as mapping the images to the intra-personal subspace using the whitening process (2).

The learned metric on the intra-personal subspace should best reflect the geometry induced by the similarity and dissimilarity of face images: the distance defined on the intra-personal subspace between similar image-pairs is small while the distance between dissimilar image-pairs is large. The metric learning methods [11, 22, 25, 27] used different objective functions to achieve this goal.

However, the above methods mainly have two limitations: (**L1**) Although these methods can be regarded as metric learning over the intra-personal subspace, they mainly focused on the discrimination of the metric and do not explicitly take into account its robustness. Hence, the learned metrics may not be robust to intra-personal variations; (**L2**) Despite the fact that the bilinear similarity function s_M and CS_M outperform metric learning using d_M for face verification [14], the above methods only used the distance metric d_M. These limitations could degenerate their final verification performance. Our proposed method Sub-SML addressed the above limitations by introducing a new similarity metric and a novel regularization framework for learning similarity metrics.

4. Experiments

In this section, we evaluate our proposed method on the Labeled Faces in the Wild (LFW) database [10]. There are 13233 face images of 5749 people in this database, and 1680 of them appear in more than two images. It is commonly regarded to be a challenging dataset for face verification since the faces were detected from images taken from Yahoo! News and show large variations in pose, expression, lighting, and age etc.

The images were prepared in two ways: "aligned" using commercial face alignment software by [20] and "funneled" available on the LFW website [10]. We use two facial descriptors on the "aligned" images: local Binary Patterns (LBP) [16] and three-Patch Local Binary Patterns (TPLBP) [24]. On the "funneled" images, we use SIFT descriptors [8] which are computed at 9 facial key points. Both original values and square roots of these descriptors are tested as suggested in [8, 24].

The images are divided into ten folds where the identities are mutually exclusive. In each fold, 300 similar and

Method	d	Original	Square Root
PCA	100	0.7598 ± 0.0031	0.7730 ± 0.0023
Intra-PCA	100	0.8132 ± 0.0046	0.8253 ± 0.0033
Sub-ML	100	0.8153 ± 0.0037	0.8252 ± 0.0029
Sub-SL	100	0.8247 ± 0.0036	0.8305 ± 0.0058
Sub-SML	100	0.8452 ± 0.0045	0.8527 ± 0.0052
PCA	200	0.7640 ± 0.0057	0.7787 ± 0.0027
Intra-PCA	200	0.8232 ± 0.0034	0.8345 ± 0.0024
Sub-ML	200	0.8220 ± 0.0042	0.8330 ± 0.0026
Sub-SL	200	0.8417 ± 0.0042	0.8460 ± 0.0041
Sub-SML	200	0.8540 ± 0.0042	0.8632 ± 0.0046
PCA	300	0.7723 ± 0.0053	0.7855 ± 0.0035
Intra-PCA	300	0.8218 ± 0.0027	0.8295 ± 0.0023
Sub-ML	300	0.8218 ± 0.0033	0.8265 ± 0.0038
Sub-SL	300	0.8348 ± 0.0047	0.8403 ± 0.0068
Sub-SML	300	0.8555 ± 0.0061	0.8622 ± 0.0027

(a)

Method	d	Original	Square Root
PCA	100	0.7843 ± 0.0033	0.7855 ± 0.0028
Intra-PCA	100	0.8307 ± 0.0037	0.8332 ± 0.0045
Sub-ML	100	0.8335 ± 0.0031	0.8350 ± 0.0041
Sub-SL	100	0.8368 ± 0.0041	0.8340 ± 0.0037
Sub-SML	100	0.8447 ± 0.0056	0.8397 ± 0.0053
PCA	200	0.8043 ± 0.0046	0.8043 ± 0.0031
Intra-PCA	200	0.8455 ± 0.0063	0.8460 ± 0.0061
Sub-ML	200	0.8452 ± 0.0068	0.8457 ± 0.0059
Sub-SL	200	0.8563 ± 0.0055	0.8508 ± 0.0052
Sub-SML	200	0.8608 ± 0.0049	0.8628 ± 0.0055
PCA	300	0.8047 ± 0.0051	0.8098 ± 0.0038
Intra-PCA	300	0.8423 ± 0.0055	0.8445 ± 0.0043
Sub-ML	300	0.8435 ± 0.0056	0.8432 ± 0.0043
Sub-SL	300	0.8500 ± 0.0052	0.8510 ± 0.0058
Sub-SML	300	0.8673 ± 0.0053	0.8688 ± 0.0061

(b)

Table 1: Performance of Sub-SL,Sub-ML, Sub-SML across different PCA dimension d : (a) SIFT descriptor and (b) LBP descriptor.

