Multiple-Shot Person Re-identification via Riemannian Discriminative Learning

Yuheng Lu[1,2], Ruiping Wang[1,2,3](✉), Shiguang Shan[1,2,3], and Xilin Chen[1,2,3]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
yuheng.lu@vipl.ict.ac.cn, {wangruiping,sgshan,xlchen}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Cooperative Medianet Innovation Center, Beijing, China

Abstract. This paper presents a Riemannian discriminative learning framework for multiple-shot person re-identification. Firstly, image regions are encoded into covariance matrices or a Gaussian extension as robust feature descriptors. Since these matrices lie on some specific Riemannian manifolds, we introduce a manifold averaging strategy to fuse the feature descriptors from multiple images for a holistic representation, and exploit Riemannian kernels to implicitly map the averaged matrices to a Reproducing Kernel Hilbert Space (RKHS), where conventional discriminative learning algorithms can be conducted. In particular, we apply kernel variants of two typical methods, i.e., the Linear Discriminant Analysis (LDA) and Metric Learning to Rank (MLR), to demonstrate the flexibility of the framework. Extensive experiments on five public datasets exhibit impressive improvements over existing multiple-shot re-identification methods as well as representative single-shot approaches.

1 Introduction

Person re-identification, a task of recognising pedestrian appearance in different time and locations captured with a multi-camera network without field of view overlap, has attracted wide interest in the field of surveillance. Applications in security, medical guardianship, tracking, and even online image retrieval on clothes [1] demonstrate the great while growing significance of the problem.

Current practices on person re-identification have mainly concentrated on two stages, extracting distinctive while stable features [2–9], and/or learning discriminative cross-view metrics [10–16]. Due to the enormous challenges in the task including vast variations in (i) pedestrian viewpoint and pose, (ii) environment illumination and occlusion, and (iii) camera position, configuration and resolution, person re-identification still remains an open problem.

Among all the literatures endeavouring to tackle this problem, most focus on the single-shot scenario [10,12,14–30], which refers to querying one single image at a time in a gallery constituted by single images, where one identity

C.-S. Chen et al. (Eds.): ACCV 2016 Workshops, Part III, LNCS 10118, pp. 408–425, 2017.
DOI: 10.1007/978-3-319-54526-4_30

is presented by usually one image (*Single-vs-Single*, or *SvS*), or *independent* multiple images (*Single-vs-Multiple*, or *SvM*), as shown in Fig. 1(a,b). Recently, approaches based on the multi-shot scenario [2–9,11,13,31–39], where multiple images of the same person captured in the same camera are grouped as the probe to match the gallery formed also by multiple-image *groups* (*Multiple-vs-Multiple*, or *MvM*), have started to blossom (Fig. 1(c)). Owing to the practical accessibility and wider variation coverage of the extra information, which usually brings higher and more robust performance, the MvM case is addressed in this paper. According to different matching schemes, we categorise multi-shot person re-identification methods into four classes: (i) closest point-based approaches [3,4] where a pair of closest points are exhaustively searched for to calculate the distance between each group pair; (ii) score voting-based methods [31,32] that average all similarity scores of all image pairs of two groups as the group similarity; (iii) set structure-based approaches [11,34] which model the distribution structure of each image group (or set) and measure the similarity of the set models for matching; and (iv) signature-based ones [2,6,9] that generate a signature for each group to facilitate subsequent matching, converting the problem into the single-shot case.

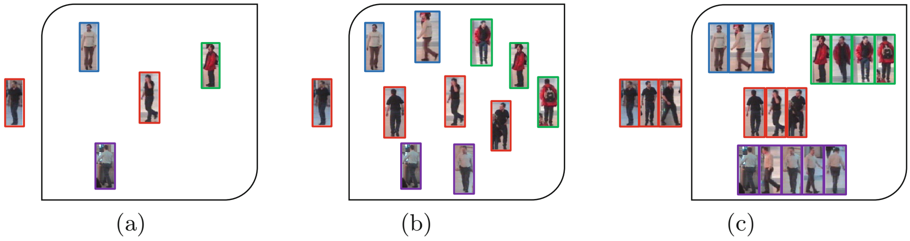


Fig. 1. Three cases of person re-identification matching scenario: (a) Single-vs-Single (SvS), (b) Single-vs-Multi (SvM) and (c) Multi-vs-Multi (MvM). Images outside and inside the boxes are the probe and gallery respectively. Images corresponding to different identities are bordered with different colors. (Color figure online)

Considering that signature generating is not only flexible in feature modeling and concise in multi-frame encapsulation, but also naturally compatible with a wide variety of subsequent processes applicable in the well-studied single-shot cases, we propose a signature-based approach, following the majority [2,5–7,9,13,35–39] of the multi-shot community. Specifically, having noticed the encouraging performance of region covariance matrix (a.k.a. Symmetric Positive Definite (SPD) matrix) [6,7,28,30,40], we adopt it as our feature representation of images and further extend it into a pixel-oriented local Gaussian descriptor, which is also written in the form of SPD matrix under the framework of information geometry [41,42]. Considering the fact that both the region covariance matrix and the Gaussian extension lie on the SPD Riemannian manifold, conventional operations that work in Euclidean space are not directly applicable.

Hence, we utilize Riemannian metrics to integrate the feature descriptors (i.e. SPD matrices) of multiple images and measure their similarities, and further exploit the corresponding Riemannian kernel functions to map points on Riemannian manifold to a high-dimensional Hilbert space for performing discriminative learning. We name our approach Multi-shot Riemannian Discriminative Learning (MRDL), and show the flowchart in Fig. 2.

MRDL has the following three main advantages. (i) As a multi-shot approach, it models signatures to fuse rich information from multiple images based on robust feature descriptors, i.e., the region covariance and Gaussian matrix which characterize correlations between feature dimensions. (ii) MRDL performs discriminative learning on the manifold, strengthening the discrimination power of the approach compared to other methods that conduct unsupervised matching, and retaining the geometric structure of the manifold space by operating on the SPD Riemannian manifold. (iii) What presented in this paper is a framework with remarkable extensibility. Along with abundant types of features applicable in the front end and various manifold distance metrics replaceable in the middle part, the rear end is also an arena for the performance of quantities of discriminative learning algorithms.

The rest of this paper is organised as follows. Related works of both the overall person re-identification field and the multi-shot branch are introduced in Sect. 2. Section 3 presents the region covariance and Gaussian descriptors we employed. Then in Sect. 4, we deliver the proposed method MRDL, including the kernel derivation of two typical kinds of discriminative learning algorithms: linear discriminant analysis and metric learning. Subsequently, extensive experiments are conducted on five benchmark datasets in Sect. 5. Lastly, we summarise our work and discuss possible future research directions in Sect. 6.

