Person Reidentification via Discrepancy Matrix and Matrix Metric

Zheng Wang, Ruimin Hu, *Senior Member, IEEE*, Chen Chen, *Member, IEEE*, Yi Yu, Junjun Jiang, *Member, IEEE*, Chao Liang, and Shin'ichi Satoh, *Member, IEEE*

***Abstract*—Person reidentification (re-id), as an important task in video surveillance and forensics applications, has been widely studied. Previous research efforts toward solving the person re-id problem have primarily focused on constructing robust vector description by exploiting appearance's characteristic, or learning discriminative distance metric by labeled vectors. Based on the cognition and identification process of human, we propose a new pattern, which transforms the feature description from characteristic vector to discrepancy matrix. In particular, in order to well identify a person, it converts the distance metric from vector metric to matrix metric, which consists of the intradiscrepancy projection and interdiscrepancy projection parts. We introduce a consistent term and a discriminative term to form the objective function. To solve it efficiently, we utilize a simple gradient-descent method under the alternating optimization process with respect to the two projections. Experimental results on public datasets demonstrate the effectiveness of the proposed pattern as compared with the state-of-the-art approaches.**

***Index Terms*—Discrepancy matrix, matrix metric, metric projection, person reidentification (re-id).**

I. Introduction

PERSON reidentification (re-id) is the task of visually matching images of the same person, obtained from different cameras distributed over nonoverlapping scenes [1]. It has drawn significant attentions in recent years [2]–[9] due to its important applications in video surveillance [10]–[12]. Although face [13]–[16] and gait [17], [18] may be more reliable biometrics to identify a person, they are not always available due to low resolution and pose variations of an individual in typical surveillance scenario [19]–[21]. Therefore, the appearance of individuals is mainly exploited for person re-id. Generally, given a probe person image taken from camera A, the re-id algorithm aims to search for images of the same person from the gallery captured by camera B. Previous research efforts for solving the re-id problem have primarily focused on the following two aspects.

A. Feature Description

Many approaches have been proposed to develop discriminative visual descriptions that are robust to distinguish different persons in various cameras, such as ensemble of localized features (ELF) [22], symmetry-driven accumulation of local features (SDALF) [23], salient color-name-based color descriptor (SCNCD) [24], local maximal occurrence (LOMO) [25], Gaussian of Gaussian (GoG) descriptor [26], and deep convolutional neural network (CNN) approaches [27]–[31]. In general, a feature vector $x \in \mathbb{R}^{N_f \times 1}$ is always used to describe an image I by these methods [32], where N_f denotes feature dimension.

B. Distance Metric

There are also many efforts toward learning optimal matching metrics under which instances belonging to the same person are closer than those belonging to different persons, such as probabilistic relative distance comparison (PRDC) [19], keep it simple and straightforward metric learning (KISSME) [33], locally adaptive decision functions (LADF) [34], local Fisher discriminant analysis (LFDA) [35], and cross-view quadratic discriminant analysis (XQDA) [25]. Generally, most of these methods would learn a distance metric $\mathbf{M} \in \mathbb{R}^{N_f \times N_f}$, then the distance of an image pair (I_p^A, I_q^B) is calculated by $d(x_p^A, x_q^B) = (x_p^A - x_q^B)^\top \mathbf{M}(x_p^A - x_q^B)$, where the superscripts A and B stand for the camera label, and the subscripts p and q stand for the person ID. Actually, by performing eigenvalue decomposition on $\mathbf{M}$ with $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$, the distance can be rewritten as (1). With this definition, it is easy to see that the essence of the metric learning-based method is to seek a projection that

Manuscript received December 11, 2016; revised May 21, 2017, August 28, 2017, and August 30, 2017; accepted September 12, 2017. This work was supported in part by the National Nature Science Foundation of China under Grant U1611461, Grant 61231015, Grant 61671336, and Grant 61501413, in part by the National High Technology Research and Development Program of China under Grant 2015AA016306, in part by the Technology Research Program of Ministry of Public Security under Grant 2016JSYJA12, in part by the Hubei Province Technological Innovation Major Project under Grant 2016AAA015, and in part by the Nature Science Foundation of Jiangsu Province under Grant BK20160386. This paper was recommended by Associate Editor L. Shao. *(Corresponding author: Ruimin Hu.)*

Z. Wang, R. Hu, and C. Liang are with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China, also with the Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China, and also with the Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China (e-mail: wangzwhu@whu.edu.cn; hurm1964@gmail.com; cliang@whu.edu.cn).

C. Chen is with the Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816 USA (e-mail: chenchen870713@gmail.com).

Y. Yu and S. Satoh are with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: yiyu@nii.ac.jp; satoh@nii.ac.jp).

J. Jiang is with the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: junjun0595@163.com).

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/TCYB.2017.2755044

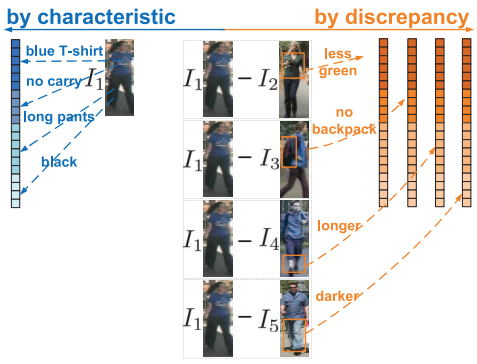


Fig. 1. Toy example to make a comparison of descriptions by characteristic and discrepancy, respectively. We suppose that different parts of a vector denote different properties of the corresponding person. Left: person in I_1 can be described as a feature vector by the characteristic of its appearance. Right: person can be described as a combination of four vectors by discrepancies with $I_2 - I_5$.

transforms the original image features into a new feature space

$$d\left(x_p^A, x_q^B\right) = \left\|\mathbf{L}\left(x_p^A - x_q^B\right)\right\|^2. \tag{1}$$

Almost all of the state-of-the-art approaches follow the same routine. They represent each person image using a feature vector, no matter hand-crafted or deep-learned, based on the person's own appearance. Then, they compare image pairs by their feature vector distances [36]. However, when describing a person, we can exploit not only the characteristic of his/her own appearance, but also the appearance relationships with the others. Recently, An *et al.* [37] proposed a reference descriptor (RD) [37] method. They introduced a reference set and selected typical identities from the reference set to reconstruct each person. The reconstruction weights were used to describe the corresponding person. In our opinion, the RD method focuses on discovering global and coarse-grained reconstruction relationships with reference identities as a whole, and ignores its local difference relationship with each part of each reference identity, which is fine-grained and may contain important information. We name the difference relationship as discrepancy. To describe a person, his/her appearance discrepancies with the reference set are exploited in this paper. Fig. 1 shows a toy example. We suppose that different parts of a vector denote different properties of the corresponding person. As Fig. 1(left) shows, in a traditional manner, the person in I_1 is described as wearing a blue T-shirt and long pants, without any carryings. The upper body is blue and the lower body is black. These characteristics are extracted as a feature vector. On the other hand, we can also represent the person by the discrepancies with other persons. Let us see Fig. 1(right). We provide another way to describe I_1 as follows: the color of the T-shirt of I_2 shifts to green comparing with that of I_1, I_1 does not hold a backpack but I_3 holds, the pants of I_1 are longer than those of I_4, and the color of the pants of I_1 is darker than that of I_5. So the toy example shows that a person can be described not only by his/her characteristic, but also by his/her discrepancies with the others. Generally, we utilize the feature vector to represent a person. To this end, each discrepancy of two images can be denoted as the difference of the feature vectors, i.e., one subtracts another. In this toy case, the image I_1 is described by a combination of four discrepancy vectors (or a discrepancy matrix) instead.

If each discrepancy between two persons is represented by a vector, all the discrepancies with a reference set will form a matrix to describe the corresponding person. Comparing to the reference description vector [37], the discrepancy matrix captures more diverse characteristics, and it also reduces the effect caused by the variation of camera conditions through introducing a couple of reference sets from two cameras. The details are presented in Section III-B. In the discrepancy matrix pattern, the distance of each image pair should be calculated based on a pair of discrepancy matrices. Therefore, existing metric learning approaches, which focus on generating vector metric, are obviously inappropriate for the proposed description. It is easy to recognize that two discrepancy matrices from the same person should be similar to each other, while those from different persons should be dissimilar. To this end, an effective matrix distance metric learning method is proposed, including intradiscrepancy and interdiscrepancy projections.

The contributions of this paper are as follows.

1) *From Characteristic Vector to Discrepancy Matrix:* We propose a new idea to describe a person image, which is represented by the discrepancies with a set of images rather than the characteristic of the image itself. It is proved that compared with characteristic vector, discrepancy matrix is more discriminative and effective for person re-id.
2) *From Vector Metric to Matrix Metric:* We propose a matrix metric for the re-id task, which consists of intradiscrepancy projection part and interdiscrepancy projection part. The matrix metric is learned by simultaneously considering the consistency constraint (pulling two discrepancy matrices from the same person close) and the discrimination constraint (pushing two discrepancy matrices from different persons far away) in the training stage.
3) *A New Pattern:* We provide a new re-id pattern which reidentifies and ranks the images by discrepancy matrix and matrix metric (DM^3). Extensive experimental evaluations on benchmark datasets demonstrate the effectiveness of the proposed pattern. It is also worth noting that the proposed re-id pattern is independent of the choices of feature descriptors. Combined with the state-of-the-art feature extraction methods, for example deep learning method, better results can be obtained.

The rest of this paper is organized as follows. In Section II, a brief review of related work for re-id is given. In Section III, we present the motivation of our approach and formally define the new problem. In Section IV, we illustrate the details of the proposed matrix metric learning. Section V reports experimental results and analysis. Finally, Section VI concludes this paper. Table I summarizes the notations used in this paper.