Method	SIFT	LBP
Xing [25]	0.7593 ± 0.0059	0.7462 ± 0.0045
DML-eig [27]	0.8127 ± 0.0230	0.8228 ± 0.0041
SILD [11]	0.8085 ± 0.0061	0.8007 ± 0.0135
ITML [7]	0.7812 ± 0.0045	0.7998 ± 0.0039
Sub-ITML	0.8145 ± 0.0046	0.8398 ± 0.0048
LDML [8]	0.7750 ± 0.0050	0.8065 ± 0.0047
Sub-LDML	0.8105 ± 0.0048	0.8227 ± 0.0058
CSML [14]	–	0.8557 ± 0.0052
KISSME [12]	0.8308 ± 0.0056	0.8337 ± 0.0054
Sub-SML	$\mathbf{0.8555 \pm 0.0061}$	$\mathbf{0.8673 \pm 0.0053}$

Table 2: Comparison of Sub-SML with other metric learning methods on the single descriptor in the restricted setting of LFW. Sub-ITML and Sub-LDML denote ITML and LDML over the intra-personal subspace. The result of CSML on LBP is copied from [14] and the notation '–' means that the result on SIFT was not reported.

300 dissimilar image-pairs are provided. It has two different training settings. In the restricted setting, only 600 similar/dissimilar pairs are available while the identity of images is unknown. In the unrestricted setting, the identity information of images is provided. The performance is reported using mean verification rate (standard error) and ROC curve.

In particular, on each test, for Sub-SML, PCA is applied to reduce the noise of face images and the resultant Eigenfaces are further mapped to the intra-personal subspace by using $\tilde{x} = L_{\mathcal{S}}^{-1}x$, where $L_{\mathcal{S}}$ is given by equation (3). The covariance matrix to extract PCA components is computed only from the 9-fold training set. Also, similar image-pairs from the 9-fold training set are used to compute the intra-personal covariance matrix $C_{\mathcal{S}}$. Image vectors $\tilde{x}$ are then L2 normalized to 1 (*i.e.* $\|\tilde{x}\| = 1$) before being fed into Sub-SML. Interestingly, we observed in our experiment that L2 normalization usually improves the performance of most of metric learning methods. On each test, the trade-off parameter γ and the PCA dimension d in Sub-SML are tuned via three-fold cross validation over the remaining 9-fold training sets.

4.1. Image Restricted setting

We first evaluate our method in the restricted setting of the LFW dataset.

Effectiveness of Sub-SML. We conduct experiments to show that Sub-SML has effectively addressed limitations of existing metric learning methods listed as (L1) and (L2) at the end of Section 3. In particular, we show the effectiveness of Sub-SML in two main aspects: the generalized similarity metric $f_{(M,G)}$ combining d_M and s_G, and Sub-SML as a metric learning method over the intra-personal subspace. To this end, we conduct the following two comparisons.

Firstly, we compare Sub-SML with the following two formulations, where only the distance metric d_M or the bilinear similarity metric s_G is used as the similarity metric. More specifically, we compare Sub-SML with the formulation called Sub-ML given by

$$\begin{aligned} \min_{M \in \mathbb{S}^d} \quad & \sum_{t \in \mathcal{P}} \xi_t + \frac{\gamma}{2}\|M - I\|_F^2, \\ \text{s.t.} \quad & y_{ij}[-d_M(\tilde{x}_i, \tilde{x}_j)] \geq 1 - \xi_{ij}, \\ & \xi_t \geq 0, \ \ \forall t = (i,j) \in \mathcal{P}, \end{aligned} \tag{17}$$