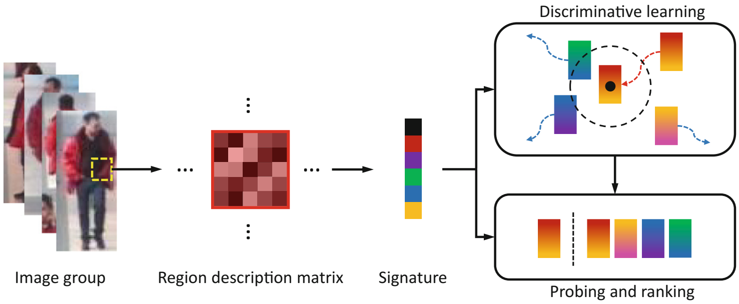


Fig. 2. Workflow of the proposed method. We firstly compute the region description matrices (covariance or Gaussian) of each image, and those from the same region of different images are fused to form the signature for the image group. Afterwards, group descriptors of training samples are processed in the learning phase to find a discriminative subspace for classification. In the testing phase, signature of a probe image group is projected into the learned space and matched with the gallery samples.

2 Related Works

In the early years, discriminative and robust feature extraction [3,4,8,9,18–20,22] was a dominating research topic in the area of person re-identification. Quite a few methods sought for representative and robust features with sophisticated designs, such as spatio shape and appearance context modeling [3], human body symmetry and asymmetry based local feature accumulation [4], fisher vector encoding [8], and local maximal occurrence representation [18]. Meanwhile, supervised learning was also adopted to guide extracting more discriminative features, including support vector machine (SVM) based ranking [19], weak classifiers boosting [20], deep neural network training [22], and bag-of-words based patch description [9]. On the other hand, learning discriminative metrics [10,11,13] has rapidly grown to share the person re-identification empire in recent years. Zheng *et al.* [10] use triplets to learn distance metric. Wu *et al.* [11] adopt a structural SVM based algorithm to rank image sets. Pedagadi *et al.* [13] apply locality preserving strategy to Fisher discriminant analysis.

Practically, one can usually obtain more than one frame from a video track for each person. Owing to the additional information compared with single images, multiple frames alleviate the disadvantages brought by variations in viewpoint, pose and occlusion. As mentioned in the previous section, we categorise approaches specifically designed for or applicable to multi-shot scenario into four classes: closest point-based, score voting-based, set structure-based and signature-based approaches.

The first class of **closest point**-based methods [3,4,8] select the minimum distance of all sample pairs between two image sets as the set distance. On the contrary, works on **score voting** usually average similarities of all possible sample pairs between the sets [31,32], or reconstruction residuals of each probe sample from the gallery group [33], to get the set distance. For **set structure** modeling, Wu *et al.* [11,34] formulate image sets as affine hulls, and then conduct discriminative metric learning [11] or locality constrained collaborative representation with l-2 regularization [34] on the hulls. Finally, most of the large number of **signature**-based approaches compute the mean in the vectorized feature space [5,9,13,35–37] or on the SPD manifold of region covariance matrices [6,7], while the others model spatio-temporal appearance [2,38] or body action [39] to generate a holistic representation for each image group. Among them, mean Riemannian covariance grid (MRCG) [6] is the most related one to our proposed method. However, MRCG is unsupervised, while ours performs discriminative learning on the Riemannian manifold, introducing great discrimination power to the learned space, which will be detailed in the next section.

3 Region Covariance and Gaussian Descriptor

Suppose we have a dataset with I identities, the i-th of which includes T_i image groups. The t-th group is composed of $F_{i,t}$ image frames, and each image is spatially partitioned into S local regions for more precise modeling since different body parts may show sharply contrasted appearances. Tuzel *et al.* [40] proposed the region covariance matrix as a fast and robust descriptor, which aggregates the covariances of all pixel-level feature pairs inside a spatial region. Let $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n]$ be the data matrix of an image region with n pixels, where $\mathbf{p}_k \in \mathbb{R}^d$ denotes the d-dimensional feature descriptor of the k-th pixel, the region is encoded by the $d \times d$ covariance matrix:

$$\mathbf{C} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{p}_k - \bar{\mathbf{p}})(\mathbf{p}_k - \bar{\mathbf{p}})^T, \tag{1}$$

where $\bar{\mathbf{p}}$ is the mean of $\mathbf{p}_k (k = 1, ..., n)$.

Since the SPD covariance matrix $\mathbf{C}$ captures the second-order statistics of the features, the first-order statistics, feature mean, is lost. Considering that $\mathbf{C}$, which encodes the pixel feature variation pattern, and $\bar{\mathbf{p}}$ that captures the general position of the features in the original space are complementary, we embed them in a more informative Gaussian representation: $(\bar{\mathbf{p}}, \mathbf{C})$. Specifically, the mean vector $\bar{\mathbf{p}}$ and $d \times d$ covariance matrix $\mathbf{C}$ are mapped together into the space of $(d+1) \times (d+1)$ SPD matrices. Under the framework of information geometry [41,42], the embedding is fulfilled through two mappings, from affine transformation $(\bar{\mathbf{p}}, \mathbf{C}^{1/2})$ to a simple Lie group, and subsequently to the SPD matrix space:

$$\mathbf{G} = |\mathbf{C}|^{-\frac{1}{d+1}}\begin{pmatrix} \mathbf{C} + \bar{\mathbf{p}}\bar{\mathbf{p}}^T & \bar{\mathbf{p}} \\ \bar{\mathbf{p}}^T & 1 \end{pmatrix}. \tag{2}$$

We name $\mathbf{G}$ as the region Gaussian matrix. Considering local appearances of the same person usually shift horizontally in different cameras, each pedestrian image is divided into six horizontal *stripes* equally, as in [10,37]. For the s-th $(s = 1, ..., 6)$ stripe, a region Gaussian matrix $\mathbf{G}^s$ is calculated through Eq. 2.

Then, for the t-th image group of the i-th identity, the region descriptor of the s-th stripe of the f-th frame is denoted as $\mathbf{G}_{i,t}^{f,s}$, as shown in Fig. 3. Considering averaging is the most widely applied strategy [5–7,9,13,35–37] in the multi-shot community, the SPD matrices of multiple images in the same area are averaged to generate the image group signature $S_{i,t} = \{\overline{\mathbf{G}}_{i,t}^{s}, s = 1, ..., 6\}$, which will be explained in detail later in Sect. 4.1.

