# Relevance Metric Learning for Person Re-Identification by Exploiting Listwise Similarities

Jiaxin Chen, Zhaoxiang Zhang, *Senior Member, IEEE*, and Yunhong Wang, *Member, IEEE*

***Abstract*— Person re-identification aims to match people across non-overlapping camera views, which is an important but challenging task in video surveillance. In order to obtain a robust metric for matching, metric learning has been introduced recently. Most existing works focus on seeking a Mahalanobis distance by employing sparse pairwise constraints, which utilize image pairs with the same person identity as positive samples, and select a small portion of those with different identities as negative samples. However, this training strategy has abandoned a large amount of discriminative information, and ignored the relative similarities. In this paper, we propose a novel relevance metric learning method with listwise constraints (RMLLCs) by adopting listwise similarities, which consist of the similarity list of each image with respect to all remaining images. By virtue of listwise similarities, RMLLC could capture all pairwise similarities, and consequently learn a more discriminative metric by enforcing the metric to conserve predefined similarity lists in a low-dimensional projection subspace. Despite the performance enhancement, RMLLC using predefined similarity lists fails to capture the relative relevance information, which is often unavailable in practice. To address this problem, we further introduce a rectification term to automatically exploit the relative similarities, and develop an efficient alternating iterative algorithm to jointly learn the optimal metric and the rectification term. Extensive experiments on four publicly available benchmarking data sets are carried out and demonstrate that the proposed method is significantly superior to the state-of-the-art approaches. The results also show that the introduction of the rectification term could further boost the performance of RMLLC.**

***Index Terms*— Person re-identification, metric learning, list-wise similarities, alternating iterative optimization.**

Manuscript received October 24, 2014; revised March 27, 2015 and June 18, 2015; accepted July 23, 2015. Date of publication August 7, 2015; date of current version September 18, 2015. This work was supported in part by the Hong Kong, Macao, and Taiwan Science and Technology Cooperation Program of China under Grant L2015TGA9004, in part by the Foundation for Innovative Research Groups through the National Natural Science Foundation of China under Grant 61421003 and Grant 61375036, in part by the Beijing Natural Science Foundation under Grant 4132064, in part by the Program for New Century Excellent Talents in University, in part by the Beijing Higher Education Young Elite Teacher Project, and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianfei Cai. *(Corresponding author: Zhaoxiang Zhang.)*

The authors are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: chenjiaxinx@gmail.com; zxzhang@buaa.edu.cn; yhwang@buaa.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.2015.2466117

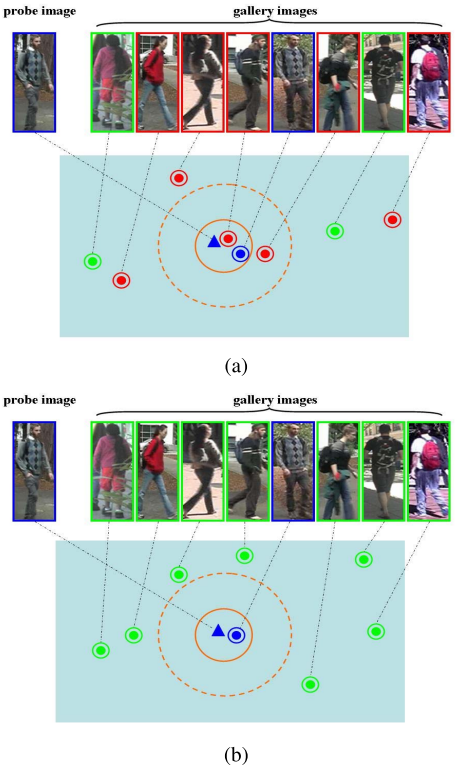



Fig. 1. Samples from the VIPeR dataset [14]: (a) shows metric learning with pairwise constraints, and (b) demonstrates metric learning with listwise constaints.

## I. Introduction

GIVEN one single shot or multiple shots of a target pedestrian captured by one camera, person re-identification aims to identify the target among a set of gallery candidates obtained from non-overlapping camera networks. This task is critical in surveillance applications such as tracking, person identification, re-acquisition and verification [53]. However, it is challenging due to the large within-class dissimilarity and small between-class dissimilarity, caused by unconstrained variations such as viewpoint, illumination, occlusion, pose, background clutter, intrinsic camera parameters (see Fig. 1).

In order to handle these variations, extensive work has been done, which could be roughly divided into three categories: feature-based methods concentrating on developing robust feature representations; model-based methods focusing on

developing models to directly tackle changes of illumination, pose, view point and etc.; metric learning based methods focusing on constructing discriminative (dis)similarity measures. Due to its promising performance for solving the person re-identification problem, metric learning based methods have drawn a lot of research interests recently. In this paper, we will mainly focus on this kind of approaches.

State-of-the-art metric learning methods [25], [29], [39], [41], [44], [63] focus on learning a Mahalanobis distance, by which the (dis)similarity of an image pair is measured by their distance, i.e., pairs with smaller distance are considered to be more similar. Different tactics have been proposed in order to learn a discriminative Mahalanobis distance, almost all of which take two common strategies. Firstly, the Mahalanobis distance is always employed as the final similarity metric. Secondly, pairwise constraints are often adopted for training: image pairs of the same person should have smaller Mahalanobis distances than those belonging to different persons.

However, these two strategies may yield two potential shortcomings. As to the first strategy, the Mahalanobis distance is intrinsically an Euclidean distance in a mapped space after an affine transformation, and it ignores other useful geometry information for measuring similarities, such as the angle between two vectors. Actually, the Mahalanobis distance is more sensitive to outliers than other distance such as the cosine metric.

As to the second strategy, for a set of gallery images and a probe image, pairwise constraints contain binary (dis)similarity labels: image pairs belonging to the same person are labeled "similar" and used as positive samples, while image pairs belonging to different persons are labeled "dissimilar" and treated as negative samples. These binary labels only reflect incomplete similarities, since it is unable to provide relative similarity information: images from a group of people are more similar to the probe image than images from another group. Considering the situation where two different pedestrians have similar appearances (e.g., they have similar dresses), they should be labeled "similar", or at least be treated to be more similar than other pairs. If we use the pairwise constraints, and assign them "dissimilar" labels, the generalization performance of the learnt distance would be deteriorated. Meanwhile, in the testing phase, person re-identification is often casted as an retrieval problem by comparing the probe image with the whole gallery set, and return the $r$ most similar ones. Therefore, it seems to be unreasonable to train distance functions by using pairwise constraints. In contrast, the listwise similarity of an image, consisting of its similarities to the remaining images, could capture all available similarity information and reflect relative relevance. Additionally, it treats the similarity information of a given image in totality, and therefore sounds more reasonable to be used to train a metric for the person re-identification application.

Moreover, existing metric learning approaches, such as KISSME [20], PCCA [41], PRDC [63], LFDA [44], MtMCML [39], generally utilize sparse pairwise constraints, yielding a loss of useful discriminative information. As shown in Fig. 1 (a), the leftmost image is the probe, while the rest images constitute the gallery set. Most existing methods construct the similar image pair by selecting the probe image and the image with the same identity in the gallery set (with blue box). Generally, there are totally $(N-1)$ dissimilar image pairs with respect to a single probe image. However, existing work sparsely samples $k$ (in Fig.1 (a), $k=2$) dissimilar pairs (with green boxes). The rest $(N-k-1)$ dissimilar images are then simply abandoned. After training, similar images (denoted by the blue triangular and blue circle) departs from the selected dissimilar images (denoted by green circles) with a large margin (denoted by the orange circle). While the unselected dissimilar images (denoted by red circles) scatter without control, some of which may locate closely to the probe image, and therefore deteriorate the generative performance of the learnt metric. Conversely, listwise similarities contain similarity information of any image pair in the training data. As shown in Fig. 1 (b), by forcing the training samples to conserve listwise similarities (in the rest of this paper, we will call this learning strategy the listwise constraints) in the projection feature space, all dissimilar images are forced (denoted with green circles) to keep a certain distance(i.e., a margin denoted by the orange circles) from the similar images to the probe. Thus, intuitively, by learning metric from listwise constraints would yield a better generalization performance than that by using sparse pairwise constraints.

Though those ignored information could be considered by the existing metric learning methods by simply constructing $(N-1)*N$ dissimilar image pairs, the computation cost would be increased quadratically, which is unaffordable. For instance, the computation complexity of PRDC [63], one of the state-of-the-art metric learning methods, is $O(N^3)$ when $k=1$. And for $k=N-1$, the computational cost would be $O(N^4)$. Moreover, as validated in [44], the performance of metric learning methods can not be consistently enhanced by increased (dis)similarity pairs. When $k$ exceeds 20, the performance would fall down. This phenomenon is caused by the unbalanced negative (dissimilar) and positive (similar) training pairs, when $k$ is sufficiently large.

In order to address the limitations mentioned above, we propose a novel metric learning method called Relevance Metric Learning with Listwise Constraints (RMLLC). Unlike existing approaches [25], [41], [63] RMLLC firstly cast the person re-identification as an image retrieval task, and measures the similarity of two feature vectors by using their inner product, or, relevance, which implicitly contains the angle information. This relevance metric can be seen as a complement to traditional Mahalanobis distance metrics. To overcome the shortcomings of sparse pairwise constraints, we predefine a list of similarity scores for each probe image with respect to the gallery set, where similar pairs are initialized as 1, and dissimilar pairs are initialized as 0. Subsequently, we incorporate the listwise similarities to learn an optimal relevance metric, by forcing similarities measured with learnt metric to conserve the predefined similarity lists. Through this way, we formulate an optimization problem, which encodes the listwise constraints to metric learning. Benefitted from

the relevance metric, the final optimization problem could be solved efficiently, after being reformulated to a convex problem. Since the predefined binary listwise similarities could not capture the relative relevance, we employ a rectification term to optimize the predefined similarity lists, and propose an improved algorithm called RMLLC with rectification (RMLLC(R)). The original problem is then turned to an optimization problem with respect to two unknown variables: the metric and the rectification similarity matrix. We introduce an efficient alternating iterative algorithm to jointly learn these two matrices, and prove its convergence as well.

To evaluate the performance of RMLLC together with its improvement RMLLC(R), we conduct extensive experiments on four benchmark datasets for person re-identification in various scenarios: the VIPeR dataset, the GRID dataset, the iLIDS (MCTS) dataset and the CAVIAR4REID dataset. The results show that: (1) by jointly optimizing the learnt metric and the listwise similarities, RMLLC(R) obtains an improvement on RMLLC; (2) by incorporating the listwise constraints and encoding all available similarity information for training, we can obtain substantial improvements on matching rate, which is comparable or superior to existing state-of-the-art person re-identification approaches. Besides, we also provide experimental results on the cross view gait recognition application. The significant boost in performance compared with the current best results demonstrates the applicability of the proposed method in other identification tasks.

