Large Scale Similarity Learning Using Similar Pairs for Person Verification

Yang Yang, Shengcai Liao, Zhen Lei, Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{yang.yang, scliao, zlei, szli}@nlpr.ia.ac.cn

Abstract

In this paper, we propose a novel similarity measure and then introduce an efficient strategy to learn it by using only similar pairs for person verification. Unlike existing metric learning methods, we consider both the *difference* and *commonness* of an image pair to increase its discriminativeness. Under a pair-constrained Gaussian assumption, we show how to obtain the Gaussian priors (i.e., corresponding covariance matrices) of dissimilar pairs from those of similar pairs. The application of a log likelihood ratio makes the learning process simple and fast and thus scalable to large datasets. Additionally, our method is able to handle heterogeneous data well. Results on the challenging datasets of *face verification* (LFW and PubFig) and *person re-identification* (VIPeR) show that our algorithm outperforms the state-of-the-art methods.

1 Introduction

Person verification, i.e., verifying whether two unseen images contain the same person or not, has attracted increasing attention in computer vision. There exist two main clues in the images - face and human body, based on which, the problem of person verification can be further classified into two subproblems: *face verification* and *person re-identification*. Both of them are challenging due to variations in illumination, viewpoint, pose and expression. A general framework of addressing these two subproblems includes feature extraction and matching, which solves the issues of (1) how to extract efficient and robust features and (2) how to measure the similarity between an image pair based on the extracted features, respectively. This paper is mainly dedicated to the latter - similarity learning.

Recently, learning a similarity measure (Köstinger et al. 2012; Gal Chechik and Bengio 2010; Nguyen and Bai 2010; Bohne et al. 2014; Cao, Ying, and Li 2013) has been well studied and utilized to address the task of person verification. Among them, metric learning aims at learning a Mahalanobis metric while similarity metric learning is to learn a bilinear similarity metric or a cosine similarity metric. However, a single metric is inappropriate to handle heterogeneous data. To overcome this limitation, many approaches based on more metrics are put forward.

In this paper, we propose a novel similarity measure which can be further rewritten as a combination of a Mahalanobis metric and a bilinear similarity metric. With more metrics, it is able to handle heterogeneous data well. We also present an efficient strategy to jointly learn the similarity measure by using only similar pairs. Different from triplets (Schultz and Joachims 2003) or quadruplets (Law, Thome, and Cord 2013), we employ pairwise constraints because it is easier to specify labels in the form of equivalence constraints (Köstinger et al. 2012). To be specific, given an image pair, we first introduce the concepts of *difference* and *commonness*, which are defined by the subtraction between the pair and the summation of them, respectively. Under a pair-constrained Gaussian assumption (detailed in section 3), we then show how to calculate the Gaussian priors (or 'priors' for brevity) of dissimilar pairs from those of similar pairs. Inspired by KISS metric (KISSME) (Köstinger et al. 2012), we employ a log likelihood ratio to directly compute our similarity measure in terms of priors of labeled similar pairs. The time complexity of our method is $O(Nd^2 + d^3)$, where d is the dimension of PCA-reduced features and N is the number of similar pairs. Therefore, our method is scalable to large-scale data as long as d is small. Considering that large scale learning sometimes refers to the regime where learning is limited by computational resources rather than availability of data (Gal Chechik and Bengio 2010) and that our method has low time complexity, we then name our approach as large scale similarity learning (LSSL) using similar pairs. We validate the performances of LSSL on challenging datasets of *face verification* (LFW and PubFig) and *person re-identification* (VIPeR). Experimental results show that LSSL is both fast and accurate.

In summary, the main contributions are two-fold: (1) We propose a novel similarity measure and introduce a fast and efficient method to learn it; (2) Benefiting from the consideration of both *difference* and *commonness* of an image pair and from a pair-constrained Gaussian assumption, we point out how to deduce priors of dissimilar pairs from those of similar pairs. The latter contribution is interesting and important because it is useful for those based on Bayesian rule (Moghaddam, Jebara, and Pentland 2000) (e.g., KISSME) and avoids dealing with dissimilar pairs.

The rest of this paper is organized as follows. We review the related work on similarity learning and give a brief introduction to KISSME in section 2. The proposed method and experimental results are shown in sections 3 and 4, respectively. In section 5, we conclude the paper.

2 Related Work

According to the literature survey (Bellet, Habrard, and Sebban 2014), learning a global Mahalanobis metric has dominated the metric learning literature and competitive results are obtained. Based on the learned matrix M, the distance or similarity between a d-dimensional pair (x_i, y_j) is:

$$d_M(x_i, y_j) = (x_i - y_j)^T M (x_i - y_j) \tag{1}$$

where $M \in \mathcal{R}^{d \times d}$ is a positive semi-definite matrix. On the basis of labeled pairs, how to learn M gives rise to different metric learning methods. The first Mahalanobis distance learning approach - Xing (Xing et al. 2002) optimizes M by maximizing the sum of distances between dissimilar points under the constraint of maintaining a small overall distances between similar points. Afterwards, Weinberger et al. (Weinberger, Blitzer, and Saul 2006) introduces one of the most widely used Mahalanobis distance learning method named Large Margin Nearest Neighbors (LMNN) by strengthening the correlation of target neighbors while keeping instances from different classes far away. Without regularization, LMNN is prone to overfitting during training. To overcome this problem, Davis *et al.* (Davis et al. 2007) propose Information Theoretic Metric Learning (ITML) which guarantees the closeness of the possible solution to a given distance metric prior. In contrast to previous methods, KISSME which learns the Mahalanobis metric from equivalence constraints in (Köstinger et al. 2012), does not rely on complex optimization and is orders of magnitudes faster. In (Law, Thome, and Cord 2014), a linear regularization term is incorporated in the objective function, which minimizes the k smallest eigenvalues of the Mahalanobis metric. Under the regularization, the rank of a learned Mahalanobis metric is explicitly controlled and the recognition on both controlled and real datasets are greatly improved. Instead of the Mahalanobis metric, other similarity metric for verification problems have two main forms: the bilinear similarity metric $s_M(x_i, y_j) = x_i^T M y_j$ (Gal Chechik and Bengio 2010) and the cosine similarity metric $CS_M(x_i, y_j) = x_i^T M y_j / (\sqrt{x_i^T M x_i}\sqrt{y_i^T M y_j})$ (Nguyen and Bai 2010).

To address the limitations of a global Mahalanobis metric or a similarity metric in dealing with heterogeneous data (Bellet, Habrard, and Sebban 2014; Bohne et al. 2014), many local metric learning methods are proposed recently. A generalized similarity measure is proposed in (Cao, Ying, and Li 2013) - Sub-SML, which combines the similarity function and distance function. Then, the learned metrics preserve the discriminative power. In (Li et al. 2013), a joint model bridging a global distance metric and a local decision rule is proposed to achieve better performance than metric learning methods. Large Margin Local Metric Learnig (LMLML) (Bohne et al. 2014) introduces a novel local metric learning method that first computes a Gaussian Mixture Model from the labeled data and then learns a set of local metrics by solving a convex optimization problem. It is flexible and can be applied to a wide variety of scenarios.

2.1 A Brief Introduction to KISSME

In consideration of the fact that the solution of our method is inspired by KISSME, we briefly introduce it in this subsection. Additionally, in experiments, we also show how to improve KISSME based on our method.

In a statistical inference perspective, KISSME aims at learning a global Mahalanobis metric (defined by Eq. 1) from equivalence constraints. As there is a bijection between the set of Mahalanobis metric and that of multivariate Gaussian distribution [1] , the Mahalanobis metric can be directly computed in terms of the covariance matrix without optimization. To seek their connection, the log likelihood ratio defined by Eq. 2 is employed:

$$s(z) = 2 \log \frac{P(z|\mathcal{H}_\mathcal{S})}{P(z|\mathcal{H}_\mathcal{D})} = C + z^T(\Sigma_{z\mathcal{D}}^{-1} - \Sigma_{z\mathcal{S}}^{-1})z \tag{2}$$

where $C = d \times log \frac{|\Sigma_{z\mathcal{D}}|}{|\Sigma_{z\mathcal{S}}|}$ is a constant (here, d is the dimension of z). In KISSME, z refers to the *difference* of an image pair (x_i-y_j) and is assumed to follow two different Gaussian distributions (one is based on $\mathcal{H}_\mathcal{S}$ which represents the hypothesis of a similar pair while the other on $\mathcal{H}_\mathcal{D}$ denoting the hypothesis of a dissimilar pair).

It can be seen that a higher value of $s(z)$ indicates that the pair is similar with a high probability. After stripping the constant C which just provides an offset, M in Eq. 1 can be written as $\Sigma_{z\mathcal{D}}^{-1} - \Sigma_{z\mathcal{S}}^{-1}$. To make M be a positive semi-definite matrix, the authors of (Köstinger et al. 2012) further re-project it onto the cone of positive semi-definite matrixes, i.e., clipping the spectrum of M by eigenanalysis. Though simple, KISSME achieves surprisingly good results in *person re-identification* (Yang et al. 2014).

3 Large Scale Similarity Learning

In this section, we first propose a novel similarity measure. Then, we demonstrate how to learn it using only similar pairs based on a statistical inference perspective. A pair-constrained Gaussian assumption is made in the following. Under this assumption, we further show how to preprocess the features. Finally, we discuss the parameter setting and the benefit of PCA and compare with a previous work.

Pair-constrained Gaussian Assumption Let us assume that we have a data set of N d-dimensional similar pairs $\{(x_1, y_1), ..., (x_N, y_N)\}$. Any instance x_i (or y_i), $i = 1, 2, ..., N$ is represented by

$$x_i = \mu_i + \epsilon_{i1} \;\; (or \;\; y_i = \mu_i + \epsilon_{i2}) \tag{3}$$

where μ_i is an implicit variable which refers to the i-th similar pair (x_i, y_i) while ϵ_{i1} (or ϵ_{i2}) is also an implicit variable denoting the variation of x_i (or y_j) within the similar pair. The independent variables μ and ϵ (subscript omitted) are supposed to follow two different Gaussian distributions $\mathcal{N}(0, S_\mu)$ and $\mathcal{N}(0, S_\epsilon)$, where S_μ and S_ϵ are two unknown covariance matrixes.