Similarly, MRCG computes region covariance matrices for densely sampled patches on each image. Afterwards, a covariance grid is produced for each image group as signature for direct patch-level matching. The main difference between MRCG, an unsupervised robust feature extraction method, and our MRDL is that we design a way to perform *discriminative learning* on the Riemannian manifold with each object represented by *multiple* covariance matrices. Besides, instead of on a dense patch grid, we extract region covariances on stripes, which not only takes severe horizontal appearance shifts caused by viewpoint and pose changes into consideration, but also greatly reduces model complexity. Furthermore, we merge feature mean with the covariance matrix, resulting in a more powerful local feature representation, the region Gaussian matrix. Extensive experiments in Sect. 5 exhibit impressive improvements in our approach.

4 Multi-shot Riemannian Discriminative Learning

4.1 Symmetric Positive Definite Matrix Manifold

Residing on the symmetric positive definite matrix manifold (SPD matrix manifold), which is a special Riemannian manifold, the covariance/Gaussian matrix distinguishes itself from the common feature descriptors in Euclidean space with the manifold characterising non-linear data distributions.

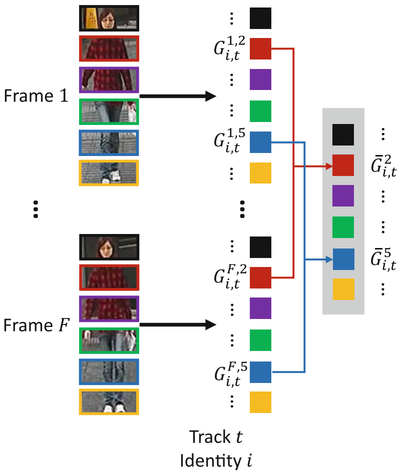


Fig. 3. Image group signature generation. Firstly, the SPD matrices (colored square) of each image stripe are computed, then those of the same region from all images in a group are averaged to generate the signature. Best viewed in color.

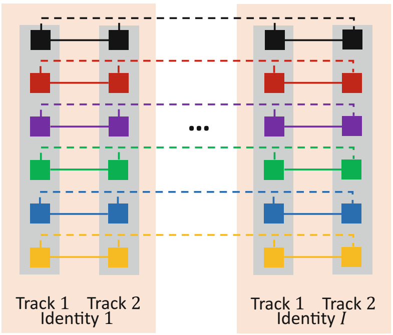


Fig. 4. Graph embedding technique in the learning phase. Positive pairs are constructed with the *same-position* matrices of *same*-identity image tracks (solid lines); negative pairs are formed also with the *same-position* matrices but of *different*-identities tracks (dash lines connected *only*). Best viewed in color.

Two different distance metrics are widely applied in Riemannian geometry: the log-Euclidean distance (LED) [43] and affine-invariant distance (AID) [44]:

$$d_{LED}(\mathbf{C}_1, \mathbf{C}_2) = \left\| \log(\mathbf{C}_1) - \log(\mathbf{C}_2) \right\|_F \tag{3}$$

$$d_{AID}(\mathbf{C}_1, \mathbf{C}_2) = \sqrt{\textstyle\sum_{k=1}^{d} \ln^2 \lambda_k(\mathbf{C}_1, \mathbf{C}_2)}, \tag{4}$$

where $\log(\mathbf{C})$ denotes the matrix logarithm operation, $\|\centerdot\|_F$ denotes the Frobenius norm, and $\lambda_k(\mathbf{C}_1, \mathbf{C}_2)$ $(k = 1, ..., d)$ are the generalized eigenvalues of $\mathbf{C}_1, \mathbf{C}_2$. Note that in this section, to keep simplicity, $\mathbf{C}$ can denote either covariance matrix (Eq. 1) or Gaussian matrix (Eq. 2).

With the metrics above, one can find the mean SPD matrix on a manifold $\mathcal{M}$ with m samples by using the Karcher or Fréchet mean, as in [6]:

$$\overline{\mathbf{C}} = \arg\min_{\mathbf{C}\in\mathcal{M}} \sum_{k=1}^{m} d^2(\mathbf{C}, \mathbf{C}^k), \tag{5}$$

where d can be either Eq. 3 or Eq. 4. We utilize this averaging method to generate signatures in the previous section.

4.2 Kernel Linear Discriminant Analysis

Although the SPD matrices lie on the manifold, to fully activate the discriminative power beneath the samples, we can also extend the kernel algorithms in Euclidean space to SPD matrix manifold with appropriate kernel functions. Here we investigate three Riemannian kernel functions derived from LED and AID, the log-Euclidean trace (LET) kernel [45], the log-Euclidean Gaussian (LEG) kernel [46] and the affine-invariant Gaussian (AIG) kernel [46]:

$$k_{LET}(\mathbf{C}_1, \mathbf{C}_2) \quad = \mathrm{tr}[\log(\mathbf{C}_1) \cdot \log(\mathbf{C}_2)] \tag{6}$$

$$k_{LEG}(\mathbf{C}_1, \mathbf{C}_2) = \exp(-d^2_{LED}(\mathbf{C}_1, \mathbf{C}_2)/2\sigma^2) \tag{7}$$

$$k_{AIG}(\mathbf{C}_1, \mathbf{C}_2) = \exp(-d^2_{AID}(\mathbf{C}_1, \mathbf{C}_2)/2\sigma^2). \tag{8}$$

where σ is the Gaussian bandwidth. Note that AIG is not strictly positive definite. But since AID is a true geodesic distance and AIG could yield good results, it is compared in our experiments.

Let $k(\mathbf{C}_1, \mathbf{C}_2) = \langle \phi(\mathbf{C}_1), \phi(\mathbf{C}_2) \rangle$ denote a kernel function where $\phi(\cdot)$ maps points on $\mathcal{M}$ to a high dimensional Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ for richer representations and inner products. For the three kernels mentioned above, LET performs explicit mapping while LEG and AIG map implicitly. Suppose we have N training samples on $\mathcal{M}$. We first perform kernel linear discriminant analysis (KLDA) by solving the following optimization similar to [47]:

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^T \mathbf{K}\mathbf{L}_B \mathbf{K}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{K}\mathbf{L}_W \mathbf{K}\boldsymbol{\alpha}}, \tag{9}$$

where $\mathbf{K}$ is the kernel Gram matrix: $\mathbf{K}_{jl} = k(\mathbf{C}_j, \mathbf{C}_l)$ $(j, l = 1, ..., N)$, and $\mathbf{L}_B$ and $\mathbf{L}_W$ are the Laplacian matrices of the between-class and within-class graph matrices $\mathbf{E}_B$, $\mathbf{E}_W$ respectively:

$$\mathbf{E}_{B,jl} = \begin{cases} 1/m_{jl}, & \text{if } \mathbf{C}_j \text{ and } \mathbf{C}_l \text{ are in the negative sets of each other} \\ 0, & \text{else} \end{cases} \tag{10}$$

$$\mathbf{E}_{W,jl} = \begin{cases} 1/n_{jl}, & \text{if } \mathbf{C}_j \text{ and } \mathbf{C}_l \text{ are in the positive sets of each other} \\ 0, & \text{else} \end{cases} \tag{11}$$

and m_{jl}, n_{jl} here indicate the corresponding set sizes. For each sample, a positive set and a negative set are formed by all same-class samples and different-class samples respectively. The optimization objective $\boldsymbol{\alpha}$ defines a discriminative projection direction in the RKHS space: $\mathbf{w} = \sum_{k=1}^{N} \boldsymbol{\alpha}_k \phi(\mathbf{C}_k)$.