An early version of this work appeared in [5], which also adopted listwise similarities to train a relevance metric. We have made the following improvements on this early work: (1). an efficient alternating iterative optimizing method was developed to learn an optimal metric and similarity lists in order to capture the relative relevance information, which was ignored in [5]; (2). additional analysis on the performance of the proposed method, such as the computation cost, was discussed; (3). more extensive experimental evaluation together with its applicability on other identification tasks such as the cross view gait recognition were provided.

## II. Related Work

Many recent works have addressed the person re-identification problem, most of which primarily focus on the feature-based methods, the model-based methods and the metric learning approaches.

### A. Feature-Based Methods

This category mainly focuses on pursing discriminative image representation robust to view, illumination, pose, background variations. In [10], Farenzena et. al augmented maximally stable color regions with histograms and recurrent local color patches. Bazzani et al. [4] extracted a collection of recurrent stable color pathes, and used histograms for image representation. Ma et al. [37] constructed local descriptors encoded by Fisher Vectors using pixel-wise intensity. They further developed a robust representation to background and illumination variations by combining Biologically Inspired Features (BIF) and Covariance descriptors in [38]. Kviatkovsky et al. [25] investigated the intra-distribution structure of color descriptors, which is invariant under certain illumination changes. Wu et al. [52] introduced a viewpoint-invariant descriptor, taking into account the viewpoint of the human by using what they called a pose prior learned from training data. Recently, some research works on learning reliable and effective mid-level features for person re-identification have been done. Li and Wang [27] proposed a deep learning framework to automatically learn robust features, by encoding the part displacement, photometric, pose and viewpoint transforms across camera views. However, in order to gain satisfying performance, it requires larger scale training data, which is often unavailable in practice. Alternatively, Zhao et al. [61] learnt mid-level filters from automatically discovered patch clusters, which had achieved promising cross-view invariance.

### B. Model-Based Methods

Feature-based approaches try to extract features that are invariant to appearance variations caused by changes of illumination, viewpoint and etc. However, as shown in Fig. 1, the appearances captured from different pedestrians are far more similar to those from the same pedestrian. Thus, it is generally very difficult to design a universal feature representation method that could handle all kinds of variations, and reliably identify images of the same pedestrian while distinguishing images belong to different pedestrians. To address limitations of the feature-based methods, several approaches have been proposed to directly explore the unknown information such as viewpoint, pose, illumination, and camera parameters. Javed et al. [18] proposed to exploit inter-camera Brightness Transfer functions (BTF) to handle appearance variations caused by intrinsic camera parameters. Jungling and Arens [19] investigated the possibility of acquiring more distinctiveness by determining the pose prior to feature extraction. Zhao et al. [59] extracted saliency regions by building dense correspondence to handle misalignment caused by large viewpoint and pose variations. Li and Wang [27] learned a mixture of cross-view transforms to handle the viewpoint changes. Very recently, Chen et al. [6] presented a mirror representation for developing a view-specific mapping. In practice, it is often desirable to precisely obtain the position of each parts of a human body (such as head, torso, legs), which could overcome the occlusion and pose variation problems. Xu et al. [36] employed assembly of part templates to handle the articulation of human body, and improved the re-identification performance. In [49], Wang et al. explored generative probabilistic topic model to simultaneously discover localised person foreground appearance saliency, and remove noisy background clutters. Zheng et al. [64] proposed a transfer local relative distance comparison model to address the open-world person re-identification by one-shot group-based verification.

### C. Metric Learning Approaches

In practice, illumination, viewpoint and other variations change continuously, and labeled training samples are usually unavailable. It is therefore rather difficult to reliably model those variations by using model-based approaches.

Recently, metric learning methods have been introduced, and have significantly improved the performance of person re-identification. This kind of approaches aims to construct a robust metric for similarity measurement. In [63], Zheng proposed a relative distance learning method from a probabilistic prospective. Köstinger et al. [20] presented a scalable distance metric learning method from equivalence constraints, called KISSME. However, KISSME was not stable for a small size training set, and performed poorly. Tao et al. subsequently proposed two improvements on KISSME. In [54], they presented a regularized smoothing KISS metric learning by integrating smoothing and regularization techniques for robustly estimating covariance matrices. In [55], they replaced the maximum likelihood (ML) estimator on covariance matrix for the classical KISSME by the minimum classification error (MCE), which was considered to be more reliable than ML estimation with increasing training samples. Mignon and Jurie [41] learned a distance metric from sparse pairwise similarity constraints. Pedagadi et al. [44] utilized local Fisher Discriminant Analysis (LFDA) to map high dimensional features into a more discriminative low dimensional space. Xiong et al. [57] further extended LFDA and several other metric learning methods by using kernel tricks and different regularizers.

However, all the aforementioned metric learning methods make decisions based on a fixed threshold, which is insufficient and sub-optimal for the re-identification problem. To address this problem, Li et al. [29] jointly learnt a locally adaptive decision function (LADF) and a Mahalanobis distance metric. In [39], the authors designed multiple Mahalanobis distance metrics by constructing one metric for any camera pair, in order to cope with the complicated conditions existing in a typical camera networks. They assumed that the multiple metrics are related, and formulated person re-identification over camera network as a multitask distance metric learning problem. Besides metric learning based matching strategy, Lisanti et al. [30] matched candidate targets by making use of soft and hard re-weighting to redistribute energy among the most relevant contributing elements, which could be seen as an iterative extension to sparse discriminative classifiers.

Several existing metric learning approaches are closely related to this work. In [42], Nguyen et al. proposed a cosine similarity metric learning framework, and measured the similarity of face pairs through the cosine similarity, which was closely related to the inner product similarity. However, our method formulated the problem in a very different way, and had encoded listwise similarities. Loy et al. [32] also cast the person re-identification as an image retrieval task, and considered the listwise similarities. But the authors assumed that the metric had been learnt by using any prevailing metric learning with pairwise constraints beforehand, and just introduced an unsupervised manifold ranking to rerank the listwise similarities. In contrast, our method utilized the listwse similarities to train a discriminative metric supervisedly. Our work could be used as a start point of [32].

This work also could be seen as an attempt to take into account the relative similarity. Several previous works from other applications have tried to address this issue. Lu et al. [35] proposed a new neighborhood repulsed metric learning (NRML) method for kinship verification, which was a representative work for exploiting the different importance of negative training pairs. The main idea was to pull intra-class samples (with a kinship relation) as close as possible and simultaneously repulsed interclass samples lying in a neighborhood as far as possible. By virtue of this training principle, Lu et al. [35] has significantly boosted the performance of kinship verification. Liu et al. [33] developed a metric learning algorithm by minimizing squared residuals from relative comparisons, where pair-wise constraints were not natural to obtain. While, these two approaches could not be directly used to exploit the relative similarity information in metric learning for person re-identification, due to the unavailability of relative similarity information in practical person re-identification.

The main contributions of this work lie on three folds: 1). we propose a novel relevance metric learning method which could effectively encode all available pairwise similarities, by adopting listwise constraints; 2). we develop a scheme to automatically exploit relative similarities, which are often unavailable in practice, by introducing a rectification term to predefined binary similarity lists; 3). we develop an efficient alternating iterative optimization algorithm to jointly learn an optimal similarity metric and listwise similarities, and prove its convergence as well.

## III. Relevance Metric Learning for Person Re-Identification With Listwise Constraints

### A. Problem Formulation

In this section, we treat the person re-identification as an image retrieval task. Without loss of generality, we suppose that the training data consists of $N$ $D-$dimensional feature vectors, denoted by $X = (x_1, x_2, \cdots, x_N)^{\mathrm{T}} \in \mathbb{R}^{N\times D}$. Considering that $X$ is a gallery set, and any feature vector $x_j$ is a probe, we further assume that we have a list of similarity scores $L(j) = (l(1, j), \cdots, l(N, j))^{\mathrm{T}} \in \mathbb{R}^N$, where $l(i, j)$ is the similarity score between the gallery $x_i$ and the probe $x_j$. $l(i, j) > l(k, j)$, if $x_i$ is more similar to $x_j$ than $x_k$, indicating that $L(j)$ provides the global similarity information of the probe $x_j$ with respect to the gallery set $X$. For all $1 \leq i \leq N$, we can obtain a similarity score matrix $L$, of which the $j-$th column is $L(j)$. In practice, $L$ is always a hidden variable, and is unavailable. However, we know the pairwise constraints whether an image pair is similar or dissimilar. Thus, we can construct $L$ by simply setting $l(i, j) = 1$ if $x_i$ and $x_j$ is similar, and $l(i, j) = 0$ otherwise. Alternatively, we could obtain $L$ by constructing the affinity matrix with learnt distance metrics. Then, the predefined $L$ can be seen as a reliable approximation to its true value.

For the $i-$th gallery image $x_i$ and the $j-$th probe image $x_j$, we could estimate their similarity score by using the inner product: $s(i, j) =< x_i, x_j >= x_i^{\mathrm{T}} x_j$. When $x_i$ and $x_j$ are normalized, i.e., $\|x_i\|_2 = \|x_j\|_2 = 1$, $s(i, j)$ is the cosine of the angle between $x_i$ and $x_j$. Thus, the metric utilized in this work could reflect the "angle" information of feature

vectors, and is complementary to the prevailing Mahalanobis metrics, which measures the similarity of two feature vectors by "distance". Like all the other metric learning methods, we intend to learn a projection matrix $W \in \mathbb{R}^{D\times d}$, and map the $D-$dimensional feature into a $d-$dimensional feature space, where the similarity of $x_i$ and $x_j$ is rewritten as

$$s_W(i, j) = (W^{\mathrm{T}}x_i)^{\mathrm{T}}(W^{\mathrm{T}}x_j) = x_i^{\mathrm{T}}WW^{\mathrm{T}}x_j.$$

Thereafter, for a gallery set $\{x_i\}_{i=1}^N$ and the probe $x_j$, the estimated similarity list is $S(j) = (s_W(1, j), \cdots, s_W(N, j))^{\mathrm{T}}$. The optimal $W$ should be the one which could conserve the predefined similarity matrix $L$, i.e., minimize the error between the estimated similarity list $S(i)$ and the predefined similarity list $L(i)$. Here, we utilize the sum of square error, which therefore yield the following optimization problem:

$$\min_W \quad \sum_{j=1}^{N}(S(j) - L(j))^{\mathrm{T}}(S(j) - L(j)). \tag{1}$$

By rearranging Eq.(1) using matrices, we can reformulate Eq.(1) as follows:

$$\min_W \quad \|XWW^{\mathrm{T}}X^{\mathrm{T}} - L\|_F^2, \tag{2}$$

where the $j$-th column of $S_W$ is $S(j)$.