[1] e.g., Gaussian distribution of d-dimensional similar pairs: $P(z|\mathcal{H}_\mathcal{S}) = \mathcal{N}(0, \Sigma_{z\mathcal{S}}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma_{z\mathcal{S}}|^{d/2}} \exp\{-\frac{1}{2} z^T \Sigma_{z\mathcal{S}}^{-1} z\}$.

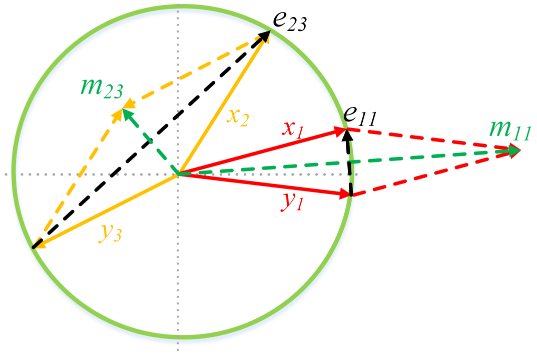

Figure 1: An illustration of *difference* and *commonness* in 2-dimensional Euclidean space. Each feature is l_2-normalized. (x_1, y_1) (red) denotes a similar pair while (x_2, y_2) (yellow) represents a dissimilar pair. The green circle is a unit one.

3.1 The Proposed Similarity Measure

For an image pair (x_i, y_j), each of which is a column and normalized with the l_2-norm (explained in the subsection of this section: Date Preprocessing), the *difference* e_{ij} and *commonness* m_{ij} are defined by Eq. 4. For brevity, we will omit the subscript when we do not treat the variable as a specific instance in the following.

$$\begin{cases} e = x - y \\ m = x + y \end{cases} \tag{4}$$

Fig. 1 illustrates the *difference* and *commonness* in 2-dimensional Euclidean space. (x_1, y_1) (red) denotes a similar pair while (x_2, y_3) (yellow) represents a dissimilar pair. We find that for a similar pair (x_1, y_1), the value of $\|e_{11}\|_2$ is small but that of $\|m_{11}\|_2$ is high. Meanwhile, for a dissimilar pair (x_2, y_3), the value of $\|e_{23}\|_2$ is high but that of $\|m_{23}\|_2$ is small. Therefore, if we combine e and m, we can expect more discriminativeness than those metric learning methods which only consider the *difference* of an image pair.

We then propose a new similarity measure which is defined by subtracting the dissimilarity score $e^T Be$ from similarity score $m^T Am$:

$$r(x, y) = m^T Am - \lambda e^T Be \tag{5}$$

where A and B are two matrixes, parameterizing the similarity and dissimilarity scores, respectively, and the parameter λ is used to balance the effects between similarity and dissimilarity scores. In (Gal Chechik and Bengio 2010), it is pointed out that although adding a positive semi-definite constraint is useful for reducing overfitting in the condition of little training data, if there are sufficient training data, which is the most case in modern applications, the process of adding the constraint consumes additional computation time and its benefit is limited. With this view, we do not impose the positive semi-definite constraint on the learned A or B.

3.2 Jointly Learn A and B from Similar Pairs

Inspired by KISSME (Köstinger et al. 2012), we employ the log likelihood ratio (defined in Eq. 2) to compute A and B directly. Under the pair-constrained Gaussian assumption, it is easy to know that based on $\mathcal{H}_{\mathcal{S}}$ and $\mathcal{H}_{\mathcal{D}}$, m follows two Gaussian distributions while e follows another two Gaussian distributions. Since $s(z) = z^T(\Sigma_{z\mathcal{D}}^{-1} - \Sigma_{z\mathcal{S}}^{-1})z$ (discarding the constant C) indicates that the pair is similar with a high probability, $s(z)$ can be taken as a similarity score while $-s(z)$ as a dissimilarity score, e.g., $s(m) = m^T Am$ and $s(e) = -e^T Be$. Thus, we can obtain

$$\begin{cases} A = \Sigma_{m\mathcal{D}}^{-1} - \Sigma_{m\mathcal{S}}^{-1} \\ B = \Sigma_{e\mathcal{S}}^{-1} - \Sigma_{e\mathcal{D}}^{-1}. \end{cases} \tag{6}$$