In order to focus on close sample pairs that should be paid more attention to during the classification, inspired by [13], we take advantage of the locality structure [48] of samples by utilizing the affinity matrix $\mathbf{A}$, which is obtained through a local scaling method [49]: assign the distance with the q-th nearest neighbor as the distance scaling factor θ_k for each point k:

$$\mathbf{A}_{jl} = \exp\left(- \frac{d^2(\mathbf{C}_j, \mathbf{C}_l)}{\theta_j \theta_l} \right). \tag{12}$$

Here q can be set proportional to N. Instead of applying the affinity matrix on both $\mathbf{E}_B, \mathbf{E}_W$, we perform neighbor emphasizing by only penalizing close negative pairs since positive sample pairs are way less in number and all expected to be drawn near: $\mathbf{E}_B^{'} = \mathbf{E}_B \cdot \mathbf{A}$. The affinity $\mathbf{A}$ functions as graph weight here.

While, as described above, a signature of an image group contains 6 independent SPD matrices. To consider matrices semantically irrelevant in body parts separately, we relate only the ones of the same stripe by introducing a graph embedding technique [50] to generate the positive/negative sets, as illustrated in Fig. 4. For any local matrix, a positive pair is only constructed with the *same-position* matrix from other image groups (tracks) with the *same* identity (solid lines connected), while negative pairs are formed with also the *same-position* matrices of other tracks with *different* identities (dash lines connected only). As another way to look at the embedding technique, the 6 stripes are treated as 6 different classes, making a training task with I identities has $c = 6I$ classes.

The problem in Eq. 9 can be tackled by solving the generalized eigenvalue problem: $\mathbf{K}\mathbf{L}_B\mathbf{K}\boldsymbol{\alpha} = \lambda \mathbf{K}\mathbf{L}_W\mathbf{K}\boldsymbol{\alpha}$. Once the $(c-1)$ leading eigenvectors $\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_{c-1}$ are obtained, the coefficient matrix of $(c-1)$ projection directions is naturally settled: $\boldsymbol{\mathcal{A}} = [\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_{c-1}] \in \mathbb{R}^{N \times (c-1)}$.

In the testing phase, the SPD matrices in a signature $S_{test} = \{\mathbf{C}_{test}^s, s = 1, ..., 6\}$ are projected into the same $(c-1)$-dimensional subspace through $\boldsymbol{\mathcal{A}}^T \mathbf{K}_s$, where $\mathbf{K}_s = [k(\mathbf{C}_{train}^1, \mathbf{C}_{test}^s), ..., k(\mathbf{C}_{train}^N, \mathbf{C}_{test}^s)]^T$, and matched with points of *corresponding* spatial position. In the end, the distance between two signatures $S_{i1,t1}$, $S_{i2,t2}$ are obtained by averaging the 6 SPD matrix pair distances:

$$d(S_{i1,t1}, S_{i2,t2}) = \frac{1}{6} \sum_{s=1}^{6} d(\mathbf{C}_{i1,t1}^s, \mathbf{C}_{i2,t2}^s). \tag{13}$$

4.3 Kernel Metric Learning to Rank

To further exhibit the flexibility of our framework, we reformulate a metric learning algorithm to learn another discriminative subspace. Considering that person re-identification is usually formulated as a ranking problem [9,11,16,19,38], the metric learning to rank [51] algorithm, which is specifically designed for ranking and also applied in [11], is adopted with the kernel variant. We refer to it as kernel metric learning to rank (KMLR). It learns a metric matrix with respect to which, the ranking list of gallery samples for each probe resembles the corresponding ground truth list as much as possible.

Given two points in the RKHS space $\boldsymbol{\phi}_j, \boldsymbol{\phi}_l$ and any metric matrix $\mathbf{M} = \boldsymbol{\Phi}\boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{A}}^T\boldsymbol{\Phi}^T = \boldsymbol{\Phi}\mathbf{W}\boldsymbol{\Phi}^T$ where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_N]$ and $\mathbf{W} = \boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{A}}^T$, distance between them is represented with Frobenius inner products:

$$\begin{aligned}\|\boldsymbol{\phi}_j - \boldsymbol{\phi}_l\|_{\mathbf{M}}^2 &= (\boldsymbol{\phi}_j - \boldsymbol{\phi}_l)^T\mathbf{M}(\boldsymbol{\phi}_j - \boldsymbol{\phi}_l) = (\mathbf{K}_j - \mathbf{K}_l)^T\mathbf{W}(\mathbf{K}_j - \mathbf{K}_l) \\ &= \mathrm{tr}(\mathbf{W}(\mathbf{K}_j - \mathbf{K}_l)(\mathbf{K}_j - \mathbf{K}_l)^T) = \langle \mathbf{W}, (\mathbf{K}_j - \mathbf{K}_l)(\mathbf{K}_j - \mathbf{K}_l)^T \rangle_F,\end{aligned} \tag{14}$$

where $\mathbf{K}_{\boldsymbol{\cdot}} = [\langle \boldsymbol{\phi}_1, \boldsymbol{\phi}_{\boldsymbol{\cdot}} \rangle, ..., \langle \boldsymbol{\phi}_N, \boldsymbol{\phi}_{\boldsymbol{\cdot}} \rangle]^T$.