TABLE I
SUMMARY OF NOTATIONS

Symbol	Description
A (B)	Camera A (B)
o_i	Person with the ID number i
M	Number of the training person images in each camera
N	Number of the testing person images in camera B
x_i^A (x_i^B)	Feature vector of person o_i in camera A (B)
N_f	Number of dimensions of feature vector
N_r	Number of reference person images in each camera
f_i^A (f_i^B)	Feature vector of reference person o_i in camera A (B)
x_p^A	Feature vector of the probe person o_p in camera A
x_q^B	Feature vector of the gallery person o_q in camera B
$\mathbf{X}_i^A$ ($\mathbf{X}_i^B$)	Discrepancy matrix of person o_i in camera A (B)
$\mathbf{X}_p^A$	Discrepancy matrix of the gallery person o_p in camera A
$\mathbf{X}_q^B$	Discrepancy matrix of the gallery person o_q in camera B
$\mathbf{L}$	Feature projection of traditional methods
$\mathbf{L}_1$	Intra-discrepancy projection
$\mathbf{L}_2$	Inter-discrepancy projection
E_{con}	Consistent term
E_{dis}	Discriminative term
E_{spr}	Sparse term
μ	Weight for the sparse term
s_k	A triple sample ($\mathbf{X}_i^A, \mathbf{X}_i^B, \mathbf{X}_j^B$), $i \neq j$
$\mathbf{Z}_i$	Difference of $\mathbf{X}_i^A$ and $\mathbf{X}_i^B$
$\mathbf{U}_k$	Difference of $\mathbf{X}_i^A$ and $\mathbf{X}_i^B$ in s_k
$\mathbf{V}_k$	Difference of $\mathbf{X}_i^A$ and $\mathbf{X}_j^B$ in s_k
$\mathbf{l}_m$	The m-th row of $\mathbf{L}_2$
$\mathbf{D}$	A diagonal matrix
d_{mm}	The m-th diagonal element of $\mathbf{D}$
$l_\beta(z)$	The generalized logistic loss function
$g(z)$	The derivative of logistic loss function $l_\beta(z)$
λ_1 (λ_2)	Step length at each gradient update
ε	A small positive value

II. RELATED WORK

In this section, we give a brief review of the related work on person re-id. Current re-id research can be generally categorized into two classes: 1) feature description-based and 2) distance metric-based approaches.

Feature description approaches aim to construct discriminative visual descriptions. Generally, this kind of approaches can be divided into hand-crafted-based and deep learning-based. The hand-crafted descriptions in re-id task are designed by exploiting special appearance characteristics of pedestrians. Wang *et al.* [38] studied an appearance model to capture the spatial distribution of the appearance. Gray and Tao [22] performed viewpoint invariant description using an ELF. Farenzena *et al.* [23] described the appearance image with segmented regions by using symmetry and asymmetry perceptual principles. Ma *et al.* [39] combined biologically inspired features and covariance descriptors (BiCov). Layne *et al.* [40] learned a selection and weighting of mid-level semantic attributes to describe people. Kviatkovsky *et al.* [41] used shape context descriptors to represent the intradistribution structure, which are invariant in different lighting conditions. Zhao *et al.* [42] assigned salience to each patch in an unsupervised manner. Yang *et al.* [24] proposed an SCNCD to represent person image. Eiselein *et al.* [43] fused multiple basic features, such as color histograms, SURF [44], and designed an efficient person descriptor which is fast and meets the practical need of low runtimes. Liao *et al.* [25] analyzed the horizontal occurrence of local features, and maximized the occurrence to make a stable representation against viewpoint changes. Matsukawa *et al.* [26] modeled each person image region as a set of multiple Gaussian distributions in which each Gaussian represents the appearance of a local patch. Recently, deep-learned descriptions are emerged for the re-id task. Li *et al.* [27] utilized a unified deep architecture to learn a filter for re-id. Ding *et al.* [28] presented a scalable distance driven feature learning framework based on the deep neural network for re-id. Zhang *et al.* [29] developed deep bit-scalable hashing codes to represent raw images. Wang *et al.* [30] combined four CNNs, each of which embeds images from different scale or different body part. Overall, all of these methods focus on the person's own appearance, and represent each person image as a feature vector.

Besides the characteristics of the person image itself, the relationships with other identities can also be exploited. An *et al.* [37] utilized a reference set to describe a person. However, they select some typical reference features to construct a vector, while we exploit discrepancies rather than original reference features to form a matrix for the re-id problem.

The distance metric approaches pay attention to find a proper distance measure. Hirzer *et al.* [45] and Dikmen *et al.* [46] employed LMNN [47] to learn the optimal metric for re-id. Zheng *et al.* [19] learned a Mahalanobis distance metric with a PRDC. Köstinger *et al.* [33] used Gaussian distribution to fit pair-wise samples and got a simpler metric function. Tao *et al.* [48] presented a regularized smoothing KISS metric learning by seamlessly integrating smoothing and regularization techniques for robustly estimating the covariance matrices. Mignon and Jurie [49] introduced pairwise constrained component analysis (PCCA) to learn distance metric from sparse pairwise similarity/dissimilarity constraints in high-dimensional input space. Pedagadi *et al.* [35] combined unsupervised principle component analysis (PCA) dimensionality reduction and LFDA defined by a training set to perform metric learning. Li *et al.* [34] proposed to learn a decision function that can be viewed as a joint model of a distance metric and a locally adaptive thresholding rule. Wang *et al.* [50] transformed the metric learning problem to a feature projection matrix learning problem that projects image features of one camera to the feature space of the other camera. Liao *et al.* [25] learned a discriminant low-dimensional subspace by XQDA. Wang *et al.* [1] investigated consistencies between two cameras and adjusted the metric for each query-gallery pair. Zhang *et al.* [51] proposed to learn a discriminative null space for person re-id, by minimizing the within-class scatter to the extreme and maximizing the relative between-class separation simultaneously. Zheng and Shao [52] learned the distance metric in the Hamming space for fast person re-id.

All of these methods attempt to obtain a proper distance metric for feature vectors. In contrast, our proposed approach exploits discrepancy matrix. Therefore, traditional approaches, which focus on generating vector metric rather than matrix metric, cannot be used in our problem. To this end, a matrix metric learning method is proposed.

III. Problem Statement and Motivation

In this section, we first review the traditional pattern for the re-id task. Then, we present the motivation of DM^3 learning in the proposed method.

A. Feature and Vector Metric

For the traditional re-id problem, a set of labeled persons $O = \{o_1, o_2, \ldots, o_M\}$ is associated with two cameras, where M is the number of persons. We denote the representative description of person o_i captured by camera A (or B) as x_i^A (or x_i^B), $x_i^A, x_i^B \in \mathbb{R}^{N_f \times 1}$. Then, $\{x_1^A, \ldots, x_i^A, \ldots, x_M^A\}, 1 \le i \le M$ and $\{x_1^B, \ldots, x_j^B, \ldots, x_M^B\}, 1 \le j \le M$, respectively, represent the two labeled training sets captured by A and B. Based on these two sets, a uniform distance metric $\mathbf{L}$ is learned.

Let x_p^A stand for a testing probe data from camera A, and $\{x_{M+1}^B, \ldots, x_q^B, \ldots, x_{M+N}^B\}, M+1 \le q \le M+N$ represent the test data from B, where N is the number of testing data in camera B. Then, for each testing probe data x_p^A, the distance between the testing probe data and every testing data x_q^B can be calculated by exploiting (1) (as the left column of Fig. 2 shows). After obtaining all of the distances, the ranking list is generated.

B. Discrepancy Matrix

The general pattern described above exploits a person's own characteristic to describe its appearance. Most of the feature descriptors transform the characteristics of each person image to a discriminative feature vector. We can also represent the person by the discrepancies with other persons. Given a set of reference image features $\{f_1, f_2, \ldots, f_{N_r}\}, f \in \mathbb{R}^{N_f \times 1}$, the description of image I is to construct a series of feature vector differences with the set of images as $[x-f_1; x-f_2; \cdots; x-f_{N_r}]$. We denote the description as $\mathbf{X} \in \mathbb{R}^{N_f \times N_r}$, where N_r is the number of images in the reference set. It should be noted that [37] and [53] utilized a reference set to describe a person as well, and their focus is on selecting some typical images from the reference set and exploiting the reconstruction parameters to produce a vector descriptor. However, our attention is paid on the discrepancies with all the images in the reference set to generate a matrix descriptor.

On the other hand, scale zooming, illumination change, and capture equipment difference between two cameras make the original feature description not robust enough. However, the effects of external environment on different persons may be similar in the same camera. Therefore, the difference between two feature descriptions may reduce these effects. Based on this consideration, we construct the discrepancy matrix description by feature differences with the reference image set from the same camera. Specifically, in our pattern, two sets of reference images are selected, which produce image feature descriptions $\{f_1^A, f_2^A, \ldots, f_{N_r}^A\}$ from camera A, and $\{f_1^B, f_2^B, \ldots, f_{N_r}^B\}$ from camera B. By definition, if the subscripts of two feature descriptions are the same, the features are extracted from the same person from two different cameras. Hence, as the right column of Fig. 2 shows, for the image I_p^A from camera A, its discrepancy matrix description is $\mathbf{X}_p^A = [x_p^A - f_1^A; x_p^A - f_2^A; \cdots; x_p^A - f_{N_r}^A]$. While for the image I_q^B from camera B, its discrepancy matrix description is $\mathbf{X}_q^B = [x_q^B - f_1^B; x_q^A - f_2^B; \cdots; x_q^B - f_{N_r}^B]$. In this way, the feature description is transformed from characteristic vector to discrepancy matrix. After obtaining discrepancy matrix description, we calculate the distance between each description pair. Following [54], we use the Frobenius norm to measure the distance of different matrices. For example, for the image pair (I_p^A, I_q^B), the distance is $d(\mathbf{X}_p^A, \mathbf{X}_q^B) = \|(\mathbf{X}_p^A - \mathbf{X}_q^B)\|_F^2$.

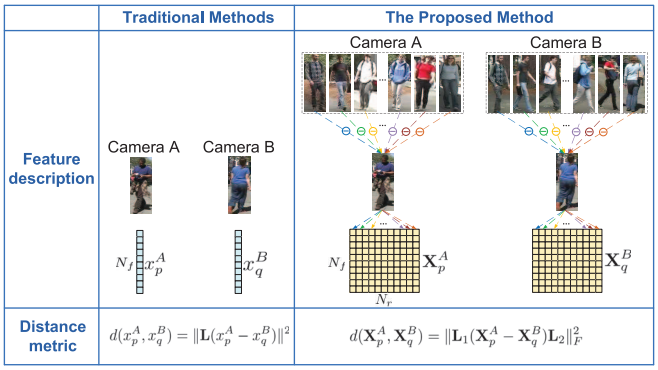


Fig. 2. Comparison of traditional methods and the proposed method. For both feature description and distance metric parts, the proposed method is different from traditional methods. The feature description is transformed from vector (x_p^A, x_q^B) to matrix $(\mathbf{X}_p^A, \mathbf{X}_q^B)$. N_f is the number of the dimension of feature vector, and N_r is the number of the reference images. The distance metric is transformed from feature projection ($\mathbf{L}$) to intradiscrepancy projection $\mathbf{L}_1$ and interdiscrepancy projection $\mathbf{L}_2$.

