Neurocomputing 307 (2018) 25–37

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning

Husheng Dong[a,b], Ping Lu[b], Shan Zhong[c], Chunping Liu[a,d,e], Yi Ji[a], Shengrong Gong[c,a,f,*]

[a] School of Computer Science and Technology, Soochow University, Suzhou, China
[b] Suzhou Institute of Trade and Commerce, Suzhou, China
[c] Changshu Institute of Science and Technology, Changshu, China
[d] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China
[e] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China
[f] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

ARTICLE INFO

Article history:
Received 29 October 2016
Revised 26 February 2018
Accepted 8 April 2018
Available online 28 April 2018

Communicated by Dr Jiwen Lu

Keywords:
Person re-identification
Feature representation
Metric learning
Generalized similarity

ABSTRACT

To solve the challenging person re-identification problem, great efforts have been devoted to feature representation and metric learning. However, existing feature extractors are either stripe-based or dense-block-based, the fine details and coarse appearance are not well integrated. What is more, the metrics are generally learned independently from distance view or bilinear similarity view. Few works have exploited the mutual complementary effects of their combination. To address these issues, we propose a new feature representation termed enhanced Local Maximal Occurrence (eLOMO) which fuses a new overlapping-stripe-based descriptor with the Local Maximal Occurrence (LOMO) extracted from dense blocks. Such integration makes eLOMO resemble the coarse-to-fine recognition mechanism of human vision system, thus it can provide a more discriminative descriptor for re-identification. Besides, we show the advantages of learning generalized similarity by combining the Mahalanobis distance and bilinear similarity together. Specifically, we derive a logistic metric learning method to jointly learn a distance metric and a bilinear similarity metric, which exploits both the distance and angle information from training data. Taking advantage of learning in the intra-class subspace, the proposed method can be solved efficiently by coordinate descent optimization. Experiments on four challenging datasets including VIPeR, PRID450S, QMUL GRID, and CUHK01, show that the proposed method outperforms the state-of-the-art approaches significantly.

1. Introduction

Person re-identification is the task of matching individuals across disjoint camera views over distributed spaces, which plays an important role in intelligent video surveillance. Although it is assumed that people do not change clothes in different camera views, person re-identification still remains a challenging problem due to large appearance variations caused by illumination, pose, viewpoint, and occlusion.

Great efforts have been devoted for years to tackle person re-identification along two directions. One is to design robust visual descriptors against cross-view variations, and the other is to learn a discriminant similarity/distance functissson to determine whether an image pair belongs to the same person or not. For visual descriptors, a number of feature representations have been proposed, such as Symmetry-Driven Accumulation of Local Features (SDALF) [1], Mid-level Filter (MLF) [2], Biologically Inspired Features (BIF) [3], Salient Color Names (SCN) [4], Local Maximal Occurrence (LOMO) [5], and the Gaussian of Gaussian (GOG) descriptor [6]. Most of them are extracted from either horizontal stripes or dense blocks. Although impressive advancement has been made, designing a more robust yet discriminative descriptor remains an open problem.

As for similarity/distance function learning, a number of metric learning algorithms have been devised [5,7–14]. Some of them, like [10,11,13,14], focus on learning a Mahalanobis distance metric from distance constraints. While some other works, like [7,12], seek for a bilinear similarity metric by utilizing the angle information between instances in high-dimensional feature space. However, most

* Corresponding author at: Changshu Institute of Science and Technology, Changshu, China..
E-mail address: shrgong@cslg.edu.cn (S. Gong).

https://doi.org/10.1016/j.neucom.2018.04.013
0925-2312/

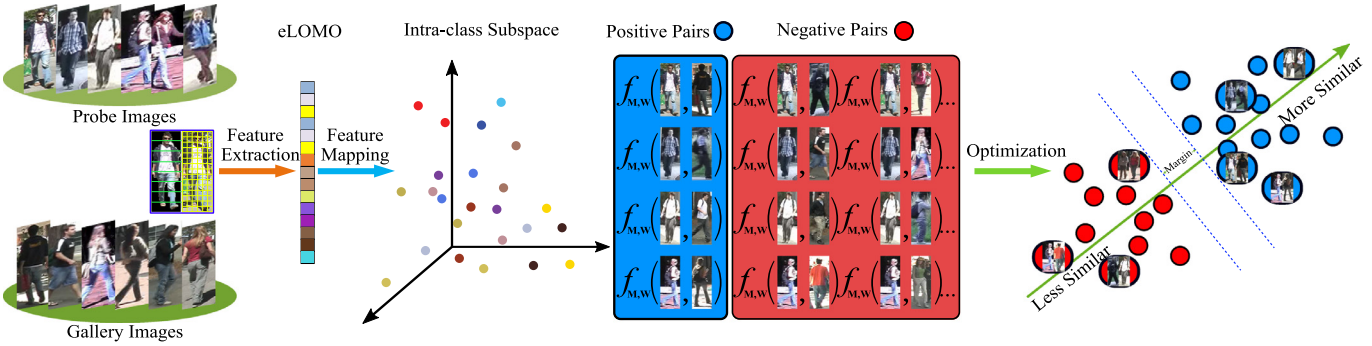


Fig. 1. The pipeline of the proposed method. The eLOMO features are extracted from every image first, and then they are mapped into the intra-class subspace. The generalized similarity function $f_{M,W}$ is trained by maximizing intra-class similarities and minimizing inter-class similarities.

of the works fail to exploit the mutual complementary effects of their combination. Only considering either of them may lead to a less discriminative similarity measurement.

In this paper, we propose an efficient feature representation termed enhanced Local Maximal Occurrence (eLOMO), and a Generalized Similarity Metric Learning (GSML) method for person re-identification. The eLOMO is an integration of a new overlapping-stripe-based descriptor with the existing LOMO [5] feature. The stripe-based descriptor can better exploit coarse appearance information from larger regions, while LOMO is good at capturing the fine details of dense blocks. Thus the fusion of them can lead to a coarse-to-fine representation which is in line with the human recognition mechanism. To learn a discriminant similarity function, we combine the Mahalanobis distance and the bilinear similarity together, such that the distance and angle information of training data are exploited simultaneously. The proposed method is formulated as a logistic metric learning problem with Positive Semi-definite (PSD) constraints, and we derive an efficient coordinate descent algorithm to solve it based on the Accelerated Proximal Gradient (APG) optimization method. To suppress the large intra-class variations of cross-view appearances, we project samples into the intra-class subspace before learning. The pipeline of the proposed method is shown in Fig. 1.

We conduct extensive experiments to validate the efficacy of the proposed method. Experimental results show that the proposed method achieves significant improvements over existing approaches on four challenging person re-identification datasets, namely VIPeR [15], PRID450S [16], QMUL GRID [17], and CUHK01 [2].

The rest of this paper is organized as follows. In Section 2 we briefly review the related works and discuss their differences with our method. Section 3 introduces the eLOMO feature representation. Section 4 presents the GSML in detail. The experimental results and the analysis of our method are reported in Section 5. Finally, we draw some conclusions in Section 6.

2. Related work

Given one probe image containing an individual of interest, the task of person re-identification is to find its true match (or usually the best match) from a large number of gallery images. Existing works for solving this problem generally follow a two-step paradigm. Firstly, a robust and distinctive feature representation is extracted for every pedestrian image. Secondly, the similarity/distance for each probe-gallery image pair is measured by a certain metric, which is then used to rank the gallery images for each probe. Majority of existing methods focus on either designing feature representations or learning discriminative metrics. Here we only briefly review some related works, more comprehensive surveys can be found in [18–21].

2.1. Feature representations for person re-identification

Many approaches try to build distinctive feature representation for describing pedestrian appearance in different environments. In order to achieve discrimination and robustness against different variations, they are generally extracted from horizontal stripes [4,15,22,23] or dense blocks [5,8,9,24,25]. For example, the Ensemble of Local Features (ELF) [15] and the SCN [4] are extracted from six non-overlapping horizontal stripes. As an extension of ELF, the ELF18 [22] is computed from 18 non-overlapping stripes. From overlapping stripes, Lisanti et al. [23] computed the weighted local features of color histogram, Local Binary Patterns (LBP), and Histogram of Gradient (HOG). In general, these stripe-based descriptors are robust to the cross-view body misalignment problem and can well capture the holistic appearance information.

Compared to stripe-based descriptors, the features computed from dense blocks can better capture fine details from relatively small patches. By computing Gabor filters and covariance from dense grids, Ma et al. [3] proposed the BIF feature. Zhao et al. [2] tried to learn Mid-level filters from the clusters of dense patches. Recently, Liao et al. [5] proposed the LOMO descriptor which has shown impressive robustness against viewpoint changes. By describing each pedestrian image as a set of hierarchical Gaussian distributions represented by the means and covariance, Matsukawa et al. [6] designed the GOG descriptor which is also computed from dense blocks in essence. However, just as one coin has two sides, a disadvantage of these dense-block-based descriptors is that they are not good at describing the holistic appearance, although some of them have considered computing features from multi-scale spaces (e.g. LOMO and GOG).

To combine the advantages of both stripe-based and dense-block-based features, we argue that feature representations should be computed from stripes (larger regions) and dense blocks (small patches) simultaneously. In this work, we fuse the successful LOMO with the features extracted from a pyramid space of two-level overlapping stripes. As a consequence, the fine details from dense blocks and coarse appearance from larger regions are well integrated to boost the discrimination.

With the blossom of deep learning, there are also some works try to learn features by the powerful deep models, such as [26–28]. Even with the generic metric of $L2$ norm, impressive performance can be achieved. However, they require very large number of training data and can easily suffer the over-fitting

risk. Especially on small datasets with limited training samples (e.g. VIPeR [15] and GRID [17]), they are still inferior to metric learning approaches.

2.2. Metric learning for person re-identification

Metric learning has been widely applied for re-identification problem, and the metrics are generally learned from either distance view [9–11,14,24] or bilinear similarity view [7,12]. From the distance view, Mignon et al. [11] proposed the Pairwise Constrained Component Analysis (PCCA) to learn a projection matrix from sparse pairwise constraints. Li et al. [24] proposed the Locally Adaptive Decision Functions (LADF) to jointly learn the distance metric and locally adapted thresholds. To address the problem of imbalanced training data, Liao and Li [10] proposed to learn the metric from a logistic formulation by applying asymmetric weighting strategy. Using triplet-wise constraints, Zheng et al. [14] proposed to learn the metric from relative distance comparison. To tackle the cross-view misalignment problem, Sun et al. [29] integrated the metric with some latent variables during learning. By exploiting easily-available negative samples, Zhou et al. [30] proposed to learn a global metric with some fine-tuned local metrics. Without heavy optimization, the Keep It Simple and Straightforward MEtric learning (KISSME) [9] algorithm derived a closed-form solution of the metric which can be computed very efficiently. As an extension of KISSME, the Cross-view Quadratic Discriminant Analysis (XQDA) [5] learned a more effective metric accompanied by a discriminative subspace. There are also some works try to handle the non-linear appearance transformation patterns across camera views via the kernel trick [31] or neural networks [32,33], which are usually called non-linear metric learning.

