Zero-Shot Person Re-identification via Cross-View Consistency

Zheng Wang, Ruimin Hu, *Senior Member, IEEE*, Chao Liang, Yi Yu, Junjun Jiang, *Member, IEEE*, Mang Ye, Jun Chen, and Qingming Leng

***Abstract*—Person re-identification, aiming to identify images of the same person from various cameras configured in different places, has attracted much attention in the multimedia retrieval community. In this problem, choosing a proper distance metric is a crucial aspect, and many classic methods utilize a uniform learnt metric. However, their performance is limited due to ignoring the zero-shot and fine-grained characteristics presented in real person re-identification applications. In this paper, we investigate two consistencies across two cameras, which are *cross-view support consistency* and *cross-view projection consistency*. The philosophy behind it is that, in spite of visual changes in two images of the same person under two camera views, the support sets in their respective views are highly consistent, and after being projected to the same view, their context sets are also highly consistent. Based on the above phenomena, we propose a data-driven distance metric (DDDM) method, re-exploiting the training data to adjust the metric for each query-gallery pair. Experiments conducted on three public data sets have validated the effectiveness of the proposed method, with a significant improvement over three baseline metric learning methods. In particular, on the public VIPeR dataset, the proposed method achieves an accuracy rate of 42.09% at rank-1, which outperforms the state-of-the-art methods by 4.29%.**

***Index Terms*—Cross-view consistency, data-driven distance metric, person re-identification (re-id).**

Manuscript received March 03, 2015; revised September 12, 2015; accepted November 21, 2015. Date of publication December 03, 2015; date of current version January 15, 2016. This work was supported by the National Nature Science Foundation of China under Grant 61231015, Grant 61172173, Grant 61303114, Grant 61170023, Grant 61501413, and Grant 61562048, by the National High Technology Research and Development Program of China under Grant 2015AA016306, by the Internet of Things Development Funding Project of Ministry of Industry in 2013 under Grant 25, by the Technology Research Project of Ministry of Public Security under Grant 2014JSYJA016, by the Major Science and Technology Innovation Plan of Hubei Province under Grant 2013AAA020, by the Nature Science Foundation of Hubei Province under Grant 2014CFB712, by the Nature Science Foundation of Jiangxi Province under Grant 20151BAB217013, and by the Fundamental Research Funds for the Central Universities under Grant 2042014kf0250. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. K. Selcuk Candan.

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

Y. Yu is with the National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: yi.yu.yy@gmail.com).

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

Q. Leng is with the School of Information Science and Technology, Jiujiang University, Jiujiang 332005, China (e-mail: lengqingming@126.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/TMM.2015.2505083

I. Introduction

PERSON re-identification (re-id) is the task of visually matching images of the same person, obtained from different cameras distributed over non-overlapping locations of potentially substantial distances and time differences [3]. It underpins many crucial applications such as person retrieval [4], long-term multi-camera tracking [5], [6], path modeling [7], movement analysis [8], video summarization [9] and forensic search [10]. Since classical biometric cues, such as face and gait, may be unreliable or even infeasible in the uncontrolled surveillance environment [1], the appearance of the individual is mainly exploited for re-id. Generally speaking, re-id can be regarded as a pedestrian-oriented image retrieval problem [13]. Given a probe person image taken from one camera, the algorithm is expected to search images of the same person from the gallery captured by another [14] or multiple cameras [15]. It aims at generating a ranking list, in which top galleries are more likely to be the same person as the probe.

Previous research efforts for solving the re-id problem have primarily focused on the following two aspects.

- *Feature representation* aims at constructing discriminative visual descriptions that are robust and can easily distinguish different persons in various cameras. However, due to low resolution, partial occlusion, motion blur, view change, and illumination variation in various cameras [18], designing discriminative and robust feature representation is extremely challenging, if not impossible, under realistic conditions [1], [11].
- *Distance measure* aims at seeking a proper distance measure by metric learning based on a group of labeled training data. However, the learnt uniform metric usually has a high tendency to overfit to the training data, yielding insufficient results during testing [3].

The goal of this paper is to design and obtain a more effective metric for the re-id problem. Because constructing a discriminative feature representation to adapt to different camera conditions is almost impossible, metric learning based methods have gradually become a main stream procedure in solving the re-id problem. In addition, if we could find excellent visual descriptions for person appearance, the learnt metric could still promote the results [3].

This paper considers the re-id task of comparing samples from two cameras where the probe image comes from one camera and the retrieved gallery images come from the other, the same as previous metric learning literatures [1], [13], [14],

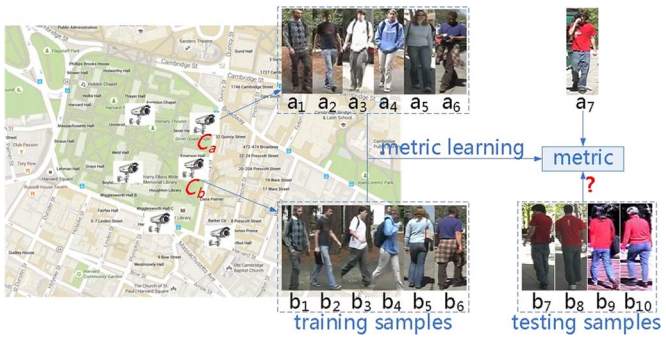


Fig. 1. Example illustrating the characteristics of the re-id task. Training samples $a_1 - a_6$ from camera C_a and $b_1 - b_6$ from camera C_b, where each column represents the same person, are utilized to learn a uniform metric. The learnt metric constructs the relationship between the two cameras [14], and is used to measure the distances between novel person image a_7 and other novel images $b_7 - b_{10}$. It illustrates the characteristic of zero-shot. We can also see that images of different persons $b_7 - b_{10}$ look very similar and are hard to be distinguished. It illustrates the characteristic of fine-grained.