In order to avoid overfitting, we introduce a regularization term $tr(W^{\mathrm{T}}W)$, by which Eq.(2) is turned to the following:

$$\min_W \quad \left(\|XWW^{\mathrm{T}}X^{\mathrm{T}} - L\|_F^2 + \gamma\,\mathrm{tr}(W^{\mathrm{T}}W)\right), \tag{3}$$

where $\gamma$ is the regularization weight parameter to make a trade-off between the error and the regularization term.

It should be noted that Eq.(3) could be formulated as a kernel alignment [8], [26], or low rank kernel learning problem [15], [17], [21] by defining a kernel $k(x_i, x_j) = x_i^T WWx_j$, and treating $L$ as an ideal kernel $K^*$. Compared to these existing works, our formulation in Eq. (3) explicitly learns a low rank kernel by the parameter matrix $W \in R^{N\times d}$ with small $d$, which may better generalize to unseen testing data for small scale training data. Moreover, Eq.(3) adopts the sum of square error, and could be solved more efficiently, which could be seen in the next subsection and TABLE VI.

### *B. Optimization Method*

In the previous section, we formulate the learning problem in Eq.(3). Since $\|XWW^{\mathrm{T}}X^{\mathrm{T}} - L\|_F^2$ is noncovex with respect to $W$, Eq. (3) is not a convex optimization problem. In this section, we will relax it to a convex problem, and develop a close form approximation solution.

Generally, the predefined similarity matrix $L$ is a real symmetric matrix, and we have the following eigenvalue decomposition: $L = U\Lambda U^{\mathrm{T}}$, $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_N)$, where $\lambda_i$ is the $i-$th largest eigenvalue of $L$ with the $i-$th column of $U$ being the corresponding eigenvector. Based on this decomposition, we have the following equation

$$\begin{aligned}\|XWW^{\mathrm{T}}X^{\mathrm{T}} - L\|_F^2 &= \|XWW^{\mathrm{T}}X^{\mathrm{T}} - U\Lambda U^{\mathrm{T}}\|_F^2 \\ &= \|U^{\mathrm{T}}XWW^{\mathrm{T}}X^{\mathrm{T}}U - \Lambda\|_F^2. \end{aligned}\tag{4}$$

**Algorithm 1** Relevance Metric Learning With Listwise Constraints (RMLLC) for Person Re-Identification

**Input:** training samples $\{x_i\}_{i=1}^N$, predefined similarity score matrix $L$, and $d$.

**Output:** the projection matrix $W$.

1: Do the eigenvalue decomposition of $L$: $L = U\Lambda U^{\mathrm{T}}$;

2: Calculate the low rank approximation of $U^{\mathrm{T}}XW$: $U^{\mathrm{T}}XW \approx \left[\Lambda_+^{1/2}\right]_d$;

3: Solve the following subproblem

$$W = \underset{W}{\mathrm{argmin}} \left(\left\|U^{\mathrm{T}}XW - \left[\Lambda_+^{1/2}\right]_d\right\|_F^2 + \gamma\mathrm{tr}(W^{\mathrm{T}}W)\right). \quad (*)$$

[ ***:** The closed form solution to this subproblem is $W = (X^{\mathrm{T}}X + \gamma I_D)^{-1}X^{\mathrm{T}}U\,[\Lambda_+]_d^{1/2}$.]

The second equation holds, because the Frobenius norm is invariant to unitary transformations [13].

Since $U^{\mathrm{T}}XWW^{\mathrm{T}}X^{\mathrm{T}}U$ is a semi-positive definite matrix, $\Lambda_+ = \mathrm{diag}([\lambda_1]_+, \cdots, [\lambda_N]_+)$ is the optimal approximation of $U^{\mathrm{T}}XWW^T X^{\mathrm{T}}U$, where

$$[\lambda_i]_+ = \begin{cases}\lambda_i, & \text{if } \lambda_i \geq 0;\\ 0, & \text{if } \lambda_i < 0.\end{cases}$$

Thus, the optimization problem

$$W = \underset{W}{\mathrm{argmin}} \quad \|U^{\mathrm{T}}XWW^{\mathrm{T}}X^{\mathrm{T}}U - \Lambda\|_F^2$$

could be relaxed to

$$W = \underset{W}{\mathrm{argmin}} \quad \|U^{\mathrm{T}}XW - \left[\Lambda_+\right]_d^{1/2}\|_F^2,$$

where $\left[\Lambda_+\right]_d \in \mathbb{R}^{N\times d}$ consists of the first $d$ columns of $\Lambda_+$. Finally, Eq.(3) can be relaxed to the following optimization problem.

$$\min_W \quad \left(\left\|U^{\mathrm{T}}XW - \left[\Lambda_+\right]_d^{1/2}\right\|_F^2 + \gamma\,\mathrm{tr}(W^{\mathrm{T}}W)\right). \tag{5}$$

Since the first term of the equation above is convex, and the regularization term $tr(W^{\mathrm{T}}W)$ is also convex with respect to $W$, Eq.(5) is a convex optimization problem.

Actually, Eq.(5) is readily to be solved, since it is a standard least square problem, and has a closed form solution: $W = (X^{\mathrm{T}}X + \lambda I_N)^{-1}X^{\mathrm{T}}U\left[\Lambda_+\right]_d^{1/2}$, where $I_N$ is the identity matrix with order $N$. We summarize the solution to Eq.(3) in Algorithm 1.

### *C. Kernel RMLLC*

The formulation in Algorithm 1 can be generalized to support non-linear mapping by using kernels. Following the method of Globerson and Roweis [11], we firstly map the data into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ via a feature map $\phi$ with corresponding kernel function $k(x, y) =< \phi(x), \phi(y) >_{\mathcal{H}}$. Thereafter, the data is mapped to $\mathbb{R}^d$ by a linear projection $M : \mathcal{H} \to \mathbb{R}^d$. The project matrix $W$ can be expressed as $W = \Phi M$, where $\Phi_i = \phi(x_i)$. Thus, we can reformulate Eq.(3) as an optimization problem over $M$

**Algorithm 2** Kernel Relevance Metric Learning With List-Wise Constrained (RMLLC) for Person Re-Identification

**Input:** training samples $\{x_i\}_{i=1}^N$, predefined similarity score matrix $L$, and given rank $d$.

**Output:** the projection matrix $M$.

1: Do the eigenvalue decomposition of $L$: $L = U\Lambda U^{\mathrm{T}}$;

2: Calculate the low rank approximation of $U^{\mathrm{T}}KM$: $U^{\mathrm{T}}KM \approx \left[\Lambda_+^{1/2}\right]_d$;

3: Solve the regularized subproblem

$$M = \underset{M}{\operatorname{argmin}} \left( \left\| U^{\mathrm{T}}KM - \left[\Lambda_+^{1/2}\right]_d \right\|_F^2 + \gamma \mathrm{tr}(M^{\mathrm{T}}M) \right). \quad (*)$$

[*: The closed form solution: $M = (K^{\mathrm{T}}K + \gamma I_N)^{-1}K^{\mathrm{T}}U\,[\Lambda_+]_d^{1/2}$.]

(a) (b)

Fig. 2. Samples from the VIPeR dataset [14]: image pair in (a) is more similar than image pari in (b).

rather than $W$. Since $\Phi^{\mathrm{T}}\Phi MM^{\mathrm{T}}\Phi\Phi^{\mathrm{T}} = KMM^{\mathrm{T}}K^{\mathrm{T}}$, where $K_{i,j} = k(x_i, x_j)$, Eq.(3) could be rewritten as

$$\min_M \quad \left( \|KMM^{\mathrm{T}}K^{\mathrm{T}} - L\|_F^2 + \gamma\, \mathrm{tr}(M^{\mathrm{T}}M) \right). \quad (6)$$

We can see that Eq.(6) has the same form as Eq.(3) except that $X$ is replaced by $K$. Therefore, Eq.(6) can be solved using the same approach. We list the solution to Eq.(6) in Algorithm 2.

After training, we obtain the mapping matrix $M$, and the overall transformation matrix $W = \Phi M$. Given a gallery $x^{(g)}$, and an input probe $x^{(p)}$, their similarity $s(x^{(g)}, x^{(p)})$ can be calculated by using the learnt $M$:

$$\begin{aligned} s(x^{(g)}, x^{(p)}) &= \phi(x^{(g)}) \cdot \Phi MM^{\mathrm{T}}\Phi^{\mathrm{T}} \cdot \phi(x^{(g)}) \\ &= \Phi^{(g)}MM^{\mathrm{T}}{\Phi^{(p)}}^{\mathrm{T}}, \end{aligned}$$

where $\Phi_i^{(g)} = k(x^{(g)}, x_i)$, and $\Phi_i^{(p)} = k(x^{(p)}, x_i)$.

## IV. Relevance Metric Learning With Listwise Constraints by Jointly Optimizing Metrics and Listwise Similarities

### A. Problem Formulation

In the previous section, since the groundtruth listwise similarities are always unavailable, we predefine a binary principle similarity matrix $L = (l_{ij})_{N\times N}$ as follows:

$$l_{ij} = \begin{cases} 1, & \text{if } I^{(i)} \text{ and } I^{(j)} \text{ denote the same person;} \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

This predefinition has incorporated the pairwise "similar", "dissimilar" label information, i.e., image pairs for the same person are assigned 1, and those for different persons are assigned 0. Though this hard assignment contains the dominant discriminative similarity information, it might fail to capture relative similarity information. For instance, both of the two image pairs shown in Fig. 2 belong to different persons, while the image pair (a) is apparently more similar than the image pair (b). Thus, it is more reasonable to set $s_a > s_b$ than to use the hard assignment where $s_a = s_b = 0$.

To overcome this shortcoming of the hard assignment, we introduce a rectification term, the matrix $G$, to define a new listwise similarity matrix $\tilde{L}$ as follows:

$$\tilde{L} = L + \beta G. \quad (8)$$

Here, $\tilde{L}$ consists of two components: $L$ is the listwise similarity matrix by hard assignment from Eq.(7), which contains the principle discriminative similarity information. $G$ is the rectification matrix capturing the relative similarity information, and $\beta$ is the trade-off parameters leveraging the effect of $G$.