Different from above methods that learn metrics from distance constraints, some other works tried to learn metrics with the discriminative angle information between instances in high-dimensional feature space. Chen et al. [7] proposed to learn a discriminative metric from listwise inner product similarity constraints. Ustinova et al. [34] learned bilinear similarities via multi-regional Convolutional Neural Networks (CNN). To fuse global and local similarity scores, Lisanti et al. [35] tried to learn multi-channel Kernel Canonical Component Analysis (KCCA) for computing pairwise cosine similarities. Although such works of exploiting angle information between sample pairs seem relatively sparse in person re-identification, they are much more popular in other fields like face recognition and cross-media retrieval. For example, Nguyen and Bai [12] tried to learn a cosine metric to measure the similarity of face pairs. Kang et al. [36] learned a low-rank bilinear similarity metric for cross-modal image retrieval. To learn the similarity metric in high dimensions, Liu et al. [37] introduced the Alternate Direction Method of Multiplier (ADMM) algorithm for optimization.

In aforementioned works, the discriminative information are exploited separately from distance constraints or bilinear similarity constraints. The mutual complementary effects are ignored. In this paper, we propose an efficient algorithm termed GSML to learn a generalized similarity by combining the functions of Mahalanobis distance and bilinear similarity. Therefore, our GSML can jointly exploit the distance and angle information from training samples to enhance the discrimination of learned metrics.

The proposed algorithm is mostly related to the Metric Learning by Accelerated Proximal Gradient (MLAPG) [10], Spatially Constrained Similarity function on Polynomial feature map (SCSP) [38], and Subspace Similarity Metric Learning (Sub-SML) [39]. Since a similarity metric is learned besides the distance metric, our GSML can be viewed as an extension of MLAPG. Although the similarity function in SCSP can also be viewed as a model of combining the distance and similarity metrics, it is achieved by applying a polynomial kernel map [40] for sample pairs. In addition, MLAPG and SCSP learn metrics in the Principle Component Analysis (PCA) subspace directly, while the metrics in GSML are jointly learned in the intra-class subspace, which differs from them obviously. As for Sub-SML, though sharing the superiority of learning generalized similarity, GSML is quite different from it in the formulation. The Sub-SML takes a form of the Supported Vector Machine (SVM) framework, while our GSML is formulated as a logistic metric learning problem.

3. Enhanced local maximal occurrence representation

Similar to the coarse-to-fine recognition mechanism of human vision system, a discriminative feature representation for visual learning should also take both fine details and holistic appearance information into consideration. The advantage is that they can work co-operatively to capture the invariance of pedestrian appearance in different camera views. As a result, it will greatly help to identify the interested target. Although some descriptors like LOMO and GOG have considered computing local features from multi-scale spaces, there is still space to improve. Here, we fuse the successful LOMO feature representation with a new stripe-based descriptor to achieve this goal.

3.1. Review of LOMO

The LOMO feature representation consists of two elementary features of joint HSV histogram and Scale Invariant Local Ternary Pattern (SILTP). To deal with cross-view illumination variations, the Retinex algorithm is applied first to preprocess pedestrian images before extracting HSV histogram. Then, both HSV histogram and SILTP are extracted from 50% overlapped dense blocks with size of 10×10 pixels. To deal with viewpoint changes in different cameras, an operation of maximizing the local occurrence pattern is executed. Such an operation can favor in capturing local information and achieving some viewpoint robustness. To obtain multi-scale information, LOMO further builds a three-scale pyramid space by 2×2 average pooling operation and extracts features on each scale. The final descriptor is the concatenation of features obtained on all scales with total dimensions of 26960 for a 128×48 image.

Although LOMO has already been widely used in person re-identification with impressive performance [5,10,41,42], it is not a good feature extractor for capturing holistic appearance information of larger regions due to the computation from dense blocks. To enhance its discrimination, we fuse it with the features extracted from horizontal stripes.

3.2. Stripe-based pyramid features

We extract the stripe-based features from a two-level pyramid space of overlapping stripes obtained similar to [23]. To decrease the interference of background clutter, we first use Deep Decompositional Network (DDN) [43] to estimate the foregrounds for each image if no masks provided. Then, each foreground image is equally divided into 8 horizontal stripes. After that, we crop the foreground image by abandoning 1/2 stripe height from the top and bottom. The left foreground is re-divided into 7 stripes again. So there are totally 15 horizontal stripes with the same size obtained.

From each stripe, we extract four elementary features: $8 \times 8 \times 8$-bin joint histograms of HSV and RGB, two scales of $\mathrm{SILTP}^{0.3}_{4,3}$ and $\mathrm{SILTP}^{0.3}_{4,5}$ histograms [5], and the SCN [4] feature. The joint HSV histogram and SILTP are extracted with the same settings in LOMO, and the joint RGB histogram is computed in the same way as HSV histogram with only difference in the color space. The

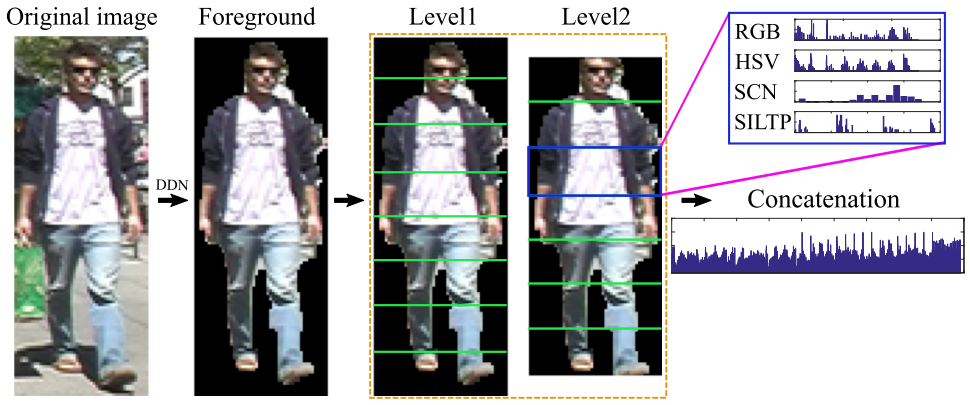


Fig. 2. Illustration of the SPF extraction procedure.

Table 1
Comparison of LOMO and SPF descriptors.

Descriptor	Pre-processing	Color	Texture	Regions	Multi-scale space	Foreground
LOMO	Retinex (HSV)	HSV histogram	SILTP	10×10 dense blocks	2×2 average pooling	No
SPF	Retinex (HSV)	HSV/RGB histogram, SCN	SILTP	Horizontal stripes	Two-level stripe space	DDN

SCN is computed using the same 16 standard colors as in [4]. Different from LOMO, there are no operations of maximizing local occurrence pattern after computing the local features, because each stripe is computed as a whole. At last, we concatenate the four elementary features obtained from all stripes, and obtain a descriptor with total dimensions of $(8^3 + 8^3 + 3^4 \times 2{+}16) \times 15{=}18030$. To obtain some robustness to noise, we normalize each elementary feature to unit length except SCN as it has already be normalized. Since the features are extracted from a stripe-based pyramid space, we call the obtained descriptor Stripe-based Pyramid Features (SPF). The extraction procedure of SPF descriptor is shown Fig. 2.

Table 1 shows the comparison of LOMO and SPF, from which we can find the two descriptors share some common characteristics. Both employ color histogram and SILTP to capture pedestrian appearance, and the Retinex is applied before computing joint HSV histogram. So SPF can be considered as a stripe-based variant of LOMO. The main difference between LOMO and SPF lies in their extraction way. LOMO is computed from dense blocks on different scale spaces, while SPF is extracted from two-level overlapping stripes. Besides, the additional RGB histogram and SCN in SPF can provide richer color information than the only HSV histogram in LOMO. Our motivation of composing SPF descriptor is to capture the holistic information from larger regions, such that the fusion of it and LOMO feature can capture complementary aspects of pedestrian appearance.

In Fig. 3, we show an example of the joint HSV histograms computed from one pedestrian image stripe via the LOMO and SPF extractors. It can be found that among the obtained histograms there are more bins in the LOMO patterns, while the histogram of SPF seems much 'cleaner'. This indicates the LOMO extractor can capture more details than SPF, while the latter is good at describing the holistic appearance. This is because there are generally about four color patterns in the cropped foreground stripe, and they correspond to the four clusters of bins in the SPF histogram. Inspired by the characteristics of human vision system, we believe the fusion of SPF and LOMO can provide a descriptor to perform coarse-to-fine recognition. In consequence, the fused descriptor is supposed to be much more discriminative than each of them alone. Since SPF can be viewed as a stripe-based variant of LOMO,

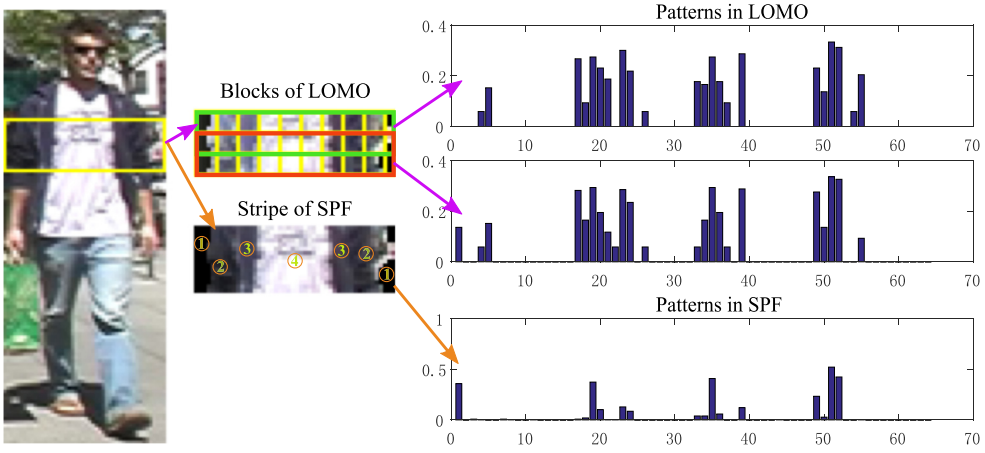


Fig. 3. Example of the coarse-to-fine representation effect in the fusion of LOMO and SPF. Four better visualization, $4 \times 4 \times 4$-bin joint histograms are extracted. The numbers in circle indicate different color patterns.

we call the fused descriptor enhanced Local Maximal Occurrence (eLOMO).

4. Generalized similarity metric learning in intra-class subspace

4.1. Problem formulation

Let {**X, Z, Y**} be a given cross-view training set, where $\mathbf{X} \in \mathbb{R}^{d\times n}$ and $\mathbf{Z} \in \mathbb{R}^{d\times m}$ are the feature matrices of probe set and gallery set, with n and m samples in a d-dimensional feature space respectively; $\mathbf{Y} \in \mathbb{R}^{n\times m}$ is the matching label matrix between **X** and **Z**, with $y_{ij} = 1$ if $(\boldsymbol{x}_i, \boldsymbol{z}_j)$ is a positive pair (i.e. $\boldsymbol{x}_i$ and $\boldsymbol{z}_j$ represent the same person), and $y_{ij} = -1$ otherwise. The re-identification task is to learn a similarity function $f(\boldsymbol{x}_i, \boldsymbol{z}_j)$ to measure the similarity between each pair $\{(\boldsymbol{x}_i, \boldsymbol{z}_j)\}_{i,j=1}^{n,m}$. However, it is impracticable to directly learn function f in the original d-dimensional space, because d is usually as large as thousands, or even tens of thousands in person re-identification task. For example, the dimension of our visual descriptor eLOMO on VIPeR [15] dataset amounts to as high as 44,990 (26,960+18,030=44,990). To lower the computational cost, PCA is commonly applied to reduce feature dimension and remove noise [9,10,40].

