Person Re-Identification by Dual-Regularized KISS Metric Learning

Dapeng Tao, Yanan Guo, Mingli Song, *Senior Member, IEEE*, Yaotang Li, Zhengtao Yu, and Yuan Yan Tang, *Fellow, IEEE*

***Abstract*— Person re-identification aims to match the images of pedestrians across different camera views from different locations. This is a challenging intelligent video surveillance problem that remains an active area of research due to the need for performance improvement. Person re-identification involves two main steps: feature representation and metric learning. Although the keep it simple and straightforward (KISS) metric learning method for discriminative distance metric learning has been shown to be effective for the person re-identification, the estimation of the inverse of a covariance matrix is unstable and indeed may not exist when the training set is small, resulting in poor performance. Here, we present dual-regularized KISS (DR-KISS) metric learning. By regularizing the two covariance matrices, DR-KISS improves on KISS by reducing overestimation of large eigenvalues of the two estimated covariance matrices and, in doing so, guarantees that the covariance matrix is irreversible. Furthermore, we provide theoretical analyses for supporting the motivations. Specifically, we first prove why the regularization is necessary. Then, we prove that the proposed method is robust for generalization. We conduct extensive experiments on three challenging person re-identification datasets, VIPeR, GRID, and CUHK 01, and show that DR-KISS achieves new state-of-the-art performance.**

Manuscript received December 18, 2015; revised March 12, 2016; accepted April 5, 2016. Date of publication April 12, 2016; date of current version April 26, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61572486, Grant 61572428, Grant 61472168, Grant 61273244, Grant U1509206, and Grant 11361074, in part by the Guangdong Natural Science Funds under Grant 2014A030310252, in part by the Key Project of Yunnan Nature Science Foundation under Grant 2013FA130, in part by the Research Grants of University of Macau under Grant MYRG2015-00049-FST, Grant MYRG2015-00050-FST, and Grant RDG009/FST-TYY/2012, in part by the Science and Technology Development Fund (FDCT) of Macau under Grant 100-2012-A3, 026-2013-A, in part by the Macau-China Join Project under Grant 008-2014-AMJ, in part by the Shenzhen Technology Project under Grant JCYJ20140901003939001, and in part by the Science and Technology Innovation Talents Fund Projects within the Ministry of Science and Technology under Grant 2014HE001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaochun Cao.

D. Tao is with the School of Information Science and Engineering, Yunnan University, Kunming 650091, China (e-mail: dapeng.tao@gmail.com).

Y. Guo and Y. Li are with the School of Mathematics and Statistics, Yunnan University, Kunming 650091, China (e-mail: yananguo.ynu@qq.com; liyaotang@ynu.edu.cn).

M. Song is with the College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: brooksong@ieee.org).

Z. Yu is with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650091, China (e-mail: ztyu@hotmail.com).

Y. Y. Tang is with the Faculty of Science and Technology, University of Macau, Macau 999078, China, and also with the College of Computer Science, Chongqing University, Chongqing 400000, China (e-mail: yytang@umac.mo; yytang@cqu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2553446

***Index Terms*— Person re-identification, metric learning, regularization technique, intelligent video surveillance.**

I. Introduction

PERSON re-identification refers to the task of matching images of pedestrians across different camera views at different locations, and the technique is particularly popular for video surveillance [46]. However, person re-identification remains a challenging problem due to the real-world problems of background clutter, occlusions, small target size, and large intra-class variability in illumination, viewpoint, and pose.

Person re-identification typically considers two critical stages: feature representation and metric learning. In the first, effective features are extracted to represent a person, and ideally these should be robust to viewpoint changes and illumination. A discriminant and efficient matching model is then applied for high-performance searching. Figure 1 shows the diagram of whole person re-identification system. Many efforts have been made to improve re-identification performance [29], which can be grouped into two main categories: 1) extracting robust features for effective representation, and 2) developing discriminant matching models to measure image similarity.

Low-level visual features such as color, shape, and texture are widely used for feature representation [37], [53]. For example, Gray and Tao [13] proposed AdaBoost to select robust HSV histograms from a set of texture and color features, in which the HSV histograms were robust to perspective and resolution variability. Local features provide robust feature representations over regions of interest or points [7], [8], [38]. Based on a nonparametric method, local binary patterns (LBP) [39] can be used to depict the local geometric structures of points of interest. By introducing scale-invariant local comparison tolerance, scale-invariant local ternary patterns (SILTP) [25] improved LBP to deliver robustness to image noise and invariance to intensity scale changes. Ma *et al.* [32] proposed a global representation by converting local descriptors into Fisher vectors. Since pedestrians viewed by different cameras may appear in different viewpoints, some schemes [40] have proposed dividing a person image into several horizontal stripes so that a single histogram can be computed for each stripe. Köstinger *et al.* [21] partitioned images into regular grids from which texture and color features were extracted from overlapping blocks, while Cheng *et al.* [4] employed a pictorial structure model and considered color displacement and part-based color information.

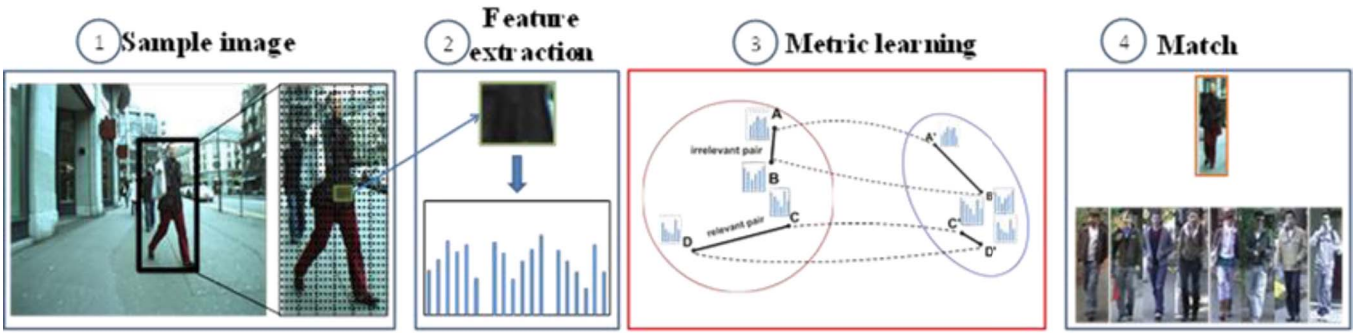

Fig. 1. The diagram of whole person re-identification system. Features are first extracted from each sample, and then a metric learning model is trained, finally the matching rank is found according to the query target.

Liao *et al.* [26] proposed local maximal occurrence (LOMO) to represent each image as a high-dimensional feature and in doing so not only extracted HSV color histograms and SILTP features from the image to form high-level descriptors but also analyzed the horizontal occurrence of local geometric features and maximized the occurrence against viewpoint changes.

Combined features are usually of high dimension. Thus, dimension reduction is required to obtain a concise yet effective feature for subsequent matching. Principal component analysis (PCA) [19] is an extremely popular unsupervised linear algorithm that is effective for eliminating Gaussian noise. Other typical unsupervised algorithms include locality preserving projections (LPP) [16], locally linear embedding (LLE) [41], and Laplacian eigenmaps (LE) [1]. Linear discriminant analysis (LDA) [9] is a traditional supervised algorithm that aims to optimally classify samples drawn from Gaussians with different means but identical covariance. Geometric mean-based subspace selection (GMSS) [44] is LDA improvements that use the geometric mean instead of the arithmetic mean.

Here we focus on how best to learn a similarity function or robust distance to optimize matching. Typical metric learning algorithms include relative distance comparison (RDC) [54], large margin nearest neighbor (LMNN) [47], and information-theoretic metric learning (ITML) [5]. Although state-of-the-art approaches have achieved top-level performance, there is still plenty of room for improvement. KISS metric learning is one of the most efficient and effective metric learning algorithms, exhibiting excellent performance for person re-identification and face recognition [21]. However, KISS suffers from the small sample size problem when estimating the inverse of covariance matrices, which can result in poor retrieval precision in practice.

In this paper, motivated by theoretical justifications, we introduce regularization techniques to improve KISS metric learning for person re-identification. Usually, large eigenvalues in the true covariance matrix are highly biased in the estimated covariance matrix, and the intrapersonal and extrapersonal covariance matrices contain different samples. The utilization of the estimated covariance matrix in subsequent operations such as classification suffers due to the overestimation of large eigenvalues. To obtain a more robust estimate of the two covariance matrices for KISS metric learning, we introduce regularization [10] to the two covariance matrices to reduce this bias. Furthermore, we prove why the regularization is necessary and the proposed method is robust for generalization. We refer to the improved KISS method as dual-regularized KISS (DR-KISS). The main contributions of the proposed DR-KISS algorithm are summarized as follows:

1. We introduce a recently developed regularization [10] to the two covariance matrices to reduce the issue that large eigenvalues in the true covariance matrix are highly biased.

2. We theoretically show that this regularization is necessary and the proposed method is robust for generalization.

The DR-KISS procedure for person re-identification can be summarized as follows: 1) LOMOs are extracted from each sample; 2) PCA is then conducted to obtain a low-dimensional feature representation for each sample; 3) DR-KISS is trained; and 4) the matching rank is found according to the query target.

The remainder of the paper is organized as follows. Previous studies related to improving metric learning for person re-identification are briefly reviewed in Section II. In Section III, we detail DR-KISS. We present our experimental results on representative datasets (VIPeR [12], GRID [30], and CUHK 01 [24]) in Section IV, and we conclude in Section V.