For a similar pair $(x_i, y_i), i = 1, 2, ..., N$, we have $x_i = \mu_i + \epsilon_{i1}$ and $y_i = \mu_i + \epsilon_{i2}$. Then $e_{ii} = \epsilon_{i1} - \epsilon_{i2}$ and $m_{ii} = 2\mu_i + \epsilon_{i1} + \epsilon_{i2}$. Under the pair-constrained Gaussian assumption, we know that μ_i, ϵ_{i1} and ϵ_{i2} are independent. Then, we have $cov(e, e) = 2S_\epsilon$ and $cov(m, m) = 4S_\mu + 2S_\epsilon$ (subscript omitted). Thus, for similar pairs, $P(m|\mathcal{H}_{\mathcal{S}}) = \mathcal{N}(0, \Sigma_{m\mathcal{S}})$ and $P(e|\mathcal{H}_{\mathcal{S}}) = \mathcal{N}(0, \Sigma_{e\mathcal{S}})$, where

$$\begin{cases} \Sigma_{m\mathcal{S}} = 4S_\mu + 2S_\epsilon \\ \Sigma_{e\mathcal{S}} = 2S_\epsilon. \end{cases} \tag{7}$$

For a dissimilar pair $(x_i, y_j), i \neq j, i, j = 1, 2, ..., N$, we have $x_i = \mu_i + \epsilon_{i1}$ and $y_j = \mu_j + \epsilon_{j2}$. Then $e_{ij} = \mu_i - \mu_j + \epsilon_{i1} - \epsilon_{j2}$ and $m_{ij} = \mu_i + \mu_j + \epsilon_{i1} + \epsilon_{j2}$. Under the pair-constrained Gaussian assumption, we know that μ_i, μ_j, ϵ_{i1} and ϵ_{j2} are independent. Then, we have $cov(e, e) = 2S_\mu + 2S_\epsilon$ and $cov(m, m) = 2S_\mu + 2S_\epsilon$ (subscript omitted). Thus, for dissimilar pairs, $P(m|\mathcal{H}_{\mathcal{D}}) = \mathcal{N}(0, \Sigma_{m\mathcal{D}})$ and $P(e|\mathcal{H}_{\mathcal{D}}) = \mathcal{N}(0, \Sigma_{e\mathcal{D}})$, where

$$\begin{cases} \Sigma_{m\mathcal{D}} = 2S_\mu + 2S_\epsilon \\ \Sigma_{e\mathcal{D}} = 2S_\mu + 2S_\epsilon. \end{cases} \tag{8}$$

By comparing Eq. 7 with Eq. 8, we observe that $\Sigma_{m\mathcal{D}} + \Sigma_{e\mathcal{D}} = \Sigma_{m\mathcal{S}} + \Sigma_{e\mathcal{S}}$ and $\Sigma_{m\mathcal{D}} = \Sigma_{e\mathcal{D}}$. Thus, $\Sigma_{m\mathcal{D}}$ and $\Sigma_{e\mathcal{D}}$ can be directly calculated by the priors of labeled similar pairs by Eq. 9:

$$\Sigma_{m\mathcal{D}} = \Sigma_{e\mathcal{D}} = \frac{1}{2}(\Sigma_{m\mathcal{S}} + \Sigma_{e\mathcal{S}}) = \Sigma. \tag{9}$$

With it, it is interesting to note that (1) we can neglect the dissimilar pairs and thus saving the time to deal with them and (2) KISSME can be improved by rewriting M as $\Sigma^{-1} - \Sigma_{e\mathcal{S}}^{-1}$. We will compare their performances in experiments.

Learning Priors of Similar Pairs When computing $\Sigma_{m\mathcal{S}}$ and $\Sigma_{e\mathcal{S}}$, we do not rely on how to estimate S_μ and S_ϵ both of which have implicit meanings. Instead, we choose the maximum likelihood estimate (MLE) to compute them based on the labeled similar pairs:

$$\begin{cases} \Sigma_{m\mathcal{S}} = \frac{1}{N}\sum_{i=1}^{N} m_{ii} m_{ii}^T \\ \Sigma_{e\mathcal{S}} = \frac{1}{N}\sum_{i=1}^{N} e_{ii} e_{ii}^T \end{cases} \tag{10}$$

where $m_{ii} = x_i + y_i, e_{ii} = x_i - y_i, i = 1, 2, ..., N$.

Learning Priors of Dissimilar Pairs from Those of Similar Pairs Based on Eqs. 9 and 10, we can derive priors of dissimilar pairs - $\Sigma_{m\mathcal{D}}$ and $\Sigma_{e\mathcal{D}}$ from the learned priors of similar pairs - $\Sigma_{m\mathcal{S}}$ and $\Sigma_{e\mathcal{S}}$:

$$\Sigma_{m\mathcal{D}} = \Sigma_{e\mathcal{D}} = \Sigma = \frac{1}{2N}\sum_{i=1}^{N}(m_{ii} m_{ii}^T + e_{ii} e_{ii}^T). \tag{11}$$

Thus, A and B in Eq. 6 are jointly learned:

$$\begin{cases} A = \Sigma^{-1} - \Sigma_{m\mathcal{S}}^{-1} \\ B = \Sigma_{e\mathcal{S}}^{-1} - \Sigma^{-1} \end{cases} \tag{12}$$

where $\Sigma_{m\mathcal{S}}$ and $\Sigma_{e\mathcal{S}}$ are defined by Eq. 10 while Σ in Eq. 11.