Another problem in learning function f is the large intra-class variations cross camera views, which may seriously affect the matching accuracy. To reduce the influence of such handicap, we follow the idea in [44] to map samples into the intra-class subspace before learning. Let $\mathbf{X}' \in \mathbb{R}^{d'\times n}$ and $\mathbf{Z}' \in \mathbb{R}^{d'\times m}$ be the feature matrices after PCA, where $d'(d' \le d)$ is the reduced feature dimension. We first compute the intra-class covariance matrix as

$$\boldsymbol{\Sigma}_{\mathcal{P}} = \sum_{(i,j)\in\mathcal{P}} (\boldsymbol{x}'_i - \boldsymbol{z}'_j)(\boldsymbol{x}'_i - \boldsymbol{z}'_j)^{\top}, \tag{1}$$

where $\mathcal{P}$ is the index set of positive pairs, $\boldsymbol{x}'_i$ and $\boldsymbol{z}'_j$ are the ith and jth samples in $\mathbf{X}'$ and $\mathbf{Z}'$ respectively. Let $\boldsymbol{\Lambda}_{\mathcal{P}} = \{\lambda_1, \ldots, \lambda_k\}$ be the top-leading k ($k \le d'$) eigenvalues of $\boldsymbol{\Sigma}_{\mathcal{P}}$, and $\mathbf{V}_{\mathcal{P}} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] \in \mathbb{R}^{d'\times k}$ be a matrix consisting of the corresponding eigenvectors, we can map $\boldsymbol{x}'_i$ and $\boldsymbol{z}'_j$ into the k-dimensional intra-class subspace by the whitening process:

$$\begin{aligned} \widetilde{\boldsymbol{x}}_i &= diag\big(\lambda_1^{-1/2}, \ldots, \lambda_k^{-1/2}\big)\mathbf{V}_{\mathcal{P}}^{\top}\boldsymbol{x}'_i, \\ \widetilde{\boldsymbol{z}}_j &= diag\big(\lambda_1^{-1/2}, \ldots, \lambda_k^{-1/2}\big)\mathbf{V}_{\mathcal{P}}^{\top}\boldsymbol{z}'_j. \end{aligned} \tag{2}$$

In this way, the eigenvectors with large eigenvalues will be penalized, because they are weighted by the reciprocal of eigenvalues. As a result, the variance of features will be reduced, and this is equivalent to reduce the intra-class variations, which will greatly help to separate each positive pair from negative ones. In practice, $\boldsymbol{x}'_i$ and $\boldsymbol{z}'_j$ are already in a much lower d'-dimensional space after PCA, there is no need to choose a smaller k anymore. In this work, we set the dimension of intra-class subspace to $k = d'$. In this case, $\boldsymbol{\Sigma}_{\mathcal{P}}$ is invertible and we can obtain $\boldsymbol{\Sigma}_{\mathcal{P}} = \mathbf{L}_{\mathcal{P}}\mathbf{L}_{\mathcal{P}}^{\top}$ by Cholersky decomposition. Then we have $\mathbf{L}_{\mathcal{P}} = \mathbf{V}_{\mathcal{P}} diag(\lambda_1^{1/2}, \ldots, \lambda_k^{1/2})$, and Eq. (2) can be rewritten as $\widetilde{\boldsymbol{x}}_i = \mathbf{L}_{\mathcal{P}}^{-1}\boldsymbol{x}'_i$ and $\widetilde{\boldsymbol{z}}_j = \mathbf{L}_{\mathcal{P}}^{-1}\boldsymbol{z}'_j$.

With all training samples mapped to intra-class subspace, we formulate the similarity of a pair $(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j)$ as

$$f_{\mathbf{M},\mathbf{W}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j) = s_{\mathbf{W}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j) - d^2_{\mathbf{M}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j), \tag{3}$$

where $d^2_{\mathbf{M}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j) = (\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{z}}_j)^{\top}\mathbf{M}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{z}}_j)$ is the Mahalanobis distance function [45], and $\mathbf{M} \succeq \mathbf{0}$ is a PSD matrix to guarantee the validity of $d^2_{\mathbf{M}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j)$. In this case, we can also obtain $d^2_{\mathbf{M}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j) = \big\|\mathbf{L}_{\mathbf{M}}^{\top}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{z}}_j)\big\|_2^2$ with Cholersky decomposition $\mathbf{M} = \mathbf{L}_{\mathbf{M}}\mathbf{L}_{\mathbf{M}}^{\top}$. This means the Mahalanobis distance is equivalent to the Euclidean distance in subspace $\mathbf{L}_{\mathbf{M}}$. In Eq. 3, $s_{\mathbf{W}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j) = \widetilde{\boldsymbol{x}}_i^{\top}\mathbf{W}\widetilde{\boldsymbol{z}}_j$ is the bilinear similarity function [36], and we also impose the constraint of $\mathbf{W} \succeq \mathbf{0}$ which leads to $\mathbf{W} = \mathbf{L}_{\mathbf{W}}\mathbf{L}_{\mathbf{W}}^{\top}$. Then we can obtain $s_{\mathbf{W}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j) = (\mathbf{L}_{\mathbf{W}}^{\top}\widetilde{\boldsymbol{x}}_i)^{\top}(\mathbf{L}_{\mathbf{W}}^{\top}\widetilde{\boldsymbol{z}}_j)$, which is the inner product similarity of $\widetilde{\boldsymbol{x}}_i$ and $\widetilde{\boldsymbol{z}}_j$ in subspace $\mathbf{L}_{\mathbf{W}}$. Noting that $\mathbf{L}_{\mathbf{W}}$ and $\mathbf{L}_{\mathbf{M}}$ has the same dimensions here. Therefore, $f_{\mathbf{M},\mathbf{W}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j)$ is a combination of the inner product similarity and the negative Mahalanobis distance, which is a generalized similarity parameterized by **W** and **M**. The higher value of $f_{\mathbf{M},\mathbf{W}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j)$, the more similar $\widetilde{\boldsymbol{x}}_i$ and $\widetilde{\boldsymbol{z}}_j$ are.

To show the superiority of learning generalized similarity, let's consider a simple example in 1-dimensional space: the samples of class 1 are uniformly distributed in $[-2, -0.5]$, and the class 2 in $[0.5, 2]$. Although the two classes are separable, we cannot achieve this with only distance constraint. This is because $\max_{y_{ij}=1} \|x_i - z_j\|_2 = 1.5$ and $\min_{y_{ij}\neq 1} \|x_i - z_j\|_2 = 1$. From the Probability Density Function (PDF) of $\Delta = x - z$ shown in Fig. 4 (a), we can see there are inevitable errors to classify the points. When the generalized similarity is considered, i.e., let $f(x_i, z_j) = x_i z_j - \|x_i - z_j\|_2$, we have $-8 \le f(x_i, z_j | y_{ij=-1}) \le -1.25$ and $-0.5 \le f(x_i, z_j | y_{ij=1}) \le 4$. The two classes can be easily separated with zero error. In 2-dimensional space, the problem seems a bit sophisticated as shown in Fig. 4 (b). By using Eq. 3 to learn a decision boundary of $xy = 0.2$, we can still separate them perfectly. Based on this example, we believe the advantages of learning generalized similarity also exist in high-dimensional space.

To learn $f_{\mathbf{M},\mathbf{W}}$, we choose the log-logistic loss function similar as in [10,11]:

$$l_{\mathbf{M},\mathbf{W}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j) = \log\big(1 + \exp\big(y_{ij}\big(\delta - f_{\mathbf{M},\mathbf{W}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j)\big)\big)\big), \tag{4}$$

where $\delta \ge 0$ is a constant positive bias which is set to $E[f_{\mathbf{I},\mathbf{I}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j)]$ ($\mathbf{I}$ is the $d' \times d'$ identity matrix) in this work. Intuitively, the logistic formulation of Eq. (4) provides a soft margin between positive and negative sample pairs. It will penalize the positive pairs with low similarities and the negative pairs that have high similarities. Besides, Eq. (4) is a differentiable function which will greatly benefit the optimization. Then, the total loss over whole training set is

$$\mathcal{L}(\mathbf{M}, \mathbf{W}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \beta_{ij} l_{\mathbf{M},\mathbf{W}}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j), \tag{5}$$

where $\beta_{ij} = 1/N^+$ if $y_{ij} = 1$, and $1/N^-$ otherwise, N^+ and N^- are the numbers of positive and negative pairs respectively. Such an asymmetric weighting strategy can lead to more robust metrics due to the imbalance of sample pairs [10]. Minimizing the total loss with respect to **M** and **W** will encourage the separation of positive pairs from negative ones. As a result, our generalized similarity metric learning problem can be formulated as

$$\min_{\mathbf{M},\mathbf{W}} \mathcal{L}(\mathbf{M}, \mathbf{W}), \quad s.t. \mathbf{M} \succeq \mathbf{0}, \mathbf{W} \succeq \mathbf{0}. \tag{6}$$

4.2. Optimization

There are two coupled metrics to be optimized in Eq. (6), which leads to a non-convex optimization problem. Here we solve it by the coordinate descent method, i.e., we optimize one metric with the other fixed. Although only local optimal solutions can be obtained in this way, we find the learned metrics are still very discriminative in experiments.

Considering the optimization for **M** or **W** alone, the problem is structured as a nonlinear but convex and smooth objective function, with the PSD constraint. This is especially proper to be solved by the APG optimization algorithm [46,47]. The APG belongs to the first-order gradient descent optimization methods with a fast convergence of $O(1/t^2)$, where t is the iteration number [46].