To testify the effectiveness of the discrepancy matrix description, we made a preliminary experiment to compare it with the feature vector description. The experiment included ten runs. For each run, 100 pairs of images were randomly selected from the VIPeR dataset [55], respectively, from camera A and camera B, and were set as two reference image sets. Another 100 pairs of images were randomly selected from the rest of the dataset. Each image was represented as a feature vector, respectively, using the hand-crafted GoG descriptor [26] and the deep-learned fine-tuned CNN (FTCNN) descriptor [56], and then the discrepancy matrix of each image was also constructed following the method described in the previous paragraph. For the vector description, we calculated the Euclidean distance of each pair of feature vectors. For the matrix description, we computed the Frobenius norm of the matrix difference of each pair of discrepancy matrices. We used the CMC curve [38], which describes the expectation of finding the true match within the first r ranks, to evaluate the average results of ten runs. The results are shown in Table II. From this table, we can easily conclude that the discrepancy matrix description is more effective than the feature vector description. We attribute the improvement to the reduction of the effect caused by the variation of camera conditions. The reason is as follows. In terms of theory, it is reasonable to assume that although big viewpoint warps and occlusions exist sometimes, from cameras A to B, the visual appearance of each person will encounter consistent cross-camera imaging variations in a period of time. Kviatkovsky *et al.* [41] proved that under different illumination conditions, each color will get a constant shift in the log-chromaticity color space.

TABLE II
PRELIMINARY EXPERIMENT COMPARING THE DISCREPANCY MATRIX DESCRIPTION AND THE FEATURE VECTOR DESCRIPTION. PERSON RE-ID MATCHING RATES (%) AT DIFFERENT RANKS ON THE VIPeR DATASET

Method (rank@)	1	2	3	4	5	6	7	8	9	10
hand-crafted feature vector	21.3	29.7	34.6	39.4	43.9	47.5	49.9	53.3	55.6	57.6
hand-crafted discrepancy matrix	**22.9**	**33.4**	**39.5**	**43.7**	**46.5**	**49**	**51.9**	**54.7**	**56.9**	**59.7**
deep-learned feature vector	29.3	37.9	44.5	49.5	53.5	57.6	61.0	62.9	65.6	67.7
deep-learned discrepancy matrix	**31.8**	**43.1**	**49.8**	**54.2**	**57.9**	**61.1**	**63.2**	**65.3**	**68.3**	**69.7**

Wang *et al.* [1] demonstrated that the transformation for each person is consistent across two different cameras. Inspired by these two methods, we assume that feature differences caused by cross-camera imaging conditions are the same, and define this uniform feature difference as v for each person. Then, for a pair of reference images (I_i^A, I_i^B), their feature difference can be formulated as $f_i^A - f_i^B = v + \sigma_i$, where σ_i stands for the bias error. Meanwhile, for a pair of test images, their feature difference can be formulated as $f_p^A - f_q^B = v + \sigma$. As we know, the traditional feature distance is determined by the difference $v + \sigma$. It means that the cross-camera imaging variation v acts as a key factor, especially when the imaging variation is very large. Introducing v will make different image pairs hard to be distinguished. Whereas, the distance of discrepancy matrices relies on $\mathbf{X}_p^A - \mathbf{X}_q^B = [(x_p^A - x_q^B) - (f_1^A - f_1^B); (x_p^A - x_q^B) - (f_2^A - f_2^B); \cdots; (x_p^A - x_q^B) - (f_{N_r}^A - f_{N_r}^B)] = [\sigma - \sigma_1; \sigma - \sigma_2; \cdots; \sigma - \sigma_{N_r}]$, where the cross-camera imaging variation is removed. Without considering v, the bias error σ from the same person is always smaller than that from different persons, which makes re-id relatively easier. Hence, it would help improve the performance of discrepancy matrix description.

In addition, experiments also show that the samples whose results rank at the first ten places are different for these two types of description (i.e., vector description and matrix description), and the difference ratio is 12.84%. This indicates that the discriminative abilities of discrepancy matrix description and feature vector description are different, probably due to the image relationships introduced by discrepancy matrix description besides the person's own feature representation.

C. Matrix Metric

A standard nontrained metric, without considering the differences of the elements or the relationships of the elements, may not be proper. A trained metric is necessary to make the distance of two matrix descriptions from the same person small, and that from different persons large. For traditional vector metric learning, as (1) demonstrates, $\mathbf{L}$ is a projection matrix. If $\mathbf{L} \in \mathbb{R}^{N_0 \times N_f}$, feature vectors will be projected to $\mathbb{R}^{N_0 \times 1}$. Following this rule, if we introduce a projection $\mathbf{L} \in \mathbb{R}^{N_0 \times N_f}$, after left multiplication $\mathbf{L}\mathbf{X}_p^A$ or $\mathbf{L}\mathbf{X}_q^B$, discrepancy matrices will be projected to $\mathbb{R}^{N_0 \times N_r}$.

As we know, the effectiveness of the left multiplication, which multiplies weights on the entries of each discrepancy, is different from that of the right multiplication. Actually, the right multiplication works on the entries of different discrepancies. Considering this difference, we introduce an intradiscrepancy projection matrix $\mathbf{L}_1 \in \mathbb{R}^{N_1 \times N_f}$ as the left-multiplier, as well as an interdiscrepancy projection matrix $\mathbf{L}_2 \in \mathbb{R}^{N_r \times N_2}$ as the right-multiplier. This two projections $\mathbf{L}_1$ and $\mathbf{L}_2$ together form the matrix metric.

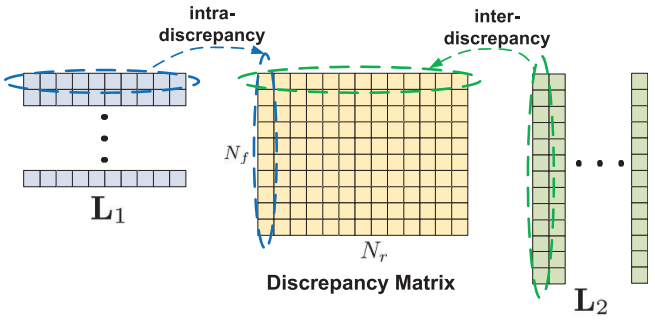


Fig. 3. Illustration of the two projections. Based on the principle of matrix multiplication, the row of $\mathbf{L}_1$ and the column of discrepancy matrix (intradiscrepancy) are combined, and the column of $\mathbf{L}_2$ and the row of discrepancy matrix (interdiscrepancy) are combined.

To make the functions of these two projections clear, we draw Fig. 3 to illustrate how left multiplication and right multiplication work. Each column of discrepancy matrix stands for a discrepancy generated from a corresponding reference image. Based on the principle of matrix multiplication, the row of $\mathbf{L}_1$ and the column of discrepancy matrix are combined in element-wise. It means that $\mathbf{L}_1$ works on each discrepancy. Whereas, $\mathbf{L}_2$ works on different discrepancies. Fig. 3 also shows that the column of $\mathbf{L}_2$ and the row of discrepancy matrix (the same dimension of different discrepancies) are combined in element-wise. In this paper, we introduce the interdiscrepancy projection, and do not convert the discrepancy matrix to vector. The reasons are as follows.

1) *Easy to Understand With Independent Physical Meanings:* If we do not introduce $\mathbf{L}_2$ as the right-multiplier, each discrepancy will be treated equally by $\mathbf{L}_1$. However, we consider that not all, but some of the reference images will be effective for the distance measurement. Hence, we sparsely select a few typical discrepancies by exploiting the interdiscrepancy projection $\mathbf{L}_2$ to improve the performance. In Section V-C, we analyze the sparsity of the interdiscrepancy projection. From the experiment, we can see that if a person image pair exists large variations, such as different illuminations and background changes, it is of a small possibility to be selected by the projection matrix $\mathbf{L}_2$. $\mathbf{L}_2$ demonstrates its ability on the selection of different discrepancies. While, $\mathbf{L}_1$ focuses on weighting different dimensions of each discrepancy. We consider that these two projections have independent physical meanings.

2) *Less Parameters and Constraints for $\mathbf{L}_1$:* If the matrix is converted to vector by concatenating the columns, traditional vector-based metric learning methods could be utilized directly, in addition, the interdiscrepancy projection $\mathbf{L}_2$ is not needed in this case. We agree that designing a metric for the converted discrepancy vector may work. However, the dimension of the vector will increase dramatically after reshaping the matrix to the vector. If we attempt to give different contributions to different discrepancies, more constraints should be introduced to $\mathbf{L}_1$. Then, it will make the intradiscrepancy projection $\mathbf{L}_1$ difficult to learn.
3) *Convenient to Add Constraints:* The form of the converted discrepancy vector would destroy the structural information of the data matrix, and the independence of each discrepancy would be broken. On the contrary, if we retain the structure of matrix, $\mathbf{L}_1$ and $\mathbf{L}_2$ will be exploited simultaneously. To learn the matrix metric, we can optimize the two projections alternatively. Meanwhile, we can add constraints to the two projections independently. For example, we introduce a sparse term to make the projection $\mathbf{L}_2$ sparse.

For the image pair $(\mathbf{X}_p^A, \mathbf{X}_q^B)$, the new distance is calculated using (2). In this way, the distance metric is transformed from vector metric to matrix metric

$$d\left(\mathbf{X}_p^A, \mathbf{X}_q^B\right) = \left\|\mathbf{L}_1\left(\mathbf{X}_p^A - \mathbf{X}_q^B\right)\mathbf{L}_2\right\|_F^2. \tag{2}$$

As illustrated in the right column of Fig. 2, different from traditional methods, the proposed method exploits matrix instead of vector in feature description, and consists of a couple of projections (the left-multiplier and right-multiplier) instead of a single projection in the distance metric part.

IV. Proposed Matrix Metric Learning

This section presents our matrix metric learning method. We begin with the matrix metric learning for the re-id problem. Then, a new objective function consisting of consistent and discriminative terms is put forward. Considering that not all the reference persons are useful for discrepancy, a sparse term is introduced into the objective function as well. Meanwhile, we exploit the alternating optimization and the gradient-descent method to learn the metric, and a stochastic sampling-based solution method is designed to accelerate the optimization process.