II. Related Work

Metric learning has received increasing attention for person re-identification [5], [22], [34], [42], [47], [52], [54]. From a statistical inference perspective, Köstinger *et al.* [21] proposed a simple yet effective method, the KISS metric, to learn the distance metric. Zheng *et al.* [54] proposed probabilistic relative distance comparison (RRDC), a discriminative model that maximized the probability that distances corresponding to the correct matches were smaller than the incorrect matches. Weinberger *et al.* [47] proposed the large-margin nearest neighbor metric (LMNN) to improve on traditional kNN classification. However, by using the k closest within-class samples, LMNN is time consuming. As an LMNN variant, Dikmen *et al.* [6] improved on previous results by introducing a reject option for unfamiliar matches (LMNN-R). Vapnik [45] proposed a metric learning model called pairwise-constrained component analysis (PCCA) to learn a low-dimensional mapping in which the distances between data points complied with some sparse training pairwise constraints. From the

TABLE I
IMPORTANT NOTATIONS USED IN THIS PAPER AND THEIR DESCRIPTION

Notation	Description	Notation	Description
$\mathbf{x}_i$	Sample	$\mathbf{x}_{ij}$	Difference of the two samples
H_I / H_E	hypothesis that the feature vector pair is similar/ dissimilar	Ω_I / Ω_E	The intrapersonal/ extrapersonal variation
$\delta(\mathbf{x}_{ij})$	the ratio between the two posteriors	γ_I / γ_E	The intrapersonal/ extrapersonal parameter
$f(\mathbf{x}_{ij} \mid \theta_t)$	The probability density functions of $\mathbf{x}_{ij}$ with parameter θ_t	$\mathbf{\Sigma}_I / \mathbf{\Sigma}_E$	The covariance matrix of intrapersonal/ extrapersonal difference $\mathbf{x}_{ij}$
$\hat{\mathbf{\Sigma}}_I / \hat{\mathbf{\Sigma}}_E$	The estimated of $\mathbf{\Sigma}_I / \mathbf{\Sigma}_E$	N_I / N_E	The number of Ω_I / Ω_E

information theory perspective, Davis *et al.* [5] proposed information-theoretic metric learning (ITML) based on the Mahalanobis distance metric. However, most of these methods perform poorly when there are insufficient training samples or when the view conditions are highly variable.

As well as the distance metric learning-based matching schemes, some other matching schemes have been exploited to improve person re-identification performance. Jiang *et al.* [20] proposed cross-domain SVM, in which each training sample was weighted according to its proximity to the test sample. Rank support vector machines (RankSVM) [17] and Ensemble RankSVM [40] have also been widely used for person re-identification, while Hamdoun *et al.* [14] collected points of interest from surveillance video shots to estimate the appearance model.

III. DUAL-REGULARIZED KISS

A. Review of KISS Metric Learning

Based on the assumption that different pairwise samples are Gaussian distributed, the KISS metric learning (KISS) has acquired the state of the art performance for person re-identification on the several benchmark databases.

For convenience, Table I lists frequently used notations and descriptions in this paper. Consider a feature vector pair $\mathbf{x}_i$ and $\mathbf{x}_j$ representing two samples, respectively. We denote the feature vector pair difference by $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$: $\mathbf{x}_{ij}$ is the "intrapersonal difference" if $\mathbf{x}_i$ and $\mathbf{x}_j$ are samples from the same person corresponding to hypothesis H_I that the feature vector pair is similar; $\mathbf{x}_{ij}$ is the "extrapersonal difference" if $\mathbf{x}_i$ and $\mathbf{x}_j$ are samples from the different people corresponding to the hypothesis H_E that the pair of feature vectors are dissimilar. We can therefore define two variation classes: the extrapersonal variation Ω_E and the intrapersonal variation Ω_I. The logarithm of the ratio between the two posteriors is defined as:

$$\delta(\mathbf{x}_i, \mathbf{x}_j) = \log\left(\frac{p(\mathbf{x}_i, \mathbf{x}_j \mid H_E)}{p(\mathbf{x}_i, \mathbf{x}_j \mid H_I)}\right). \tag{1}$$

For classification, a large $\delta(\mathbf{x}_i, \mathbf{x}_j)$ value indicates that $\mathbf{x}_i$ and $\mathbf{x}_j$ represent different people, while a small value indicates that $\mathbf{x}_i$ and $\mathbf{x}_j$ represent the same people. Thus, we have

$$\delta(\mathbf{x}_{ij}) = \log\left(p(\mathbf{x}_{ij} \mid H_E) / p(\mathbf{x}_{ij} \mid H_I)\right), \tag{2}$$

which can be rewritten as

$$\delta(\mathbf{x}_{ij}) = \log\left(f(\mathbf{x}_{ij} \mid \theta_E) / f(\mathbf{x}_{ij} \mid \theta_I)\right), \tag{3}$$

where $f(\mathbf{x}_{ij} \mid \theta_t)$ are the probability density functions of $\mathbf{x}_{ij}$ with parameter θ_t under hypothesis H_t. It is assumed that $\mathbf{x}_{ij}$ satisfies a Gaussian distribution.

For intrapersonal hypothesis H_I, we have:

$$f(\mathbf{x}_{ij} \mid H_I) = \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}_I|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{x}_{ij}^T \mathbf{\Sigma}_I^{-1} \mathbf{x}_{ij}\right); \tag{4}$$

and for extrapersonal hypothesis H_E, we have:

$$f(\mathbf{x}_{ij} \mid H_E) = \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}_E|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{x}_{ij}^T \mathbf{\Sigma}_E^{-1} \mathbf{x}_{ij}\right), \tag{5}$$

where d is the feature vector dimensionality, $\mathbf{\Sigma}_I$ is the covariance matrix of intrapersonal difference $\mathbf{x}_{ij}$, and $\mathbf{\Sigma}_E$ is the covariance matrix of extrapersonal difference $\mathbf{x}_{ij}$.

Given (4), (3) can be rewritten as

$$\delta(\mathbf{x}_{ij}) = \frac{1}{2}\mathbf{x}_{ij}^T \left(\mathbf{\Sigma}_I^{-1} - \mathbf{\Sigma}_E^{-1}\right) \mathbf{x}_{ij} + \frac{1}{2}\log\left(\frac{|\mathbf{\Sigma}_I|}{|\mathbf{\Sigma}_E|}\right). \tag{6}$$

By dropping the constant terms, we have

$$\delta(\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T \left(\mathbf{\Sigma}_I^{-1} - \mathbf{\Sigma}_E^{-1}\right) \mathbf{x}_{ij}. \tag{7}$$

Define y_{ij} as the indicated variable of $\mathbf{x}_i$ and $\mathbf{x}_j$: $y_{ij} = 1$ if $\mathbf{x}_{ij}$ belong to Ω_I ($\mathbf{x}_i$ and $\mathbf{x}_j$ are the same person), and $y_{ij} = 0$ if $\mathbf{x}_{ij}$ belong to Ω_E. Let N_I and N_E denote the number of Ω_I and Ω_E, respectively. It should be obvious that N_I is usually much smaller than N_E. Thus, the small sample size problem often happens. The covariance matrices are estimated as follows:

$$\begin{aligned}
\hat{\mathbf{\Sigma}}_I &= \frac{1}{N_I}\sum_{y_{ij}=0} \mathbf{x}_{ij}\mathbf{x}_{ij}^T = \frac{1}{N_I}\sum_{y_{ij}=0} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \\
\hat{\mathbf{\Sigma}}_E &= \frac{1}{N_E}\sum_{y_{ij}=1} \mathbf{x}_{ij}\mathbf{x}_{ij}^T = \frac{1}{N_E}\sum_{y_{ij}=1} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T.
\end{aligned} \tag{8}$$

From (8) we can easily see that $\mathbf{\Sigma}_I$ and $\mathbf{\Sigma}_E$ are semi-positive definite matrices. Therefore, their eigenvalues are non-negative.

Let $M = \hat{\mathbf{\Sigma}}_I^{-1} - \hat{\mathbf{\Sigma}}_E^{-1}$. Therefore, the derived distance function between $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$d\left(\mathbf{x}_i, \mathbf{x}_j\right) = \delta\left(\mathbf{x}_{ij}\right) = \mathbf{x}_{ij}^T M \mathbf{x}_{ij}, \tag{9}$$

where M is the KISS metric matrix.

B. DR-KISS Metric Learning

Although KISS significantly improves person re-identification, there is plenty of scope to improve its stability and efficiency. Specifically, accurately estimating the covariance matrices in (6) is crucial for improving person re-identification performance. Given limited training samples, it is known that the Gaussian distribution model suffers from estimation error. Therefore, the large eigenvalues of the true covariance matrix are always highly biased in the estimated covariance matrix (we provide theoretical justification for this claim in Section C), which results in an ill-posed problem.

There are many statistical techniques that provide robust estimations. Through analyzing the generalization error [27], [45], we introduce a regularization method to improve covariance matrix estimation in KISS, because the regularization greatly improves the generalization ability of the proposed method by directly minimizing the upper generalization error bound; and it also reduces the effect of overestimation of large eigenvalues in an estimated covariance matrix, as justified in [10]. Furthermore, regularization can also be implemented to avoid the problem that the covariance matrix may be reversible. The use of regularization consequently improves the effectiveness of KISS for person re-identification.