Optimize M with fixed W. For convenience, the optimization with respect to **M** can be rewritten as

$$\min_{\mathbf{M}} \mathcal{L}(\mathbf{M}) = g(\mathbf{M}) + h(\mathbf{M}), \tag{7}$$

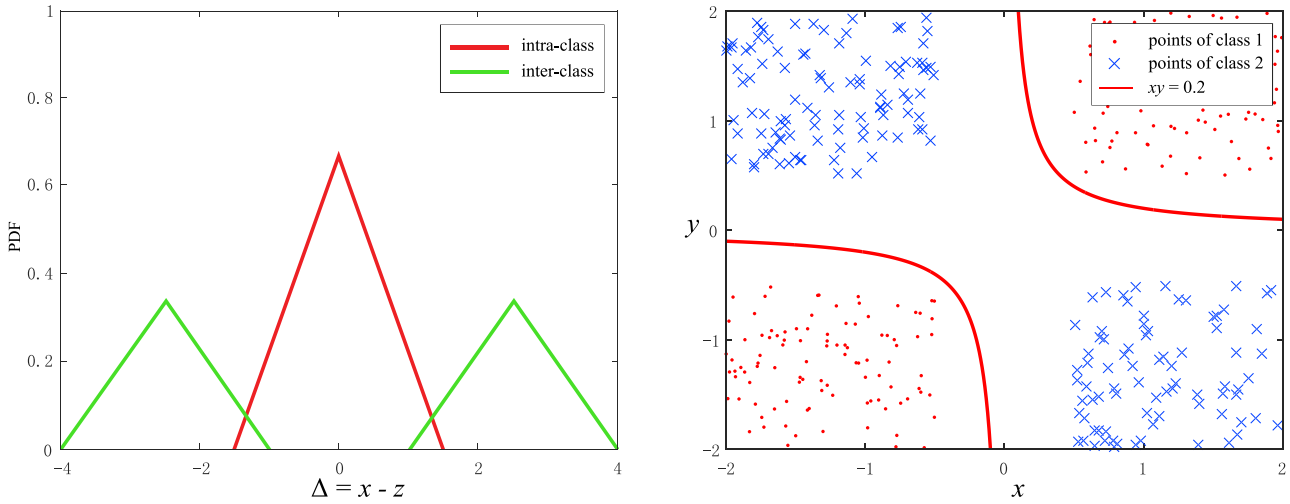


Fig. 4. (a) Distribution of $\Delta = x - z$ in case of samples of class 1 are uniformly distributed in $[-2, -0.5]$ and class 2 in $[0.5, 2]$. (b) In 2-dimensional space, the two classes can be perfectly separated by a decision boundary of $xy = 0.2$.

with

$$g(\mathbf{M}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \beta_{ij} l_{\mathbf{M},\mathbf{W}}\left(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j\right), \tag{8}$$

$$h(\mathbf{M}) = \mathbf{I}_{\mathbb{S}_+^{d'}}(\mathbf{M}), \tag{9}$$

where $\mathbb{S}_+^{d'}$ indicates the set of symmetric PSD matrices with size of $d' \times d'$, and $\mathbf{I}_{\mathbb{S}_+^{d'}}(\mathbf{M})$ is the indicator function defined as

$$\mathbf{I}_{\mathbb{S}_+^{d'}}(\mathbf{M}) = \begin{cases} 0 & \mathbf{M} \in \mathbb{S}_+^{d'} \\ +\infty & otherwise \end{cases}. \tag{10}$$

To find the optimal solution $\mathbf{M}^*$, APG first constructs a proximal operator of $g(\mathbf{M})$ at current aggregation point $\mathbf{V}_{\mathbf{M}}^t (t \geq 1)$ to obtain a point $\mathbf{M}_t$, and then optimizes $h(\mathbf{M})$, both decrease the objective value $\mathcal{L}(\mathbf{M})$ [46]. After that, APG constructs the next aggregation point $\mathbf{V}_{\mathbf{M}}^{t+1}$ by an extrapolation of $\mathbf{M}_{t-1}$ and $\mathbf{M}_t$. By repeating these operations, the $\mathbf{M}^*$ will be found within a given tolerance.

The proximal operator of $g(\mathbf{M})$ at aggregation point $\mathbf{V}_{\mathbf{M}}^t$ can be constructed as

$$P_{\eta_t}\left(\mathbf{M}, \mathbf{V}_{\mathbf{M}}^t\right) = g\left(\mathbf{V}_{\mathbf{M}}^t\right) + \left\langle \mathbf{M} - \mathbf{V}_{\mathbf{M}}^t, \nabla g\left(\mathbf{V}_{\mathbf{M}}^t\right)\right\rangle + \frac{1}{2\eta_t}\left\|\mathbf{M} - \mathbf{V}_{\mathbf{M}}^t\right\|_F^2, \tag{11}$$

where $\langle \cdot, \cdot \rangle$ is the operation of matrix inner product, $\|\cdot\|_F$ is the Frobenius norm, and η_t is the step size. When $\eta_t \in (0,\ 1/L_{\mathbf{M}}]$, P_{η_t} is an upper bound of $g(\mathbf{M})$ [46], where $L_{\mathbf{M}}$ is a Lipschitz constant of $\nabla g(\mathbf{M})$. $\nabla g\left(\mathbf{V}_{\mathbf{M}}^t\right)$ is the gradient of $g(\mathbf{M})$ at point $\mathbf{V}_{\mathbf{M}}^t$, which is computed as

$$\nabla g\left(\mathbf{V}_{\mathbf{M}}^t\right) = \frac{\partial g(\mathbf{V})}{\partial \mathbf{V}}\bigg|_{\mathbf{V}_{\mathbf{M}}^t} = \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}^t \left(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{z}}_j\right)\left(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{z}}_j\right)^\top, \tag{12}$$

where

$$h_{ij}^t = \frac{\beta_{ij} y_{ij}}{1 + \exp\left(y_{ij}\left(f_{\mathbf{V}_{\mathbf{M}}^t,\mathbf{W}}\left(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j\right) - \delta\right)\right)}, \tag{13}$$

and $f_{\mathbf{V}_{\mathbf{M}}^t,\mathbf{W}}\left(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j\right) = s_{\mathbf{W}}\left(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j\right) - d^2_{\mathbf{V}_{\mathbf{M}}^t}\left(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j\right)$. Eq. (12) needs heavy computation due to the outer product. To accelerate the computation, it can be rewritten in the matrix operation form as

$$\nabla g\left(\mathbf{V}_{\mathbf{M}}^t\right) = \widetilde{\mathbf{X}}\mathbf{R}_t\widetilde{\mathbf{X}}^\top - \widetilde{\mathbf{X}}\mathbf{H}_t\widetilde{\mathbf{Z}}^\top - \left(\widetilde{\mathbf{X}}\mathbf{H}_t\widetilde{\mathbf{Z}}^\top\right)^\top + \widetilde{\mathbf{Z}}\mathbf{C}_t\widetilde{\mathbf{Z}}^\top, \tag{14}$$

where $\widetilde{\mathbf{X}} = [\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, \ldots, \widetilde{\boldsymbol{x}}_n]$ and $\widetilde{\mathbf{Z}} = [\widetilde{\boldsymbol{z}}_1, \widetilde{\boldsymbol{z}}_2, \ldots, \widetilde{\boldsymbol{z}}_m]$ are the feature matrices have been embedded into the intra-class subspace, $\mathbf{H}_t$ is a $n \times m$ matrix with entries of h_{ij}^t. $\mathbf{R}_t$ and $\mathbf{C}_t$ are two diagonal matrices, and their main diagonal entries are the row sum and column sum of $\mathbf{H}_t$, respectively.

In the tth iteration of APG, the minimization of Eq. (8) is identical to the problem of

$$\min_{\mathbf{M}}\ P_{\eta_t}\left(\mathbf{M}, \mathbf{V}_{\mathbf{M}}^t\right). \tag{15}$$

By adding $\frac{\eta_t}{2}\left\|\nabla g\left(\mathbf{V}_{\mathbf{M}}^t\right)\right\|_F^2$ and removing $g\left(\mathbf{V}_{\mathbf{M}}^t\right)$ that are both independent of $\mathbf{M}$, the solution of Eq. (15) is equivalent to

$$\widetilde{\mathbf{M}}_t = \arg\min_{\mathbf{M}} \frac{1}{2\eta_t}\left\|\mathbf{M} - \left(\mathbf{V}_{\mathbf{M}}^t - \eta_t \nabla g\left(\mathbf{V}_{\mathbf{M}}^t\right)\right)\right\|_F^2. \tag{16}$$

It can be found this is a least square problem and the solution is

$$\widetilde{\mathbf{M}}_t = \mathbf{V}_{\mathbf{M}}^t - \eta_t \nabla g\left(\mathbf{V}_{\mathbf{M}}^t\right). \tag{17}$$

Considering the optimization of $h(\mathbf{M})$, which is an indicator function and the solution can be obtained by projecting $\widetilde{\mathbf{M}}_t$ onto the PSD cone [46]:

$$\mathbf{M}_t = \mathbf{U}_{\widetilde{\mathbf{M}}_t} \mathbf{\Lambda}_{\widetilde{\mathbf{M}}_t}^+ \mathbf{U}_{\widetilde{\mathbf{M}}_t}^\top, \tag{18}$$

where $\mathbf{U}_{\widetilde{\mathbf{M}}_t} \mathbf{\Lambda}_{\widetilde{\mathbf{M}}_t} \mathbf{U}_{\widetilde{\mathbf{M}}_t}^\top$ is the eigenvalue decomposition of $\widetilde{\mathbf{M}}_t$, and $\mathbf{\Lambda}_{\widetilde{\mathbf{M}}_t}^+ = \max\left\{\mathbf{\Lambda}_{\widetilde{\mathbf{M}}_t}, 0\right\}$.

When $\mathbf{M}_t$ is obtained, APG constructs the aggregation forward matrix $\mathbf{V}_{\mathbf{M}}^{t+1}$ to accelerate the proximal gradient descent as

$$\mathbf{V}_{\mathbf{M}}^{t+1} = \mathbf{M}_t + \frac{\theta_t - 1}{\theta_{t+1}}(\mathbf{M}_t - \mathbf{M}_{t-1}), \tag{19}$$

where $\theta_{t+1} = (1 + \sqrt{4\theta_t^2 - 1})/2$ and $\theta_0 = 1$ following [48].

Optimize W with fixed M. Similar to the optimization of $\mathbf{M}$, we split Eq. (6) into two parts when only $\mathbf{W}$ is considered:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = g(\mathbf{W}) + h(\mathbf{W}), \tag{20}$$

where

$$g(\mathbf{W}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \beta_{ij} l_{\mathbf{M},\mathbf{W}}\left(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j\right), \tag{21}$$

$$h(\mathbf{W}) = \mathbf{I}_{\mathbb{S}_+^{d'}}(\mathbf{W}). \tag{22}$$

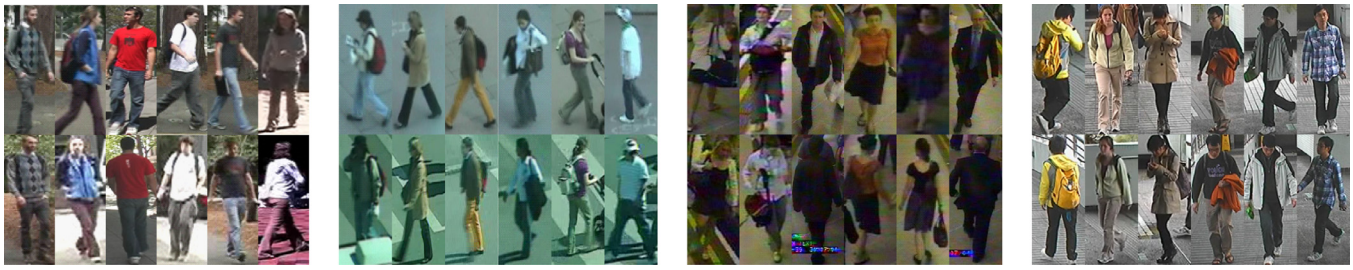


Fig. 5. Example image pairs from VIPeR, PRID450S, QMUL GRID, and CUHK01 datasets. Images in the same column are captured from the same person by different cameras.