From Eq.(8), $G$ appears to be a noise term to the incomplete similarity term $L$, based on which our formulation then seems to be a metric learning problem with label noises as described in [48]. While, this perspective may not totally be true. In [48], the authors defined the noisy label as "incorrect labels", and utilized the conditional probability to model the noisy label. However, in Eq.(8) we don't assume that the incomplete similarity list is incorrect. Conversely, we believe that the predefined similarity matrix $L$ could reflect the principle similarity information. Our argument is that $L$ is not fine enough, since it ignores the relative similarity information. Thus, $L$ should be modestly rectified around the predefined one. Based on this motivation, we introduce the rectification term $G$.

By following [47], we further introduce the constraint $G^{\mathrm{T}}G = I$ to guarantee that $G$ is well defined to represent the similarity. We finally present the optimization problem for jointly learning the projection matrix $W$ and the rectification term $G$:

$$(W, G) = \underset{W\in R^{N\times d}, G\in S^N, G^{\mathrm{T}}G=I_N}{\operatorname{argmin}} f(W, G), \quad (9)$$

where $f(W, G) = \|KWW^{\mathrm{T}}K^{\mathrm{T}} - (L + \beta G)\|_F^2 + \gamma \|W\|_F^2$. The objective function $f(W, G)$ is not convex with respect to $W$ and $G$. Here, we attempt to archive a local minima by alternatively optimizing $W$ with fixed $G$ and optimizing $G$ with fixed $W$:

$$G_t = \underset{G\in S^N, G^{\mathrm{T}}G=I}{\operatorname{argmin}} f(W_{t-1}, G); \quad (10)$$

$$W_t = \underset{W\in R^{N\times d}}{\operatorname{argmin}} f(W, G_t). \quad (11)$$

By using this alternating iterative optimization strategy, we can deduce that:

$$f(W_{t-1}, G_{t-1}) \geq f(W_{t-1}, G_t) \geq f(W_t, G_t),$$

which could guarantee that the objective function decreases until converging to a local minima. Note that the first inequality holds due to (10), and the second inequality holds due to (11). The outline of the solution to (9) could be seen in Algorithm 3.

### B. Solutions to Subproblems

In Algorithm 3, Two subproblems, i.e., subproblem (10) and subproblem (11), remain unsolved. In this part, we will develop solutions to these two problems.

**Algorithm 3** RMLLC With Rectification (RMLLC(R)) for Person Re-Identification

**Input:** training samples $\{x_i\}_{i=1}^N$, predefined similarity score matrix $L$, projection dimension $d$, trade-off parameters $\beta$, $\gamma$.
**Output:** the projection matrix $W$.

1: Initialization: $t = 1$, $G_0 = I_N$, and

$$W_0 = \underset{W \in R^{N\times d}}{\operatorname{argmin}} \left( \|KWW^{\mathrm{T}}K^{\mathrm{T}} - L\|_F^2 + \gamma \|W\|_F^2 \right).$$

**while** not convergence **do**

2: Solve the following subproblem according to Eq.(15) and Eq.(16).

$$G_t = \underset{G\in S^N, G^{\mathrm{T}}G=I}{\operatorname{argmin}} f(W_{t-1}, G). \tag{12}$$

3: Let $f_W(W) = \|KWW^{\mathrm{T}}K^{\mathrm{T}} - (L + \beta G_t)\|_F^2 + \gamma \|W\|_F^2$. Calculate the gradient of $f_W(W)$ with respect to $W$:

$$\begin{aligned} \nabla f_W(W) = \; & 4K^{\mathrm{T}}KWW^{\mathrm{T}}K^{\mathrm{T}}K^{\mathrm{T}}KW \\ & -4K^{\mathrm{T}}(L + \beta G_t)KWW^{\mathrm{T}} + 2\gamma W. \end{aligned}$$

Calculate the step length $\alpha_t$ by using backtracking line search.
Update $W$: $W_t = W_{t-1} - \alpha_t \nabla f_W(W_{t-1})$.

4: $t := t + 1$.

**end while**

Inspired by [47], we firstly derive a close form solution to (12). Specifically, we attemp to solve the following problem:

$$G_t = \underset{G\in S^N, G^{\mathrm{T}}G=I_N}{\operatorname{argmin}} f(W_{t-1}, G),$$

where $f(W, G) = \|KWW^{\mathrm{T}}K^{\mathrm{T}} - (L + \beta G)\|_F^2 + \gamma \|W\|_F^2$. To simplify the optimization problem above, we denote that $R = KW_{t-1}W_{t-1}^{\mathrm{T}}K^{\mathrm{T}} - L$. (12) then turns to the problem of finding $G^*$ such that

$$G^* = \underset{G\in S^N, G^{\mathrm{T}}G=I_N}{\operatorname{argmin}} \|R - \beta G\|_F^2. \tag{13}$$

Based on the fact that $G^{\mathrm{T}}G = I_N$, we have the following

$$\begin{aligned} \|R - \beta G\|_F^2 &= \beta^2 tr(G^{\mathrm{T}}G) - 2\beta tr(G^{\mathrm{T}}R) + tr(R^{\mathrm{T}}R) \\ &= -2\beta tr(G^{\mathrm{T}}R) + \beta^2 N + tr(R^{\mathrm{T}}R) \\ &= -2\beta tr(G^{\mathrm{T}}R) + \text{const}. \end{aligned} \tag{14}$$

Thus, (13) is equivalent to

$$G^* = \underset{G\in S^N, G^{\mathrm{T}}G=I_N}{\operatorname{argmax}} tr(G^{\mathrm{T}}R). \tag{15}$$

Assuming that the eigenvalue decomposition of $R$ is $R = U_R \Lambda_R U_R^{\mathrm{T}}$, then

$$tr(G^{\mathrm{T}}R) = tr(G^{\mathrm{T}}U_R\Lambda_R U_R^{\mathrm{T}}) = tr(U_R^{\mathrm{T}}G^{\mathrm{T}}U_R\Lambda_R),$$

where $\Lambda_R = \operatorname{diag}(\lambda_1, \cdots, \lambda_N)$, $\lambda_i$ is the $i$-th eigenvalue of $R$. Let $Z = (z_{ij})_{N\times N} = U_R^{\mathrm{T}}G^{\mathrm{T}}U_R$, then $tr(G^{\mathrm{T}}R) = tr(Z\Lambda_R) = \Sigma_{i=1}^N \lambda_i z_{ii}$. On the other hand, $ZZ^{\mathrm{T}} = U_R^{\mathrm{T}}G^{\mathrm{T}}U_R U_R^{\mathrm{T}}GU_R = I_N$, which yields that $|z_{ii}| \le 1$. Therefore, $tr(G^{\mathrm{T}}R)$ is maximized if and only if when

$$z_{ij}^* = \begin{cases} -1, & \text{if } i = j, \text{ and } \lambda_i < 0; \\ 1, & \text{if } i = j, \text{ and } \lambda_i \ge 0; \\ 0, & \text{if } i \ne j. \end{cases} \tag{16}$$

Denoting that $Z^* = (z_{ij}^*)_{N\times N}$, then

$$G^* = U_R Z^* U_R^{\mathrm{T}} \tag{17}$$

is the close form solution to subproblem (12).

As to the following subproblem (11)

$$W_t = \underset{W\in R^{N\times d}}{\operatorname{argmin}} f(W, G_t),$$

it is a nonlinear and unconvex problem. Thus, it's difficult to find the global optima, or a close form solution. Here, we adopt the steepest descend gradient method [43]. Denoting that $f_W(W) = f(W, G_t) = \|KWW^{\mathrm{T}}K^{\mathrm{T}} - (L + \beta G_t)\|_F^2 + \gamma \|W\|_F^2$, it is not difficult to derive that the gradient of $f_W(W)$ with respect to $W$ is:

$$\begin{aligned} \nabla f_W(W) = \; & 4K^{\mathrm{T}}KWW^{\mathrm{T}}K^{\mathrm{T}}K^{\mathrm{T}}KW \\ & -4K^{\mathrm{T}}(L + \beta G_t)KWW^{\mathrm{T}} + 2\gamma W. \end{aligned}$$

In order to determine a proper step length in each iteration, we use the backtracking line search, which could grantee the convergence [43]. Considering the slow convergent rate of the steepest descend gradient method, we utilize $W_{t-1}$ as the begin point, and iterate just once to accelerate the convergence rate of the whole alternating iterative algorithm. Specifically, in the $t$-th iteration, we calculate $W_t$ by using the following updating formula:

$$W_t = W_{t-1} - \alpha_t \nabla f_W(W_{t-1}). \tag{18}$$

By combining (17) and (18), we propose Algorithm 3 to solve the original joint optimization problem. Despite that the local optima could not be achieved through the one step updating, we can proof that the proposed algorithm converges to a local minima. The proof is straightforward, we therefore omit it here.

## V. Experimental Results and Analysis

### A. Datasets

In order to evaluate the performance, we test our methods on four most challenging and widely used benchmark datasets: the VIPeR dataset [14], QMUL underGround Re-IDentification (GRID) dataset [31], iLIDS Multiple-Camera Tracking Scenario (iLIDS MCTS) dataset [62], and CAVIAR4REID dataset [7].

**VIPeR dataset** contains 632 pedestrians captured from two non-overlapping cameras in an outdoor academic environment. Each pedestrian has one image per camera view. This dataset is very challenging due to large variations in viewpoint, illumination and pose, but with less occlusions. **GRID dataset** is the most difficult person re-identification dataset, to the best of our knowledge. This dataset is captured from 8 disjoint camera views in a busy underground station, with severe inter-object occlusion, large viewpoint variations

TABLE I
TOP $r$ RANKED MATCHING RATE (%) OF RMLLC (R) AND RMLLC ON FOUR BENCHMARKS

| Dataset | VIPeR | | | | iLIDS (MCTS) | | | |
|---|---|---|---|---|---|---|---|---|
| | p=316 | | | | p=30 | | | |
| Methods | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 |
| RMLLC (R) | **31.27** | **62.12** | **75.31** | **86.71** | **56.53** | 82.45 | 91.39 | 97.18 |
| RMLLC | 29.68 | 59.81 | 74.27 | 85.89 | 55.39 | **84.48** | **93.16** | **98.73** |
| Methods | p=432 | | | | p=50 | | | |
| RMLLC (R) | **24.14** | **49.58** | **64.49** | 77.55 | **46.53** | **73.01** | 84.54 | 93.25 |
| RMLLC | 22.99 | 49.4 | 63.08 | **78.1** | 45.19 | 72.88 | **84.57** | **93.45** |
| Methods | p=532 | | | | p=80 | | | |
| RMLLC (R) | **17.56** | **37.99** | **51.82** | **66.02** | **35.13** | **59.69** | **72.57** | **85.23** |
| RMLLC | 16.9 | 37.80 | 50.34 | 65.28 | 31.85 | 57.13 | 69.68 | 81.91 |
| Dataset | GRID | | | | CAVIAR | | | |
| Methods | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 |
| RMLLC (R) | **17.12** | **35.04** | **44.24** | **55.36** | **41.17** | 73.51 | 84.99 | 94.37 |
| RMLLC | 16.8 | 34.24 | 43.84 | 54.80 | 40.16 | **73.80** | **86.17** | **95.48** |

and poor image quality. It contains 250 pedestrian image pairs. Each pair contains two images of the same individual seen from different camera views. In addition, there are 775 extra individual images that do not belong to any of the paired images. In **iLIDS (MCTS) dataset**, which is captured indoor at a busy airport arrival hall, there are 119 people with totally 476 images captured by multiple nonoverlapping cameras, four images per person on average. **CAVIAR4REID dataset** contains 1220 images of 72 individuals from 2 cameras in a shopping mall, where each person has 10-20 images. The images in this dataset also have severe occlusions, illumination and resolution changes.