Since the two covariance matrices $\hat{\mathbf{\Sigma}}_I$ and $\hat{\mathbf{\Sigma}}_E$ in (8) are semi-positive definite matrices, they can be diagonalized and written as:

$$\hat{\mathbf{\Sigma}}_I = \mathbf{\Phi}_I \mathbf{\Lambda}_I \mathbf{\Phi}_I^T, \tag{10}$$

$$\hat{\mathbf{\Sigma}}_E = \mathbf{\Phi}_E \mathbf{\Lambda}_E \mathbf{\Phi}_E^T, \tag{11}$$

where $\Lambda_I = \mathrm{diag}\left[\lambda_1, \lambda_2, \cdots, \lambda_m\right]$, with λ_i being the first i^{th} eigenvalue of $\hat{\mathbf{\Sigma}}_I$, and $\mathbf{\Phi}_I = [\phi_1, \phi_2, \cdots, \phi_m]$ with ϕ_i corresponding to λ_i being an eigenvector of $\hat{\mathbf{\Sigma}}_I$. $\Lambda_E = \mathrm{diag}\left[\mu_1, \mu_2, \cdots, \mu_m\right]$, with μ_i being the first i^{th} eigenvalue of $\hat{\mathbf{\Sigma}}_E$, and $\mathbf{\Phi}_E = [\varphi_1, \varphi_2, \cdots, \varphi_m]$, with φ_i corresponding to λ_i being an eigenvector of $\hat{\mathbf{\Sigma}}_E$. Furthermore, we know that $\mathbf{\Phi}_I$ and $\mathbf{\Phi}_E$ are orthogonal matrices.

According to the regularization technique [10] and Theorem 1 in Section C, which provides a theoretical justification for the regularization technique from the statistical learning theory [45] viewpoint and also explicitly explains why the small sample size problem can be best reduced, the two covariance matrices in (8) are modified[1] and interpolated using an identity matrix via the intrapersonal and extrapersonal parameters; that is,

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{I,\gamma_I} &= (1-\gamma_I)\,\hat{\mathbf{\Sigma}}_I + \gamma_I \alpha_I \mathbf{I} \\ &= (1-\gamma_I)\,\mathbf{\Phi}_I \mathbf{\Lambda}_I \mathbf{\Phi}_I^T + \gamma_I \alpha_I \mathbf{\Phi}_I \mathbf{\Phi}_I^T \\ &= \mathbf{\Phi}_I \left[(1-\gamma_I)\,\mathbf{\Lambda}_I + \gamma_I \alpha_I \mathbf{I}\right] \mathbf{\Phi}_I^T, \end{aligned} \tag{12}$$

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{E,\gamma_E} &= (1-\gamma_E)\,\hat{\mathbf{\Sigma}}_E + \gamma_E \alpha_E \mathbf{I} \\ &= (1-\gamma_E)\,\mathbf{\Phi}_E \mathbf{\Lambda}_E \mathbf{\Phi}_E^T + \gamma_E \alpha_E \mathbf{\Phi}_E \mathbf{\Phi}_E^T \\ &= \mathbf{\Phi}_E \left[(1-\gamma_E)\,\mathbf{\Lambda}_E + \gamma_E \alpha_E \mathbf{I}\right] \mathbf{\Phi}_E^T, \end{aligned} \tag{13}$$

where $\alpha_I = (1/d)\,\mathrm{tr}\left(\hat{\mathbf{\Sigma}}_I\right)$, $\alpha_E = (1/d)\,\mathrm{tr}\left(\hat{\mathbf{\Sigma}}_E\right)$, $0 < \gamma_I < 1$, and $0 < \gamma_E < 1$. The intrapersonal parameter γ_I and extrapersonal parameter γ_E can shrink $\hat{\mathbf{\Sigma}}_{I,\gamma_I}$ and $\hat{\mathbf{\Sigma}}_{E,\gamma_E}$ toward an identity matrix, respectively. Since the two covariance matrices represent the variability of the intrapersonal and extrapersonal differences, respectively, it is reasonable to use different trade-off parameters to regularize the two covariance matrices. The shrunken estimates of the two covariance matrices suppress the larger estimate of the large eigenvalues. In this way, performance is improved in practice [10], [15].

We define

$$\begin{aligned}\hat{\mathbf{\Lambda}}_{I,\gamma_I} &= (1-\gamma_I)\,\mathbf{\Lambda}_I + \gamma_I \alpha_I \mathbf{I} \\ &= \mathrm{diag}[(1-\gamma_I)\,\lambda_1 + \gamma_I \alpha_I, (1-\gamma_I)\,\lambda_2 + \gamma_I \alpha_I, \\ &\qquad \cdots, (1-\gamma_I)\,\lambda_m + \gamma_I \alpha_I] \end{aligned} \tag{14}$$

$$\begin{aligned}\hat{\mathbf{\Lambda}}_{E,\gamma_E} &= (1-\gamma_E)\,\mathbf{\Lambda}_E + \gamma_E \alpha_E \mathbf{I} \\ &= \mathrm{diag}[(1-\gamma_E)\,\mu_1 + \gamma_I \alpha_E, (1-\gamma_E)\,\mu_2 + \gamma_E \alpha_E, \\ &\qquad \cdots, (1-\gamma_E)\,\mu_m + \gamma_E \alpha_E] \end{aligned} \tag{15}$$

Replacing $\hat{\mathbf{\Sigma}}_I$ and $\hat{\mathbf{\Sigma}}_E$ with $\hat{\mathbf{\Sigma}}_{I,\gamma_I}$ and $\hat{\mathbf{\Sigma}}_{E,\gamma_E}$ in equation (12), respectively, we obtain

$$\begin{aligned}\delta\left(\mathbf{x}_{ij}\right) &= \mathbf{x}_{ij} \left(\hat{\mathbf{\Sigma}}_{I,\gamma_I}^{-1} - \hat{\mathbf{\Sigma}}_{E,\gamma_E}^{-1}\right) \mathbf{x}_{ij}^T \\ &= \mathbf{x}_{ij} \left(\mathbf{\Phi}_I \hat{\mathbf{\Lambda}}_{I,\gamma_I}^{-1} \mathbf{\Phi}_I^T - \mathbf{\Phi}_E \hat{\mathbf{\Lambda}}_{E,\gamma_E}^{-1} \mathbf{\Phi}_E^T\right) \mathbf{x}_{ij}^T \\ &= \left[\mathbf{\Phi}_I^T \mathbf{x}_{ij}\right]^T \hat{\mathbf{\Lambda}}_{I,\gamma_I}^{-1} \left[\mathbf{\Phi}_I^T \mathbf{x}_{ij}\right] \\ &\quad - \left[\mathbf{\Phi}_E^T \mathbf{x}_{ij}\right]^T \hat{\mathbf{\Lambda}}_{E,\gamma_E}^{-1} \left[\mathbf{\Phi}_E^T \mathbf{x}_{ij}\right]. \end{aligned} \tag{16}$$

By substituting (14) and (15) into (16), we obtain

$$\begin{aligned}\delta\left(\mathbf{x}_{ij}\right) &= \sum_{n=1}^{m} \frac{1}{(1-\gamma_I)\,\lambda_n + \gamma_I \alpha_I} \left(\phi_n^T \mathbf{x}_{ij}\right)^2 \\ &\quad - \sum_{n=1}^{m} \frac{1}{(1-\gamma_E)\,\mu_n + \gamma_I \alpha_E} \left(\varphi_n^T \mathbf{x}_{ij}\right)^2. \end{aligned} \tag{17}$$

Given (17), it is easy to match by ranking reference images $\mathbf{x}_j$ according to $\delta\left(\mathbf{x}_{ij}\right)$ given a query image $\mathbf{x}_i$. A reference image corresponding to smaller $\delta\left(\mathbf{x}_{ij}\right)$ ranks near the top.

C. Theoretically Analyses

This paper mainly introduces a regularization technique for improving the performance of person re-identification motivated by that it has the small sample size problem for

[1]From statistical learning theory, the covariance matrices in (8) are the solutions of the empirical risk minimization algorithms in (18) and (19). As motivated by Theorem 1, to improve the generalization ability and solve the small sample size problem, we would like to penalize the trace norm to be small as illustrated in (20), whose solutions are the ones presented in (12) and (13).

estimating the inverse of the covariate matrices, which may result in retrieval precision not performing robustly in practice. We therefore provide theoretically analyses for supporting the motivations. Specifically, we first prove why the regularization is necessary. Then, we prove that the proposed method is also robust for generalization.

Let us first analyze the theoretical motivation for our proposed estimations in (12) and (13). The expected covariance matrices Σ_I and Σ_E are usually unavailable for practical applications. They are then approximated by solving the following problem:

$$\begin{aligned}\hat{\Sigma}_I &= \arg\min_{\Sigma} \frac{1}{N_I} \sum_{y_{ij}=0} \ell(x_i, x_j, \Sigma) \\ &= \arg\min_{\Sigma} \frac{1}{N_I} \sum_{y_{ij}=0} \left\| \left(x_i - x_j\right)\left(x_i - x_j\right)^T - \Sigma \right\|_F^2, \quad (18) \\ \hat{\Sigma}_E &= \arg\min_{\Sigma} \frac{1}{N_E} \sum_{y_{ij}=1} \ell(x_i, x_j, \Sigma) \\ &= \arg\min_{\Sigma} \frac{1}{N_E} \sum_{y_{ij}=1} \left\| \left(x_i - x_j\right)\left(x_i - x_j\right)^T - \Sigma \right\|_F^2, \quad (19)\end{aligned}$$

where $\ell\left(x_i, x_j, \Sigma\right) = \left\| \left(x_i - x_j\right)\left(x_i - x_j\right)^T - \Sigma \right\|_F^2$ is the prediction loss with respect to the sample pair $\left(x_i, x_j\right)$.