Method	Accuracy
Combined b/g samples based methods, aligned [23]	0.8683 ± 0.0034
LDML combined, funneled [8]	0.7927 ± 0.0060
DML-eig combined, funneled & aligned [27]	0.8565 ± 0.0056
HTBI Features, aligned [18]	0.8813 ± 0.0058
CSML + SVM, aligned [14]	0.8800 ± 0.0037
Sub-SML combined, funneled & aligned	$\mathbf{0.8973 \pm 0.0038}$

Table 3: Comparison of Sub-SML with other state-of-the-art methods in the restricted setting of LFW.

and the formulation called Sub-SL given by

$$\begin{aligned} \min_{G \in \mathbb{S}^d} \quad & \sum_{t \in \mathcal{P}} \xi_t + \frac{\gamma}{2}\|G - I\|_F^2, \\ \text{s.t.} \quad & y_{ij}[s_G(\tilde{x}_i, \tilde{x}_j)] \geq 1 - \xi_{ij}, \\ & \xi_t \geq 0, \ \ \forall t = (i,j) \in \mathcal{P}. \end{aligned} \tag{18}$$

As baselines, PCA and Intra-PCA denote the methods using the Euclidean distance over the PCA-reduced subspace and the intra-personal subspace, respectively. It is worth mentioning that, when $\|x_i\| = \|x_j\| = 1$ and M and G are identity matrices, $s_G(x_i, x_j) = (2 - d_M(x_i, x_j))/2 = (f_{(M,G)}(x_i, x_j) - 2)/3$. Hence, in this special case the verification rate using the Euclidean distance is the same as that using $f_{(M,G)}$.

Table 1 reports the comparison results on the SIFT descriptor and LBP descriptor. We can observe from Table 1a that, across different PCA dimensions, Intra-PCA is much better than PCA, which shows the effectiveness of removing intra-personal variations by mapping Eigenfaces into the intra-personal subspace using the whitening process given by equation (2). From Table 1a, we can further see that Sub-ML and Sub-SL are only comparable with or slightly improve Intra-PCA while the performance of Sub-SML is much better than Intra-PCA. Taking the PCA dimension 300 for instance, Sub-SML yields 85.55%, which is better than 82.18% of Sub-ML and 83.48% of Sub-SL. Similar observation can be made on the LBP descriptor as shown in Table 1b. These observations show the effectiveness of learning the generalized similarity metric $f_{(M,G)}$ compared with only learning the distance metric d_M or the bilinear similarity metric s_G.

Secondly, we compare with other metric learning methods such as the method in [25] denoted by Xing, ITML [7], LDML [8], SILD [11], and DML-eig [27]. For fairness of comparison, we also compare with their variants where image-vectors were processed by PCA and further mapped to the intra-personal subspace before being fed into metric learning methods. As shown in Section 3, Xing, SILD and DML-eig implicitly incorporate the above processing steps. For simplicity, we refer to such variants of ITML and LDML as Sub-ITML and Sub-LDML, respectively.

From Table 2 we can see that, on the SIFT descriptor, Sub-SML significantly outperforms the other methods such as ITML, LDML, Sub-ITML and Sub-LDML by obtaining 85.55% verification rate. Furthermore, Sub-SML achieves 86.73% on the LBP descriptor which is better than 85.57% by CSML [14]. These are the best results, to the best of our knowledge, reported so far for SIFT and LBP on the restricted setting of LFW dataset. This observation validates the effectiveness of Sub-SML as a similarity metric learning method over the intra-personal subspace. In addition, we can observe that Sub-ITML and Sub-LDML improve the performance of ITML and LDML, respectively, which shows the effectiveness of the mapping to the intra-personal subspace mentioned in Section 2.

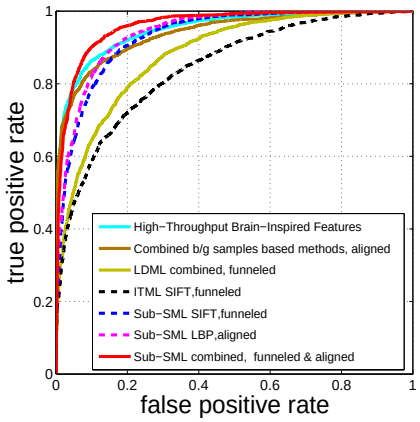

Figure 2: ROC curve of Sub-SML and other state-of-the-art methods in the restricted setting of the LFW database.