A. Definition

The new distance between two images is defined as (2). Compared with (1), where the projection transformation $\mathbf{L}$ is applied, the proposed metric consists of the intradiscrepancy projection part and the interdiscrepancy projection part. Generally, with the intradiscrepancy projection and the interdiscrepancy projection part, each pair of discrepancy sets from the same person are pulled close, and those from different persons are pushed apart [as Fig. 4(left) illustrates]. Given two sets of descriptions $\{\mathbf{X}_1^A, \ldots, \mathbf{X}_i^A, \ldots, \mathbf{X}_M^A\}$ and $\{\mathbf{X}_1^B, \ldots, \mathbf{X}_j^B, \ldots, \mathbf{X}_M^B\}$ as training sets, the essence of metric learning is to find optimal $\mathbf{L}_1$ and $\mathbf{L}_2$ under the supervised information generally containing two pair-wise constraints, i.e., similar constraint and dissimilar constraint.

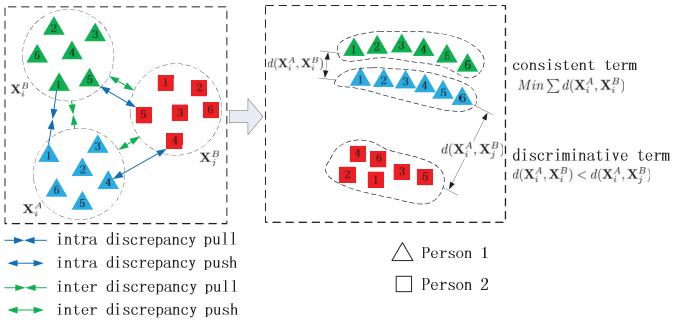


Fig. 4. Illustration of the proposed matrix metric learning. It consists of a consistent term, which makes combinations of discrepancies (matrices) of the same person close, and a discriminative term, which makes combinations of discrepancies of different persons apart.

B. Objective Function for Matrix Metric Learning

Motivated by Wang *et al.* [50], we formulate the objective function with two terms, and learn the matrix metric, including $\mathbf{L}_1$ and $\mathbf{L}_2$. The first term projects $\mathbf{X}_i^A$ and $\mathbf{X}_i^B$, which are generated from the same person, close to each other, thereby the inconsistency of two cameras is effectively eliminated. We call it the consistent term. The second term projects $\mathbf{X}_i^A$ away from $\mathbf{X}_j^B$, where $i \neq j$. It holds the discriminative ability of the metric, and we refer to it as the discriminative term [Fig. 4(right)].

Specifically, the consistent term can be defined by the sum of matrix distances of all similar pairs

$$E_{\text{con}}(\mathbf{L}_1, \mathbf{L}_2) = \frac{1}{M}\sum_{i=1}^{M} d\left(\mathbf{X}_i^A, \mathbf{X}_i^B\right). \tag{3}$$

Intuitively, this term in the objective function penalizes large discrepancy matrix distance between images of the same people.

Before introducing the discriminative term, we denote the triple set as $T = \{(\mathbf{X}_i^A, \mathbf{X}_i^B, \mathbf{X}_j^B)_k | i \neq j, k = 1, \ldots, S\}$, where s is the size of the set. Then, for each triple sample s_k, the following inequality $d(\mathbf{X}_i^A, \mathbf{X}_i^B) < d(\mathbf{X}_i^A, \mathbf{X}_j^B)$ needs to be satisfied. We define an error function for one triple sample as $e(s_k) = d(\mathbf{X}_i^A, \mathbf{X}_i^B) - d(\mathbf{X}_i^A, \mathbf{X}_j^B)$. With this error function, the formulation of the discriminative term is defined as

$$E_{\text{dis}}(\mathbf{L}_1, \mathbf{L}_2) = \frac{1}{S}\sum_{k=1}^{S} l_\beta(e(s_k)) \tag{4}$$

where $l_\beta(z) = (1/\beta)\log(1 + e^{\beta z})$ is the generalized logistic loss function (refer to [50]). It is easy to see that this term in the objective function penalizes triple samples invading the inequality. Here, we choose the logistic loss function instead of the hinge loss function for two reasons. First, the hinge loss is not differentiable at zero, while logistic loss function has derivatives everywhere which makes the solution simpler. Second, the logistic loss gives a soft approximation to hinge loss and is more flexible.

In addition, as discussed in Section III-C, $\mathbf{L}_2$ is the interdiscrepancy projection, acting as the right-multiplier controls the weights of different discrepancies. It selects typical reference persons which are useful for discrepancies. We consider that not all the reference persons are useful for discrepancy, and some of them have more discriminative power and bring less noise. Hence, discrepancies should be sparsely selected [57]. We utilize $\ell_{2,1}$-norm to improve discrepancy selection [58], [59]. By solving the $\ell_{2,1}$-norm minimization problem, $\mathbf{L}_2$ will be sparse in each column. The formulation of the sparse term is defined as

$$E_{\text{spr}}(\mathbf{L}_2) = \|\mathbf{L}_2\|_{2,1}. \tag{5}$$

Finally, we combine E_{con}, E_{dis}, and E_{spr} terms into a single objective function for learning matrix metric as (6), where the weights for the consistent term and the discriminative term are set equal for simplicity in this paper, and μ is the weight for the sparse term

$$E(\mathbf{L}_1, \mathbf{L}_2) = E_{\text{con}}(\mathbf{L}_1, \mathbf{L}_2) + E_{\text{dis}}(\mathbf{L}_1, \mathbf{L}_2) + \mu E_{\text{spr}}(\mathbf{L}_2). \tag{6}$$

C. Optimization Algorithm

With the above objective function, the optimal metric can be learned by solving the following optimization problem:

$$\left(\mathbf{L}_1^*, \mathbf{L}_2^*\right) = \arg\min_{\mathbf{L}_1, \mathbf{L}_2} E(\mathbf{L}_1, \mathbf{L}_2). \tag{7}$$

Due to the coupled variables and joint nonconvexity of the proposed model, the global optimality cannot be guaranteed. To solve the model efficiently, we present an alternating optimization process to learn $\mathbf{L}_1$ and $\mathbf{L}_2$ iteratively. Similar to [60], we fix one of the projections and optimize the other one, then in turn. As we know, the sparse term (5) is convex as demonstrated in [58]. In addition, the consistent term (3) and the discriminative term (4) are based on the distance function (2). When we fix $\mathbf{L}_1$ or $\mathbf{L}_2$, the original distance function degrade from a four-order polynomial to a two-order polynomial. Consequently, (7) can be solved using a simple gradient-descent method under the alternating optimization process with respect to $\mathbf{L}_1$ and $\mathbf{L}_2$. We exploit a simple stochastic strategy with randomly selected samples to accelerate the iteration speed and meanwhile keep the optimization accuracy. Specifically, for each positive example, we randomly select k ($k \ll M$) negative samples. A simple gradient descent method can be exploited to learn $\mathbf{L}_1$ and $\mathbf{L}_2$. The gradients of the objective function are given as

$$\frac{\partial E(\mathbf{L}_1, \mathbf{L}_2)}{\partial \mathbf{L}_1} = \frac{\partial E_{\text{con}}(\mathbf{L}_1, \mathbf{L}_2)}{\partial \mathbf{L}_1} + \frac{\partial E_{\text{dis}}(\mathbf{L}_1, \mathbf{L}_2)}{\partial \mathbf{L}_1} \tag{8}$$

and

$$\frac{\partial E(\mathbf{L}_1, \mathbf{L}_2)}{\partial \mathbf{L}_2} = \frac{\partial E_{\text{con}}(\mathbf{L}_1, \mathbf{L}_2)}{\partial \mathbf{L}_2} + \frac{\partial E_{\text{dis}}(\mathbf{L}_1, \mathbf{L}_2)}{\partial \mathbf{L}_2} + \mu \frac{\partial E_{\text{spr}}(\mathbf{L}_2)}{\partial \mathbf{L}_2} \tag{9}$$

where

$$\frac{\partial E_{\text{con}}(\mathbf{L}_1, \mathbf{L}_2)}{\partial \mathbf{L}_1} = \frac{2}{M} \sum_{i=1}^{M} \mathbf{L}_1 \mathbf{Z}_i \mathbf{L}_2 \mathbf{L}_2^\top \mathbf{Z}_i^\top \tag{10}$$

Algorithm 1 Learning the Matrix Metric $\mathbf{L}_1$ and $\mathbf{L}_2$

Input: The training data: Positive samples with pair form $\{(\mathbf{X}_i^A, \mathbf{X}_i^B)\}$, and Negative Samples with triple form $\{(\mathbf{X}_i^A, \mathbf{X}_i^B, \mathbf{X}_j^B)_k\}$.
Output: The optimal matrix $\mathbf{L}_1^*$ and $\mathbf{L}_2^*$.

```
Initialize L_1 and L_2;
for n = 1 to MaxIter do
    Fix L_2^n;
    Compute ∇E(L_1) = ∂E(L_1,L_2)/∂L_1 as Eq. 8, Eq. 10, and Eq. 12;
    Choose a proper step λ_1 as [61];
    Compute L_1^{n+1} = L_1^n − λ_1 ∇E(L_1);
    Fix L_1^{n+1};
    Compute ∇E(L_2) = ∂E(L_1,L_2)/∂L_2 as Eq. 9, Eq. 11, and Eq. 13;
    Choose a proper step λ_2 as [61];
    Compute L_2^{n+1} = L_2^n − λ_2 ∇E(L_2);
    if converge then
        break;
    end if
end for
```

$$\frac{\partial E_{\text{con}}(\mathbf{L}_1, \mathbf{L}_2)}{\partial \mathbf{L}_2} = \frac{2}{M} \sum_{i=1}^{M} \mathbf{Z}_i^\top \mathbf{L}_1^\top \mathbf{L}_1 \mathbf{Z}_i \mathbf{L}_2 \tag{11}$$

$$\frac{\partial E_{\text{dis}}(\mathbf{L}_1, \mathbf{L}_2)}{\partial \mathbf{L}_1} = \frac{2}{S} \sum_{k=1}^{S} g(e(s_k)) \times \left(\mathbf{L}_1 \mathbf{U}_k \mathbf{L}_2 \mathbf{L}_2^\top \mathbf{U}_k^\top - \mathbf{L}_1 \mathbf{V}_k \mathbf{L}_2 \mathbf{L}_2^\top \mathbf{V}_k^\top\right) \tag{12}$$

$$\frac{\partial E_{\text{dis}}(\mathbf{L}_1, \mathbf{L}_2)}{\partial \mathbf{L}_2} = \frac{2}{S} \sum_{k=1}^{S} g(e(s_k)) \times \left(\mathbf{U}_k^\top \mathbf{L}_1^\top \mathbf{L}_1 \mathbf{U}_k \mathbf{L}_2 - \mathbf{V}_k^\top \mathbf{L}_1^\top \mathbf{L}_1 \mathbf{V}_k \mathbf{L}_2\right) \tag{13}$$

and

$$\frac{\partial E_{\text{spr}}(\mathbf{L}_2)}{\partial \mathbf{L}_2} = 2\mathbf{D}\mathbf{L}_2. \tag{14}$$

Here, $g(z) = (1 + e^{-\beta z})^{-1}$ is the derivative of logistic loss function $l_\beta(z)$. In above formulations, $\mathbf{Z}_i = \mathbf{X}_i^A - \mathbf{X}_i^B$, $\mathbf{U}_k = \mathbf{X}_i^A - \mathbf{X}_i^B$, $\mathbf{V}_k = \mathbf{X}_i^A - \mathbf{X}_j^B$. Following [58], $\mathbf{D}$ is a diagonal matrix with the mth diagonal element as $d_{mm} = (1/2\|\mathbf{l}_m\|_2)$, where $\mathbf{l}_m$ denotes the mth row of $\mathbf{L}_2$.