It is easy to verify that the solutions to problems (18) and (19) have the closed forms that are the same as listed in equations (8):

$$\begin{aligned}\hat{\mathbf{\Sigma}}_I &= \frac{1}{N_0} \sum_{y_{ij}=0} \mathbf{x}_{ij}\mathbf{x}_{ij}^T = \frac{1}{N_I} \sum_{y_{ij}=0} \left(\mathbf{x}_i - \mathbf{x}_j\right)\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \\ \hat{\mathbf{\Sigma}}_E &= \frac{1}{N_1} \sum_{y_{ij}=1} \mathbf{x}_{ij}\mathbf{x}_{ij}^T = \frac{1}{N_E} \sum_{y_{ij}=1} \left(\mathbf{x}_i - \mathbf{x}_j\right)\left(\mathbf{x}_i - \mathbf{x}_j\right)^T\end{aligned}$$

The above solutions are widely used for estimating the covariance matrices, e.g., [21]. We therefore want to know if the empirical estimations $\hat{\mathbf{\Sigma}}_I$ and $\hat{\mathbf{\Sigma}}_E$ general well. By exploiting the law of large numbers, it can be easily verified that the empirical estimations will asymptotically converge to the expected covariance matrices as the sample size goes to infinity. However, in practice, we often have a small size of sample for applications. Then, we are interested in analyzing the generalization bounds [27], [45] of the algorithms in (18) and (19) with finite sample sizes N_I and N_E:

$$\sup_{\Sigma} \left(E_{\Omega_I}\left[\frac{1}{N_I} \sum_{y_{ij}=0} \ell\left(x_i, x_j, \Sigma\right)\right] - \frac{1}{N_I} \sum_{y_{ij}=0} \ell\left(x_i, x_j, \Sigma\right) \right)$$

and

$$\sup_{\Sigma} \left(E_{\Omega_E}\left[\frac{1}{N_E} \sum_{y_{ij}=1} \ell\left(x_i, x_j, \Sigma\right)\right] - \frac{1}{N_E} \sum_{y_{ij}=1} \ell\left(x_i, x_j, \Sigma\right) \right)$$

where $E_x\left[\cdot\right]$ represents the expectation operation with respect to the variable x. We further assume that there are n_E person examples in total and each example has n_I observations (or pictures). Then, $N_I = |\Omega_I| = n_E \times n_I \times (n_I - 1)$ and $N_E = n_I \times n_E \times n_I \times (n_E - 1)$. We have

Theorem 1: Let the observations are bounded, i.e., $\left\|(x-y)(x-y)^T\right\|_F \leq M, \forall x, y$ are the observations. For any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\begin{aligned}&\sup_{\Sigma} \left(E_{\Omega_I}\left[\frac{1}{N_I} \sum_{y_{ij}=0} \ell\left(x_i, x_j, \Sigma\right)\right] - \frac{1}{N_I} \sum_{y_{ij}=0} \ell\left(x_i, x_j, \Sigma\right) \right) \\ &\quad \leq \frac{8M}{\sqrt{n_I \times n_E}} \mathrm{Tr}\left(\Sigma\right) + M\sqrt{\frac{2\log 1/\delta}{n_I \times n_E}}\end{aligned}$$

and

$$\begin{aligned}&\sup_{\Sigma} \left(E_{\Omega_E}\left[\frac{1}{N_E} \sum_{y_{ij}=1} \ell\left(x_i, x_j, \Sigma\right)\right] - \frac{1}{N_E} \sum_{y_{ij}=1} \ell\left(x_i, x_j, \Sigma\right) \right) \\ &\quad \leq \frac{8M}{\sqrt{n_I \times n_E \times n_I}} \mathrm{Tr}\left(\Sigma\right) + M\sqrt{\frac{2\log 1/\delta}{n_I \times n_E \times n_I}}\end{aligned}$$

See the proof in the Appendix.

We know that estimating (the inverse of) the covariance matrices (Σ_I and Σ_E) suffers the small sample size problem. Our results in Theorem 1 also provide theoretical justification for this phenomenon, the generalization bound will be large because when $n_I n_E$ is small because the term $\frac{8M}{\sqrt{n_I \times n_E}} \mathrm{Tr}\left(\Sigma\right)$ is large. According to Theorem 1, for every Σ, $\frac{1}{N_I} \sum_{y_{ij}=0} \ell\left(x_i, x_j, \Sigma\right) - \frac{8M}{\sqrt{n_I \times n_E}} \mathrm{Tr}\left(\Sigma\right) - M\sqrt{\frac{2\log 1/\delta}{n_I \times n_E}}$ is upper bounded by $E_{\Omega_I}\left[\frac{1}{N_I} \sum_{y_{ij}=0} \ell\left(x_i, x_j, \Sigma\right)\right]$ with a high probability. To encourage $E_{\Omega_I}\left[\frac{1}{N_I} \sum_{y_{ij}=0} \ell\left(x_i, x_j, \Sigma\right)\right]$ to be small and overcome the small sample size problem, we are theoretically motivated to minimize $\frac{1}{N_I} \sum_{y_{ij}=0} \ell\left(x_i, x_j, \Sigma\right) - \frac{8M}{\sqrt{n_I \times n_E}} \mathrm{Tr}\left(\Sigma\right)$. We therefore modify the learning algorithms in (18) as follows:

$$\begin{aligned}\hat{\Sigma}_{I,\lambda} &= \arg\min_{\Sigma} \frac{1}{N_I} \sum_{y_{ij}=0} \ell\left(x_i, x_j, \Sigma\right) - \lambda \mathrm{Tr}(\Sigma), \\ \hat{\Sigma}_{E,\lambda} &= \arg\min_{\Sigma} \frac{1}{E_I} \sum_{y_{ij}=1} \ell\left(x_i, x_j, \Sigma\right) - \lambda \mathrm{Tr}(\Sigma). \quad (20)\end{aligned}$$

By solving problems (20), we can obtain the closed form solutions that are the same as that in equations (12) and (13):

$$\begin{aligned}\hat{\Sigma}_{I,\lambda_I} &= \frac{1}{N_I} \sum_{y_{ij}=0} \left(x_i - x_j\right)\left(x_i - x_j\right)^T + \lambda_I I, \\ \hat{\Sigma}_{E,\lambda_E} &= \frac{1}{N_E} \sum_{y_{ij}=1} \left(x_i - x_j\right)\left(x_i - x_j\right)^T + \lambda_E I,\end{aligned}$$

where $\lambda_I = \frac{\gamma_I \alpha_I}{1-\gamma_I}$ and $\lambda_E = \frac{\gamma_E \alpha_E}{1-\gamma_E}$.

The above theoretical analyses provide theoretical justification for our proposed DR-KISS algorithm.

Note that our proposed DR-KISS algorithm is simple and efficient. Compared with the traditional estimation methods of covariance matrices, e.g., [21], our proposed method is also robust, such as label noise [28], [35], for person re-identification. That is to say, when Ω_I (or Ω_E) has noise or observations from the same (or different) person, our

TABLE II
STATISTICS OF DATASETS

Datasets	individuals	images	single/multi-shot
VIPeR	632	1264	single-shot
GRID	900	1275	single-shot
CUHK 01	971	3884	multi-shot

proposed method will provide better covariance estimations and thus identification performances. We prove our assertion by analyzing the robustness [49] of the proposed DR-KISS algorithm. We have the following theorem to measure the robustness:

Theorem 2: Let $\mathcal{N}(\eta, \text{support}(\Omega), \ell_2)$ be the covering number of support (Ω). Assume that the observations are upper bounded, *i. e.*, for all $x \in X, \|x\| \leq \Lambda_X$. If x and y are in the same subset divided by the covering number, which means that $\|x - y\| \leq \eta$, for any query z, we have

$$|d(x,z) - d(y,z)| \leq 2\Lambda_X \eta \sqrt{\sum_{i=1}^{n} \frac{1}{\left((\Lambda_I)_{ii} + \lambda_I\right)^2} + \sum_{i=1}^{n} \frac{1}{\left((\Lambda_E)_{ii} + \lambda_E\right)^2}}.$$

where $\hat{\mathbf{\Sigma}}_I = \mathbf{\Phi}_I \mathbf{\Lambda}_I \mathbf{\Phi}_I^T$ and $(\mathbf{\Lambda}_I)_{ii}$ is the i^{th} diagonal element of $\mathbf{\Lambda}_I$, and the same as Λ_E.

See the proof in the Appendix.

According to Theorem 2, we can see that if the two observations are close to each, the output by our proposed method will become closer than the traditional ones. Thus, our method is more robust to noise as claimed.

IV. EXPERIMENTS

To evaluate DR-KISS, we conducted experiments on three widely used public datasets: VIPeR [12], QMUL underGround Re-IDentification (GRID) [30], and CUHK 01 (CUHK Campus) [24]. The VIPeR dataset is widely used for benchmark evaluation of person re-identification algorithms, and it contains 632 individuals and each individual has two images. The GRID dataset contains 1275 images collected from 900 individuals, and the CUHK Campus dataset contains 971 individuals and each individual has four images. Since these datasets each have their own characteristics, they are suitable for evaluating different person re-identification schemes. Table II shows some properties of three datasets.

LOMO was used to represent each image as a high-dimensional feature. LOMO extracts HSV color histograms and SILTP [25] from the image to form high-level descriptors. LOMO employs the Retinex transformation and a scale invariant operator, analyzes the horizontal occurrence of local features, and maximizes the occurrence to obtain a robust representation to handle viewpoint changes; it is, therefore, ideally suited to person re-identification. In this paper, we learn a distance metric on set A and test the learned metric on set B. Performance is measured by calculating and comparing average cumulative match characteristic (CMC) curves. A CMC curve plots the probability of identification against the returned $1:n$ candidate list size. It shows the probability that a given user appears in different sized candidate lists. The faster the CMC curve approaches 1, indicating that the user always appears in the candidate list of specified size, the better the matching algorithm. The experimental details and results are described in the following sections.