### B. Feature Descriptors and Kernels

In our experiments, we employed a mixture of color and texture histogram features similar to [57]. Specifically, we divided a person image into six equal horizontal stripes to roughly capture the head, torso and leg of a human body. For each stripe, 8 color channel (RGB, YCbCr, HS) color features were computed, where each channel was represented by a 16D histogram vector. To represent the texture patterns, Local Binary Patterns (LBP) were computed for each region. The histograms were then normalized in each channel and concatenated to form the final feature vector. As for kernels, in this paper we uniformly use RBF $\chi^2$ kernel [41].

### C. Evaluation Setting

We follow the evaluation protocol dictated in [10] and [14], and report Cumulative Matching Characteristic (CMC) curves. In all our experiment, we adopt a single-shot experiment setting. Specifically, we randomly select $p$ pedestrians for testing and the rest for training. The test images are divided into a gallery set and a probe set. The gallery set consists of one image for each pedestrian, and the remaining images form the probe set. For VIPeR dataset, the number of traning/testing persons are set to 316/316, 200/432 and 100/532, respectively, in order to test the performance of comparison approaches using training data of different scales. For GRID dataset, in order to make a fair comparison, we use the training/test partitions provided in [32]. For iLIDS MCTS dataset, the numbers of persons for training/testing are set to 89/30, 69/50 and 39/80, respectively. Finally, for CAVIAR4REID dataset, the number of training/testing is set to 36/36. To obtain a reasonable statistical significance, we report the average CMC curves over 10 rounds.

### D. RMLLC vs. RMLLC With Rectification

We firstly compare RMLLC with RMLLC(R) on the aforementioned four benchmark datasets. In TABLE I, we report the matching rate of RMLLC and RMLLC(R) at rank 1, 5, 10, 20, on VIPeR, GRID, iLIDS MTCS, and CAVIAR4REID datasets, respectively. As can be seen, RMLLC(R) obtains 1.59%, 1.15%, 0.66% higher matching rate than RMLLC on VIPeR for $p = 316$, 432, and 532 at rank 1, respectively, together with a 0.32% improvement on GRID dataset. For iLIDS (MCTS), RMLLC(R) outperforms RMLLC with 1.14%, 1.34%, 3.28% at rank 1 for $p = 30$, 50, and 80, respectively. And on CAVIAR4REID dataset, the rank 1 matching rate of RMLLC(R) is 1.01% higher than RMLLC.

The consistent improvement of RMLLC(R) is due to the introduction of the rectification matrix $G$, which could automatically explore more accurate similarity information and relative relevance by optimizing the predefined listwise similarity matrix $L$. From a statistical learning perspective, the predefined binary similarity matrix $L$ provides a decision margin. The rectification matrix $G$ serves as a set of slack variables, which allows training samples to violate the decision margin, making RMLLC(R) more generative to unseen test data.

Since RMLLC(R) could consistently boost the performance of RMLLC, we will only demonstrate the performance of RMLLC(R) in the rest of this paper, and denote it as the proposed method. It should be noted that RMLLC without rectification is also superior to state-of-the-art approaches in most situations. The superiority could be easily seen by combing TABLE I and TABLE II-V, which indicates that the learnt metric by only using predefined listwise similarities could yield significant improvements.

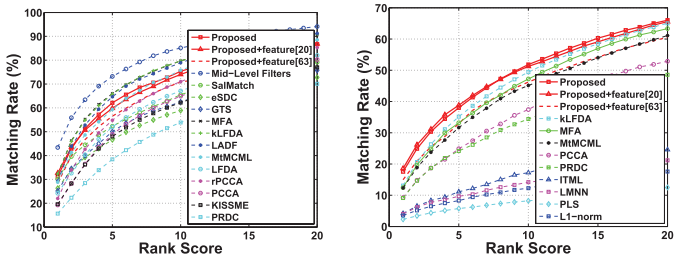
### E. Proposed vs. State-of-the-Art

In this subsection, we will compare our method to other state-of-the-art approaches.

TABLE II

TOP $r$ RANKED MATCHING RATE (%) ON THE VIPeR DATASET FOR $p = 316$, $p = 432$ AND $p = 532$ PEDESTRIANS IN THE TESTING SET

| Methods | p=316 | | | | p=432 | | | | p=532 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 |
| Proposed | 31.27 | 62.12 | 75.31 | 86.71 | 24.14 | **49.58** | 64.49 | 77.55 | 17.56 | 37.99 | **51.82** | **66.02** |
| Proposed+feature[20] | 32.65 | 60.06 | 74.02 | 86.84 | **24.77** | 49.51 | 63.26 | 76.85 | **18.59** | **38.82** | 51.24 | 65.55 |
| Proposed+feature[63] | 28.43 | 56.51 | 70.83 | 82.05 | 21.06 | 44.82 | 59.63 | 71.74 | 14.96 | 32.92 | 45.76 | 60.54 |
| Mid-level Filter[61] | **43.49** | **73.2** | **85.1** | **94.1** | - | - | - | - | - | - | - | - |
| SalMatch[60] | 30.16 | 52.31 | 65.54 | 79.15 | - | - | - | - | - | - | - | - |
| GTS[49] | 25.15 | 50.03 | 62.5 | 75.76 | - | - | - | - | - | - | - | - |
| ISR[30] | 27 | - | 61 | 73 | - | - | - | - | - | - | - | - |
| MFA[57] | 32.2 | 66.0 | 79.7 | 90.6 | 18.73 | 46.57 | 63.31 | 78.19 | 12.44 | 33.2 | 47.22 | 63.35 |
| kLFDA[57] | 32.3 | 65.8 | 79.7 | 90.9 | 22.01 | 48.89 | **64.64** | **81.05** | 13.09 | 35.15 | 49.4 | 65.01 |
| LADF [29] | 29.34 | 65 | 79 | 91 | - | - | - | - | - | - | - | - |
| LADF+our feautre | 27.0 | 60.9 | 75.4 | 87.3 | - | - | - | - | - | - | - | - |
| MtMCML[39] | 28.83 | 59.34 | 75.82 | 88.51 | 20.39 | 46.83 | 62.13 | 77.92 | 12.33 | 31.64 | 45.13 | 61.11 |
| LFDA[44] | 24.18 | 52 | 67.12 | 82 | - | - | - | - | - | - | - | - |
| LFDA+our feature | 19.7 | 46.7 | 52.1 | 77.0 | - | - | - | - | - | - | - | - |
| MCE-KISS[55] | 28.2 | - | 72.1 | - | - | - | - | - | 13.6 | - | 49.0 | - |
| RS-KISS[54] | 24.5 | - | 66.6 | - | - | - | - | - | 9.8 | - | 40.5 | - |
| KISSME[20] | 20 | 47 | 62 | 78 | 12 | 33 | 50 | 64 | 6 | 18 | 31 | 48 |
| rPCCA[57] | 22.0 | 54.8 | 71.0 | 85.3 | - | - | - | - | - | - | - | - |
| PCCA[41] | 19.27 | 48.89 | 64.91 | 80.28 | - | - | - | - | 9.27 | 24.89 | 37.43 | 52.89 |
| PRDC[63] | 15.66 | 38.42 | 53.86 | 70.09 | 12.64 | 31.97 | 44.28 | 59.95 | 9.12 | 24.19 | 34.4 | 48.55 |
| RankSVM[45] | 16.27 | 38.23 | 53.73 | 69.87 | 10.63 | 29.7 | 42.31 | 58.26 | 8.87 | 22.88 | 32.69 | 45.98 |
| LMNN[51] | 6.23 | 19.65 | 32.63 | 52.25 | 5.14 | 13.13 | 20.3 | 33.91 | 4.04 | 9.68 | 14.19 | 21.18 |
| ITML[9] | 11.61 | 31.39 | 45.76 | 63.86 | 8.38 | 24.54 | 36.81 | 52.29 | 4.19 | 11.11 | 17.22 | 24.59 |
| PLS[53] | 2.72 | 7.53 | 10.92 | 17.34 | 2.43 | 6.6 | 9.33 | 13.84 | 2.31 | 5.75 | 8.21 | 12.5 |
| L1-norm | 4.18 | 11.65 | 16.52 | 22.37 | 3.8 | 9.81 | 13.94 | 19.44 | 3.55 | 8.29 | 12.27 | 17.59 |




(a) $p = 316$ (b) $p = 532$

Fig. 3. CMC as the function of $r$ on the VIPeR datset for the proposed method and the state-of-the-art approaches. (a). $p = 316$; (b). $p = 532$. (Best viewed in color).

TABLE III

TOP $r$ RANKED MATCHING RATE (%) ON THE GRID DATASET FOR $p = 125$ PEDESTRIANS IN THE PROBE TESTING SET

| Methods | r=1 | r=5 | r=10 | r=20 |
|---|---|---|---|---|
| Proposed | **17.12** | **35.04** | 44.24 | 55.36 |
| Proposed+feature[63] | 11.68 | 27.04 | 37.20 | 48.88 |
| MtMCML[39] | 14.08 | 34.64 | **45.84** | **59.84** |
| rankSVM[62] | 10.24 | 24.56 | 33.28 | 43.68 |
| PRDC[63] | 9.68 | 22.00 | 32.96 | 44.32 |
| L1-norm | 4.40 | 11.68 | 16.24 | 24.80 |
| Bhat. | 4.88 | 14.24 | 20.32 | 26.24 |

*1) VIPeR Dataset:* The proposed method is evaluated and compared with tremendous state-of-the-art metric learning methods, including PLS [53], LMNN [51], ITML [9], RankSVM [45], PRDC [63], PCCA [41] together with its improvement rPCCA [57], KISSME [20] and its improvements RS-KISS [54] and MCE-KISS [55], LFDA [44] and its improvement kLFDA [57], LADF [29], MtMCML [39] together with MFA [57]. Performance of L1 norm is also reported as the baseline for comparison purpose only. We also compare with non-metric learning approaches, such as the most state-of-the-art SalMatch [60], GTS [49], Mid-level Filters [61].