Based on Eqs. 4, 5 and 12, we can further reformulate our similarity measure as a combination of a bilinear similarity metric and a Mahalanobis metric:

$$r(x, y) = x^T M_b y - (x - y)^T M_d (x - y) \tag{13}$$

with

$$\begin{cases} M_b = 4(\Sigma^{-1} - \Sigma_{m\mathcal{S}}^{-1}) \\ M_d = \Sigma_{m\mathcal{S}}^{-1} + \lambda\Sigma_{e\mathcal{S}}^{-1} - (1 + \lambda)\Sigma^{-1} \end{cases} \tag{14}$$

where M_b and M_d parameterize the bilinear similarity metric and the Mahalanobis metric, respectively. Therefore, with more metrics, our LSSL is able to handle the heterogeneous data better than those approaches which are only based on a single metric. In experiments, we will compare the performances of the bilinear similarity metric, Mahalanobis metric and LSSL which are learned in our methods. The learning scheme of LSSL is described in Algorithm 1.

Algorithm 1 Large Scale Similarity Learning Using Similar Pairs

```
Input:  A data set of $N$ $d$-dimensional similar training
        pairs $\{(x_1, y_1), ..., (x_N, y_N)\}$ after PCA.
Output: $A \in \mathcal{R}^{d \times d}$ and $B \in \mathcal{R}^{d \times d}$.
1: normalize each data with the $l_2$-norm;
2: compute $\Sigma_{m\mathcal{S}}$ and $\Sigma_{e\mathcal{S}}$ by Eq. 10;
3: compute $\Sigma$ by Eq. 11;
4: compute $A$ and $B$ by Eq. 12.
```

3.3 Data Preprocessing

According to the pair-constrained Gaussian assumption, μ and ϵ, which are two independent variables, follow two Gaussian distributions $\mathcal{N}(0, S_\mu)$ and $\mathcal{N}(0, S_\epsilon)$. Then, variables of e and m are independent and follow two Gaussian distributions with zero mean. Based on them, we show how to preprocess the features.

Zero Mean Suppose that for any similar pair $(x_i, y_i), i = 1, ..., N$, there is a corresponding 'negative' similar pair $(-x_i, -y_i), i = 1, 2, ..., N$. Then, variable e or m follows a zero-mean distribution since $e_{ii} + (-e_{ii}) = m_{ii} + (-m_{ii}) = 0, i = 1, 2, ..., N$. Thus, zero-mean distribution of e and m always holds. Based on Eq. 10, if we discard the priors of the 'negative' similar pairs, the results of $\Sigma_{e\mathcal{S}}$ and $\Sigma_{m\mathcal{S}}$ remain unchanged. So, in practice, we do not calculate the priors of the 'negative' similar pairs.

Independence Based on the pair-constrained Gaussian assumption, two random variables e and m are independent and follow two Gaussian distributions with zero mean. In such a case, independence is equivalent to orthogonality (Gareth James and Tibshirani 2013). Then we make e and m orthogonal by normalizing each feature using the l_2-norm. This is because

$$\begin{aligned} \langle e, m \rangle = \langle x - y, x + y \rangle = \langle x, x \rangle - \langle y, y \rangle + \\ (\langle x, y \rangle - \langle y, x \rangle) = 1 - 1 + 0 = 0 \end{aligned} \tag{15}$$

where $\langle ., . \rangle$ denotes the inner product. In this way, the independence between e and m is guaranteed. Note that whatever was done to preprocess the features [2] , the last step is normalizing each feature with the l_2-norm (see Algorithm 1).

3.4 Discussion

Parameter Setting In Eq. 14, when we compute M_d, there is a parameter λ balancing the similarity and dissimilarity scores. It is not very sensitive and takes value in [0.9,1.5]. In experiments, we set λ to 0.9, 1.2 and 1.5 for LFW, PubFig and VIPeR, respectively.

The Benefit of PCA It is hard to obtain reliable estimations of priors of high-dimension data without sufficient data. In view of this, we use PCA to reduce the dimension and represent the Gaussian distributed data efficiently.

Comparison with a previous work In (Li et al. 2013), the LADF learns a global Mahalanobis metric as a similarity measure and the idea of similarity/dissimilarity is reflected in the local decision rule when they make decisions. However, our LSSL aims to learn a discriminative similarity measure, which is intuitively defined in Eq. 5.

4 Experiments

We evaluate the performances of the proposed LSSL on three publicly available challenging datasets: Labeled Faces in the Wild (LFW) (Huang et al. 2007), Public Figures Face Database (PubFig) (Kumar et al. 2009) and Viewpoint Invariant Pedestrian Recognition (VIPeR) (Gray, Brennan, and Tao 2007). The first two datasets focus on *face verification*, while the last one on *person re-identification*.

Firstly, we fairly compare LSSL with existing similarity learning approaches on LFW dataset. In addition, we further test the performance of LSSL on PubFig dataset in attribute space. Finally, we show that LSSL with improved features achieves a new state-of-the-art result on VIPeR dataset.