Fig. 2. Samples from the VIPeR dataset. Each column represents the same person from different camera views. The different variations in pose, viewpoint, illumination, shooting location, and image quality are easily appreciated.

A. Experiments on VIPeR

The widely used VIPeR dataset [12] includes 1,264 outdoor images of 632 subjects taken from two different views. All images are normalized to 128×48. Most image pairs have viewpoint changes near or over 90 degrees, making it difficult to match the same person from two different views; furthermore, other variations such as light conditions, image quality, and shooting locations are also considered. Thus, the VIPeR dataset is particularly challenging. Example images are shown in Figure 2.

In our experiments we adopted two widely-used experimental protocols to evaluate performance: (i) by randomly selecting 100 subjects as the training set and the remaining 532 subjects as the test set; and (ii) by randomly selecting 316 subjects as the training set and the remaining 316 subjects as the test set. The single-shot evaluation approach was adopted, and the test set was partitioned into a gallery set and a probe set. By calculating the average CMC curves and estimating the correct match in the top n matches, we obtained a ranking for each example in the gallery set with respect to the probe set. This procedure was repeated 10 times in the presented CMC curves.

Seven representative metric learning approaches were compared to our algorithm: Euclidean distance (L2), Mahalanobis metric (MM), PolyMap [2], LADF [22], PRDC [54], PCCA [36], and XQDA [26]. The L2 and MM distances are two baseline algorithms that have been used in many published person re-identification studies. PolyMap, LADF, PRDC, PCCA, and XQDA are current state-of-the-art metric learning algorithms. PCA was first conducted to reduce feature dimensionality; thus, learning was accelerated and signal noise reduced. DR-KISS outperformed the other metric learning approaches (Figure 3), indicating that DR-KISS successfully

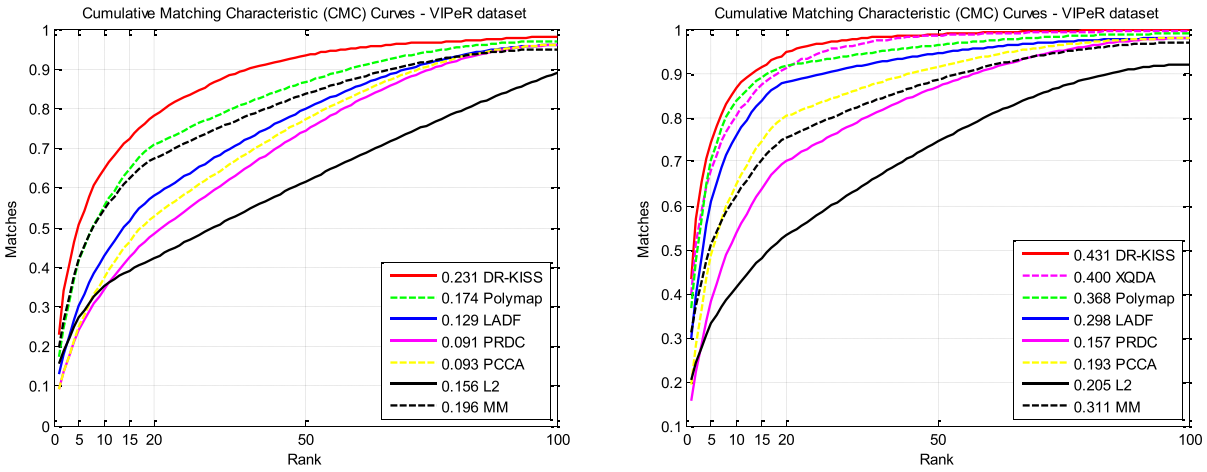

Fig. 3. Performance comparison using CMC curves. In each subfigure, the x coordinate is the rank score and y coordinate is the corresponding matching rate. We compare DR-KISS with Polymap, LADF, PRDC, PCCA, L2, and MM on the VIPeR dataset. Only the top 100 ranking positions are depicted. The compared results are from [26]. $p = 100$. (b) $p = 316$.

TABLE III

A COMPARISON OF TYPICAL ALGORITHMS AND DR-KISS ON PERSON RE-IDENTIFICATION MATCHING RATES ON THE VIPeR DATASET. SOME RESULTS ARE FROM [26]

	p=100				p=316			
RANK	1	5	10	20	1	5	10	20
DR-KISS	**0.231**	**0.507**	**0.647**	**0.782**	**0.431**	**0.744**	**0.868**	**0.948**
XQDA [26]	-	-	-	-	0.400	0.681	0.805	0.911
PolyMap [2]	0.174	0.416	0.553	0.708	0.368	0.704	0.837	0.917
LADF [22]	0.129	0.303	0.427	0.580	0.298	0.611	0.759	0.881
PRDC [54]	0.091	0.242	0.344	0.485	0.157	0.384	0.539	0.701
PCCA [36]	0.093	0.249	0.374	0.529	0.193	0.489	0.649	0.803
L2	0.156	0.272	0.351	0.423	0.205	0.332	0.415	0.532
MM	0.196	0.421	0.545	0.674	0.311	0.512	0.623	0.754

TABLE IV

AVERAGE TRAINING TIME

Size of training set	DR-KISS	RS-KISS	KISS
p = 316	1.76s	1.1ms	1.2ms

learns an effective metric. In each subfigure, the x coordinate represents the rank score and y coordinate represents the corresponding matching rate. Only the top 100 ranking positions are depicted. The performance of all the methods according to the first 20 ranks is shown in Table III, our proposed algorithm achieves new state-of-the-art performance, with 23.10% and 43.10% rank-1 accuracy for p = 100 and p = 316, respectively, where p is the training subject. We selected p = 316 on VIPeR dataset and applied DR-KISS, RS-KISS [43] and KISS to update the metric, respectively. Table IV reports mean time of the experiment. We conduct all experiments on an i5-2500K 3.30 GHz computer with a 8-GB memory.

B. Experiments on GRID

The GRID dataset [30] contains 250 pedestrian image pairs; there are two images of the same individual from different camera angles. In addition, 775 individual images are included to enlarge the gallery set; specifically, these images do not belong to the primary pedestrian set. Images were captured from eight disjointed camera views in a bustling underground station. There are snapshots of figures from each camera view, with sample images selected to make up the final dataset. GRID is a challenging dataset due to variations in color, pose, lighting, and image quality (caused by low resolution). Example images are shown in Figure 4.

We randomly selected 125 image pairs for training, with the remaining 125 image pairs and 775 extra background images used for testing. The single-shot evaluation protocol was adopted, and the test set was partitioned into a gallery set and a probe set consisting of the 775 extra background images. As above, CMC curves were calculated.

We compared six representative and state-of-the-art metric learning approaches with our method (Figure 5), namely LCRML [3], PolyMap [2], MRank-RankSVM [29], PRDC [54], MtMCML [33], and XQDA [26]. PCA was again first conducted to reduce feature dimensionality. DR-KISS metric learning outperformed the other metric learning approaches, achieving new state-of-the-art performance (Table V): 20.60% rank-1 accuracy for $p = 125$, where p is the training subject.

Fig. 4. Samples from the GRID dataset. Each column depicts the same person samples from different camera views in the first two rows. The third row depicts extra background samples. Minor variations in pose, viewpoint, image quality, illumination, and shooting location are noted.

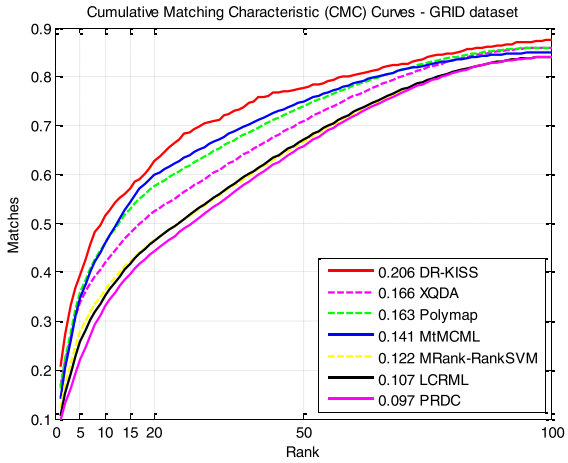

Fig. 5. CMC curves for performance comparison. The x coordinate is the rank score and the y coordinate is the corresponding matching rate. We compared DR-KISS with XQDA, Polymap, MtMCML, MRank-RankSVM, LCRML, and PRDC on the GRID dataset. The top 100 ranking positions are depicted. Some results are from [26].

C. Experiments on CUHK 01

The CUHK 01 dataset (or CUHK Campus dataset) [24] contains 971 individuals captured by two disjointed camera views in a campus environment. Specifically, camera view A captured the back or frontal view of individuals, while camera view B captured the side views. Four images of each individual were partitioned equally in camera views A and B. Compared to the other datasets, CUHK 01 contains higher resolution images.

We randomly selected 485 image pairs for training, and the remaining 486 image pairs were used for testing. There are four images of the same person for the multi-shot evaluation (the same as used in the other methods), and the multi-shot evaluation was repeated 10 times.