By defining the opposite of the latter part as a kernel feature map for sample pair $(\boldsymbol{\phi}_j, \boldsymbol{\phi}_l)$: $\boldsymbol{\varphi}_{j,l} = -(\mathbf{K}_j - \mathbf{K}_l)(\mathbf{K}_j - \mathbf{K}_l)^T$, we use the partial order feature [52] to present a ranking list $\mathbf{y}_k$ with respect to a sample $\boldsymbol{\phi}_k$:

$$\boldsymbol{\psi}_k(\mathbf{y}_k) = \sum_{j \in \mathcal{X}_k^+} \sum_{l \in \mathcal{X}_k^-} y_k^{jl} \left(\frac{\boldsymbol{\varphi}_{k,j} - \boldsymbol{\varphi}_{k,l}}{|\mathcal{X}_k^+| \cdot |\mathcal{X}_k^-|} \right), \tag{15}$$

where $\mathcal{X}_k^+$ and $\mathcal{X}_k^-$ are the positive and negative sample set w.r.t. $\boldsymbol{\phi}_k$, $|\boldsymbol{\cdot}|$ denotes the set size, and

$$y_k^{jl} = \begin{cases} +1 \text{ if } \boldsymbol{\phi}_j \text{ ranks prior to } \boldsymbol{\phi}_l \\ -1 \text{ if } \boldsymbol{\phi}_j \text{ ranks posterior to } \boldsymbol{\phi}_l \end{cases}. \tag{16}$$

Then with a ground truth ranking $\mathbf{y}_k^*$ for $\boldsymbol{\phi}_k$, the optimization objective is formed by introducing the structural SVM framework:

$$\begin{aligned}&\mathbf{W}^* = \arg\min_{\mathbf{W}} \big\{ \mathrm{tr}(\mathbf{W}) + \beta \cdot \xi \big\}, \\ &\text{s.t. } \frac{1}{N} \sum_{k=1}^{N} \big(\underbrace{\langle \mathbf{W}, \boldsymbol{\psi}_k(\mathbf{y}_k^*) \rangle_F}_{(i)} - \underbrace{\langle \mathbf{W}, \boldsymbol{\psi}_k(\mathbf{y}_k) \rangle_F}_{(ii)} \big) \geq \frac{1}{N} \sum_{k=1}^{N} \underbrace{\Delta(\mathbf{y}_k^*, \mathbf{y}_k)}_{(iii)} - \underbrace{\xi}_{(iv)}, \\ &\quad \forall \mathbf{y}_k \neq \mathbf{y}_k^*, \; \mathbf{W} \succeq 0, \; \xi \geq 0\end{aligned} \tag{17}$$

where ξ is the slack variable and β is the trade-off constant. $\Delta(\mathbf{y}_k^*, \mathbf{y}_k)$, the loss function of $\mathbf{y}_k$ w.r.t. $\mathbf{y}_k^*$, is set to the difference of AUC (area under the ROC curve) scores and serves as the SVM margin. Intuitively, the expression Eq. 17 finds an optimal coefficient matrix $\mathbf{W}$ by which after projected from the original RKHS space, the actual rankings of every sample to all relevant samples obtained simply by Euclidean distances resemble the ground truth rankings close enough (term (*i*)), closer than all the other possible rankings (term (*ii*)) with margins (term (*iii*)), relaxable with a slack variable (term (*iv*)). It can be solved by the cutting-plane algorithm [53] in an iterative manner.

The same graph embedding and matching scheme in Sect. 4.2 are applied here.

5 Experiments

5.1 Feature Representation

As introduced before, our MRDL is flexible with pixel-level feature choices. However, to make better comparisons, we applied the most widely used [10,11,19,20,31,33–35,37] color and texture mixture, RGB+YUV+HS with Schmid and Gabor filters (RGB+YUV+HS+SG). As the same configurations in the literature, the 13-channel Schmid filters have parameters τ and σ set to (2,1), (4,1), (4,2), (6,1), (6,2), (6,3), (8,1), (8,2), (8,3), (10,1), (10,2), (10,3) and (10,4) respectively. Also, the 16-channel Gabor filters use parameters γ, λ, θ and σ^2 set to (0.3,4,0,2), (0.3,8,0,2), (0.4,4,0,1), (0.4,8,0,1), (0.3,4,π/2,2), (0.3,8,π/2,2), (0.4,4,π/2,1), (0.4,8,π/2,1) respectively, each producing a magnitude and phase as responses. In addition, the pixel spatial position (x, y) is appended. Together with the 8 color channels, each pixel is presented by a 39-dimensional feature vector. Thus, the region covariance matrix is of size 39×39 and the region Gaussian matrix 40×40. Meanwhile, we compared the feature with two other ones, RGB+Gradient used in [6] and RGB+YUV+HS+LBP [14], in the experiments.

5.2 Datasets and Evaluation Protocols

We conducted experiments on five benchmark person re-identification datasets: i-LIDS [27], CAVIAR (CAVIAR4REID) [23], PRID2011 [24], CUHK01 [25] and SPRD [54]. The famous **i-LIDS** is constructed with 119 identities with 476 image frames from 2 cameras in an airport arrival hall. Each identity is represented by 2 to 8 (mostly 4) frames of normalized size 64×128 pixels. This dataset is characterized by large illumination changes and severe occlusions, as shown in Fig. 5(a). **CAVIAR** (Fig. 5(b)), a classic multi-shot dataset, consists of 72 identities (50 overlapping ones) in 2 cameras in a shopping center, 10 frames for each person in each view of size varying from 17×39 to 72×144. The images were selected manually to maximize variations in resolution changes, light conditions, occlusions and pose changes, making CAVIAR much more difficult to conquer. As a video-based dataset, **PRID2011** is recorded by 2 cameras outside a building with 385 identities in view A and 749 in view B, and we only make use of the 200 identities appear in both views. The images in each video track vary from 5 frames to hundreds with a unified size of 64×128. Figure 5(c) demonstrates the viewpoint change and stark difference in illumination, background and camera characteristics. Despite of the frame number, **CUHK01** is a pretty large set of 971 identities, 2 views and 2 frames each (60×160) captured in a campus. Images in camera A are all of front and back view, and those in camera B are of lateral view (Fig. 5(d)). As a recently released multi-shot dataset, **SPRD** (Fig. 5(e)) contains image sequences of 37 identities taken from 24 real surveillance cameras. This dataset undergoes huge variations in track and frame number, image size, pose, view, illumination, occlusion, background, even within

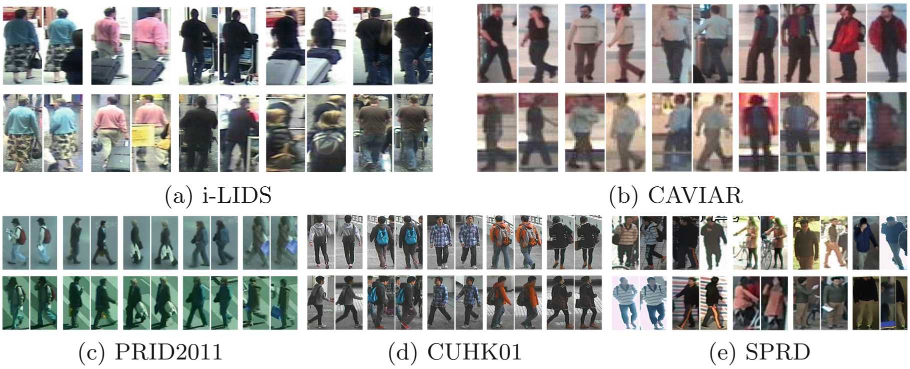

(a) i-LIDS
(b) CAVIAR
(c) PRID2011
(d) CUHK01
(e) SPRD

Fig. 5. Sample images of five introduced person re-identification datasets: (a) i-LIDS, (b) CAVIAR, (c) PRID2011, (d) CUHK01, (e) SPRD. In each dataset except SPRD, the first row presents images from camera A and the second from camera B. For SPRD, images from two random cameras are shown for each person. Each identity in each view is exampled by two frames.

the same sequence. The reason why VIPeR [26], the most well-known dataset, is not adopted is that it is single-shot-based with only 1 frame per view.