TABLE II shows the matching rate for $p = 316$, $p = 432$ and $p = 532$ on VIPeR dataset at ranks 1, 5, 10 and 20, respectively. Since not all approaches provide results for the same settings, we use "-" to indicate that the approach does not give a result for the corresponding setting. We also show the CMC curves for various comparison methods in the range of the first 20 ranks in Fig. 3.

Since existing metric learning methods report their performance using various kinds of features, we also provide experimental results of our method using the same features as comparison approaches besides the one described in subsection *B*, in order to make fair comparisons. MFA, kLFDA, LADF (see row 12 in TABLE II), LFDA (see row 15), rPCCA, PCCA report the results by using the same feature as ours (see row 2). As can be seen from TABLE II and Fig. 3, our method outperforms LFDA,

TABLE IV

TOP $r$ RANKED MATCHING RATE (%) ON THE iLIDS (MCTS) DATASET FOR $p = 30$, $p = 50$ AND $p = 80$ PEDESTRIANS IN THE TESTING SET

| Methods | p=30 | | | | p=50 | | | | p=80 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 | r=1 | r=5 | r=10 | r=20 |
| Proposed | **56.53** | **82.45** | **91.39** | **97.18** | **46.53** | **73.01** | **84.54** | **93.25** | **35.13** | **59.69** | **72.57** | **85.23** |
| Proposed+feautre[63] | 55.26 | 80.74 | 90.12 | 96.23 | 44.32 | 70.63 | 80.95 | 90.09 | 33.21 | 56.18 | 68.07 | 80.22 |
| GTS[49] | - | - | - | - | 42.39 | 61.35 | 71.04 | 82.21 | - | - | - | - |
| SDC_ocsvm[59] | - | - | - | - | 36.81 | 58.10 | 69.69 | 82.94 | - | - | - | - |
| SDC_knn[59] | - | - | - | - | 33.31 | 57.55 | 68.22 | 83.13 | - | - | - | - |
| LADF[29] | 37.17 | 72.55 | 87.27 | 97.06 | 26.51 | 58.88 | 74.17 | 89.23 | 15.27 | 39.95 | 55.25 | 72.98 |
| LFDA[44] | 42.46 | 75.75 | 88.10 | 97.04 | 36.12 | 64.81 | 78.13 | 89.48 | 28.27 | 52.3 | 65.26 | 78.92 |
| KISSME[20] | 41.14 | 70.71 | 84.22 | 96.28 | 33.44 | 60.09 | 73.98 | 87.09 | 21.71 | 46.10 | 58.36 | 72.78 |
| PRDC[63] | 44.05 | 72.74 | 84.69 | 96.29 | 37.83 | 63.70 | 75.09 | 88.35 | 32.60 | 54.55 | 65.89 | 78.3 |
| Adaboost[14] | 35.58 | 66.43 | 79.88 | 93.22 | 29.62 | 55.15 | 68.14 | 82.35 | 22.79 | 44.41 | 57.16 | 70.55 |
| LMNN[51] | 33.68 | 63.88 | 78.17 | 92.64 | 27.97 | 53.75 | 66.17 | 82.33 | 23.70 | 45.42 | 57.32 | 70.92 |
| ITML[9] | 36.37 | 67.99 | 83.11 | 95.55 | 28.96 | 53.99 | 70.50 | 86.67 | 21.67 | 41.80 | 55.12 | 71.31 |
| MCC[11] | 40.24 | 73.64 | 85.87 | 96.65 | 31.28 | 59.30 | 75.62 | 88.34 | 12.00 | 33.66 | 47.96 | 67.00 |
| Xing's[56] | 31.80 | 62.62 | 77.29 | 90.63 | 27.04 | 52.28 | 65.35 | 80.70 | 23.18 | 45.24 | 56.90 | 70.46 |
| PLS[53] | 25.76 | 57.36 | 73.57 | 90.31 | 22.10 | 46.04 | 59.95 | 78.68 | 18.32 | 38.23 | 49.68 | 64.95 |
| L1-norm | 35.31 | 64.62 | 77.37 | 91.35 | 30.72 | 54.95 | 67.99 | 82.98 | 26.73 | 49.04 | 60.32 | 72.07 |
| Bhat. | 31.77 | 61.43 | 74.19 | 89.53 | 28.42 | 51.06 | 64.32 | 78.77 | 24.76 | 45.35 | 56.12 | 69.31 |

TABLE V

TOP $r$ RANKED MATCHING RATE (%) ON THE CAVIAR4REID DATASET FOR $p = 36$ PEDESTRIANS IN THE TESTING SET

| Methods | r=1 | r=5 | r=10 | r=15 | r=20 |
|---|---|---|---|---|---|
| Proposed | **41.17** | **73.51** | **84.99** | **90.84** | 94.37 |
| MFA[57] | 38.4/40.2 | 69.0/70.2 | 83.6/83.9 | - | 95.1/95.1 |
| kLFDA[57] | 36.2/39.1 | 64.0/66.8 | 78.7/80.9 | - | 92.2/93.4 |
| LFDA[44] | 31.7/35.2 | 56.1/59.9 | 70.4/73.7 | - | 86.9/88.8 |
| LADF [29] | 25.8/28.9 | 61.4/62.5 | 78.6/79.2 | - | 93.6/93.3 |
| rPCCA[57] | 34.0/35.1 | 67.5/67.2 | 83.4/83.1 | - | 95.8/**95.6** |
| PCCA[41] | 33.4/33.9 | 67.2/67.8 | 83.1/83.5 | - | 95.7/**95.6** |
| KISSME[20] | 31.4/34.9 | 61.9/64.7 | 77.8/79.7 | - | 92.5/93.3 |
| LMNN[51] | 18.64 | 36.71 | 48.98 | 61.77 | - |
| ITML[9] | 27.21 | 56.76 | 75.23 | 85.71 | - |
| PS[7] | 15.49 | 48.02 | 70.89 | 85.17 | - |
| SDALF[7] | 12.25 | 44.78 | 65.93 | 82.46 | - |

rPCCA and PCCA for different values of $p$ at rank 1. It should be noted that LFDA and LADF also report results by using other features (see row 11 and 14, respectively). While, our method still achieves higher rank 1 matching rate. Though the performance of our method is about 1% lower than MFA and kLFDA at rank 1, our method yields 2.13% and 5.41% higher rank 1 matching rate than kLFDA and MFA for $p = 432$, respectively. And when $p = 532$, RMLLC(R) boosts the performance of kLFDA and MFA by 4.45% and 5.12% at rank 1, indicating that our method has better scalability to the size of training data. MtMCML, PRDC, RankSVM, LMNN, ITML, PLS, L1-norm report their performance by using the feature adopted in [63]. TABLE II also lists the performance of our method by using the same feature (see row 4). As can be seen, our method is significantly superior to PRDC, RankSVM, LMNN, ITML, PLS, but yields slightly lower performance than MtMCML when $p = 316$. However, the performance of our method at rank 1 is 0.67% and 2.63% higher than MtMCML when $p = 432$ and $p = 532$.

Therefore, our method has its advantages when the training data decreases. KISSME and RS-KISS provide experimental results by using the feature utilized in [20]. We also list the performance of our method by using the same feature in row 3 of TABLE II. It is obvious that our method is significantly superior to KISSME and RS-KISS.

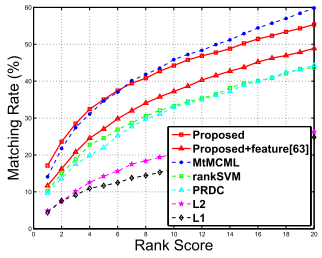
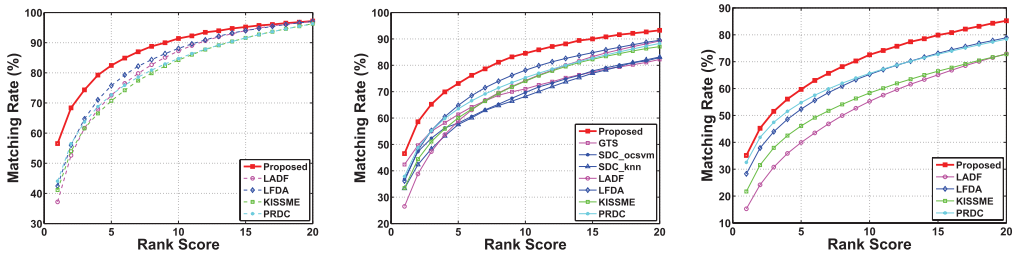



Fig. 4. CMC as the function of $r$ on the GRID dataset for the proposed method and the state-of-the-art approaches. (Best viewed in color).

Moreover, as shown in TABLE II and Fig. 3, the proposed method also yields better performance than most state-of-the-art non-metric learning approaches. For rank 1 ($p = 316$), RMLLC(R) is 1.11% higher than SalMatch, 3.27% higher than ISR, and 6.12% higher than GTS. While, it should be noted that Mid-Level filters combining LADF obtains a 43.39% rank 1 matching rate for $p = 316$, which is significantly superior to other metric learning methods including ours. The significant improvement is due to the middle level representation proposed in [61], which could remove the influences caused by background clutter and pose variation. In contrast, these influences are difficult to be alleviated by using the low level features that are widely used by current metric learning approaches.

*2) GRID Dataset:* In this dataset, the state-of-the-art performances were obtained by using MtMCML [39], PRDC [63] and rankSVM [45]. We also report two baseline methods by directly using L1 norm [46] and Bhat. norm [16]. In all our experiments, we utilize the same train/test splits utilized by the aforementioned comparison approaches. Fig. 4 shows the CMC curves for these approaches in the range of the first 20 ranks. For all the comparison approaches, we reproduce the figures in [39]. We also report the comparison results at rank 1, 5, 10 and 20 in TABLE III.