Overall, the above comparison results suggest that our proposed method Sub-SML has effectively overcome limitations of existing metric learning methods listed as (L1) and (L2) at the end of Section 3.

Comparison with the state-of-the-art methods. Now we compare Sub-SML with previously published results by combining different descriptors followed the procedure in [8, 24]. Specifically, we first generate the similarity scores by Sub-SML from three descriptors SIFT, LBP and TPLBP and their square roots (six scores). And then we train a Support Vector Machine (SVM) on the vector fused by the six scores to make prediction. Note that each of these published results uses its own learning technique and different feature extraction approaches. Table 3 lists the comparison re-

sults and Figure 2 depicts the ROC curve comparison, from which we observe that Sub-SML outperforms existing results. In particular, it achieves $\mathbf{89.73}\%$, which outperforms the current state-of-the-art result 88.13% obtained by using High-Throughput Brain-Inspired (HTBI) Features [18] [1].

4.2. Image Unrestricted setting

Here, we evaluate Sub-SML on the unrestricted setting of LFW, where the label information allows us to generate more image-pairs during training.

Firstly, we study the performance of Sub-SML when using an increasing number of image-pairs: 1000, 1500 and 2000 pairs per cross-validation fold (instead of the 600 provided in the restricted setting), where the image-pairs (half similar image-pairs and half dissimilar ones) are randomly generated by following the procedure in [10]. Table 4 shows the comparison results on the SIFT descriptor against state-of-the-art metric learning methods such as ITML [7], LDML [8], and their variants Sub-ITML and Sub-LDML. We observe that, across the number of pairs per fold, the performance of Sub-SML is significantly better than other methods, which shows its effectiveness as a similarity metric learning method over the intra-personal subspace. In addition, we observe that Sub-ITML and Sub-LDML respectively improve the performance of ITML and LDML, which again verifies the effectiveness of removing intra-personal variations using the whitening process given by equation (2). We did not directly compare our method with LMNN in Table 4, since LMNN needs the information of triplets. However, we notice that the performance of Sub-SML on SIFT listed in Table 4 is much better than the best performance 80.50% of LMNN as reported in [8].

Secondly, we compare Sub-SML with existing state-of-the-art results on the unrestricted setting of LFW using single and multiple descriptors. Table 5 presents the comparison results and Figure 3 depicts the ROC curves comparison. In particular, we see from Table 5 that Sub-SML 86.42% on the SIFT descriptor outperforms PLDA [17] 86.20% and LDML 83.20%. As for the LBP descriptor, Sub-SML is competitive with that of PLDA. By further combining three descriptors and their square roots following the procedure [8, 24], Sub-SML using 2000 image-pairs achieves 90.75%, which outperforms 90.07% of PLDA and is competitive with 90.90% of Joint Bayesian [4]. The performance of Sub-SML may be further improved by including more image-pairs.

5. Conclusion

In this paper we introduced a novel regularization framework of learning a similarity metric for unconstrained face verification. We formulate its learning objective by incorporating the robustness to large intra-personal variations and the discrimination power of novel similarity metrics, a property most existing metric learning methods do not hold. Our formulation is a convex optimization problem which guarantees the existence of its global solution. Our proposed method has achieved the state-of-the-art performance on the benchmark LFW dataset.

[1] Recently, Cui et al. [6] obtains 89.35% in their CVPR 2013 paper which was achieved, however, by using spatial face region descriptors and a multiple metric learning method.

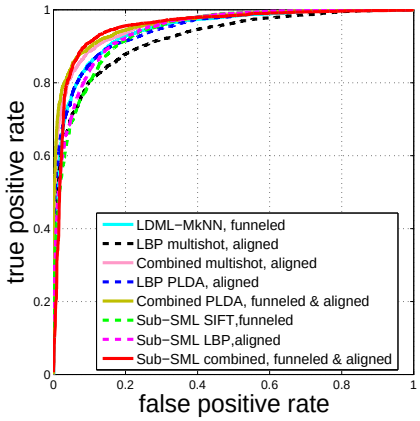

Figure 3: ROC curve of Sub-SML and other state-of-the-art methods in the unrestricted setting of LFW.

Acknowledgment

This work was supported by EPSRC under grant EP/J001384/1. The corresponding author is Yiming Ying.

References

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. on Imaging Sciences*, 2009.