4.1 Datasets and Setup

LFW Labeled Faces in the Wild (LFW) is a widely used dataset and can be considered as the current state-of-the-art face recognition benchmark. There are 13,233 unconstrained face images of 5,749 individuals collected from the web. It is very challenging because there are large variations in pose, lighting, facial expression, age, gender, ethnicity and general imaging and environmental conditions. For the *face verification* task, persons who appear in testing have not been seen in training.

We follow the standard 'restricted' evaluation protocol. The images are divided into 10 folds that are used for cross-validation and there are 300 similar pairs and 300 dissimilar pairs in each fold. The subjects in the 10 folds are mutually exclusive. In the restricted setting, no inference on the identity of the image is allowed. Image pairs of 9 folds are used for training while the remaining fold is used for testing. The average result over 10 trials is reported.

PubFig Public Figures Face Database (PubFig) is also a challenging large-scale real-world dataset. It contains

[2] There is a slight improvement if we preprocess the features by subtracting the mean of all samples.

Method	Accuracy(%)
ITML	78.9 ± 0.5
Xing	74.4 ± 0.6
KISSME	80.6 ± 0.6
LDML	77.5 ± 0.5
PCCA-$\mathcal{X}^2_{RBF}$	83.8 ± 0.4
Fantope Regularization	83.5 ± 0.5
LMLML(Intra-PCA)	85.4 ± 0.5
Sub-SML(Intra-PCA)	85.6 ± 0.6
LSSL(PCA)	85.5 ± 0.5
LSSL(Intra-PCA)	**86.2** ± **0.4**

Table 1: Comparison of LSSL with existing similarity learning approaches on the LFW in the restricted setting. Results are shown in terms of mean and standard error (best in bold).

58,797 images of 200 persons collected from Google images and Flickr. 10 folds are used for cross-validation with 1000 similar pairs and 1000 dissimilar pairs in each fold. Images in each fold are sampled from 14 individuals. For the verification task, testing is conducted on individuals who have not been seen in the training. Similar to LFW, we also choose the 'restricted' setting. The identity of the subject is unknown in training. We also report the average result over 10 runs.

VIPeR Viewpoint Invariant Pedestrian Recognition (VIPeR) is one of the most popular dataset for *person re-identification*. It consists of 1264 images from 632 pedestrians with the resolution of 128×48. For each person, a pair of images are taken from two disjoint cameras with different views. It suffers from viewpoint, pose and illumination changes. To compare our approach with others, we randomly choose 316 image pairs for training and the remaining 316 image pairs are used for testing. In the stage of testing, images from one camera are employed as a probe set while those from the other camera as a gallery set. Then, we switch the probe and gallery. The average of the results is regarded as one run. We report the final average results over 100 runs.

4.2 *Face Verification*: LFW

Features To fairly compare with other approaches, we use the commonly used SIFT descriptors (Lowe 2004) (downloaded from the website of (Guillaumin, Verbeek, and Schmid 2009)) to represent the 'funneled' face images. We then employ PCA to project the original 3456 dimensional features to a 100 dimensional subspace.

In Table 1, we compare LSSL with existing similarity learning approaches: ITML, Xing, KISSME, LDML (Guillaumin, Verbeek, and Schmid 2009), PCCA-$\mathcal{X}^2_{RBF}$ (Mignon and Jurie 2012), Fantope Regularization (Law, Thome, and Cord 2014), LMLML (Intra-PCA) and Sub-SML(Intra-PCA). The results of the first four metric learning methods were generated with their original codes which are publicly available. Other results are copied from the public reports or released by the corresponding authors, all of which are based on the same SIFT descriptors under the restricted setting of LFW. LSSL(PCA) only uses PCA to reduce the dimension while LSSL(Intra-PCA) further employs Intra-PCA to reduce the effect of large intra-pair variations. Note that Intra-PCA (using similar pairs for training) is also fast.

Method	Accuracy(%)
Gaussian Distribution	70.6 ± 1.8
KISSME	77.7 ± 0.9
Qwise+LMNN	77.6 ± 2.0
LMNN+Fantope	77.5 ± 1.6
LSSL	**79.2** ± **0.8**

Table 2: An overview of the state-of-the-art results (mean and standard error) on PubFig (best in bold).

From Table 1, we can see that LSSL(PCA) outperforms the other approaches except Sub-SML(Intra-PCA) while LSSL(Intra-PCA) obtains the best result with 86.2% ± 0.4% verification rate. Additionally, all of LSSL, Sub-SML and LMLML which employ more metrics, perform better than PCCA-$\mathcal{X}^2_{RBF}$ which achieves the best result of a single metric. These observations validate the effectiveness of LSSL and also demonstrate that multiple metrics show better performance than a single metric. Even though we can obtain better results by employing Intra-PCA, we also find that in some cases, it may lead to a singular covariance matrix of dissimilar pairs for KISSME, which will harm its final performance. In the following experiments, to fairly compare with KISSME, we only use PCA to preprocess the features.