As to evaluation protocols, we split the identities of all datasets into equally sized training and testing sets, except for SPRD which had been divided into 3 sessions and was evaluated in 3-fold cross validation[1] Specifically, the training identities of i-LIDS, CAVIAR, PRID2011 and CUHK01 were 59, 25, 100 and 486 respectively. All of the datasets except i-LIDS and SPRD are captured by two views, thus naturally forms two image groups for each identity. For i-LIDS, we kept at most 4 frames for each person, and equally divided the frames of each identity into two image groups randomly, as in [4,6]. For SPRD, all image tracks were used. Also, considering the huge variation in frame number of the PRID2011 tracks, we selected at most 10 frames randomly in each track, making it a moderate size to generate signatures. In the testing phase, image groups from camera A were used as probes, and those from camera B constituted the gallery. For SPRD, a random group (not necessarily from the same camera view, making the task more difficult on SPRD) of each identity was selected as the gallery, and the others as probes. All random splits were performed 10 times. The performance is measured using the Cumulative Matching Characteristic (CMC) curve, which shows the probability of finding the correct match in top r ranks.

Several image pre-processing steps were applied before feature extraction. First, except for CUHK01, all images were resized into a uniform resolution of 64×128. Besides, all color channels of each image were normalized by histogram equalization to handle global illumination changes.

[1] http://ivlab.sjtu.edu.cn/best.

5.3 Module Validation

In this subsection, we validate the effectiveness of the proposed approach[2] by step-by-step verification. For space limits, all results are listed in one table. As demonstrated in Table 1, we firstly analysed each individual component of the method with KLDA, as well as a brief investigation in the case with KMLR. The experiments were conducted with the aforementioned RGB+YUV+HS+SG as features and LEG (Eq. 7) as kernel function. The Gaussian bandwidth σ in kernel function was derived from the mean distance of training data, and the scaling factor assigner q for Eq. 12 was set to $0.1N$. Here we exhibit CMC accuracies of $r = 1, 5, 10$ for quantitative comparisons.

Table 1. Component validation, multi-shot strategy comparison, feature comparison and kernel function comparison of the proposed method. Here *Cov.* and *Gau.* stand for region covariance and Gaussian matrix as local representations. *Unsup.*, *GE* and *NE* refer to unsupervised matching, supervised learning with graph embedding and neighbor emphasizing respectively. Default item selections are bolded. CMC accuracies of $r = 1, 5, 10$ are exhibited, with the highest ones in each validation group highlighted.

Validation group	Validation item	i-LIDS			CAVIAR		
		$r = 1$	$r = 5$	$r = 10$	$r = 1$	$r = 5$	$r = 10$
Component validation	Gau.(Unsup.)	0.4500	0.6333	0.7333	0.1600	0.5600	0.7400
	Gau.+KMLR	0.5083	0.7167	0.8417	0.4200	0.7200	**0.9200**
	Gau.+KLDA(GE)	0.5750	0.7833	0.8500	**0.5000**	**0.7600**	0.8800
	Gau.+KLDA(GE+NE)	**0.5917**	**0.8167**	**0.8917**	**0.5000**	**0.7600**	**0.9200**
	Cov.+KLDA(GE+NE)	0.5833	0.7917	0.8667	0.4600	**0.7600**	**0.9200**
Multi-shot strategy (Gau.(Unsup.))	Closest-point	0.3167	0.5667	0.7250	0.0800	0.3600	0.6000
	Score-voting	0.2500	0.5667	**0.7500**	**0.1800**	0.3600	0.6400
	Signature(mean)	**0.4500**	**0.6333**	0.7333	0.1600	**0.5600**	**0.7400**
Feature	RGB+Gradient	0.4917	0.7167	0.8167	0.3600	0.7600	**0.9600**
	RGB+YUV+HS+LBP	0.5250	0.7667	0.8750	0.4200	**0.7800**	0.8800
	RGB+YUV+HS+SG	**0.5917**	**0.8167**	**0.8917**	**0.5000**	0.7600	0.9200

Validation group	Validation item	SPRD			i-LIDS			CAVIAR		
		$r = 1$	$r = 5$	$r = 10$	$r = 1$	$r = 5$	$r = 10$	$r = 1$	$r = 5$	$r = 10$
Kernel function	LET	0.2000	0.5286	0.9351	0.3667	0.5083	0.5917	0.2600	0.5600	0.7200
	LEG	0.3429	0.7714	0.9675	**0.5917**	**0.8167**	**0.8917**	**0.5000**	0.7600	0.9200
	AIG	**0.5429**	**0.8871**	**1.0000**	0.5750	0.7917	0.8417	**0.5000**	**0.8200**	**0.9400**

It can be observed that compared with the unsupervised version, the proposed discriminative learning approach greatly improves the recognition accuracies, which is extremely obvious on the difficult CAVIAR dataset. Meanwhile, MRDL with KLDA performs generally better than with KMLR, thus the rest validations in this subsection will be held on MRDL(KLDA). Neighbor emphasizing technique boosts the re-identification rate in different degrees on i-LIDS and CAVIAR, mainly after rank-5, thus it is beneficial for practical monitoring systems where usually top-10 matches are displayed to the user. On the other hand, the precisions of region Gaussian matrix are higher than those of region

[2] The source code is released on our website: http://vipl.ict.ac.cn/resources/codes.

covariance matrix mainly on top ranks, especially in the strict multi-shot case CAVIAR, proving it is better to consider mean information in local descriptors.

Afterwards, we compared the applied multi-shot strategy (signature generating by averaging) with the previously discussed closest-point and score-voting ones. It is obvious that averaging yields much better results in most cases, showing that though averaging may lose frame-specific information, it also filters out noise which is particularly common in reID datasets.

Subsequently, we applied three kinds of pixel-level features in the framework: RGB+YUV+HS+SG, RGB+Gradient [6,7] and RGB+YUV+HS+LBP [14]. We used region Gaussian matrix as descriptor and KLDA (with neighbor emphasizing) as kernel learning method. The results exhibit that RGB+YUV+HS+SG has generally stronger description ability. We attribute it to richer color channels of RGB+YUV+HS+SG, compared to RGB+Gradient, and diverse image filters that spatially spread over neighbor pixels compared to the gradients and LBPs.