We construct the proximal operator of $g(\mathbf{W})$ at current aggregation point $\mathbf{V}_{\mathbf{W}}^t (t \geq 1)$ as

$$P_{\mu_t}\left(\mathbf{W}, \mathbf{V}_{\mathbf{W}}^t\right) = g\left(\mathbf{V}_{\mathbf{W}}^t\right) + \left\langle \mathbf{W} - \mathbf{V}_{\mathbf{W}}^t, \nabla g\left(\mathbf{V}_{\mathbf{W}}^t\right)\right\rangle + \frac{1}{2\mu_t}\left\|\mathbf{W} - \mathbf{V}_{\mathbf{W}}^t\right\|_F^2, \tag{23}$$

where μ_t is the update step size. Similar to η_t in Eq. (11), when $\mu_t \in (0,\ 1/L_{\mathbf{W}}]$, P_{μ_t} is an upper bound of $g(\mathbf{W})$, where $L_{\mathbf{W}}$ is a Lipschitz constant of $\nabla g(\mathbf{W})$. The $\nabla g\left(\mathbf{V}_{\mathbf{W}}^t\right)$ is computed as

$$\nabla g\left(\mathbf{V}_{\mathbf{W}}^t\right) = \frac{\partial g(\mathbf{V})}{\partial \mathbf{V}}\bigg|_{\mathbf{V}_{\mathbf{W}}^t} = \sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij}^t \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{z}}_j^\top, \tag{24}$$

where

$$q_{ij}^t = \frac{-\beta_{ij} y_{ij}}{1 + \exp\left(y_{ij}\left(f_{\mathbf{M},\mathbf{V}_{\mathbf{W}}^t}\left(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j\right) - \delta\right)\right)}, \tag{25}$$

and $f_{\mathbf{M},\mathbf{V}_{\mathbf{W}}^t}\left(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j\right) = s_{\mathbf{V}_{\mathbf{W}}^t}\left(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j\right) - d_{\mathbf{M}}^2\left(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{z}}_j\right)$. Eq. (24) can also be rewritten in the matrix operation form of $\nabla g\left(\mathbf{V}_{\mathbf{W}}^t\right) = \widetilde{\mathbf{X}}\mathbf{Q}_t\widetilde{\mathbf{Z}}^\top$ to accelerate computation, here $\mathbf{Q}_t$ is a matrix with entries of q_{ij}^t.

Minimizing $g(\mathbf{W})$ at the tth iteration is equivalent to solve

$$\min_{\mathbf{W}}\ P_{\mu_t}\left(\mathbf{W}, \mathbf{V}_{\mathbf{W}}^t\right). \tag{26}$$

By adding $\frac{\mu_t}{2}\left\|\nabla g\left(\mathbf{V}_{\mathbf{W}}^t\right)\right\|_F^2$ and removing $g\left(\mathbf{V}_{\mathbf{W}}^t\right)$ that are irrelevant to $\mathbf{W}$, the solution of Eq. (26) is equivalent to

$$\widetilde{\mathbf{W}}_t = \arg\min_{\mathbf{W}} \frac{1}{2\mu_t}\left\|\mathbf{W} - \left(\mathbf{V}_{\mathbf{W}}^t - \mu_t \nabla g\left(\mathbf{V}_{\mathbf{W}}^t\right)\right)\right\|_F^2, \tag{27}$$

and we can obtain $\widetilde{\mathbf{W}}_t = \mathbf{V}_{\mathbf{W}}^t - \mu_t \nabla g\left(\mathbf{V}_{\mathbf{W}}^t\right)$.

As for the optimization of $h(\mathbf{W})$, we should project $\widetilde{\mathbf{W}}_t$ onto the PSD cone, which is similar to Eq. (18). That is, $\mathbf{W}_t = \mathbf{U}_{\widetilde{\mathbf{W}}_t}\boldsymbol{\Lambda}_{\widetilde{\mathbf{W}}_t}^{+}\mathbf{U}_{\widetilde{\mathbf{W}}_t}^\top$, where $\mathbf{U}_{\widetilde{\mathbf{W}}_t}\boldsymbol{\Lambda}_{\widetilde{\mathbf{W}}_t}\mathbf{U}_{\widetilde{\mathbf{W}}_t}^\top$ is the eigenvalue decomposition of $\mathbf{W}_t$, and $\boldsymbol{\Lambda}_{\widetilde{\mathbf{W}}_t}^{+} = \max\left\{\boldsymbol{\Lambda}_{\widetilde{\mathbf{W}}_t}, 0\right\}$. Next, an aggregation forward matrix $\mathbf{V}_{\mathbf{W}}^{t+1}$ can be constructed as $\mathbf{V}_{\mathbf{W}}^{t+1} = \mathbf{W}_t + \frac{\theta_t - 1}{\theta_{t+1}}(\mathbf{W}_t - \mathbf{W}_{t-1})$.

By repeating the update of $\mathbf{M}$ and $\mathbf{W}$ alternately, the optimal solution $\mathbf{M}^*$ and $\mathbf{W}^*$ will be found. Although the Lipschitz constants $L_{\mathbf{M}}$ and $L_{\mathbf{W}}$ are difficult to estimate, and they are the keys to determine step sizes of η_t and μ_t, we can estimate them by line search in each iteration [46,48]. The learning scheme of the proposed algorithm is summarized in Algorithm 1.

Algorithm 1: GSML.

Input: training set $\{\mathbf{X}, \mathbf{Z}, \mathbf{Y}\}$, intra-class space dimension d', convergence threshold ϵ, maximum iteration number T.
Output: $\mathbf{M}^* \in \mathbb{S}_+^{d'}$, $\mathbf{W}^* \in \mathbb{S}_+^{d'}$
$\mathbf{M}_0 = \mathbf{I}^{d' \times d'}$, $\mathbf{W}_0 = \mathbf{I}^{d' \times d'}$, and $\theta_0 = 1$;
Obtain $\mathbf{X}' \in \mathbb{R}^{d' \times n}$ and $\mathbf{Z}' \in \mathbb{R}^{d' \times m}$ by PCA;
Compute $\boldsymbol{\Sigma}_{\mathcal{P}}$, and apply decomposition $\boldsymbol{\Sigma}_{\mathcal{P}} = \mathbf{L}_{\mathcal{P}}\mathbf{L}_{\mathcal{P}}^\top$;
$\widetilde{\mathbf{X}} = \mathbf{L}_{\mathcal{P}}^{-1}\mathbf{X}'$, and $\widetilde{\mathbf{Z}} = \mathbf{L}_{\mathcal{P}}^{-1}\mathbf{Z}'$;
for $t = 1, 2, \ldots, T$ **do**
  with fixed $\mathbf{W} = \mathbf{W}_{t-1}$:
    Compute $\widetilde{\mathbf{M}}_t$ by Eqs. (14) and (17);
    Update $\mathbf{M}_t$ by Eq. (18);
  with fixed $\mathbf{M} = \mathbf{M}_t$:
    Compute $\widetilde{\mathbf{W}}_t$ by Eqs. (24) and (27);
    Update $\mathbf{W}_t$ as $\mathbf{W}_t := \mathbf{U}_{\widetilde{\mathbf{W}}_t}\boldsymbol{\Lambda}_{\widetilde{\mathbf{W}}_t}^{+}\mathbf{U}_{\widetilde{\mathbf{W}}_t}^\top$;
  Update θ_t and construct $\mathbf{V}_{\mathbf{M}}^{t+1}$, $\mathbf{V}_{\mathbf{W}}^{t+1}$;
  Compute $\mathcal{L}_t$ by Eq. (5);
  if $t > 1$ & $|\mathcal{L}_t - \mathcal{L}_{t-1}| < \epsilon$ **then**
    break;
Return $\mathbf{M}^* = \mathbf{M}_t$, $\mathbf{W}^* = \mathbf{W}_t$.

5. Experiments

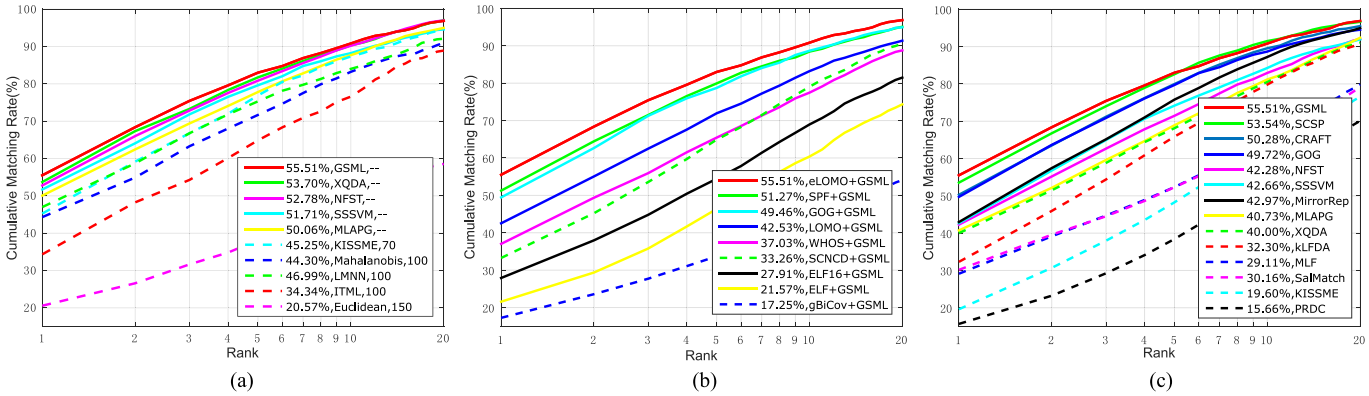
We evaluate the proposed method on four widely used person re-identification datasets including VIPeR [15], PRID450S [16], QMUL GRID [17], and CUHK01 [2]. Fig. 5 shows some image pairs randomly selected from these datasets. The performance is evaluated by the Cumulative Matching Characteristics (CMC) curve which represents the expectation of finding the right match in top r matches. To get a robust performance for comparison, we repeat the experiment procedure 10 times with random training/testing split to report the average results.

5.1. Experiments on VIPeR

The VIPeR dataset [15] has been widely applied in evaluation of person re-identification algorithms. It contains 632 pairs of pedestrian images captured by two disjoint cameras in an outdoor academic environment. There is only one image for every person in each view, and all images are scaled to 128×48 pixels. Due to large variations in viewpoint, pose and illumination, VIPeR is one of the most challenging datasets for person re-identification. Only few works have reported higher than 50% rank-1 matching rate on this dataset [28,38,41,49]. We follow the widely adopted experiment protocol to randomly select 316 persons for training and the left for testing.