4.3 *Face Verification*: PubFig

Features To represent the faces in PubFig, we adopt the 73-dimensional features provided by (Kumar et al. 2009). The 'high-level' visual features are extracted by automatically encoding the appearance in either nameable attributes such as gender, race, age etc. or 'similes'. We employ PCA to project the features to a 65 dimensional subspace.

In Table 2, we give an overview of the state-of-the-art methods reported in recent years. By use of the same features, LSSL performs better than KISSME. In addition, LSSL also outperforms other methods with different feature types and learners, including Gaussian Distribution (Parikh and Grauman 2011), Qwise+LMNN (Law, Thome, and Cord 2013) and LMNN+Fantope (Law, Thome, and Cord 2014). Thus, LSSL achieves a new state-of-the-art result (79.2% ± 0.8%) on FubFig in the restricted setting.

4.4 *Person Re-identification*: VIPeR

Features To represent the individuals in VIPeR, we extract the features based on the codes provided by the authors of salient color names based color descriptor (SCNCD) (Yang et al. 2014) and color names (Joost van de Weijer and Larlus 2009) as follows: (1) We employ ten horizontal stripes of equal size; (2) SCNCD, color names and color histograms are computed and fused; (3) The final image-foreground (Yang et al. 2014) feature representation is extracted in six different color spaces including RGB, *rgb*, $l_1l_2l_3$, HSV, YCbCr, and Lab. By doing so, the obtained features are more robust to illumination, partial occlusion and scale changes. We employ PCA to reduce the dimension of the features to 75.

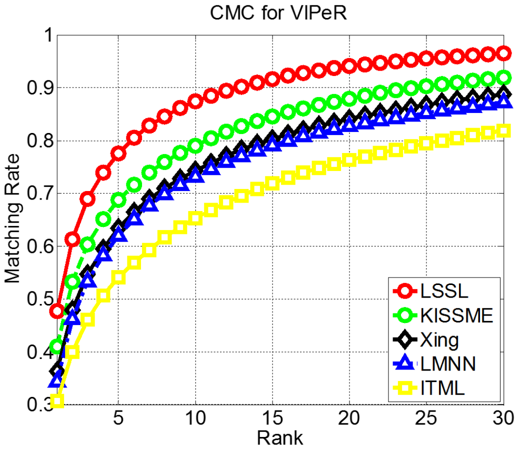

Figure 2: Comparison of LSSL with several classic metric learning methods on the VIPeR.

Rank	1	5	10	20
Final(ImgF)	37.8%	68.5%	81.2%	90.4%
CVPDL	34.0%	64.2%	77.5%	88.6%
LOMO+XQDA	40.0%	-	80.5%	91.1%
MtMCML	28.8%	59.3%	75.8%	88.5%
SSCDL	25.6%	53.7%	68.1%	83.6%
MLF	43.4%	-	84.9%	93.7%
LSSL	**47.8%**	**77.9%**	**87.6%**	**94.2%**

Table 3: Comparison with the state-of-the-art methods reported in recent 2 years on VIPeR dataset. Best in bold.

All the results are shown in the form of commonly used Cumulated Matching Characteristic (CMC) curve (Wang et al. 2007), which means that we can find the right matching people within the first n Ranks.

Using the same features, we first compare LSSL with four classic metric learning methods: Xing, LMNN, ITML and KISSME. In Fig. 2, we observe that LSSL significantly outperforms other metric learning approaches at Ranks 1-30.

Additionally, in Table 3 we also compare the performance of our method with the state-of-the-art methods (reported in recent 2 years) which focus on feature learning and/or matching, e.g., Final(ImgF) (Yang et al. 2014), MtMCML (Lianyang Ma and Tao 2014), semi-supervised coupled dictionary learning (SSCDL) (Xiao Liu and Bu 2014), Mid-level Filter (MLF) (Rui Zhao and Wang 2014), cross-view projective dictionary learning (CVPDL) (Sheng Li and Fu 2015), and LOMO+XQDA (Shengcai Liao and Li 2015). The best result in previous works is achieved by MLF (Rui Zhao and Wang 2014) with 43.4% at Rank 1. Our method achieves a new state-of-the-art result 47.8% (4.4% higher than MLF) at Rank 1.

4.5 Performance Analysis

Single Metric v.s. Joint Metrics According to Eqs. 13 and 14, the Mahalanobis metric and the bilinear similarity metric can be jointly learned in our method. Then, we conduct experiments on LFW and PubFig to validate two problems: (1) which metric is more important and (2) whether the combination of two metrics are better than a single metric.

Method	M_d	M_b	M_d and M_b
LFW	82.6 ± 0.5	81.2 ± 0.5	**85.5 ± 0.5**
PubFig	78.3 ± 0.9	73.9 ± 0.7	**79.2 ± 0.8**

Table 4: Results (mean and standard error (%)) of single metric and joint metrics on LFW and PubFig. Best in bold.

From Table 4, we can find that (1) in comparison between M_d and M_b, the Mahalanobis metric performs better than the bilinear similarity metric and (2) joint metrics are more discriminative than a single metric based on M_d (or M_b).