With the gradients, an iterative optimization algorithm can be used to learn the metric. Starting from the initial identical matrix, $\mathbf{L}_1$ and $\mathbf{L}_2$ are optimized iteratively. In the optimization progress, we fix one and update the other as follows:

$$\mathbf{L}_1^{n+1} = \mathbf{L}_1^n - \lambda_1 \nabla E(\mathbf{L}_1) \tag{15}$$

and

$$\mathbf{L}_2^{n+1} = \mathbf{L}_2^n - \lambda_2 \nabla E(\mathbf{L}_2) \tag{16}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are step lengthes automatically determined at each gradient update step. The iteration of the algorithm is terminated when it reaches the maximum iteration number (1000 in this paper) or meets the following criterion:

$$|E^{n+1} - E^n| < \varepsilon \tag{17}$$

where ε is a small positive value, i.e., $\varepsilon = 1 \times 10^{-8}$ in this paper. The complete algorithm flow is shown in Algorithm. 1.

(a) (b) (c)

Fig. 5. Some typical samples of three public datasets. Each column shows two images of the same person from two different cameras. (a) VIPeR dataset. (b) PRID 450S dataset. (c) CUHK01 dataset.

V. Experiments

A. Datasets

1) VIPeR Dataset: The widely used VIPeR dataset [55] contains 1264 outdoor images obtained from two views of 632 persons. Some example images are shown in Fig. 5(a). Each person has a pair of images taken from two different cameras, respectively. All images of individuals are normalized to a size of 128×48 pixels. View changes are the most significant cause of appearance change. Other variations are also included, such as illumination conditions and image qualities.

2) PRID 450S Dataset: The PRID 450S dataset [62] was created in co-operation with the Austrian Institute of Technology for the purpose of testing person re-id approaches. It is a more realistic dataset, which contains 450 single shot image pairs captured over two spatially disjoint camera views. All images are normalized to 168×80 pixels. It is also a challenging person re-id dataset. Different with the VIPeR dataset, this dataset has significant and consistent lighting changes. Some examples from the PRID 450S dataset are shown in Fig. 5(b).

3) CUHK01 Dataset: The CUHK01 dataset [63] is a larger dataset and contains 971 identities from two disjoint camera views. Each identity has two samples per camera view. Some example images are shown in Fig. 5(c). There are a total of 3884 images. All images are normalized to 160×60 pixels. Similar to the VIPeR dataset, view changes are the most significant cause of appearance change with most of the matched image pairs containing one front/back view and one side-view. Since a single representative image per camera view for each person is considered in this paper, we randomly selected one image from two samples per camera views for each people in our experiments for this dataset.

B. Effectiveness of Discrepancy Matrix and Matrix Metric

1) Experimental Settings: To evaluate the proposed pattern, we used the GoG descriptor [26] and demonstrated the effectiveness of the proposed method on reforming the hand-crafted feature. Meanwhile, given the success of deep learning features in computer vision applications, we also conducted experiments to show the effectiveness of the proposed method on deep learning features. The FTCNN [56] descriptor was employed to extract original feature descriptions. General parameter configurations were the same for these two feature descriptors. To accelerate the learning process and reduce noise, we conducted PCA to obtain a low-dimensional representation [33], i.e., 70 ($N_f = 70$) in the experiments. Then, the discrepancy matrix was generated using the low-dimensional feature vector. We set $N_1 = N_f$ and $N_2 = N_r$. The entire evaluation procedure was repeated ten times. CMC [38] curves were used to calculate the average performance. To fairly evaluate and show the effectiveness of the proposed method, we constructed three subsets for each dataset, including the training set, the testing set, and the reference set. The three subsets are nonoverlapped, and randomly selected from the whole dataset.

1) *VIPeR Dataset:* For the VIPeR dataset, we randomly selected 100 sample pairs as reference set ($N_r = 100$). Following the general settings, where the number of training and testing pairs are the same, 200 sample pairs are, respectively, from the rest samples ($M = 200$ and $N = 200$). The obtained results are shown in Fig. 6(a) and (d). As can be seen, the discrepancy matrix performs better than feature vector, and the proposed DM^3 method obviously outperforms the basic discrepancy matrix and feature vector over the whole range of ranks.
2) *PRID 450S Dataset:* Following the evaluation process on the VIPeR dataset, we, respectively, set $N_r = 60$, $M = 150$, and $N = 150$. The obtained results are shown in Fig. 6(b) and (e). It is evident that the discrepancy matrix performs better than feature vector, and the proposed DM^3 method clearly outperforms the basic discrepancy matrix and feature vector.
3) *CUHK01 Dataset:* Following the evaluation process on the VIPeR dataset, we, respectively, set $N_r = 100$, $M = 300$, and $N = 300$. Fig. 6(c) and (f) presents the comparison results of different methods. On this challenging CUHK01 dataset, the same conclusion as draw from the other two datasets can be achieved based on the results.

These experiments prove that, on different datasets with different feature vectors, the proposed pattern performs very well.

C. Sparsity of the Interdiscrepancy Projection

As discussed above, the role of $\mathbf{L}_2$ is to sparsely select typical references which are more useful for discrepancies. Following the set of previous experiments, we learned $\mathbf{L}_2$ with $N_r = 100$ and $M = 200$ on the VIPeR dataset. Fig. 7(a) visualized the results of $\mathbf{L}_2$. From this figure, we can see that

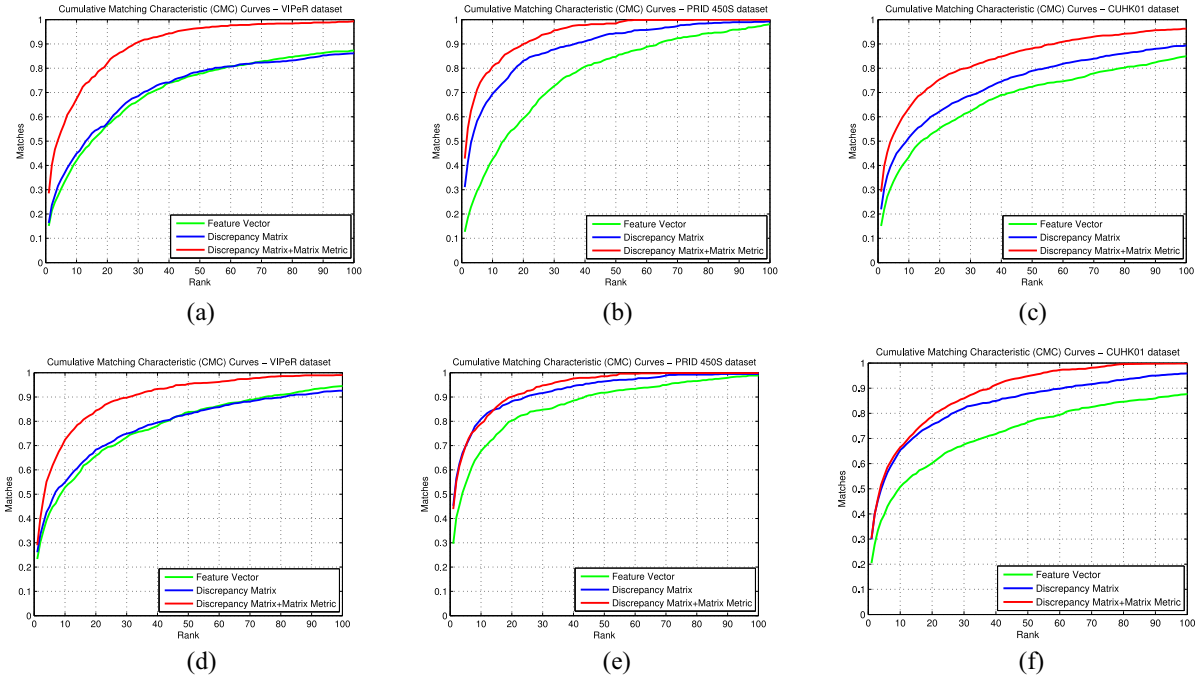


Fig. 6. Experimental results of feature vector, discrepancy matrix description and the proposed DM^3 method on three public datasets, respectively, exploiting hand-crafted feature and deep feature. Results on the (a) VIPeR dataset using the GoG descriptor, (b) PRID 450S dataset using the GoG descriptor, (c) CUHK01 dataset using the GoG descriptor, (d) VIPeR dataset using the FTCNN descriptor, (e) PRID 450S dataset using the FTCNN descriptor, and (f) CUHK01 dataset using the FTCNN descriptor.

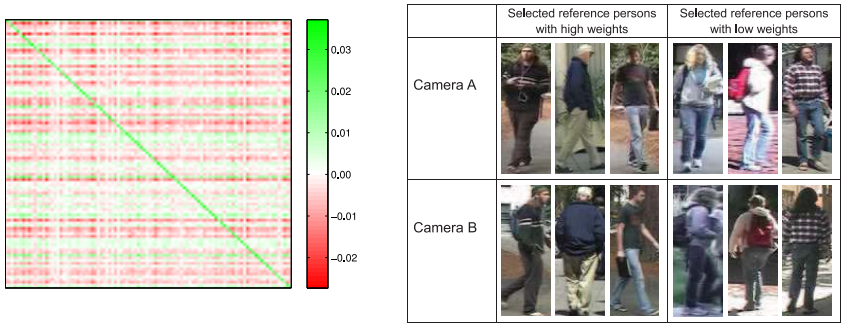


Fig. 7. Analysis on the sparsity of the interdiscrepancy projection. (a) Visualization of the interdiscrepancy projection. (b) Typical reference examples. These examples are selected depending on the values of corresponding rows of the projection matrix $\mathbf{L}_2$. The left three columns are examples, sparsely selected by the row of $\mathbf{L}_2$ with a high value (high weight). The right three columns are examples with a small possibility to be selected.