TABLE V

PERSON RE-IDENTIFICATION MATCHING RATES ON THE GRID DATASET COMPARED TO OTHER POPULAR ALGORITHMS. SOME RESULTS ARE TAKEN DIRECTLY FROM [26]

	p=125			
RANK	1	5	10	20
DR-KISS	**0.206**	**0.393**	**0.514**	**0.626**
XQDA [26]	0.166	0.338	0.418	0.524
Polymap [2]	0.163	0.358	0.460	0.576
MtMCML [33]	0.141	0.346	0.458	0.598
MRank-RankSVM [29]	0.122	0.278	0.363	0.466
LCRML [3]	0.107	0.258	0.350	0.465
PRDC [54]	0.097	0.220	0.329	0.443

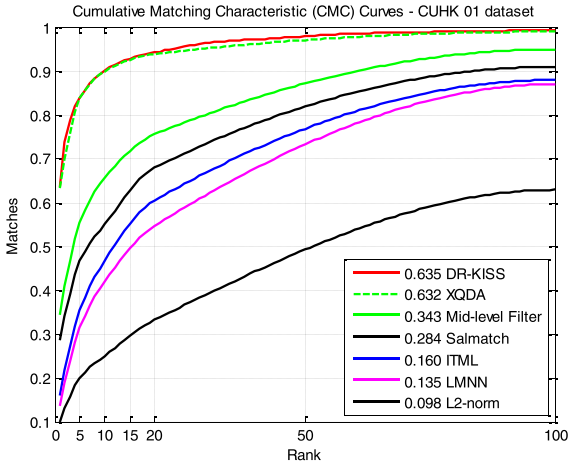

Fig. 6. Performance comparison using CMC curve. The x coordinate is the rank score and the y coordinate is the corresponding matching rate. We compared DR-KISS with XQDA, the mid-level filter, Salmatch, ITML, LMNN, and L2-norm on the CUHK 01 dataset. The top 100 ranking positions are depicted. The compared results are from [26].

The CMC curves were calculated for six comparative state-of-the-art metric learning approaches to evaluate the effectiveness of our algorithm: the mid-level Filter [51], Salmatch [50], ITML [5], LMNN [47], L2-norm, and XQDA [26]. As before, PCA was first conducted to reduce the feature dimensionality to accelerate learning and reduce noise.

DR-KISS once again outperformed the other metric learning approaches (Figure 6) to achieve a new state-of-the-art performance: 63.5% rank-1 accuracy for p = 485, where p is the training subject.

D. Experimental Results and Analysis

Figures 7 to 9 illustrate the impact of the intrapersonal parameter γ_I and the extrapersonal parameter γ_E on the results using the VIPeR and GRID datasets, which suggest how to choose the values of γ_I and γ_E in practice. Figure 7 and 8 illustrate the impact of the intrapersonal parameter γ_I and the extrapersonal parameter γ_E on the results using the VIPeR dataset with 100 and 316 training image pairs, respectively, while Figure 9 illustrates the impact of the intrapersonal parameter γ_I and the extrapersonal parameter γ_E on the GRID dataset with 125 training image pairs. When the intrapersonal parameter γ_I is set at 0, our model does underperform because

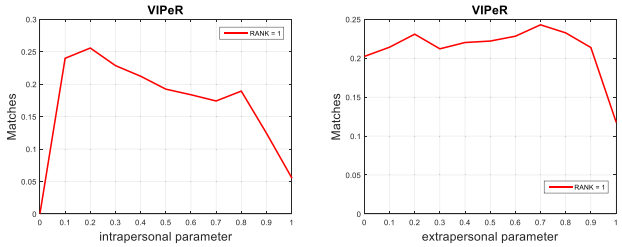

Fig. 7. Rank-1 matching rates at varying intrapersonal parameter γ_I and extrapersonal parameter γ_E on the VIPeR dataset with 100 training image pairs.

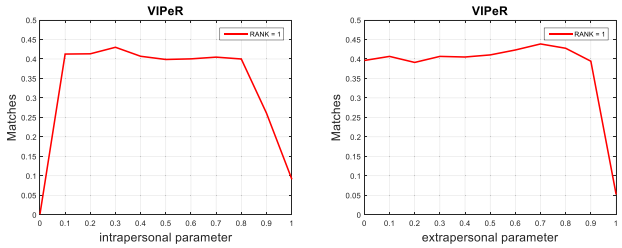

Fig. 8. Rank-1 matching rate at varying intrapersonal parameter γ_I and extrapersonal parameter γ_E on the VIPeR dataset with 316 training image pairs.

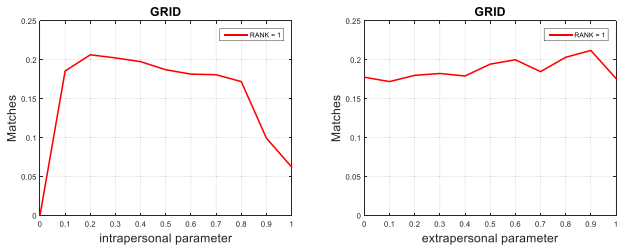

Fig. 9. Rank-1 matching rate at varying intrapersonal parameter γ_I and extrapersonal parameter γ_E on the GRID dataset with 125 training image pairs.

the corresponding estimated extrapersonal covariance matrix is reversible, resulting in an ill-posed problem. However, in most cases, DR-KISS is more robust to the extrapersonal parameter γ_E because more extrapersonal samples constitute the extrapersonal covariance matrix. When γ_I and γ_E are set at 1 or close to 1, our model performs less weakly, this is because each estimated covariance matrix has the same eigenvalues, resulting in an overfitting problem.

The main observations from our experiments can be summarized as follows:

1) DR-KISS metric learning improves KISS when the training set is small because the two KISS covariance matrices are seriously biased.
2) In most cases, introducing the regularization technique to the two covariance matrices results in DR-KISS being more robust to the extrapersonal parameter γ_E because more extrapersonal samples constitute the extrapersonal covariance matrix.

V. Conclusions

Metric learning is extremely important for person re-identification in video surveillance. Thus, a robust distance metric learning algorithm is required to improve person re-identification performance. A number of distance metric learning methods have been developed over recent years including LMNN and ITML, but they are not perfectly suited to person re-identification because the training set for learning a metric is usually very limited in practice. Although KISS performs well, it shares this problem. Some newer distance metric learning methods such as XQDA and Polymap improve person re-identification but there remains room for improvement.

In this paper, we proposed dual-regularized KISS (DR-KISS), which exploits regularization to suppress the effect of the large eigenvalues in the two estimated covariance matrices. Therefore, DR-KISS significantly improves KISS for person re-identification. Experiments on three challenging person re-identification datasets - VIPeR, GRID, and CUHK 01-fully demonstrate that our proposed approach improves on current state-of-the-art methods. However, the fixed-size feature for representing a pedestrian image [18] is essential for the success of DR-KISS for person re-identification. With this regard, we cannot straightforwardly integrate DR-KISS with other combination features concatenated by not-fixed or un-ordered size sets of different local features [11].

Appendix

We provide below the detailed proofs of the two theorems in section III. Before proving the theorem, we introduce lemma 1 and the McDiarmid's inequality [35], also known as the bounded-difference inequality.

Lemma 1: If A and B are positive semidefinite matrices, then $trace(AB) \leq trace(A)trace(B)$. If A is a positive semidefinite matrix, then $trace\,(AB) \leq \lambda_{\max}\,(B)\,trace\,(A)$. If A is a square matrix, then $trace(A) = \sum \lambda\,(A)$.

Theorem 3 (McDiarmid's Inequality): Let $X = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ be an independent distributed observations and X^i an another set of observations with the i^{th} example in X being replace by an independent and identically distributed observation x_1'. If there exists $c_1, \ldots, c_n > 0$ such that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the following conditions:

$$\left| f\,(X) - f\left(X^i\right) \right| \leq c_i, \quad \forall i \in \{1, \ldots, n\}.$$

Then, for any $X \in \mathcal{X}^n$ and $\epsilon > 0$, the following inequalities hold:

$$\Pr\left\{f\,(X) - E_x\left[f\,(X)\right] \geq \epsilon\right\} \leq \exp\left\{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right\},$$

$$\Pr\left\{E_x\left[f\,(X)\right] - f\,(X) \geq \epsilon\right\} \leq \exp\left\{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right\},$$

where $\Pr\{A\}$ denotes the probability of the event A.

We now can prove Theorem 1 and Theorem 2 as follows.

Proof of Theorem 1:

Let

$$\Phi\,(\Omega_I) = \sup_{\Sigma}\left(E_{\Omega_I}\left[\frac{1}{N_I}\sum_{y_{ij}=0}\ell\left(x_i, x_j, \Sigma\right)\right] - \frac{1}{N_I}\sum_{y_{ij}=0}\ell\left(x_i, x_j, \Sigma\right)\right).$$

We have

$$\begin{aligned}
&\left|\Phi\left(\Omega_I\right)-\Phi\left(\Omega_I^i\right)\right| \\
&=\sup_{\Sigma}\left(\frac{1}{N_I}\sum_{x-y\in\Omega_I}\ell\left(x,y,\Sigma\right)-\frac{1}{N_I}\sum_{x-y\in\Omega_I^i}\ell\left(x,y,\Sigma\right)\right) \\
&=\sup_{\Sigma}\left(\frac{1}{N_I}\sum_{x-y\in\Omega_I,x=x_i}\ell\left(x,y,\Sigma\right)\right. \\
&\qquad\left.-\frac{1}{N_I}\sum_{x-y\in\Omega_I^i,x=x_i'}\ell\left(x,y,\Sigma\right)\right) \\
&\leq\frac{1}{N_I}\left(n_I-1\right)\sup_{x,y,\Sigma}\ell\left(x_i,y,\Sigma\right)\leq\frac{2M^2}{n_I\times n_E}.
\end{aligned}$$

The first inequality holds because we have $N_I=|\Omega_I|=n_E\times n_I\times(n_I-1)$. The last equality holds because have assumed that $\forall x,y,\ (x-y)(x-y)_F^T\leq M$ Then, $\|\Sigma\|_F\leq M$ and $\ell\left(x_i,x_j,\Sigma\right)=\left\|\left(x_i-x_j\right)\left(x_i-x_j\right)^T-\Sigma\right\|_F^2\leq 2M^2$.