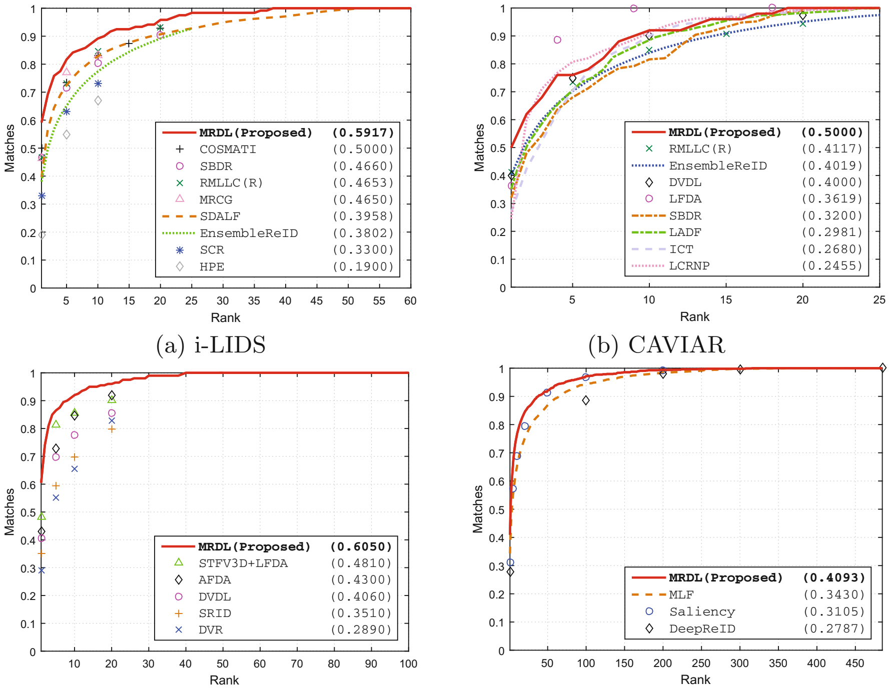
Last but not least, we tested the three kernel functions, LET (Eq. 6), LEG (Eq. 7) and AIG (Eq. 8), on three datasets with RGB+YUV+HS+SG and region Gaussian matrix. It is obvious that different kernel functions vary dramatically in the final performance, and LEG and AIG produce higher accuracies than LET, verifying that discriminative learning on the original manifold preserves Riemannian geometry better compared to in the explicitly mapped Euclidean space. While, considering AIG lacks positive definiteness, we prefer to use LEG in our work. In addition, we are optimistic that the result would be even better if more appropriate Riemannian metrics or kernel functions are applied.

5.4 Comparison Results and Analysis

We compared MRDL with both the multi-shot community and some representative single-shot methods on all datasets except SPRD (Table 2), since different methods reported results on different datasets and none reported on SPRD.

Table 2. Approach comparison chart on i-LIDS (IL), CAVIAR (CA), PRID2011 (PR) and CUHK01 (CU). A check mark denotes a certain method reported results on a corresponding dataset. These works are categorized into single-shot (Sin.) ones and multi-shot ones, which are further divided into closest point-based (CP), score voting-based (SV), set structure-based (SS) and signature-based (SG) groups.

Category		Approach	IL	CA	PR	CU	Category		Approach	IL	CA	PR	CU
Sin.	Feature extraction	SCR [28]	✓				CP		SDALF [4]	✓			
		MLF [21]				✓	SG	Mean	HPE [5]	✓			
		DeepReID [22]				✓			LFDA [13]		✓		
	Metric learning	LADF [29]		✓					MRCG [6]	✓			
		EnsembleReID[14]	✓	✓					COSMATI [7]	✓			
		RMLLC(R) [16]	✓	✓					DVDL [35]		✓	✓	
SV		ICT [32]		✓					Saliency [36]				✓
		SRID [33]			✓				AFDA [37]			✓	
SS		SBDR [11]	✓	✓				Model	STFV3D+LFDA[39]			✓	
		LCRNP [34]		✓					DVR[38]			✓	



(a) i-LIDS

(b) CAVIAR

(c) PRID2011

(d) CUHK01

Fig. 6. CMC curves of the proposed method and comparison methods on four public person re-identification datasets: (a) i-LIDS, (b) CAVIAR, (c) PRID2011, (d) CUHK01. Particularly, the rank-1 accuracies are listed behind the name of each method. Those of which the CMC curves are unavailable are represented as markers. Here the MRDL is with KLDA and region Gaussian matrix descriptor.

The CMC curves are shown in Fig. 6. With only half identities in the supervised methods, the horizontal axis of CMC curves of the unsupervised SDALF, LCRNP, HPE, MRCG and SCR are compressed by 50% for fair comparison. Also, results of methods (LFDA, LADF, DeepReID) with different testing identity numbers are rescaled in the same way.

The proposed MRDL achieves the highest rank-1 accuracy on all evaluated datasets with improvements of 9.17%, 8.83%, 12.40% and 6.63% over the state-of-the-art methods, exhibiting impressive effectiveness of the framework. It can be observed that MRDL is not only discriminative in datasets with small between-class variations, but also robust on the ones with large within-class variations in illumination, resolution, occlusion, pose and view. In addition, even the unsupervised version of our MRDL with region Gaussian matrices only is better than or comparable with most of the works on i-LIDS, demonstrating the representation ability of our local descriptors. Besides, we should note that on CAVIAR, LFDA took different frames in a group as independent training samples, which, actually as a single-shot protocol, introduces multi-modality and is easier to produce higher accuracies in large within-group variance datasets as CAVIAR. Thus, it's inappropriate to be directly compared with the multi-shot MRDL where one image group is treated as one sample. While, DVDL, STFV3D+LFDA and DVR utilized all frames in each track of the PRID2011 dataset, holding richer information since the very beginning, but are still inferior to our method in performance. Also, DeepReID on CUHK01 was trained on 871 identities, naturally possessing an edge in the learning phase. But the accuracies of our MRDL transcend those of it with a large margin.

6 Conclusion

We proposed an effective discriminative learning framework in Riemannian manifold for multiple-shot person re-identification. Different from other multi-shot approaches, our method represents local stripes as SPD matrices, averages on manifold to generate signatures, and perform kernelized learning algorithms also on the Riemannian manifold. Experiments demonstrated the impressive effectiveness, especially the improvements brought by the Riemannian discriminative learning phase, even for the simple LDA, and superiority of the method over the state of the arts on five benchmark datasets. We will further explore more suitable kernel functions and metric learning algorithms, such as those based on fixed boundary sample pairs or relative comparison triplets.