$\mathbf{L}_2$ is relatively sparse for each column, that is to say only few reference persons are selected after right multiplication.

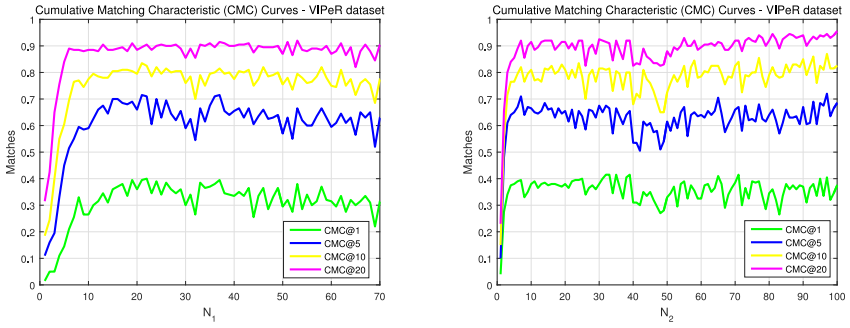
In Fig. 7(b), we list some reference examples by analyzing the projection weights. They are, respectively, the selected persons and the nonselected persons. Comparing these examples, we reckon that if a person image pair exists large variations, such as different illuminations and background changes, it is of a small possibility to be selected by the projection matrix $\mathbf{L}_2$.

With the visualization of $\mathbf{L}_2$, this experiment illustrates that some of the reference persons are more useful for discrepancy than the others, have more discriminative power, and bring less noise.

D. Evaluating Parameters of the Proposed Method

We validate the proposed approach under different parameters, including evaluating different parameter N_1 for the contribution of the intradiscrepancy projection matrix, and different parameter N_2 for the interdiscrepancy projection matrix. The experiment was conducted on the VIPeR dataset, and general configurations were the same as the former experiments.

We fixed $N_2 = 100$ and carried out experiments with different N_1 values. Then, we fixed $N_1 = 70$ and conducted experiments with different N_2 values. The results are shown in Fig. 8(a) and (b). It is obvious that when $N_1 > 6$ ($N_2 > 4$), although the performance is not stable, the results can still be improved using the proposed matrix metric learning process with different N_1 (N_2) values. To learn the projections $\mathbf{L}_1$ and $\mathbf{L}_2$ together, we exploit the alternating optimization and the gradient-descent method, where the global optimality cannot be guaranteed. Hence, it would make the metric not perfectly accurate, and the values may vary a little. Comparing with



(a) (b)

Fig. 8. Parameter analysis of the proposed method. Results of different (a) N_1s on the VIPeR dataset, where $N_2 = 100$ and (b) N_2s on VIPeR dataset, where $N_1 = 70$.

the initial results, where $N_1 = 1$, or $N_2 = 1$, we consider that the proposed matrix metric demonstrates its effectiveness with a large margin, and that the small variations are acceptable. Fig. 8(a) and (b) also shows that N_1 (N_2) should not be too small, because the metric constraints by intradiscrepancy projection matrix (interdiscrepancy projection matrix) will be reduced when the number of dimensions is low.

E. Comparison of the Discrepancy Matrix and the Discrepancy Vector

By exploiting a reference set, the discrepancy description of an image can be constructed by multiple differences. We form the description to be a matrix in this paper. Actually, it can be reshaped to a long vector as well. We evaluated these two kinds of discrepancy descriptions on the VIPeR dataset. Fifty sample pairs were randomly selected as reference set ($N_r = 50$). We used the GoG descriptor [26] as the original feature descriptor, and conducted PCA to obtain 50-D representations ($N_f = 50$). As a result, the description of discrepancy matrix was denoted as $\mathbb{R}^{50\times 50}$, while the description of discrepancy vector was denoted as $\mathbb{R}^{2500\times 1}$. For the former, two projections were used with previous configurations. For the latter, since the projection $\mathbf{L}_2$ would degrade and lose its effectiveness, we only utilized the projection $\mathbf{L}_1$. The comparison results are shown in Fig. 9. It can be seen that the discrepancy matrix performs better than reshaping it to the discrepancy vector. It might be because the independence of each discrepancy is broken, and all the discrepancies are treated equally after reshaping the matrix to the vector. When we reshape the matrix to the vector, only $\mathbf{L}_1$ is exploited. To achieve the effectiveness of $\mathbf{L}_2$, we should to weight the contribution of different discrepancies, then more constraints should be introduced to $\mathbf{L}_1$, which are not considered in this paper. On the contrary, if we retain the structure of matrix, $\mathbf{L}_1$ and $\mathbf{L}_2$ will be exploited simultaneously. Some typical discrepancies should be sparsely selected by exploiting $\mathbf{L}_2$, as discussed in Section V-C.

F. Effectiveness of Different Terms

To understand the effectiveness of different terms, we evaluated the proposed method without a discriminate term or a consistent term, following the settings in Section V-B.

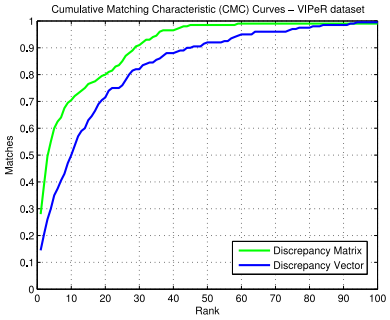


Fig. 9. Comparison of the discrepancy matrix and the discrepancy vector.

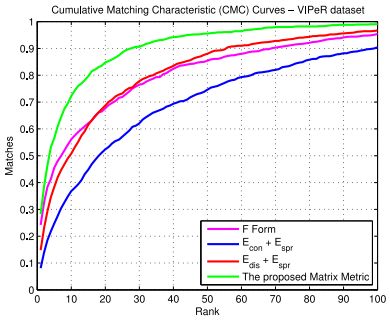


Fig. 10. Effectiveness of the proposed method without a discriminate term or a consistent term.

The obtained results are shown in Fig. 10. If the proposed method does not use the discriminate term (the blue curve in Fig. 10) or the consistent term (the red curve in Fig. 10), the results become worse, and even worse than the basic discrepancy matrix with the F-form distance.

The consistent term pulls samples of the same person close, and the discriminative term pushes samples of different persons far away. We consider that if the object function without the consistent term, the sample distribution of the same person will be dispersed, and the intraclass sample will be easily confused by the neighbor interclass sample. If the object function without the discriminative term, samples of different persons with similar appearance will be pulled together. As the experiment demonstrates, neither of these two terms can be omitted.

TABLE III
COMPARING RESULTS WITH THE STATE-OF-THE-ART PERSON RE-ID METHODS ON TOP RANKED MATCHING RATE (%) ON VIPeR DATASET

Method (rank@)	1	5	10	20
ELF [22]	12.0	-	43.0	60.0
BiCov [39]	20.6	43.2	56.1	68.0
SDALF [23]	19.9	38.4	49.4	66.0
eSDC [42]	26.3	46.4	58.6	72.8
MidFilter [64]	29.1	52.5	65.9	79.9
SCNCD [24]	37.8	68.5	81.2	90.4
RD [37]	33.3	65.1	78.3	88.5
PRDC [19]	15.7	38.4	53.9	70.1
KISSME [33]	19.6	48.0	62.2	77.0
PCCA [49]	19.3	48.9	64.9	80.3
LADF [34]	30.0	64.0	80.0	92.0
LOMO+XQDA [25]	40.0	68.5	80.5	91.0
DeepMetric [65]	28.2	59.3	73.4	86.4
DeepRanking [66]	38.4	69.2	81.3	90.4
DeepFeature+RDC [28]	40.5	60.8	70.4	84.4
DeepList [30]	40.5	69.1	80.1	91.2
LOMO+NFST [51]	42.2	71.4	82.9	92.0
(1) GoG [26]+XQDA	37.3	67.4	77.2	89.6
(2) FTCNN [56]+XQDA	31.2	59.8	74.0	83.5
(3) FTCNN+**DM**^{3}	37.3(↑6.1)	67.4(↑7.6)	80.3(↑6.3)	89.5(↑6.0)
Combine (1) and (2)	38.3	67.2	77.0	89.3
Combine (1) and (3)	**42.7**	**74.3**	**85.1**	**93.1**

TABLE IV
COMPARING RESULTS WITH THE STATE-OF-THE-ART PERSON RE-ID METHODS ON TOP RANKED MATCHING RATE (%) ON PRID 450S DATASET

Method (rank@)	1	5	10	20
SCNCD [24]	41.6	68.9	79.4	87.8
KISSME [33]	33.0	59.8	71.0	79.0
CBRA [67]	26.4	57.1	71.0	83.2
CSL [68]	44.4	71.6	82.2	89.8
Mirror [69]	55.4	79.3	87.8	93.9
DRML [70]	56.4	-	82.2	90.2
(1) GoG [26]+XQDA	51.6	76.8	88.8	94.2
(2) FTCNN [56]+XQDA	50.2	74.2	84.8	93.7
(3) FTCNN+**DM**^{3}	56.7(↑6.5)	83.1(↑8.9)	88.4(↑3.6)	94.7(↑1.0)
Combine (1) and (2)	51.8	76.9	87.0	94.2
Combine (1) and (3)	**61.0**	**85.8**	**92.0**	**96.7**

G. Comparing to the State-of-the-Art Re-Id Methods

In this section, we compared our method with the state-of-the-art re-id methods on different datasets. As we know, for most datasets, the general setting is that half of the dataset is selected as the training set, and the other half is taken as the testing set. To make a fair comparison, we randomly selected some of the samples from the training set as the reference set. Tables III–V summarize the comparing results with the state-of-the-art re-id methods, respectively, on the VIPeR, PRID 450S, and CUHK01 datasets.