Method	KISSME	iKISSME
LFW	80.6 ± 0.6	**82.2 ± 0.6**
PubFig	77.7 ± 0.9	**78.6 ± 0.9**

Table 5: Comparison of KISSME and iKISSME on LFW and PubFig (mean and standard error (%)). Best in bold.

KISSME v.s. iKISSME In section 3, we observe that KISSME (defining M as $\Sigma_{e\mathcal{D}}^{-1} - \Sigma_{e\mathcal{S}}^{-1}$) can be improved by rewriting M as $\Sigma^{-1} - \Sigma_{e\mathcal{S}}^{-1}$ where Σ is defined by Eq. 11. We name it as improved KISSME (iKISSME). To compare their performances, we conduct the experiments on LFW and PubFig. Results are shown in Table 5. It is obvious that iKISSME performs better than KISSME on both LFW and PubFig. This observation demonstrates the feasibility and effectiveness of Eq. 11 when we compute the priors of dissimilar pairs. It also reflects that we can expect a more accurate estimate of the priors of dissimilar pairs if they are selected appropriately (e.g., the priors are computed based on more dissimilar pairs).

Training Time In (Köstinger et al. 2012), KISSME has been demonstrated to be orders of magnitudes faster than comparable methods such as SVM, ITML, LDML and LMNN in traing. So we just compare the training time between LSSL and KISSME on the VIPeR dataset. All of them are evaluated on a PC with the 3.40 GHz Core I7 CPU with 8 cores. The average training time of KISSME is 66.8 ms while that of LSSL is 1.6 ms. That is, LSSL is approximately 42 times faster than KISSME. It benefits from avoiding dealing with the dissimilar pairs.

5 Conclusion

To address the task of person verification, we present a novel similarity measure and introduce an efficient method to learn it. Benefiting from the consideration of both *difference* and *commonness* of an image pair and from a pair-constrained Gaussian assumption, we show how to learn the priors of dissimilar pairs from those of similar pairs. Our proposed LSSL is very fast for training, which is important for real applications. Experimental results demonstrate the efficiency of LSSL on dealing with person verification problems.

Acknowledgments This work was supported by the Chinese National Natural Science Foundation Projects #61203267, #61375037, #61473291, #61572501, #61572536, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

References

Bellet, A.; Habrard, A.; and Sebban, M. 2014. A survey on metric learning for feature vectors and structured data. Technical report.

Bohne, J.; Ying, Y.; Gentric, S.; and Pontil, M. 2014. Large margin local metric learning. In *ECCV*.

Cao, Q.; Ying, Y.; and Li, P. 2013. Similarity metric learning for face recognition. In *ICCV*.

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

Gal Chechik, Varun Sharma, U. S., and Bengio, S. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research* 11:1109–1135.

Gareth James, Daniela Witten, T. H., and Tibshirani, R. 2013. An introduction to statistical learning. *Statistical Theory and Methods*.

Gray, D.; Brennan, S.; and Tao, H. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *Workshop on PETS*.

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

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07–49, University of Massachusetts, Amherst.

Joost van de Weijer, Cordelia Schmid, J. V., and Larlus, D. 2009. Learning color names for real-world applications. *TIP* 1512C–1523.

Köstinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *CVPR*.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and simile classifiers for face verification. In *ICCV*.

Law, M. T.; Thome, N.; and Cord, M. 2013. Quadruplet-wise image similarity learning. In *ICCV*.

Law, M. T.; Thome, N.; and Cord, M. 2014. Fantope regularization in metric learning. In *CVPR*.

Li, Z.; Chang, S.; Liang, F.; Huang, T. S.; Cao, L.; and Smith, J. R. 2013. Learning locally-adaptive decision functions for person verification. In *CVPR*.

Lianyang Ma, X. Y., and Tao, D. 2014. Person re-identification over camera networks using multi-task distance metric learning. *TIP* 23:3656C–3670.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 60(2):91–110.

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

Moghaddam, B.; Jebara, T.; and Pentland, A. 2000. Bayesian face recognition. *Pattern Recognition* 33(11):1771–1782.

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

Parikh, D., and Grauman, K. 2011. Relative attributes. In *ICCV*.

Rui Zhao, W. O., and Wang, X. 2014. Learning mid-level filters for person re-identification. In *CVPR*.

Schultz, M., and Joachims, T. 2003. Learning a distance metric from relative comparisons. In *NIPS*.

Sheng Li, M. S., and Fu, Y. 2015. Cross-view projective dictionary learning for person re-identification. In *IJCAI*.

Shengcai Liao, Yang Hu, X. Z., and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

Wang, X.; Doretto, G.; Sebastian, T.; Rittscher, J.; and Tu, P. 2007. Shape and appearance context modeling. In *ICCV*.

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

Xiao Liu, Mingli Song, D. T. X. Z. C. C., and Bu, J. 2014. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. 2002. Distance metric learning, with application to clustering with side-information. In *NIPS*.

Yang, Y.; Yang, J.; Yan, J.; Liao, S.; Yi, D.; and Li, S. Z. 2014. Salient color names for person re-identification. In *ECCV*.