5.1.1. Comparison of metric/similarity learning algorithms with eLOMO

We first evaluate the proposed GSML and several state-of-the-art metric/similarity learning algorithms with the same eLOMO descriptor, including NFST [41], SSSVM [42], XQDA [5], MLAPG [10], LMNN [13], KISSME [9], and ITML [50]. The Mahalanobis distance trained with genuine pairs and Euclidean distance are also evaluated as baselines. To reflect the optimal performance of each algorithm, we test LMNN, KISSME, ITML, Euclidian and



(a) (b) (c)

Fig. 6. Comparison of CMC curves and rank-1 matching rates on VIPeR dataset: (a) GSML with other metric learning/similarity algorithms using eLOMO descriptor, the numbers after each method name are the chosen PCA dimensions, '- -' denotes full-energy PCA (or no PCA) is applied; (b) different visual descriptors; (c) GSML with other state-of-the-art approaches.

Mahalanobis with different PCA dimensions, including 50, 70, 100, 150, 200 ~ 500 with a step size of 100, and full-energy PCA dimension. This is because some of these methods, especially the KISSME, are rather sensitive to feature dimensions. In such way, we can select the most proper dimension and report the best result for each algorithm. For SSSVM, XQDA, MLAPG, and GSML, full-energy PCA is applied because they can learn a low-rank projection matrix. For NFST, no PCA is applied since it is a kernel based metric learning algorithm, and the RBF kernel with automatically determined kernel width is chosen.

The resulting best CMC curves and rank-1 matching rates of all algorithms are shown in Fig. 6 (a), it can be found with eLOMO descriptor, all methods report encouraging results. Among them, the proposed GSML performs the best, which obtains 55.51% matching rate on rank-1. The XQDA, NFST, SSSVM, and MLAPG also report higher than 50% rank-1 matching rates, which are 53.70%, 52.78%, 51.71%, and 50.06%, respectively. The highest performance of GSML confirms the superiority of jointly learning two complementary metrics. Interestingly, although it was reported MLAPG performs better than XQDA with LOMO in [10], it is inferior to XQDA with eLOMO. We believe this maybe because the dimension of eLOMO is much higher than LOMO, the two-stage 'PCA - Metric learning' process leads to a sub-optimal metric [5]. Taking advantage of learning a generalized similarity in the intra-class subspace, our GSML can overcome this shortcoming and performs better than XQDA.

5.1.2. Comparison of visual descriptors

Next, we compare the proposed eLOMO with six available descriptors, including ELF [15], ELF18 [22], WHOS (5138-dimension) [23], gBiCov [3], SCNCD[1] [4], and GOG [6]. Since our eLOMO is the fusion of LOMO and SPF, we also test them separately. All descriptors are tested using the proposed GSML, and the results are shown in Fig. 6 (b). It can be found that eLOMO outperforms all other descriptors as well as its two components, showing much better robustness to appearance variations. Noting that the ELF, ELF18, WHOS, SCNCD, and SPF are extracted from stripes, gBiCov and LOMO are extracted from dense blocks, they are all inferior to eLOMO which is an integration of features computed from both stripes and dense blocks. Although GOG has a similar hierarchical structure with eLOMO and it is also very discriminative, it only achieves 49.46% rank-1 matching rate, while eLOMO obtains 55.51%. The result of SCNCD here is lower than that reported in [4], we think this is because its dimension has been reduced to 70 for KISSME, which may be not proper for GSML.

[1] The dimension of this descriptor has been reduced to 70 by PCA.

Table 2
Comparison of top r matching rates (%) on VIPeR dataset.

Method	r=1	r=5	r=10	r=20
GSML+eLOMO	**55.51**	**83.01**	90.82	**96.87**
SSM [49]	53.73	–	**91.49**	96.08
SpindleNet [51]	53.8	74.1	83.2	92.1
SCSP [38]	53.54	82.59	**91.49**	96.65
CRAFT [28]	50.28	79.97	89.56	95.51
GOG [6]	49.72	79.72	88.67	94.53
MPCNN [52]	47.8	74.7	84.8	91.1
NFST [41]	42.28	71.46	82.94	92.06
SSSVM [42]	42.66	74.21	84.27	91.93
MirrorRep [22]	42.97	75.82	87.28	94.84
MLAPG [10]	40.73	69.94	82.34	92.37
XQDA [5]	40.00	68.13	80.51	91.08
KLFDA [31]	32.30	65.80	79.70	90.90
MLF [2]	29.11	52.34	65.95	79.87
SalMatch [25]	30.16	52.32	65.54	79.15
KISSME [9]	19.60	49.37	62.20	77.00
PRDC [14]	15.66	38.42	53.86	70.09
MLF [2]	43.39	73.04	84.47	93.70
ME [53]	45.90	77.50	88.90	95.80
NFST (fusion) [41]	51.17	82.09	90.51	95.52

5.1.3. Comparison to the state of the art

Finally, we compare the performance of GSML+eLOMO with a number of state-of-the-art approaches, including SSM [49], SpindleNet [51], SCSP [38], CRAFT [28], GOG [6], MPCNN [52], SSSVM [42], NFST [41], MirrorRep [22], MLAPG [10], XQDA [5], KLFDA [31], MLF [2], SalMatch [25], KISSME [9], and PRDC [14]. The results are summarized in Fig. 6 (c) and Table 2. From Table 2, we can find that the proposed method achieves the highest performance. Compared to the previous best rank-1 matching rate 53.73% reported by SSM [49], our GSML achieves 55.51% with a relative improvement of 1.78%. On rank-10, our method yields the second best of 90.82%, which is slightly lower than SSM and SCSP. It should be noted that SSM applied post ranking optimization to the initial results obtained by XQDA metric (with LOMO+GOG), the SCSP fused the similarities computed from global image and local regions. While there is neither spatial constraints nor re-ranking process in our GSML, we believe that by integrating these merits, there should be further performance improvements. Compared with deep learning based models, namely SpindleNet [51], CRAFT [28], and MPCNN [52], our GSML is obviously superior to them on this dataset. Due to limited training samples, their power are difficult to be fully demonstrated.

Since our eLOMO is the fusion of LOMO and SPF, it can be viewed as a feature-level fusion model. So we also compare our method with some approaches that fuse different features [41] or

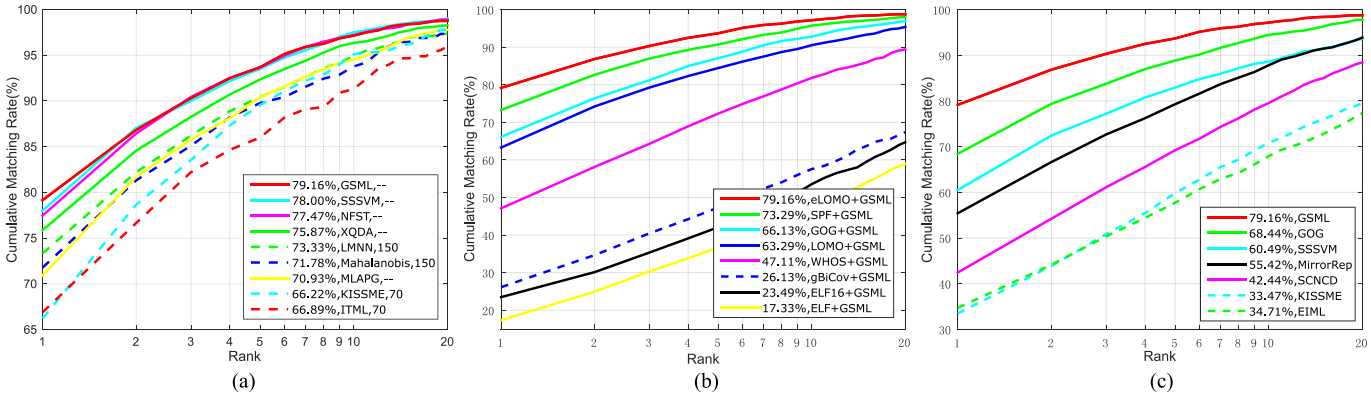


Fig. 7. Comparison of CMC curves and rank-1 matching rates on PRID450S dataset, (a) GSML with other metric/similarity learning algorithms using eLOMO descriptor, (b) different visual descriptors, (c) GSML with other state-of-the-art approaches.

different models [2,53]. The results of the compared approaches are listed in the last three rows of Table 2. It can be observed that GSML outperforms all of them. Compared to the nearest rival NFST (fusion) [41], there is an improvement of 4.34% on rank-1.

5.2. Experiments on PRID450S

The PRID450S dataset [16] includes 450 image pairs captured from two different camera-views. This is also a challenging person re-identification dataset because the images undergo serious viewpoint changes, partial occlusion and background interference. We scale all images to 128×64 for experiment and use the provided masks for extracting SPF. Similar to [42], we randomly select half of the image pairs for training and the left for testing. For GSML, full-energy PCA is applied before learning.

Performance comparison of GSML with some metric/similarity learning algorithms using the same eLOMO feature representation in shown in Fig. 7 (a). For better visualization, the Euclidean distance is not included here, while the Mahalanobis distance is kept for its remarkable performance. From the comparison, we can find that all tested algorithms obtain encouraging results with eLOMO, even the lowest rank-1 matching accuracy is as high as 66.22%. Among them, GSML performs the best with 79.16% rank-1 matching accuracy. The second best is achieved by SSSVM which gives 78% rank-1 matching accuracy. It is interesting that LMNN and Mahalanobis distance also outperform MLAPG when the feature dimension is reduced to 150. This confirms that MLAPG is not good at handling high dimensional feature representation. In contrast, our GSML can well address it by exploiting both distance and angle information from training data.

In Fig. 7 (b), we plot the CMC curves obtained by different visual descriptors. It can be found that the results are similar to that on VIPeR dataset, the eLOMO performs the best on PRID450S dataset again. On the most important rank-1, eLOMO outperforms all other descriptors by a large margin, showing remarkable discrimination on this dataset. Similar to Fig. 6 (b), we can find that SPF performs better than LOMO on both datasets, which means the holistic appearance information are more discriminative than fine details on VIPeR and PRID450S. We think there are three reasons lead to this phenomenon. First, the human bodies in images of both datasets are complete, the holistic appearance can provide enough information for identification. Second, the SPF is extracted from foreground images, thus background clutter is suppressed effectively. The third reason is the RGB histogram and SCN in SPF can capture more color information than LOMO, and color is distinctive on these datasets.

Table 3
Comparison of top r matching rates (%) on PRID450S dataset.

Method	r=1	r=5	r=10	r=20
GSML+eLOMO	**79.16**	**93.69**	**97.16**	98.76
SSM [49]	72.98	–	96.76	**99.11**
GOG [6]	68.44	88.84	94.49	97.82
XQDA [5]	61.42	–	90.84	95.33
SSSVM [42]	60.49	82.93	88.58	93.60
MirrorRep [22]	55.42	79.29	87.82	93.87
SCNCD [4]	42.44	69.22	79.56	88.44
ECM [54]	41.90	66.30	76.90	84.90
KISSME [9]	33.47	59.82	70.84	79.47
EIML [55]	34.71	57.73	67.91	77.33

On PRID450S the proposed method is compared with nine state-of-the-art approaches following the same protocol, including SSM [49], GOG [6], SSSVM [42], XQDA [5], MirrorRep [22], SCNCD [4], ECM [54], KISSME [9], and EIML [55]. The comparison results are shown in Table 3 and Fig. 7 (c), some of them are borrowed from [42]. It can be found that the proposed method outperforms all the competitors significantly. The best rank-1 matching rate reported to date is 72.98% by SSM [49], while our GSML+eLOMO achieves 79.16% with an improvement of 6.18%. Only on rank-20, the matching accuracy of GSML is slightly lower than SSM. The impressive performance of GSML shows that it has significant advantage in addressing the appearance variations in this dataset.