As can be seen from TABLE III and Fig.4, our method achieves the best rank 1 matching rate. Actually, RMLLC(R)




(a) $p = 30$ (b) $p = 50$ (c) $p = 80$

Fig. 5. CMC as the function of r on the iLIDS (MCTS) dataset for the proposed method and the state-of-the-art approaches. (a). $p = 30$; (b). $p = 50$; (c). $p = 80$. (Best viewed in color).

is 3.04% higher than MtMCML, 6.88% higher than RankSVM and 7.44% higher than PRDC at rank 1. It should be noted that the proposed method only obtains a slight performance boost at rank 5 (only 0.4%) compared with MtMCML, and performs even worse than MtMCML after rank 6. Since MtMCML and PRDC report their performance by using the feature described in [63], we therefore also provide the results of RMLLC(R) by adopting the same feature in row 3 of TABLE III. Though the performance of our method deteriorates, RMLLC(R) still outperforms PRDC, but archives lower performance than MtMCML. This is due to the fact that MtMCML explicitly learns an individual metric for each camera pair, while RMLLC(R) proposes to learn a universal metric for all cameras. However, MtMCML needs to know the identity of the camera that each image was captured by before training, which is often unavailable in practice. Moreover, MtMCML needs to learn $(k + 1) \times k$ Mahalanobis distance metrics for a surveillance network consisting of $k$ cameras, of which the computation cost would become very high for large $k$. In contrast, our method learns a universal metric over the whole camera network without any camera identity information, thus is more applicable in practice.

*3) iLIDS (MCTS) Dataset:* In this dataset, under the single-shot person re-identification evaluation protocol, the current highest rank 1 matching rate for $p = 50$ was obtained by GTS [49], while it didn't report experimental results for $p = 30$, and $p = 80$. As to $p = 30$ and 80, to the best of our knowledge, the highest performance was reported by [63] using PRDC. Thus, we compare our method with GTS, and PRDC, together with other reported approaches including SDC_knn [59], SDC_ocsvm [59], Adaboost [14], LMNN [51], ITML [11], Xing's [56], and PLS [53]. As usual we report results by using L1 norm [46] and Bhat. norm [16] as the baseline methods.

In TABLE IV, we summarize all the experimental results for $p = 30, 50, 80$, at rank 1, 5, 10, and 20. Fig. 5 provides the corresponding CMC curves in the ranges of the first 20 ranks. For GTS, SDC_knn, SDC_ocsvm, we reproduce the figures in [49]. From Fig. 5, RMLLC(R) yields an overall better performance than the best result given by the other approaches for all settings in the first 20 ranks. Specifically, as shown in TABLE IV, for rank 1 ($p = 50$), the proposed method achieves a 4.14% higher matching rate than GTS, 6.7% higher than PRDC, 10.41% higher than LFDA, 13.09% higher than KISSME, 20.02% higher than LADF. When $p = 30$, our method obtains 12.48%, 14.07%, 15.30%, 19,36% improvements over PRDC, LFDA, KISSME, LADF at rank 1, respectively. For $p = 80$, the rank 1 matching rate of RMLLC(R) is 2.53% higher than PRDC, 6,86% higher than LFDA, 13.42% higher than KISSME, and 19.86% higher than LADF. These comparison results show that our method performs significantly better than the reported approaches. It should be noted here that [63] reported experimental results of PRDC, Adaboost, LMNN, ITML, MCC, Xing's, PLS and L1 by using different features from this paper. For fair comparison, we also report the performance of our method by using the same feature in row 3 of TABLE IV. It can be seen that our methods is still significantly superior to PRDC, Adaboost and etc.

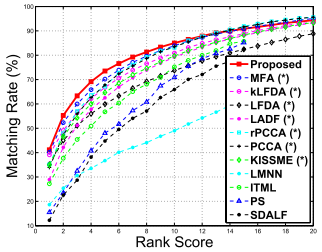



Fig. 6. CMC as the function of $r$ on the CAVIAR4REID dataset for the proposed method and the state-of-the-art approaches ($p = 36$): (*) denotes the best results using various features. (Best viewed in color).

*4) CAVIAR4REID Dataset:* In this dataset, we compare our method with the current reported state-of-the-art metric learning approaches, including MFA [57], LADF [29], LFDA [44], kLFDA [57], KISSME [20], PCCA [41], rPCCA [57] by using the same feature representation, together with baseline methods such as LMNN [51], ITML [9], PS [7] and SDALF [10]. We reproduce the results of MFA, LADF, LFDA, kLFDA, KSSME, PCCA, rPCCA from [57]. It should be noted that [57] also reports results of these methods by using different features. We select the highest ones, and lists them in Table V (after the mark "/") and Fig.6. For LMNN, ITML, PS and SDALF, we reproduce the results from [29].

From Table V, we can observe that RMLLC(R) outperforms all comparison approaches in the first 20 ranks, using the same feature representation. For instance, our method yields a 2.77% higher matching rate than the best comparison result given by MFA at rank 1, and is 4.97%, 9.47%, 15.37%, 7.17%, 7.77%,

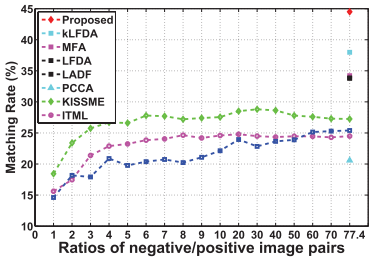



Fig. 7. Effect of different ratios of negative/positive samples.

9.77% higher than kLFDA, LFDA, LADF, rPCCA, PCCA, KISSME, respectively. By using more complicated feature representations, the rank 1 recognition accuracy of MFA could achieve as high as 40.2% (as shown in Table V after the marks "/"). However, our method is still about 1% superior to MFA, despite using relatively less discriminative feature representation.

We have seen consistent improvements by using our proposed method, compared with state-of-the-art approaches. The increased matching rate may benefit from the following three aspects: 1). the proposed method incorporates more richer similarity/dissimilarity information by using listwise constraints, which could reinforce the discriminate capability of the learnt transformation matrix $W$; 2). the kernels support nonlinear mapping, which could exploit the underlying nonlinear structure of data; 3). the mapping matrix $W$ is implicitly a low rank matrix, and the additional regularization term further alleviates the overfitting problem.

### *F. Evaluation on Ratios of Negative/Positive Training Samples*

In this subsection, we will discuss how the number of negative samples used for training effects the performance of metric learning methods.

Specifically, we use all image pairs with the same identity in the training data as positive samples, and randomly select image pairs with different identities as negative samples, of which the number is $r$ times the number of positive samples. We call $r$ the ratio of negative/positive samples. For evaluation, we choose PCCA, KISSME and ITML, considering that they are more suitable for our experimental settings, and are state-of-the-art metric learning approaches. We report the rank 1 matching rate of PCCA, KISSME, ITML on iLIDS dataset for $p = 60$, by varying the value of $r$. Since kLFDA, MFA, LFDA, LADF and our method implicitly or explicitly use all training samples, they could be seen as supervised metric learning with full supervision. We therefore also reports their results with full supervision, in order to compare their ability in utilizing negative training samples.

As shown in Fig.7, we can observe that when $r \leq 4$, i.e., the negative samples are sparsely selected for training, PCCA, KISSME and ITML all perform poorly. When $r \leq 10$, the matching rates of PCCA, KISSME, ITML roughly raise as the increase of negative training samples. As $r$ further increases to its maximum 77.4 where full supervision is used, different methods yield distinct performances. PCCA slightly fluctuates around the highest matching rate. While, the performance of KISSME slightly decreases after achieving its maximum. In contrast, the performance of ITML is consistently boosted. We can also see that our method outperforms all comparison approaches when full supervision is adopted.

These observations imply that the negative samples indeed could provide useful discriminative information, and enhance the performance of metric learning methods, though not all metric learning approaches could consistently benefit from the increased negative samples. Therefore, it is beneficial to consider all negative samples during training such as ours.

TABLE VI
COMPUTATION COST OF THE PROPOSED METHOD (SECOND)

| VIPeR, p=316 | VIPeR, p=432 | VIPeR, p=532 | GRID |
|---|---|---|---|
| 7.47 | 1.65 | 0.26 | 0.38 |
| iLIDS, p=30 | iLIDS, p=50 | iLIDS, p=80 | CAVIAR |
| 1.74 | 0.73 | 0.17 | 5.42 |

### *G. Computation Cost*

In this subsection, we will briefly analyze the computation complexity and empirical computation cost of the proposed method.

Though the proposed method incorporates the full list-wise similarity/dissimilarity information, it has relatively low computation complexity. As to Algorithm 1, the main computation cost comes from the eigen-decomposition of $L \in R^{N \times N}$ which requires $O(N^3)$ operations, and the matrix inverse of the $D \times D$ matrix $X^T X + \gamma I_D$, which requires $O(D^3) + O(D^2 N)$ operations. Thus, the overall complexity of Algorithms 1 is $O(D^3) + O(N^3)$. Similar to Algorithm 1, we can derive that the complexity of Algorithm 2 is $O(N^3)$. As to Algorithm 3, the main computation comes from step 1, and steps 2-3 in the iterative step. Step 1 could be solved by Algorithm 2, yielding $O(N^3)$ operations. Both Step 2 and step 3 require $O(N^3)$ operations. Thus, for $k$ iterations, the overall complexity of Algorithm 3 is $O\left((k+1)(N^3)\right)$. In practice, even $k \leq 10$ iterations could yield promising results. Thus, the practical complexity of Algorithm 3 is still $O(N^3)$.

We further test the empirical computation cost of the proposed method using Algorithm 3, where we fix parameters yielding results in Table I. It should be noted that the running time for kernel computation is not accounted, considering that we only focus on efficiency of the training process of the proposed method. TABLE VI summarizes the average training time of RMLLC(R) on four datasets, where all the experiments were conducted on a PC with Intel(R) Xeon(R) CPU (2 cores, 2.50GHz) and 16 GB RAM. As shown in TABLE VI, the computation cost of the proposed method is modest.