1) VIPeR Dataset: Following traditional methods, the dataset were randomly split to two sets. One for training ($M = 316$) and the other for testing ($N = 316$). Among the training set, the reference set $N_r = 100$ was randomly selected as well. The VIPeR dataset is the most popular benchmark dataset for the re-id task, and hence, a lot of recent progress reports results on this dataset. We compare our approach with the following methods: ELF [22], BiCov [39], SDALF [23], eSDC [42], midlevel filters (MidFilter) [64], SCNCD [24], RD [37], PRDC [19], KISSME [33], PCCA [49], LADF [34], LOMO descriptor and XQDA (LOMO+XQDA) [25], deep metric learning (DeepMetric) [65], DeepRanking [66], deep feature with relative distance comparison (DeepFeature+RDC) [28], deep features with adaptive listwise constraint (DeepList) [30], LOMO descriptor with the null space metric (LOMO+NFST) [51], hierarchical Gaussian descriptor (GoG) [26] with XQDA, and FTCNN [56] with XQDA. Note that, the RD method extracts a feature description with a reference set, although discrepancies with reference set are not exploited. The PRDC method attempts to learn a metric with the sample differences, although the discrepancy is focused on feature distance other than description construction. DeepFeature, DeepMetric, DeepList, DeepRanking, and FTCNN are those methods related to deep learning frameworks, which obtain good performances recently. The proposed method (DM^3) was evaluated with the FTCNN descriptor.

All the results are listed in Table III. It should be mentioned that: 1) GoG+XQDA and 2) FTCNN+XQDA stand for the traditional re-id methods, which utilize the feature vector (GoG or FTCNN) and the vector metric XQDA. Whereas, 3) FTCNN+DM^3 stands for the proposed method exploiting the discrepancy matrix (based on FTCNN) and the matrix metric DM^3. Methods 1) and 2) are evaluated following the procedure in traditional methods, such as [25]. The blue numbers in the parenthesis indicate the improvements of the proposed DM^3 over XQDA at each rank given FTCNN as the feature descriptor. Due to the difference between feature and discrepancy matrix description, vector and matrix metric, we fused the proposed method (FTCNN+DM^3) and the state-of-the-art method (GoG+XQDA) by directly merging the two ranking results. In detail, the fusion process is as follows. Given a probe image I_p^A, two ranking lists will be generated by methods 1) and 3). Then, for each image I_q^B in camera B, we, respectively, obtain its ranking number $\text{rank}_1(I_q^B|I_p^A)$ and $\text{rank}_3(I_q^B|I_p^A)$ based on the two ranking lists, and then combine the numbers together as $\text{rank}_1(I_q^B|I_p^A)+\text{rank}_3(I_q^B|I_p^A)$. After obtaining the combined numbers of all the images in camera B, we reorder them from small to large, and gain the fused ranking list. As Table III shows, the fusion result outperforms all the results generated by other methods.

2) PRID 450S Dataset: Following the evaluation process on the VIPeR dataset, the dataset were randomly split to two sets. One for training ($M = 225$) and the other for testing ($N = 225$). Among the training set, the reference set $N_r = 70$ was randomly selected as well. We compare our approach with the following methods: SCNCD [24], KISSME [33], color-based ranking aggregation (CBRA) [67], correspondence structure learning (CSL) [68], mirror representation (Mirror) [69], and diversity regularized metric learning (DRML) [70]. The proposed method (DM^3) was evaluated with an original FTCNN descriptor. All the results are listed in Table IV. The blue numbers in the parenthesis indicate the improvements of the proposed DM^3 over XQDA at each rank given FTCNN as the feature descriptor. Table IV also shows that the fusion result (FTCNN+DM^3 and GoG+XQDA) outperforms all the results generated by other methods.

3) CUHK01 Dataset: Following the evaluation process on the VIPeR dataset, the dataset were randomly split to

TABLE V
COMPARING RESULTS WITH THE STATE-OF-THE-ART PERSON RE-ID METHODS ON TOP RANKED MATCHING RATE (%) ON CUHK01 DATASET

Method (rank@)	1	5	10	20
SDALF [23]	9.9	22.6	30.3	41.0
TML [63]	20.0	43.5	56.0	69.3
SalMatch [71]	28.4	45.8	55.7	67.9
MidFilter [64]	34.3	55.1	65.0	74.9
RD [37]	31.1	-	68.5	79.1
ImprovedDeep [72]	47.5	71.0	80.0	-
(1) GoG [26]+XQDA	44.5	71.1	78.1	89.0
(2) FTCNN [56]+XQDA	41.1	63.5	73.6	85.8
(3) FTCNN+**DM**^{3}	43.7(↑2.6)	70.1(↑6.6)	77.4(↑3.8)	88.7(↑2.9)
Combine (1) and (2)	42.1	70.1	78.3	89.6
Combine (1) and (3)	**49.7**	**77.3**	**86.1**	**91.4**

two sets. One for training ($M = 485$) and the other for testing ($N = 486$). Among the training set, the reference set $N_r = 100$ was randomly selected as well. These methods used a single-shot evaluating protocol. We compare our approach with the following methods: SDALF [23], transferred metric learning (TML) [63], saliency matching (SalMatch) [71], MidFilter [64], improved deep learning architecture (ImprovedDeep) [72], and RD [37]. The proposed method (DM^3) was evaluated with an original FTCNN descriptor. All the results are listed in Table V. The blue numbers in the parenthesis indicate the improvements of the proposed DM^3 over XQDA at each rank given FTCNN as the feature descriptor. Table V also shows that the fusion result (FTCNN+DM^3 and GoG+XQDA) outperforms all the results generated by other methods.

H. Evaluation on the Large Dataset

As we know, CUHK03 [27] and Market-1501 [73] are two large datasets in person re-id task. We choose the Market-1501 dataset to evaluate our method. The Market-1501 dataset is currently the largest benchmark dataset for person re-id, which is more consistent with practical application scenario. It contains 32 668 labeled bounding boxes of 1501 identities. Following the experiment setting of [73], the dataset is split into two parts: 12 936 images with 751 identities for training and 19 732 images with 750 identities for testing. In testing, 3368 images with 750 identities are used as the probe set.

To make a fair comparison with the state-of-the-art methods, we randomly selected $N_r = 100$ images pairs from the training set as the reference set. The ID discriminative embedding (IDE) feature proposed in [74] is used as our basic feature. The IDE extractor is effectively trained on classification model including CaffeNet [75] and ResNet-50 [76]. For the convenience of description, we abbreviate the IDE trained on CaffeNet and ResNet-50 to IDE(C) and IDE(R), respectively. We compare our approach with the following methods. BoW feature with the weighted approximate rank component analysis method (BoW+WARCA) [77], LOMO descriptor with the null space metric (LOMO+NFST) [51], IDE(C) feature with XQDA, and IDE(R) feature with XQDA. The proposed method (DM^3) was evaluated with the IDE(R) feature. Besides the CMC value, we also compared the mean average precision (mAP), as described in [74]. We report the single-query evaluation results [73] for this dataset.

TABLE VI
COMPARING RESULTS WITH THE STATE-OF-THE-ART PERSON RE-ID METHODS ON THE MARKET-1501 DATASET

Method (rank@)	1	5	10	mAP
BoW+WARCA [77]	45.1	68.1	76	-
LOMO+NFST [51]	55.4	-	-	29.8
(1) IDE(C)+XQDA	61.4	81.0	87	37.4
(2) IDE(R)+XQDA	75.5	88.6	91.6	53.0
(3) IDE(R)+**DM**^{3}	73.4	87.6	91.1	51.8
Combine (1) and (3)	**75.8**	**89.1**	**92.4**	**53.2**

All the results are listed in Table VI. Table VI shows that the fusion result (IDE(R)+DM^3 and IDE(C)+XQDA) outperforms all the results generated by the other methods. However, Table VI also shows that the proposed method (DM^3) with the IDE(R) feature does not performs better than the XQDA metric. We consider it is because the person images of the Market-1501 dataset are captured from six different cameras, while the proposed method is designed under two different cameras. The advantage of our method, which focuses on removing the uniform cross-camera imaging variation, would be suppressed under the multiple cameras condition.

I. Discussion on the Running Time

During the iterative procedure, the running cost of the offline training mainly depends on gradient and loss calculation. For the gradient, both traditional and the proposed methods operate using matrix manipulation. For the loss, traditional vector-L2 distance transforms to matrix F-form distance. Using MATLAB,[1] these two operations brings no more time loss. However, two projections are used in our approach, which will double the running cost. To evaluate the offline training time, we took the VIPeR dataset as an example. We set the training set as $M = 316$, and among the training set, the reference set $N_r = 100$ was randomly selected. The average running time for the four partial gradients (10)–(13) are 0.10, 0.11, 0.36, and 0.39 s. To calculate the total loss in each step, it takes 0.203 s in average. To reduce iterations, an adaptive step length strategy is utilized, which reduces much time. It merely costs 49.8 s in average.

On the other hand, the on-line testing time of each pair of discrepancy matrices is very fast. Using (2), the time cost is 0.0002 s in average. In total, it costs 19.9 s for testing a $N = 316$ set. In addition, to evaluate the testing time of the combination result, we should also consider the XQDA [25] method, which attempts to learn a Mahalanobis distance metric in essence. After obtaining the metric, the distances of vectors are calculated by Mahalanobis distance. It is recorded that the testing time is 0.09 s in average for the $N = 316$ testing set. That is to say, the combination work will not bring in much time cost.

VI. CONCLUSION

In this paper, we proposed a new idea to describe a person image. Specifically, the feature description is transformed from characteristic vector to discrepancy matrix, and the distance

[1]The CPU and RAM of our computer are, respectively, Intel Xeon E5-2683 and 256 GB. The version of MATLAB is R2015b.

metric is transformed from vector metric to matrix metric. Our model identifying person by their discrepancies with the others is similar to human cognition process, and it is advantageous because the proposed description presumably reduces the external changes and is more fine-grained. Experimental results on public datasets demonstrate the effectiveness of the proposed pattern. In the following, we also provide a discussion on future research directions.

1) In our proposed pattern, the reference set is chosen randomly, and some typical persons are sparsely selected from the reference set by exploiting $\mathbf{L}_2$, which is learned during the metric learning process. It is proved that these selected persons are more valuable than the other persons. To this end, we can investigate what characteristics are useful and how many persons should be exploited as the reference set. If we discover the selection mechanism of reference persons, the most valuable ones will be selected before metric learning, and the metric might be more discriminative by removing the sparse term.
2) We propose the matrix metric learning method by constructing a hand-crafted objective function. As we know, the deep learning framework can be also utilized to learn the metric. For example, the proposed discriminate term uses triple samples, and its idea is similar to the triple network [78].
3) Although Section V-I shows that the combination work will not bring in much time cost, calculating distances of discrepancy matrices is much more time consuming than computing distances of feature vectors. This inspires us to study how to accelerate the process of matrix distance computation.