5.3. Experiments on QMUL GRID

The QMUL GRID dataset [17] was captured in an underground station with 8 disjoint cameras. There are 250 persons that each has one image in both the probe set and the gallery set. Besides, there are 775 extra persons that have no matching images in the probe set, which makes the matching task rather difficult on this dataset. As shown in Fig. 5, the images in this dataset are of poor quality and low resolutions. We scale all images to 128×64 for experiment. Following the experiment protocol in [5], we randomly select 125 pairs for training and the left 125 pairs along with the 775 extra images are used for testing. Full-energy PCA is performed before learning.

The comparison of the proposed method with state-of-the-art results reported under the same protocol is shown in Table 4. We can observe that our GSML+eLOMO performs well against existing methods and achieves the second best with an accuracy of 25.68% on the most important rank-1. In general, the performance of the proposed method is competitive on this dataset.

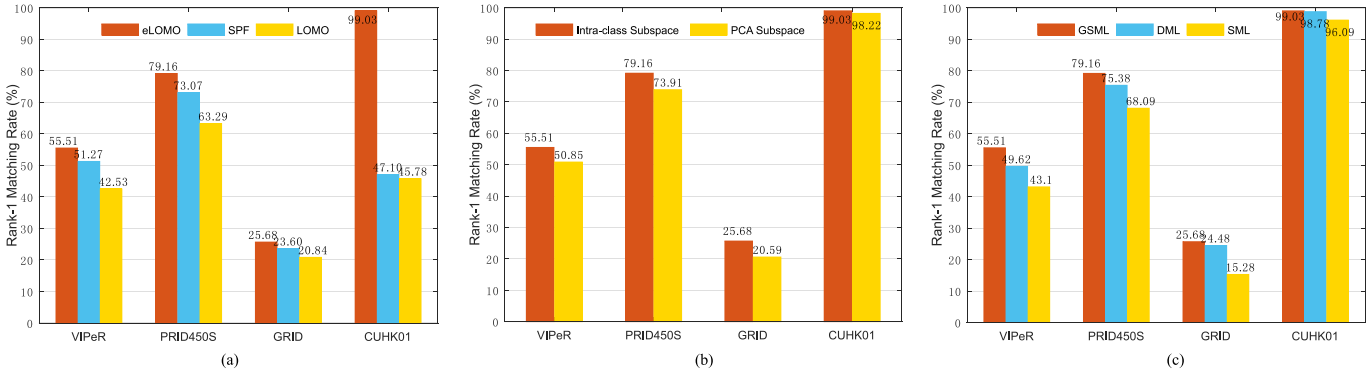


Fig. 8. Rank-1 matching rate comparison of (a) eLOMO, SPF and LOMO, (b) learning in the intra-class subspace and PCA subspace, (c) GSML, DML and SML.

Table 4
Comparison of top r matching rates (%) on GRID dataset.

Method	r=1	r=5	r=10	r=20
GSML+eLOMO	25.68	**48.16**	59.52	69.76
SSM [49]	**27.20**	–	**61.12**	**70.56**
GOG [6]	24.72	46.96	58.40	68.96
SCSP [38]	24.24	44.56	54.08	65.20
SSSVM [42]	22.40	40.40	51.28	61.20
MLAPG [10]	16.64	33.12	41.20	52.96
XQDA [5]	16.56	33.84	41.84	52.40
PolyMap [40]	16.30	35.80	46.00	57.60
MtMCML [56]	14.08	34.64	45.84	59.84
PRDC [14]	9.68	22.00	32.96	44.32

Table 5
Comparison of top r matching rates (%) on CUHK01 dataset.

Method	r=1	r=5	r=10	r=20
GSML	**99.03**	**99.96**	**100**	**100**
EDM [57]	86.59	–	–	–
SpindleNet [51]	79.9	94.4	97.1	98.6
DLPAR [58]	75.0	93.5	95.7	97.7
CRAFT [28]	74.5	91.2	94.8	97.1
GOG [6]	67.28	86.89	91.77	95.93
SSSVM [42]	65.97	–	–	–
NFST [41]	64.98	84.96	89.92	94.36
MLAPG [10]	64.24	85.41	90.84	94.92
XQDA [5]	63.21	83.89	90.04	94.16
MLF [2]	34.30	55.06	64.96	74.94
SalMatch [25]	28.45	45.85	55.67	67.95
ME [53]	53.40	76.40	84.40	90.50
NFST (fusion) [41]	69.09	86.87	91.77	95.39

5.4. Experiments on CUHK01

The CUHK01 dataset [2] was captured by two cameras in a campus environment. There are 971 persons in this dataset, and each person has two images in one view, so the total number of images is 3884. The images are normalized to 160×60 pixels. Different from previous datasets, they are of high quality and resolutions, and the images of each pedestrian in different views only take moderate variations in viewpoint and illumination. Following [10], the persons are randomly split to 485 for training and 486 for testing in experiment. Multi-shot re-identification is performed by summing the similarity scores of the same person. We set the intra-class subspace dimension to $k = 1200$ on this dataset.

Table 5 shows the comparison of our method with state-of-the-art results reported under the same evaluation protocol. Our method obtains nearly saturate matching accuracy and outperforms all competitors on this dataset. The previous best rank-1 matching rate is 86.59% reported by EDM [57], while our GSML has achieved 99.03%. So the improvement is as high as 12.44%. Note that the approaches ranked second to fifth all employed powerful deep learning models, they are still inferior to our method. After a series of experiments, we find there are two main reasons lead to the excellent performance of GSML. First and most important, the fusion of LOMO and SPF leads to a much more discriminative descriptor that well captures the appearance invariance of pedestrians in this dataset. This will be further discussed in Section 5.5.1. Due to the small range of appearance variations in CUHK01 (as can be observed from Fig. 5), the matching result should be much better than other considered datasets. Second, the generalized similarity in our GSML makes full use of the complementary distance and angle information, thus much higher re-identification accuracy is deserved.

5.5. Comparative experiments

To better understand the contributions of each part in the proposed method, we perform further comparative experiments to analyze it in the following aspects: (1) effect of fusing features extracted from both stripes and dense blocks, (2) effect of learning in the intra-class subspace, and (3) effect of learning generalized similarity. The reported results are still averaged over 10 random experiments with non-overlapping split.

5.5.1. Effect of integrating features extracted from stripes and dense blocks

We first compare the performance of eLOMO with its two components (i.e. SPF and LOMO). Fig. 8 (a) shows the rank-1 matching rates of them on four considered datasets, it can be easily found that eLOMO performs much better than SPF and LOMO. The results of eLOMO, SPF and LOMO are 55.51%, 51.27% and 42.53% on VIPeR dataset, 79.16%, 73.07% and 63.29% on PRID450S dataset, and 25.68%, 23.60% and 20.84% on GRID dataset. On CUHK01, the rank-1 matching rate of eLOMO is more than twice of SPF and LOMO, which is 99.03% vs (47.10%, 45.78%). We think such high boosting effect maybe because our eLOMO happens to well fit the characteristics of appearance patterns in CUHK01 dataset. This comparison confirms that by integrating features extracted from horizontal stripes and dense blocks, the obtained visual descriptor is much more discriminative for person re-identification. This is because the fusion of LOMO and SPF combines their merits to utilize both holistic appearance and fine details for visual matching. We believe the reasoning behind performance boosting is such fusion is more in line with the coarse-to-fine recognition mechanism of human vision system.

5.5.2. Effect of learning in the intra-class subspace

To demonstrate the effectiveness of learning metrics in the intra-class subspace, we compare the results of learning in the

Table 6
Comparison of run time on VIPeR dataset (seconds).

Method	NFST	KISSME	XQDA	SSSVM	GSML	MLAPG	ITML	LMNN
Training	1.4	3.7	3.5	8.24	25.0	62.1	195.8	269.3
Testing	0.0382	0.0184	0.0823	0.2853	0.0156	0.0031	0.0179	0.0215

intra-class subspace with that of learning in the PCA subspace directly (i.e. with and without the whitening process). The comparison results are shown in Fig. 8 (b), we can find the rank-1 matching rates obtained by learning in the intra-class subspace are obviously higher than learning in PCA subspace. The relative improvements are 4.66%, 5.25%, 5.09%, and 0.81% on VIPeR, PRID450S, GRID, and CUHK01, respectively. This indicates that by learning metrics in the intra-class subspace, the intra-class variations can be suppressed effectively. Therefore, it greatly helps to separate the matched pairs from unmatched ones. Another advantage of learning in the intra-class subspace is that the convergence is much faster. Without the whitening process, the algorithm usually has to take hundreds of iterations to converge with a tolerance of 10^{-4}, while only dozens of iterations are needed when the samples are projected into intra-class subspace. The run time of the proposed method will be discussed in Section 5.6.

5.5.3. Effect of learning generalized similarity

To study the effect of learning generalized similarity, we compare the results of GSML with that of learning a single distance metric (denoted as DML) and learning a single bilinear similarity metric (denoted as SML). The comparison is shown in Fig. 8 (c). We can find that the rank-1 matching accuracies of GSML are consistently higher than DML and SML on all datasets. The relative improvements of GSML over DML and SML are 5.89% and 12.41% on the VIPeR dataset. On PRID450S, GRID, and CUHK01, the improvements are (3.78%, 11.07%), (1.20%, 10.40%), and (0.25%, 2.94%), respectively. In experiments, we also tried to fuse the results of separately learned distance metric and inner product similarity metric, they are still inferior to GSML. This confirms that jointly learning two metrics is more effective for person re-identification. From Fig. 8 (c), we can also find that the performance of SML is inferior to DML, which demonstrates the angle information is less discriminative than distance information in person re-identification. However, much better results can be achieved by utilizing both.

5.6. Computational time

Table 6 shows a comparison of the average run time of GSML with the metric/similarity learning algorithms evaluated in Section 5.1.1. This experiment is conducted on the VIPeR dataset with eLOMO descriptor. All algorithms are implemented in MATLAB and run on a desktop PC with Intel i7-3720 @2.6GHz CPU. For fair comparison, we reduce the feature dimension to 600 for all algorithms. From Table 6 we can find it takes about 25 s to train GSML. This is slower than NFST, KISSME, and XQDA that have closed-form solutions, and SSSVM that implemented on the basis of XQDA; but much faster than MLAPG, ITML, and LMNN. Although there are two metrics need to learn in GSML, it can converge within dozens of iterations by virtue of learning in the intra-class subspace. In contrast, a large number of iterations are needed for MLAPG, ITML, and LMNN. In testing phase, our GSML only takes 0.0156 s to match the 316×316 cross-view pairs, which ranks the second and is in the same order as most algorithms.