### *H. Discussion*

We have been focusing on the proposed method in the background of person re-identification in previous sections. While, RMLLC (RMLLC(R)) is also applicable

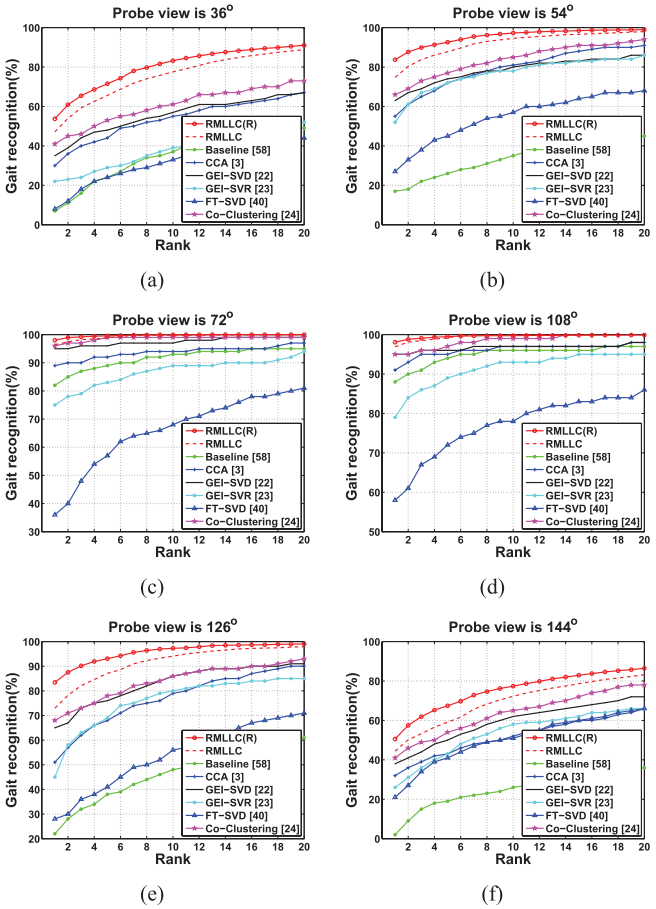



Fig. 8. CMC curves for the proposed method and comparison approaches on CASIA gait database B, when the gallery view is 90°: (a). recognition results while the probe view is 36°; (b). recognition results while the probe view is 54°; (c). recognition results while the probe view is 72°; (d). recognition results while the probe view is 108°; (e). recognition results while the probe view is 126°; (f). recognition results while the probe view is 144°. (Best viewed in color).

for other identification tasks and could yields competitive performance, especially for the task of matching people under variations such as illumination, pose, camera view. For instance, we apply our method to the cross-view gait recognition problem, which has been extensively addressed recently [3], [12], [24], [34], [40]. Fig.8 shows experimental results for the cross view gait recognition on CASIA gait database B under the gallery view 90°, from which our proposed method could obtain significantly better performance than the state-of-the-art methods such as Co-clustering [24], GEI-CCA [3], GEI-SVD [22], FT-SVD [40], GEI-SVR [23], View-rectification [12], and the baseline method, i.e., L2 [58].

## VI. Conclusions

In this paper, we formulate person re-identification as an image retrieval task, and propose a novel Relevance Metric Learning method with Listwise Constraints (RMLLC). In order to make full use of available similarity information from training data, RMLLC predefines similarity lists, and learns a similarity metric by forcing it to conserve the predefined listwise similarities after projection. Since the predefined binary listwise similarities generally fail to capture relative similarity information, we subsequently employ a rectification term to automatically exploit the relative relevance, and propose an improved method called RMLLC with rectification (RMLLC(R)). The metric and the rectification term are jointly learnt by using an efficient alternating iterative optimization method. Experimental results validate that RMLLC with rectification performs better than RMLLC, and is substantially superior to state-of-the-art approaches. Although our proposed RMLLC(R) model is originally designed to solve the person re-identification problem, it is also applicable for other person identification tasks. In this work, we apply it to the cross view gait recognition. Experimental results on the widely used CASIA gait database B for gait recognition show that our method performs significantly better than the current best results.

## References

[1] X. Bai, X. Yang, L. J. Latecki, W. Liu, and Z. Tu, "Learning context-sensitive shape similarity by graph transduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 861–874, May 2010.

[2] X. Bai, B. Wang, C. Yao, W. Liu, and Z. Tu, "Co-transduction for shape retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2747–2757, May 2012.

[3] K. Bashir, T. Xiang, and S. Gong, "Cross-view gait recognition using correlation strength," in *Proc. BMVC*, 2011, pp. 1–11.

[4] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 898–903, 2012.

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

[6] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proc. IJCAI*, 2015, pp. 3402–3408.

[7] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. BMVC*, 2011, pp. 68.1–68.11.

[8] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *Proc. Adv. NIPS*, 2001, pp. 367–373.

[9] J. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th ICML*, 2007, pp. 209–216.

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 2360–2367.

[11] A. S. Globerson and S. Roweis, "Metric learning by collapsing classes," in *Proc. NIPS*, 2006, pp. 451–458.

[12] M. Goffredo, I. Bouchrika, J. N. Carter, and M. S. Nixon, "Self-calibrating view-invariant gait biometrics," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 997–1008, Aug. 2010.

[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.

[14] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. PETS*, 2007, pp. 1–7.

[15] E.-L. Hu and J. T. Kwok, "Scalable nonparametric low-rank kernel learning using block coordinate descent," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. pp, no. 99, doi: 10.1109/TNNLS.2014.2361159, 2014.

[16] W. Hu, M. Hu, X. Zhou, T. Tan, J. Luo, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 663–671, Apr. 2006.

[17] P. Jain, B. Kulis, J. V. Davis, and I. S. Dhillon, "Metric and kernel learning using a linear transformation," *J. Mach. Learn. Res.*, vol. 13, pp. 519–547, Mar. 2012.

[18] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2005, pp. 26–33.

[19] K. Jungling and M. Arens, "View-invariant person re-identification with an implicit shape model," in *Proc. 8th IEEE Int. Conf. AVSS*, Aug./Sep. 2011, pp. 197–202.

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

[21] B. Kulis, M. A. Sustik, and I. S. Dhillon, "Low-rank kernel learning with Bregman matrix divergences," *J. Mach. Learn. Res.*, vol. 10, pp. 341–376, Feb. 2009.

[22] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," in *Proc. IEEE 12th ICCV Workshops*, Sep./Oct. 2009, pp. 1058–1064.

[23] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 974–981.

[24] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang, "Recognizing gaits across views through correlated motion co-clustering," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 696–709, Feb. 2013.

[25] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, Jul. 2013.

[26] J. T. Kwok and I. W. Tsang, "Learning with idealized kernels," in *Proc. ICML*, 2003, pp. 1–8.

[27] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 4321–4328.

[28] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 152–159.

[29] 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. CVPR*, Jun. 2013, pp. 3610–3617.

[30] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1629–1642, Aug. 2015.

[31] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1601–1608.

[32] C. C. Loy, C. Liu, and S. G. Gong, "Person re-identification by manifold ranking," in *Proc. 20th IEEE ICIP*, Sep. 2013, pp. 3567–3571.

[33] E. Y. Liu, Z. Guo, X. Zhang, V. Jojic, and W. Wang, "Metric learning from relative comparisons by minimizing squared residual," in *Proc. IEEE 12th ICDM*, Dec. 2012, pp. 978–983.

[34] J. Lu, G. Wang, and P. Moulin, "Human identity and gender recognition from gait sequences with arbitrary walking directions," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 1, pp. 51–61, Jan. 2014.

[35] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou, "Neighborhood repulsed metric learning for kinship verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 331–345, Feb. 2014.

[36] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu, "Human re-identification by matching compositional template with cluster sampling," in *Proc. IEEE ICCV*, Dec. 2013, pp. 3152–3159.

[37] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Proc. ECCV Workshop*, 2012, pp. 413–422.

[38] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image Vis. Comput.*, vol. 32, nos. 6–7, pp. 379–390, 2014.

[39] 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.

[40] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. 9th ECCV*, 2006, pp. 151–163.

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

[42] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. 10th ACCV*, 2010, pp. 709–720.

[43] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, NY, USA: Springer-Verlag, 1999.

[44] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 3318–3325.

[45] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. BMVC*, 2010, pp. 21.1–21.11.

[46] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–8.

[47] H. Wang, F. Nie, and H. Wang, "Multi-view clustering and feature learning via structured sparsity," in *Proc. 30th ICML*, 2013, pp. 352–360.

[48] D. Wang and X. Tan, "Robust distance metric learning in the presence of label noise," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1321–1327.

[49] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *Proc. BMVC*, 2014, pp. 1–11.

[50] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. 13th ECCV*, 2014, pp. 688–703.

[51] 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.

[52] Z. Wu, Y. Li, and R. J. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1095–1108, May 2014.

[53] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *Proc. 22nd Brazilian Symp. Comput. Graph. Image Process.*, 2009, pp. 322–329.

[54] 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.

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

[56] 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. NIPS*, 2002, pp. 521–528.

[57] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. 13th ECCV*, 2014, pp. 1–16.

[58] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th ICPR*, 2006, pp. 441–444.

[59] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 3586–3593.

[60] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE ICCV*, Dec. 2013, pp. 2528–2535.

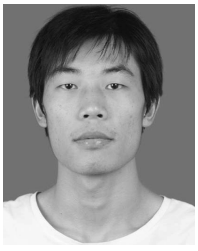
[61] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 144–151.

[62] W.-S. Zheng, S. Gong, and T. Xiang, "Associate groups of people," in *Proc. BMVC*, 2009, pp. 1–11.

[63] 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. 2013.

[64] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. pp, no. 99, doi: 10.1109/TPAMI.2015.2453984, 2015.

[65] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang, "Hierarchical Gaussianization for image classification," in *Proc. IEEE 12th ICCV*, Sep./Oct. 2009, pp. 1971–1977.



**Jiaxin Chen** received the B.S. and M.S. degrees in mathematics from Beihang University, Beijing, China, in 2009 and 2012, respectively, where he is currently pursuing the Ph.D. degree with the Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering.

His research interests include pattern recognition, computer vision, and machine learning, in particular, on person reidentification and metric learning.

**Zhaoxiang Zhang** (M'08–SM'15) received his B.S. degree in electronic science and technology from the University of Science and Technology of China, in 2004, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2009. In 2009, he joined the School of Computer Science and Engineering as an Assistant Professor and promoted to be an Associate Professor in 2012. His research interests include computer vision, pattern recognition, and machine learning. Specifically, he has focused on video surveillance and biometric analysis. He has published around 70 papers in reputable journals and conferences. He is the Associate Editor of *Neurocomputing*, involved in the Editorial Board of the Frontiers of Computer Science, the Program Committee Members of 10+ international conferences and the Reviewer of 20+ international journals. He has been granted several awards, including the MOE New Century Excellent Talents, Beijing Youth Talents, etc.

**Yunhong Wang** (M'98) received the B.S. degree in electronics engineering from Northwestern Polytechnical University, in 1989, and the M.S. and Ph.D. degrees in electronics engineering from the Nanjing University of Science and Technology, in 1995 and 1998, respectively. She was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 1998 to 2004. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, where she is currently the Director of the Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media.

Her research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing. She is a member of the IEEE Computer Society.