Note that the set Ω_{I} contains $n_E\times n_I\times(n_I-1)$ pairs but $n_E\times n_I$ of them are independent variables. (It is reasonable to assume that the observations for the same person are independent and identically distributed.) According to McDiarmid's inequality, for any given $\delta>0$, let $\delta=\exp\left\{\frac{-2\varepsilon^2}{\sum_{i=1}^{n_E\times n_I}c_i^2}\right\}$. Then, with probability at least $1-\delta$, we have

$$\begin{aligned}
\Phi\left(\Omega_I\right)&=\sup_{\Sigma}\left(E_{\Omega_I}\left[\frac{1}{N_I}\sum_{y_{ij}=0}\ell\left(x_i,x_j,\Sigma\right)\right]\right. \\
&\qquad\left.-\frac{1}{N_I}\sum_{y_{ij}=0}\ell\left(x_i,x_j,\Sigma\right)\right) \\
&\leq E_x\left[\Phi\left(\Omega_I\right)\right]+M\sqrt{\frac{2\log\frac{1}{\delta}}{n_E\times n_I}}.
\end{aligned}$$

We now upper bound $E_x\left[\Phi\left(\Omega_I\right)\right]$ as follows:

$$\begin{aligned}
E_{\Omega_I}\left[\Phi\left(\Omega_I\right)\right]&=E_{\Omega_I}\left[\sup_{\Sigma}\left(\frac{1}{N_I}\sum_{y_{ij}=0}\ell\left(x_i,x_j,\Sigma\right)\right.\right. \\
&\qquad\left.\left.-E_{\Omega_I}\left[\frac{1}{N_I}\sum_{y_{ij}=0}\ell\left(x_i,x_j,\Sigma\right)\right]\right)\right] \\
&=E_{\Omega_I}\left[\sup_{\Sigma}\left(\frac{1}{2N_I}\sum_{x-y\in\Omega_I}\ell\left(x,y,\Sigma\right)\right.\right. \\
&\qquad\left.\left.-E_{\Omega_I'}\left[\frac{1}{2N_I}\sum_{x'-y'\in\Omega_I'}\ell\left(x',y',\Sigma\right)\right]\right)\right] \\
&\leq E_{\Omega_I,\Omega_I'}\left[\sup_{\Sigma}\left(\frac{1}{2N_I}\sum_{x-y\in\Omega_I}\ell\left(x,y,\Sigma\right)\right.\right. \\
&\qquad\left.\left.-\frac{1}{2N_I}\sum_{x'-y'\in\Omega_I'}\ell\left(x',y',\Sigma\right)\right)\right],
\end{aligned}$$

where Ω_I' and Ω_I contains $n_E\times n_I$ independent and identically distributed samples.

Since $\sum_{x-y'\in\Omega_I,x=x_i'}\ell\left(x,y',\Sigma\right)$ and $\sum_{x-y\in\Omega_I,x=x_i}\ell\left(x,y,\Sigma\right)$ are independent and identically variables, $\sum_{y'\in\Omega_I\setminus x_i'}\ell\left(x_i',y',\Sigma\right)-\sum_{y\in\Omega_I\setminus x_i}\ell\left(x_i,y,\Sigma\right)$, $i=1,\ldots,n_I$ are independent and identically symmetric variables (note that if α is a symmetric variable, then $\Pr\{\alpha>0\}=Pr\{\alpha<0\}$ and thus $E_\alpha[a]=E_{\sigma,\alpha}[\sigma a]$, where σ is the Rademacher variable defined by that $\Pr\{\sigma=1\}=\Pr\{\sigma=-1\}=0.5$). Then, we have

$$\begin{aligned}
&E_{\Omega_I}\left[\Phi\left(\Omega_I\right)\right] \\
&\leq E_{\Omega_I,\Omega_I',\sigma}\left[\sup_{\Sigma}\left(\frac{1}{N_I}\sum_{i=1}^{n_En_I}\sigma_i\left(\sum_{x-y'\in\Omega_I,x=x_i'}\ell\left(x,y',\Sigma\right)\right.\right.\right. \\
&\qquad\left.\left.\left.-\sum_{x-y\in\Omega_I,x=x_i}\ell\left(x,y,\Sigma\right)\right)\right)\right] \\
&=2E_{\Omega_I,\sigma}\left[\sup_{\Sigma}\frac{1}{n_En_I\left(n_I-1\right)}\sum_{i=1}^{n_En_I}\sigma_i\sum_{x-y\in\Omega_I,x=x_i}\ell\left(x,y,\Sigma\right)\right].
\end{aligned}$$

Since the derivative of $\ell\left(x,y,\Sigma\right)$ is $2\left(x-y\right)\left(x-y\right)^T-2\Sigma$ and the largest eigenvalue of $2\left(x-y\right)\left(x-y\right)^T-2\Sigma$ is no more than $4M$, the loss function $\ell\left(x,y,\Sigma\right)$ is $4M$-Lipschitz continuous. Using the Talagrand contraction Lemma, we have

$$\begin{aligned}
&E_{\Omega_I}\left[\Phi\left(\Omega_I\right)\right] \\
&\leq 2E_{\Omega_I,\sigma}\left[\sup_{\Sigma}\frac{1}{n_En_I\left(n_I-1\right)}\sum_{i=1}^{n_En_I}\sigma_i\sum_{x-y\in\Omega_I,x=x_i}\ell\left(x,y,\Sigma\right)\right] \\
&=2E_{\Omega_I,\sigma}\left[\sup_{\Sigma}\frac{1}{n_En_I\left(n_I-1\right)}\right. \\
&\qquad\left.\times\sum_{i=1}^{n_En_I}\sigma_i\sum_{x-y\in\Omega_I,x=x_i}\mathrm{Tr}\left(\ell\left(x,y,\Sigma\right)\right)\right] \\
&\leq 8ME_{\Omega_I,\sigma}\sup_{\Sigma}\frac{1}{n_En_I\left(n_I-1\right)} \\
&\quad\times\sum_{i=1}^{n_En_I}\sigma_i\sum_{x-y\in\Omega_I,x=x_i}\mathrm{Tr}\left(\left(x-y\right)\left(x-y\right)^T-\Sigma\right) \\
&=8ME_\sigma\frac{1}{n_En_I\left(n_I-1\right)}\sum_{i=1}^{n_En_I}\sigma_i\sum_{x-y\in\Omega_I,x=x_i}\mathrm{Tr}\left(\Sigma\right) \\
&=8ME_\sigma\frac{1}{n_En_I}\sum_{i=1}^{n_En_I}\sigma_i\mathrm{Tr}\left(\Sigma\right) \\
&=8ME_\sigma\frac{1}{n_En_I}\sqrt{\left(\sum_{i=1}^{n_En_I}\sigma_i\mathrm{Tr}\left(\Sigma\right)\right)^2} \\
&\leq 8M\frac{1}{n_En_I}\sqrt{E_\sigma\left(\sum_{i=1}^{n_En_I}\sigma_i\mathrm{Tr}\left(\Sigma\right)\right)^2} \\
&=8M\frac{1}{n_En_I}\sqrt{\sum_{i=1}^{n_En_I}\left(\mathrm{Tr}\left(\Sigma\right)\right)^2}=\frac{8M}{\sqrt{n_En_I}}\mathrm{Tr}\left(\Sigma\right),
\end{aligned}$$

where $\mathrm{Tr}\left(\Sigma\right)$ denotes the trace of Σ.

The same proof method applies to the generalization bound for learning Σ_E, which ends the proof of Theorem 1.

Proof of Theorem 2:

We have

$$\begin{aligned}
&|d(x,z)-d(y,z)| \\
&= \left|(x-z)^T\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)(x-z)\right. \\
&\quad \left.-(y-z)^T\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)(y-z)\right| \\
&= \left|trace\left((x-z)^T\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)(x-z)\right)\right. \\
&\quad \left.-trace\left((y-z)^T\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)(y-z)\right)\right| \\
&\quad \left|trace\left(\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)\right.\right. \\
&\quad \left.\left.\left((x-z)(x-z)^T-(y-z)(y-z)^T\right)\right)\right|,
\end{aligned}$$

and

$$\begin{aligned}
&(x-z)(x-z)^T-(y-z)(y-z)^T \\
&\quad = xx^T-yy^T-2z^Tx+2z^Ty \\
&\quad = (x-y)\left(x^T+y^T-2z^T\right).
\end{aligned}$$

Then,

$$\begin{aligned}
&|d(x,z)-d(y,z)| \\
&\quad = \left|trace\left(\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)(x-y)\right.\right. \\
&\qquad \left.\left.\times\left(x^T+y^T-2z^T\right)\right)\right| \\
&\quad = \left|trace\left(\left(x^T+y^T-2z^T\right)\right.\right. \\
&\qquad \left.\left.\times\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)(x-y)\right)\right|.
\end{aligned}$$