Besides, we also evaluate the time of computing eLOMO descriptor. In processing 128×48 pedestrian images with foregrounds detected, it takes about 0.015 s to compute the SPF descriptor per image on average. As the LOMO feature requires 0.012 s [5], the computation of eLOMO only needs 0.027 s per image, which is quite efficient. The most time consuming part in computing eLOMO is detecting foregrounds with DDN, which requires about 0.11 s per image. Nevertheless, we can obtain them offline.

6. Conclusion

In this paper, we have proposed a discriminative and robust feature representation termed eLOMO, and an effective metric learning method called GSML for person re-identification. The eLOMO fuses the features extracted from both horizontal stripes and dense blocks, such that the fine details and holistic appearance information can be integrated together to enhance the discrimination. The proposed GSML jointly learns a Mahalanobis distance metric and a bilinear similarity metric to simultaneously exploit distance and angle information from training data. Experimental results on four person re-identification datasets have demonstrated the superiority of our method against a wide range of state-of-the-art approaches. In the future, we will extend our method by adopting spatial constraints and apply it to other visual learning tasks.

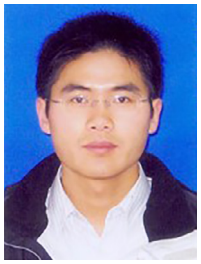
Acknowledgment

This work was partially supported by National Natural Science Foundation of China (NSFC Grant No. 61773272, 61272258, 61301299, 61572085, 61170124, 61272005), Provincial Natural Science Foundation of Jiangsu (Grant No. BK20151254, BK2015-1260), Science and Education Innovation based Cloud Data fusion Foundation of Science and Technology Development Center of Education Ministry(2017B03112), Six talent peaks Project in Jiangsu Province (DZXX-027), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (Grant No. 93K172016K08), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

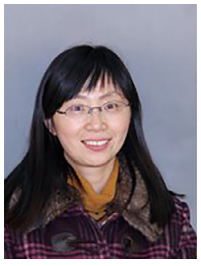
References

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2360–2367.
[2] R. Zhao, W. Ouyang, X. Wang, Learning mid-level filters for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 144–151.
[3] B. Ma, Y. Su, F. Jurie, Covariance descriptor based on bio-inspired features for person re-identification and face verification, Image Vis. Comput. 32 (6) (2014) 379–390.
[4] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, S.Z. Li, Salient color names for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 536–551.
[5] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.
[6] T. Matsukawa, T. Okabe, E. Suzuki, Y. Sato, Hierarchical gaussian descriptor for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1363–1372.
[7] J. Chen, Z. Zhang, Y. Wang, Relevance metric learning for person re-identification by exploiting listwise similarities, IEEE Trans. Image Process. 24 (12) (2015) 4741–4755.
[8] M. Hirzer, P.M. Roth, M. Köstinger, H. Bischof, Relaxed pairwise learned metric for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 780–793.

[9] M. Köestinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[10] S. Liao, S.Z. Li, Efficient PSD constrained asymmetric metric learning for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3685–3693.
[11] A. Mignon, F. Jurie, Pcca: a new approach for distance learning from sparse pairwise constraints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2666–2672.
[12] H.V. Nguyen, L. Bai, Cosine similarity metric learning for face verification, in: Proceedings of the Asian Conference on Computer Vision, 2010, pp. 709–720.
[13] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.
[14] W.-S. Zheng, S. Gong, T. Xiang, Reidentification by relative distance comparison, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 653–668.
[15] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 262–275.
[16] P.M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, H. Bischof, Mahalanobis Distance Learning for Person Re-identification, Springer London, 2014.
[17] C.C. Loy, T. Xiang, S. Gong, Multi-camera activity correlation analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1988–1995.
[18] A. Bedagkar-Gala, S.K. Shah, A survey of approaches and trends in person re-identification, Image Vis. Comput. 32 (4) (2014) 270–286.
[19] G. Doretto, T. Sebastian, P. Tu, J. Rittscher, Appearance-based person reidentification in camera networks: problem overview and current approaches, J. Ambient Intell. Humaniz. Comput. 2 (2) (2011) 127–151.
[20] S. Gong, M. Cristani, S. Yan, C.C. Loy, Person Re-identification, Springer, 2014.
[21] L. Zheng, Y. Yang, A.G. Hauptmann, Person re-identification: past, present and future, arXiv:1610.02984 (2016).
[22] Y.C. Chen, W.S. Zheng, J. Lai, Mirror representation for modeling view-specific transform in person re-identification, in: Proceedings of the International Conference on Artificial Intelligence, 2015, pp. 3402–3408.
[23] G. Lisanti, I. Masi, A.D. Bagdanov, A.D. Bimbo, Person re-identification by iterative re-weighted sparse ranking, IEEE Trans. Pattern Anal. Mach. Intell. 37 (8) (2015) 1629–1642.
[24] Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, J. Smith, Learning locally-adaptive decision functions for person verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3610–3617.
[25] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2528–2535.
[26] S.Z. Chen, C.C. Guo, J.H. Lai, Deep ranking for person re-identification via joint representation learning, IEEE Trans. Image Process. 25 (5) (2016) 2353–2367.
[27] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.
[28] Y.C. Chen, X. Zhu, W.S. Zheng, J.H. Lai, Person re-identification by camera correlation aware feature augmentation, IEEE Trans. Pattern Anal. Mach. Intell. PP (99)(2017) 1–1.
[29] C. Sun, D. Wang, H. Lu, Person re-identification via distance metric learning with latent variables, IEEE Trans. Image Process. 26 (1) (2017) 23–34.
[30] J. Zhou, P. Yu, W. Tang, Y. Wu, Efficient online local metric adaptation via negative samples for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2439–2447.
[31] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 1–16.
[32] Y. Duan, J. Lu, J. Feng, J. Zhou, Deep localized metric learning, IEEE Trans. Circuits Syst. Video Technol. PP (99)(2017) 1–1.
[33] J. Lu, J. Hu, J. Zhou, Deep metric learning for visual understanding: an overview of recent advances, IEEE Signal Process. Mag. 34 (6) (2017) 76–84.
[34] E. Ustinova, Y. Ganin, V. Lempitsky, Multi-region bilinear convolutional neural networks for person re-identification, in: Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance (AVSS), 2017.
[35] G. Lisanti, S. Karaman, I. Masi, Multichannel-kernel canonical correlation analysis for cross-view person reidentification, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 13 (2) (2017) 13.
[36] C. Kang, S. Liao, Y. He, J. Wang, W. Niu, S. Xiang, C. Pan, Cross-modal similarity learning:a low rank bilinear formulation, in: Proceedings of the Twenty–Fourth ACM International Conference on Information and Knowledge Management (CIKM), 2015, pp. 1251–1260.
[37] W. Liu, C. Mu, R. Ji, S. Ma, J.R. Smith, S.F. Chang, Low-rank similarity metric learning in high dimensions, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2015, pp. 2792–2799.
[38] D. Chen, Z. Yuan, B. Chen, N. Zheng, Similarity learning with spatial constraints for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1268–1277.
[39] Q. Cao, Y. Ying, P. Li, Similarity metric learning for face recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2408–2415.
[40] D. Chen, Z. Yuan, G. Hua, N. Zheng, J. Wang, Similarity learning on an explicit polynomial kernel feature map for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1565–1573.
[41] L. Zhang, T. Xiang, S. Gong, Learning a discriminative null space for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016a, pp. 1239–1248.
[42] Y. Zhang, B. Li, H. Lu, A. Irie, R. Xiang, Sample-specific SVM learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016b, pp. 1278–1287.
[43] P. Luo, X. Wang, X. Tang, Pedestrian parsing via deep decompositional network, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2648–2655.
[44] B. Hariharan, J. Malik, D. Ramanan, Discriminative decorrelation for clustering and classification, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 459–472.
[45] E.P. Xing, A.Y. Ng, M.I. Jordan, S. Russell, Distance metric learning, with application to clustering with side-information, Adv. Neural Inf. Process. Syst. 15 (2002) 505–512.
[46] N. Parikh, S. Boyd, Proximal algorithms, Found. Trends Optim. 1 (3) (2014) 127–239.
[47] Toh, Kim-Chuan, Yun, Sangwoon, An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems, Pac. J. Optim. 6 (3) (2010) 615–640.
[48] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, Siam J. Imaging Sci. 2 (1) (2009) 183–202.
[49] S. Bai, X. Bai, Q. Tian, Scalable person re-identification on supervised smoothed manifold, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3356–3365.
[50] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: Proceedings of the ACM International Conference on Machine learning, 2007, pp. 209–216.
[51] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 907–915.
[52] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.
[53] S. Paisitkriangkrai, C. Shen, V.D.H. Anton, Learning to rank in person re-identification with metric ensembles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1846–1855.
[54] X. Liu, H. Wang, Y. Wu, J. Yang, An ensemble color model for human re-identification, in: Proceedings of the Applications of Computer Vision, 2015, pp. 868–875.
[55] M. Hirzer, P.M. Roth, H. Bischof, Person re-identification by efficient impostor-based metric learning, in: Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance, 2012, pp. 203–208.
[56] L. Ma, X. Yang, D. Tao, Person re-identification over camera networks using multi-task distance metric learning, IEEE Trans. Image Process. 23 (8) (2014) 3656–3670.
[57] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, S.Z. Li, Embedding deep metric for person re-identification: A study against large variations, in: Proceedings of the European Conference on Computer Vision, 2016.
[58] L. Zhao, X. Li, J. Wang, Y. Zhuang, Deeply-learned part-aligned representations for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.



Husheng Dong received his M.S. degree from School of Computer Science & Technology, Soochow University in 2008, and he is pursuing the Ph.D. degree now. He is also a teacher of Suzhou Institute of Trade & Commerce. His research interest includes computer vision, image and video processing, and machine learning.



Ping Lu received her B.Eng and M.S. degree from School of Computer Science and Technology, Soochow University in 2002 and 2005, respectively. She is an associate professor at Suzhou Institute of Trade & Commerce. Her research interest includes digital image processing and pattern recognition.

Shan Zhong received her M.S. and Ph.D. from Jiang University (2007) and Soochow University (2017), respectively. She is a teacher of Changshu Institute of Technology now. Her research interests include machine learning and Deep learning.

Chunping Liu received her Ph.D. degree in pattern recognition and artificial intelligence from Nanjing University of Science & Technology in 2002. She is now a professor of School of Computer Science & Technology, Soochow University. Her research interests include computer vision, image analysis and recognition, in particular in the domains of visual saliency detection, object detection and recognition and scene understanding.

Yi Ji received her M.S. Degree from National University of Singapore, Singapore and Ph.D. degree from INSA de Lyon, France. She is now an associate professor in School of Computer Science & Technology of Soochow University. Her research areas are 3D action recognition and complex scene understanding.

Shengrong Gong received his M.S. degree from Harbin Institute of Technology in 1993, and his Ph.D. degree from Beihang University in 2001. He is the dean of School of Computer Science and Engineering, Changshu Institute of Technology, and a professor and doctoral supervisor of School of Computer Science & Technology, Soochow University. His research interests include image and video processing, pattern recognition, and computer vision.