Acknowledgement. This work is partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61390511, 61379083, 61272321, 61271445, and Youth Innovation Promotion Association CAS No. 2015085.

References

1. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person Re-identification, vol. 1. Springer, Heidelberg (2014)
2. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1528–1535. IEEE (2006)
3. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: IEEE International Conference on Computer Vision, pp. 1–8. IEEE (2007)
4. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2360–2367. IEEE (2010)
5. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-shot person re-identification by HPE signature. In: International Conference on Pattern Recognition, pp. 1413–1416. IEEE (2010)
6. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot human re-identification by mean Riemannian covariance grid. In: IEEE International Conference on Advanced Video and Signal-Based Surveillance, pp. 179–184. IEEE (2011)

7. Bąk, S., Charpiat, G., Corvée, E., Brémond, F., Thonnat, M.: Learning to match appearances by correlations in a covariance metric space. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 806–820. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33712-3_58
8. Ma, B., Su, Y., Jurie, F.: Local descriptors encoded by fisher vectors for person re-identification. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 413–422. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33863-2_41
9. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: IEEE International Conference on Computer Vision, pp. 1116–1124. IEEE (2015)
10. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 649–656. IEEE (2011)
11. Wu, Y., Minoh, M., Mukunoki, M., Lao, S.: Set based discriminative ranking for recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 497–510. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33712-3_36
12. Hirzer, M., Roth, P.M., Köstinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 780–793. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33783-3_56
13. Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B.: Local fisher discriminant analysis for pedestrian re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3318–3325. IEEE (2013)
14. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 1–16. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_1
15. Liao, S., Li, S.Z.: Efficient PSD constrained asymmetric metric learning for person re-identification. In: IEEE International Conference on Computer Vision, pp. 3685–3693. IEEE (2015)
16. Chen, J., Zhang, Z., Wang, Y.: Relevance metric learning for person re-identification by exploiting listwise similarities. IEEE Trans. Image Process. **24**, 4741–4755 (2015)
17. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3586–3593. IEEE (2013)
18. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206. IEEE (2015)
19. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: British Machine Vision Conference, pp. 21.1–21.11. BMVA Press (2010)
20. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88682-2_21
21. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 144–151. IEEE (2014)

22. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159. IEEE (2014)
23. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: British Machine Vision Conference, vol. 1, pp. 68.1–68.11. BMVA Press (2011)
24. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 91–102. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21227-7_9
25. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37331-2_3
26. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, vol. 3. IEEE (2007)
27. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: British Machine Vision Conference, pp. 23.1–23.11. BMVA Press (2009)
28. Bak, S., Corvee, E., Brémond, F., Thonnat, M.: Person re-identification using spatial covariance regions of human body parts. In: IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 435–440. IEEE (2010)
29. Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., Smith, J.R.: Learning locally-adaptive decision functions for person verification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3610–3617. IEEE (2013)
30. Ma, B., Su, Y., Jurie, F.: BiCov: a novel image representation for person re-identification and face verification. In: British Machine Vision Conference, pp. 57.1–57.11. BMVA Press (2012)
31. Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L.S., Gao, W.: Multi-task learning with low rank attribute embedding for person re-identification. In: IEEE International Conference on Computer Vision, pp. 3739–3747. IEEE (2015)
32. Avraham, T., Gurvich, I., Lindenbaum, M., Markovitch, S.: Learning implicit transfer for person re-identification. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 381–390. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33863-2_38
33. Karanam, S., Li, Y., Radke, R.: Sparse re-id: block sparsity for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 33–40. IEEE (2015)
34. Wu, Y., Mukunoki, M., Minoh, M.: Locality-constrained collaboratively regularized nearest points for multiple-shot person re-identification. In: Korea-Japan Joint Workshop on Frontiers of Computer Vision. CiteSeer (2014)
35. Karanam, S., Li, Y., Radke, R.J.: Person re-identification with discriminatively trained viewpoint invariant dictionaries. In: IEEE International Conference on Computer Vision, pp. 4516–4524. IEEE (2015)
36. Martinel, N., Micheloni, C., Foresti, G.L.: Saliency weighted features for person re-identification. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8927, pp. 191–208. Springer, Heidelberg (2015). doi:10.1007/978-3-319-16199-0_14
37. Li, Y., Wu, Z., Karanam, S., Radke, R.: Multi-shot human re-identification using adaptive fisher discriminant analysis. In: British Machine Vision Conference. BMVA Press (2015)

38. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 688–703. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_45
39. Liu, K., Ma, B., Zhang, W., Huang, R.: A spatio-temporal appearance representation for video-based pedestrian re-identification. In: IEEE International Conference on Computer Vision, pp. 3810–3818. IEEE (2015)
40. Tuzel, O., Porikli, F., Meer, P.: Region covariance: a fast descriptor for detection and classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 589–600. Springer, Heidelberg (2006). doi:10.1007/11744047_45
41. Amari, S.I., Nagaoka, H.: Methods of Information Geometry, vol. 191. American Mathematical Society, Providence (2007)
42. Lovrić, M., Min-Oo, M., Ruh, E.A.: Multivariate normal distributions parametrized as a Riemannian symmetric space. J. Multivar. Anal. **74**, 36–48 (2000)
43. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J. Matrix Anal. Appl. **29**, 328–347 (2007)
44. Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. Int. J. Comput. Vis. **66**, 41–66 (2006)
45. Wang, R., Guo, H., Davis, L.S., Dai, Q.: Covariance discriminative learning: a natural and efficient approach to image set classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2496–2503. IEEE (2012)
46. Jayasumana, S., Hartley, R., Salzmann, M., Li, H., Harandi, M.: Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 73–80. IEEE (2013)
47. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Comput. **12**, 2385–2404 (2000)
48. He, X., Niyogi, P.: Locality preserving projections. In: Neural Information Processing Systems, vol. 16, pp. 153–160. MIT (2004)
49. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: Advances in Neural Information Processing Systems, pp. 1601–1608 (2004)
50. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 40–51 (2007)
51. McFee, B., Lanckriet, G.R.: Metric learning to rank. In: International Conference on Machine Learning, pp. 775–782. ACM (2010)
52. Joachims, T.: A support vector method for multivariate performance measures. In: International Conference on Machine Learning, pp. 377–384. ACM (2005)
53. Joachims, T., Finley, T., Yu, C.N.J.: Cutting-plane training of structural svms. Mach. Learn. **77**, 27–59 (2009)
54. Zhang, C., Ni, B., Song, L., Yang, X., Zhang, W.: BEST: benchmark and evaluation of surveillance task. In: Chen, C.-S., Lu, J., Ma, K.-K. (eds.) ACCV 2016 Workshops. LNCS, vol. 10118, pp. 393–407. Springer, Heidelberg (2016)