[2] P. N. Belhumeur, J. Hespanda and D. Kiregeman. Eigenfaces vs Fisherfaces: recognition using class specific linear projection. *IEEE Trans. PAMI*, 19: 711–720, 1997.

[3] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *J. of Machine Learning Research*, 11: 1109–1135, 2010.

[4] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. *ECCV*, 2012.

[5] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively with application to face verification. *CVPR*, 2005.

[6] 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. *CVPR*, 2013.

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

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

Methods	1000	1500	2000
ITML [7]	0.7943 ± 0.0047	0.7950 ± 0.0022	0.7988 ± 0.0033
Sub-ITML	0.8188 ± 0.0044	0.8195 ± 0.0054	0.8212 ± 0.0056
LDML [8]	0.7872 ± 0.0038	0.8127 ± 0.0052	0.8093 ± 0.0036
Sub-LDML	0.8183 ± 0.0041	0.8257 ± 0.0049	0.8367 ± 0.0044
Sub-SML	$\mathbf{0.8562 \pm 0.0044}$	$\mathbf{0.8642 \pm 0.0046}$	$\mathbf{0.8613 \pm 0.0055}$

Table 4: Performance of different metric learning methods versus the number of image-pairs per fold in the unrestricted setting of LFW.

Method	Accuracy
SIFT PLDA, funneled [17]	0.862 ± 0.012
SIFT LMNN, funneled [22]	0.805 ± 0.05
SIFT LDML, funneled [8]	0.832 ± 0.0040
SIFT Sub-SML, funneled	$\mathbf{0.8642 \pm 0.0046}$
LBP multishot, aligned [20]	0.8517 ± 0.0061
LBP PLDA, aligned [17]	$\mathbf{0.8733 \pm 0.0055}$
LBP Sub-SML, aligned	0.8715 ± 0.0056
LDML-MkNN, funneled [8]	0.8750 ± 0.0040
Combined multishot, aligned [20]	0.8950 ± 0.0051
Combined PLDA, funneled & aligned [17]	0.9007 ± 0.0051
combined Joint Bayesian [4]	$\mathbf{0.9090 \pm 0.0148}$
Sub-SML combined, funneled & aligned	0.9075 ± 0.0064

Table 5: Comparison of Sub-SML with other state-of-the-art results in the unrestricted setting of LFW: the top 7 rows are based on single descriptor and the bottom 4 rows are based on multiple descriptors.

[9] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classi?cation. *ECCV*, 2012.

[10] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. *ECCV*, 2008.

[11] M. Kan, S. Shan, D. Xu, and X. Chen. Side-Information based linear discriminant analysis for face recognition. *BMVC*, 2011.

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

[13] B. Moghaddam, T. Jebara and A Pentland. Bayesian face recognition. *Pattern Recognition*, 33: 1771–1782, 2000.

[14] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. *ACCV*, 2010.

[15] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publisher, Boston, 2004.

[16] T. Ojala, M. Pietikainen and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. PAMI*, 24:971–987, 2002.

[17] P. Li, Y. Fu, U. Mohammed, J. Elder, and S. Prince. Probabilistic models for inference about identity. *IEEE Trans. PAMI*, 34(1): 144–157, 2012.

[18] N. Pinto and D. Cox. Beyond simple features: a large-scale feature search approach to unconstrained face recognition. In *International Conference on Automatic Face and Gesture Recognition*, 2011.

[19] O. Shalit, D. Weinshall and G. Chechik. Online learning in the manifold of low-rank matrices. *NIPS*, 2010.

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

[21] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE Trans. PAMI*, 26 (9): 1222–1228, 2004.

[22] K. Q. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbour classification. *NIPS*, 2006.

[23] L. Wolf, T. Hassner and Y. Taigman. Similarity scores based on background samples. In *ACCV*, 2009.

[24] L. Wolf, T. Hassner and Y. Taigman. Descriptor based methods in the wild. In *Real-Life Images workshop at the European Conference on Computer Vision*, October, 2008.

[25] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side information. *NIPS*, 2002.

[26] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. In *Technical report, Michigan State University*, 2007.

[27] Y. Ying and P. Li. Distance metric learning with eigenvalue optimization. *J. of Machine Learning Research*, 13: 1–26, 2012.