REFERENCES

[1] Z. Wang *et al.*, "Zero-shot person re-identification via cross-view consistency," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 260–272, Feb. 2016.

[2] A. J. Ma, J. Li, P. C. Yuen, and P. Li, "Cross-domain person reidentification using domain adaptation ranking SVMs," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1599–1613, May 2015.

[3] M. Ye *et al.*, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2553–2566, Dec. 2016.

[4] Z. Wang *et al.*, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *Proc. Int. Joint Conf. Artif. Intell.*, New York, NY, USA, 2016, pp. 2669–2675.

[5] S. Tan, F. Zheng, L. Liu, J. Han, and L. Shao, "Dense invariant feature based support vector ranking for cross-camera person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2016.2555739.

[6] N. Martinel, G. L. Foresti, and C. Micheloni, "Person reidentification in a distributed camera network framework," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2568264.

[7] J. García *et al.*, "Discriminant context information analysis for post-ranking person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1650–1665, Apr. 2017.

[8] J. Chen, Y. Wang, J. Qin, L. Liu, and L. Shao, "Fast person re-identification via cross-camera semantic binary transformation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3873–3882.

[9] J. Wang, Z. Wang, C. Liang, C. Gao, and N. Sang, "Equidistance constrained metric learning for person re-identification," *Pattern Recognit.*, vol. 74, pp. 38–51, Feb. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317303576

[10] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[11] R. Hou, C. Chen, and M. Shah, "Tube convolutional neural network (T-CNN) for action detection in videos," *CoRR*, vol. abs/1703.10664, 2017. [Online]. Available: http://arxiv.org/abs/1703.10664

[12] B. Zhang *et al.*, "Action recognition using 3D histograms of texture and a multi-class boosting classifier," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4648–4660, Oct. 2017.

[13] Z.-R. Lai, D.-Q. Dai, C.-X. Ren, and K.-K. Huang, "Discriminative and compact coding for robust face recognition," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1900–1912, Sep. 2015.

[14] J. Ma, J. Zhao, Y. Ma, and J. Tian, "Non-rigid visible and infrared face registration via regularized Gaussian fields criterion," *Pattern Recognit.*, vol. 48, no. 3, pp. 772–784, 2015.

[15] J. Jiang, J. Ma, C. Chen, X. Jiang, and Z. Wang, "Noise robust face image super-resolution through smooth sparse representation," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2594184.

[16] Y. Gao, J. Ma, and A. L. Yuille, "Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2545–2560, May 2017.

[17] M. Ding and G. Fan, "Multilayer joint gait-pose manifolds for human gait motion modeling," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2413–2424, Nov. 2015.

[18] D. Muramatsu, Y. Makihara, and Y. Yagi, "View transformation model incorporating quality measures for cross-view gait recognition," *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1602–1615, Jul. 2016.

[19] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 649–656.

[20] J. Jiang, R. Hu, Z. Wang, Z. Han, and J. Ma, "Facial image hallucination through coupled-layer neighbor embedding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1674–1684, Sep. 2016.

[21] J. Jiang *et al.*, "SRLSP: A face image super-resolution algorithm using smooth regression with local structure prior," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 27–40, Jan. 2017.

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

[23] 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. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, 2010, pp. 2360–2367.

[24] Y. Yang *et al.*, "Salient color names for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 536–551.

[25] 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.*, Boston, MA, USA, 2015, pp. 2197–2206.

[26] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 1363–1372.

[27] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 152–159.

[28] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, 2015.

[29] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

[30] J. Wang, Z. Wang, C. Gao, and N. Sang, "DeepList: Learning deep features with adaptive listwise constraint for person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 513–524, Mar. 2017.

[31] F. Zhu, X. Kong, Z. Liang, H. Fu, and Q. Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4806–4817, Oct. 2017.

[32] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2017.2709749.

[33] 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.*, Providence, RI, USA, 2012, pp. 2288–2295.

[34] Z. Li *et al.*, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 3610–3617.

[35] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 3318–3325.

[36] L. Zheng, S. Duffner, K. Idrissi, C. Garcia, and A. Baskurt, "Pairwise identity verification via linear concentrative metric learning," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2634011.

[37] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person reidentification with reference descriptor," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 776–787, Apr. 2016.

[38] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[39] B. Ma, Y. Su, and F. Jurie, "BiCov: A novel image representation for person re-identification and face verification," in *Proc. Brit. Mach. Vis. Conf.*, Surrey, U.K., 2012, pp. 1–11.

[40] R. Layne, T. M. Hospedales, and S. Gong, "Person re-identification by attributes," in *Proc. Brit. Mach. Vis. Conf.*, vol. 2. Surrey, U.K., 2012, pp. 1–11.

[41] 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.

[42] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 3586–3593.

[43] V. Eiselein, G. Sternharz, T. Senst, I. Keller, and T. Sikora, "Person re-identification using region covariance in a multi-feature approach," in *Proc. Int. Conf. Image Anal. Recognit.*, 2014, pp. 77–84.

[44] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[45] M. Hirzer, C. Beleznai, M. Köstinger, P. M. Roth, and H. Bischof, "Dense appearance modeling and efficient learning of camera transitions for person re-identification," in *Proc. Int. Conf. Image Process.*, Orlando, FL, USA, 2012, pp. 1617–1620.

[46] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 501–512.

[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, Dec. 2009.

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

[49] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 2666–2672.

[50] Y. Wang, R. Hu, C. Liang, C. Zhang, and Q. Leng, "Camera compensation using a feature projection matrix for person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1350–1361, Aug. 2014.

[51] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 1239–1248.

[52] F. Zheng and L. Shao, "Learning cross-view binary identities for fast person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, New York, NY, USA, 2016, pp. 2399–2406.

[53] L. An, X. Chen, S. Yang, and X. Li, "Person re-identification by multi-hypergraph fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2016.2602082.

[54] L. Yang, X. Cao, D. Jin, X. Wang, and D. Meng, "A unified semi-supervised community detection framework using latent space graph regularization," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2585–2598, Nov. 2015.

[55] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveillance*, vol. 3. 2007, pp. 1–7.

[56] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes," in *Proc. Int. Conf. Pattern Recognit.*, Cancún, Mexico, 2016, pp. 2428–2433.

[57] J. Ma, J. Zhao, J. Tian, X. Bai, and Z. Tu, "Regularized vector field learning with sparse approximation for mismatch removal," *Pattern Recognit.*, vol. 46, no. 12, pp. 3519–3532, 2013.

[58] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint *l*2, 1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2010, pp. 1813–1821.

[59] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.

[60] Y. Sun, X. Tao, Y. Li, and J. H. Lu, "Robust two-dimensional principal component analysis via alternating optimization," in *Proc. Int. Conf. Image Process.*, Melbourne, VIC, Australia, 2013, pp. 340–344.

[61] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, 2007.

[62] P. M. Roth, M. Hirzer, M. Köestinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 247–267.

[63] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comp. Vis.*, 2012, pp. 31–44.

[64] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 144–151.

[65] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 34–39.

[66] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.

[67] R. F. D. C. Prates and W. R. Schwartz, "CBRA: Color-based ranking aggregation for person re-identification," in *Proc. Int. Conf. Image Process.*, Quebec City, QC, Canada, 2015, pp. 1975–1979.

[68] Y. Shen *et al.*, "Person re-identification with correspondence structure learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 3200–3208.

[69] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, Buenos Aires, Argentina, 2015, pp. 3402–3408.

[70] W. Yao, Z. Weng, and Y. Zhu, "Diversity regularized metric learning for person re-identification," in *Proc. Int. Conf. Image Process.*, Phoenix, AZ, USA, 2016, pp. 4264–4268.

[71] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 2528–2535.

[72] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 3908–3916.

[73] L. Zheng *et al.*, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1116–1124.

[74] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016. [Online]. Available: http://arxiv.org/abs/1610.02984

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[77] C. Jose and F. Fleuret, "Scalable metric learning via weighted approximate rank component analysis," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 875–890.

[78] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity Based Pattern Recognit.*, 2015, pp. 84–92.

Zheng Wang received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2006 and 2008, respectively, and the Ph.D. degree from the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, in 2017.

His current research interests include multimedia content analysis and retrieval, computer vision, and pattern recognition.

Ruimin Hu (M'09–SM'09) received the B.S. and M.S. degrees from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1984 and 1990, respectively, and the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 1994.

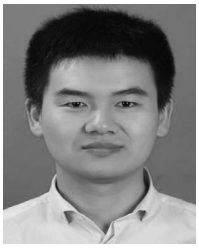
He is the Dean of the School of Computer Science, Wuhan University. His current research interests include audio/video coding and decoding, video surveillance, and multimedia data processing.

Chen Chen (M'16) received the B.E. degree from Beijing Forestry University, Beijing, China, in 2009, the M.S. degree from Mississippi State University, Starkville, MS, USA, in 2012, and the Ph.D. degree from the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX, USA, in 2016.

He is currently a Post-Doctoral Researcher with the Center for Research in Computer Vision, University of Central Florida, Orlando, FL, USA. His current research interests include compressed sensing, signal and image processing, pattern recognition, and computer vision.

Yi Yu received the Ph.D. degree in information and computer science from Nara Women's University, Nara, Japan.

She is currently an Assistant Professor with the National Institute of Informatics (NII), Tokyo, Japan. Before joining NII, she was a Senior Research Fellow with the School of Computing, National University of Singapore, Singapore. Her current research interests include large-scale multimedia data mining and pattern analysis, location-based mobile media service, and social media analysis.



Junjun Jiang (M'15) received the B.S. degree from the School of Mathematical Sciences, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2014.

He is currently an Associate Professor with the School of Computer Science, China University of Geosciences, Wuhan. Since 2016, he has been a Project Researcher with the National Institute of Informatics, Tokyo, Japan. His current research interests include image processing and pattern recognition.

Chao Liang received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012.

He is currently an Assistant Professor with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, China. His current research interests include multimedia content analysis and retrieval, computer vision, and pattern recognition.

Shin'ichi Satoh (M'04) received the B.E. degree in electronics engineering and the M.E. and Ph.D. degrees in information engineering from the University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively.

He has been a Full Professor with the National Institute of Informatics, Tokyo, Japan, since 2004. He was a Visiting Scientist with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, from 1995 to 1997. His current research interests include image processing, video content analysis, and multimedia databases.