Using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&|d(x,z)-d(y,z)| \\
&\le \left|trace\left(\left\|\left(x^T+y^T-2z^T\right)\right.\right.\right. \\
&\quad \left.\left.\left.\times\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)\right\|\|(x-y)\|\right)\right| \\
&= \left\|\left(x^T+y^T-2z^T\right)\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)\right\| \\
&\quad \times\|(x-y)\| \\
&\le \eta\left\|\left(x^T+y^T-2z^T\right)\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)\right\| \\
&= \eta\sqrt{trace\left((x+y-2z)(x+y-2z)^T\Delta\Delta^T\right)},
\end{aligned}$$

where $\Delta \triangleq \left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}$

According to lemma 1, we have

$$\begin{aligned}
&|d(x,z)-d(y,z)| \\
&\quad \le \eta\sqrt{trace\left((x+y-2z)(x+y-2z)^T\right)trace\left(\Delta\Delta^T\right)} \\
&\quad = \eta\sqrt{trace\left((x+y-2z)^T(x+y-2z)\right)trace\left(\Delta^T\Delta\right)} \\
&\quad \le \eta\sqrt{(\|x\|+\|y\|+\|2z\|)^2\,trace\left(\Delta^T\Delta\right)} \\
&\quad \le 2\Lambda_X\eta\sqrt{trace\left(\Delta^T\Delta\right)}.
\end{aligned}$$

and

$$\begin{aligned}
&trace\left(\Delta^T\Delta\right) \\
&\quad = trace\left(\left(\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)^T\right.\right. \\
&\qquad \left.\left.\times\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right)\right) \\
&\quad = \left\|\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}-\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right\|^2 \\
&\quad \le \left(\left\|\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}\right\|-\left\|\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right\|\right)^2 \\
&\quad \le \left\|\left(\hat{\mathbf{\Sigma}}_I+\lambda_I\right)^{-1}\right\|^2+\left\|\left(\hat{\mathbf{\Sigma}}_E+\lambda_E\right)^{-1}\right\|^2 \\
&\quad = trace\left(\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I I\right)^{-1}\right)^T\left(\hat{\mathbf{\Sigma}}_I+\lambda_I I\right)^{-1}\right) \\
&\qquad +trace\left(\left(\left(\hat{\mathbf{\Sigma}}_E+\lambda_E I\right)^{-1}\right)^T\left(\hat{\mathbf{\Sigma}}_E+\lambda_E I\right)^{-1}\right) \\
&\quad = trace\left(\left(\left(\hat{\mathbf{\Sigma}}_I+\lambda_I I\right)^2\right)^{-1}\right) \\
&\qquad +trace\left(\left(\left(\hat{\mathbf{\Sigma}}_E+\lambda_E I\right)^2\right)^{-1}\right) \\
&\quad = \sum_{i=1}^{n}\frac{1}{\left((\Lambda_I)_{ii}+\lambda_I\right)^2}+\sum_{i=1}^{n}\frac{1}{\left((\Lambda_E)_{ii}+\lambda_E\right)^2}
\end{aligned}$$

Hence,

$$\begin{aligned}
&|d(x,z)-d(y,z)| \\
&\quad \le 2\Lambda_X\eta\sqrt{\sum_{i=1}^{n}\frac{1}{\left((\Lambda_I)_{ii}+\lambda_I\right)^2}+\sum_{i=1}^{n}\frac{1}{\left((\Lambda_E)_{ii}+\lambda_E\right)^2}}
\end{aligned}$$

The proof ends.

REFERENCES

[1] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14. 2002, pp. 585–591.

[2] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CCPR)*, Jun. 2015, pp. 1565–1573.

[3] J. Chen, Z. Zhang, and Y. Wang, "Relevance metric learning for person re-identification by exploiting global similarities," in *Proc. 22nd Int. Conf. Pattern Recognit. (ICPR)*, 2014, pp. 1657–1662.

[4] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 1–11.

[5] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, 2007, pp. 209–216.

[6] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2011, pp. 501–512.

[7] S. R. Dubey, S. K. Singh, and R. K. Singh, "Local wavelet pattern: A new feature descriptor for image retrieval in medical CT databases," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5892–5903, Dec. 2015.

[8] M. Faraji, J. Shanbehzadeh, K. Nasrollahi, and T. B. Moeslund, "Extremal regions detection guided by maxima of gradient magnitude," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5401–5415, Dec. 2015.

[9] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[10] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[11] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 1528–1535.

[12] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, 2007, p. 5.

[13] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Marseille, France, 2008, pp. 262–275.

[14] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," in *Proc. Int. Conf. Distrib. Smart Cameras (ICDSC)*, 2008, pp. 1–6.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2001.

[16] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16. 2003, pp. 153–160.

[17] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 115–132.

[18] M. Hirzer, P. Roth, M. Koestinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 780–793.

[19] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Edu. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.

[20] W. Jiang, E. Zavesky, S.-F. Chang, and A. Loui, "Cross-domain learning methods for high-level visual concept classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2008, pp. 161–164.

[21] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2288–2295.

[22] J.-E. Lee, R. Jin, and A. K. Jain, "Rank-based distance metric learning: An application to image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[23] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3610–3617.

[24] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3594–3601.

[25] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1301–1306.

[26] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[27] T. Liu and D. Tao, "On the performance of Manhattan nonnegative matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2015.2458986.

[28] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 447–461, Mar. 2016, doi: 10.1109/TPAMI.2015.2456899.

[29] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *Proc. Int. Conf. Image Process. (ICIP)*, 2013, pp. 3567–3571.

[30] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1988–1995.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[32] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by Fisher vectors for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 413–422.

[33] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3656–3670, Aug. 2014.

[34] N. Martinel, C. Micheloni, and G. L. Foresti, "Kernelized saliency-based person re-identification through multiple metric learning," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5645–5658, Dec. 2015.

[35] C. McDiarmid, "Concentration," in *Probabilistic Methods for Algorithmic Discrete Mathematics*, M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, Eds. Berlin, Germany: Springer-Verlag, 1998, pp. 195–248.

[36] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2666–2672.

[37] B. Morago, G. Bui, and Y. Duan, "An ensemble approach to image matching using contextual features," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4474–4487, Nov. 2015.

[38] T. Mouats, N. Aouf, and M. A. Richardson, "A novel image representation via local frequency analysis for illumination invariant stereo matching," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2685–2700, Sep. 2015.

[39] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[40] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2010, pp. 1–11.

[41] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[42] D. Tao, L. Jin, Y. Wang, and X. Li, "Person reidentification by minimum classification error-based kiss metric learning," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 242–252, Feb. 2015.

[43] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing KISS metric learning," in *Proc. IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1675–1685, Oct. 2013.

[44] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.

[45] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[46] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Comput. Surv.*, vol. 46, no. 2, 2013, Art. no. 29.

[47] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[48] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, "Distance metric learning, with application to clustering with side-information," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15. 2002, pp. 505–512.

[49] H. Xu and S. Mannor, "Robustness and generalization," *Mach. Learn.*, vol. 86, pp. 391–423, Nov. 2011.

[50] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 2528–2535.

[51] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 144–151.

[52] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Hyperspectral remote sensing image subpixel target detection based on supervised metric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4955–4965, Aug. 2014.

[53] L. Zhang, Q. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding," *Pattern Recognit.*, vol. 48, no. 10, pp. 3102–3112, 2015.

[54] W. S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2012.

Dapeng Tao received the B.E. degree from Northwestern Polytechnical University, and the Ph.D. degree from the South China University of Technology. He is currently with the School of Information Science and Engineering, Yunnan University, Kunming, China, as a Professor. He has authored or co-authored over 30 scientific articles. He has served more than ten international journals, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE SIGNAL PROCESSING LETTERS, and *Information Sciences*. Over the past years, his research interests include machine learning, computer vision, and robotics.

Yanan Guo received the B.Eng. degree from Hubei Polytechnic University. She is currently pursuing the M.S. degree with Yunnan University, Kunming, China. Her research interests include machine learning and computer vision.

Mingli Song (M'06–SM'13) received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2006. He is currently a Professor with the Microsoft Visual Perception Laboratory, Zhejiang University. He has authored or co-authored over 70 scientific articles at top venues, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, *Information Sciences*, *Pattern Recognition*, the Computer Vision and Pattern Recognition Conference, the European Conference on Computer Vision, and ACM Multimedia. His research interests include pattern classification, weakly supervised clustering, color and texture analysis, object recognition, and reconstruction. He received the Microsoft Research Fellowship Award in 2004. He is also an Associate Editor of *Information Sciences*, *Neurocomputing*, and the *Journal of Visual Communication and Image Representation*.

Yaotang Li is currently a Professor of Mathematics and Statistics with Yunnan University, China. He has authored or co-authored over 80 research papers. His main research interests include numerical algebra and special matrices.

Zhengtao Yu received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His main research interests include natural language process, image processing, and machine learning.

Yuan Yan Tang (F'04) is currently a Chair Professor with the Faculty of Science and Technology, University of Macau, and a Professor/Adjunct Professor/Honorary Professor with several institutes in China, USA, Canada, France, and Hong Kong. He has authored over 400 academic papers, and has authored or co-authored over 25 monographs/books/bookchapters. His current interests include wavelets, pattern recognition, image processing, and artificial intelligence. He is a fellow of IAPR.

He is the Founder and Editor-in-Chief of the *International Journal on Wavelets, Multiresolution, and Information Processing*, and Associate Editor of several international journals. He is the Founder and Chair of the pattern recognition committee in the IEEE SMC. He has served as the General Chair, Program Chair, and Committee Member for many international conferences. He is the Founder and General Chair of the series International Conferences on Wavelets Analysis and Pattern Recognition. He is the Founder and Chair of the Macau Branch of the International Associate of Pattern Recognition (IAPR).