[30]–[36] did, because a network of multiple cameras can be expressed as multiple pairs of cameras. In this condition, each person respectively appears once in both cameras, and then numerous persons will produce multiple image pairs, a part of which are utilized to learn the metric. Most of existing metric learning methods can be summarized as learning a uniform Mahalanobis-like metric matrix $\mathbf{M}$, exploiting $\mathbf{M}$ to calculate the distances between images from the two cameras, and generating the final ranking list according to the distances [3]. Specifically, given two testing person image features x_a and x_b, their distance is defined as $d_{\mathbf{M}}(x_a, x_b) = (x_a - x_b)^\top \mathbf{M}(x_a - x_b)$, where $(\cdot)^\top$ represents the transpose of a vector or matrix.

Although the Mahalanobis-like metric learning methods have achieved good performance, there is still much room for improvement if the following specific characteristics of the re-id task are considered. It should be noted that this cannot be achieved by merely employing a uniform metric.

- *Zero-shot*. As Fig. 1 shows, the samples $a_1 - a_6$ and $b_1 - b_6$ are utilized to learn the metric in the re-id task, while this learnt metric is used to measure the distances between a_7 and $b_7 - b_{10}$. Generally, this kind of incomplete training situation, where the trained model may not be well generalized to cover novel classes (persons) not included in the training set, always appears in real applications. As a result, the training and testing samples have different and potentially unrelated classes (persons), and this is called a zero-shot problem that the learnt model must predict novel values that were omitted from the training set [45]. In this condition, using the uniform metric learnt from the training data set without any adaptation to the testing data set will cause an unknown bias, and the algorithm may fail to identify $b_7 - b_{10}$, the true target image of the same person as in a_7, due to the red clothes not involved in training samples. Therefore, *a uniform metric is improper for the zero-shot re-id task*.
- *Fine-grained*. Re-id is also a fine-grained problem. Images of the same person might not have very similar visual descriptions, whereas images of different persons could happen to be close in the original feature space. In Fig. 1, a_7, $b_7 - b_{10}$ reveal the phenomenon that images of different persons may look similar. Thus, as suggested in [3], this phenomenon leads to somehow ill-posed problems and the trained model is prone to over-fitting. The essence of the metric learning method is to seek a projection matrix that constructs a linear relationship between two cameras [14], but the established relationship is exploited to evaluate image pairs of specific individuals. As each person's appearance holds its unique characteristics, *an ad hoc procedure is extremely necessary for matching image pairs of arbitrary persons*.

Inspired by data-specific adaption research [2], [12], to cover the above issues, we introduce a data-specific adaptive metric, which is particularly depending on query-gallery pair x_a and x_b. An adaptive factor $f(x_a, x_b)$ is used to adjust the uniform metric $\mathbf{M}$ to a new metric $\mathbf{M}_{ab} = \mathbf{M} \cdot f(x_a, x_b)$. The training data is re-exploited in our method, and the relationships between new inputs and training data can be learned to obtain the factor. Then, the distance between x_a and x_b is computed as $d_{\mathbf{M}_{ab}}(x_a, x_b) = (x_a - x_b)^\top \mathbf{M}_{ab}(x_a - x_b)$. To support the above idea, we investigate two phenomena, or called consistencies, in the re-id problem, which can be exploited to adaptively adjust the uniform metric for different query-gallery pairs. We respectively name them as *cross-view support consistency* and *cross-view projection consistency*, whose details will be described in Section IV.

Based on the phenomena that an image pair from the same person would have stronger cross-view support and projection consistency than those from different persons, the proposed method learns the cross-view support and projection adaptive factors respectively, and generates a new data-specific metric for each new image pair. In this way, the ranking results are refined. The contribution of this paper is summarized as follows.

- We find two phenomena, cross-view support and projection consistencies, which are stronger for the same person than for different persons. In other words, two images of the same person should not only have similar support set across two camera views, but also hold similar context after being projected to the same camera view.
- We propose a data-specific adaptive metric method to conquer the zero-shot and fine-grained difficulties in the re-id problem. Compared to traditional metric learning methods, such as Mahalanobis, LMNN [30] and KISSME [33], which adopt a common metric for different queries, the proposed method mines the associations between test samples and the training set, and re-exploits the training samples to adjust the metric for specific image pairs. In this way, the proposed method can improve the Cumulative Match Characteristic (CMC) [21] results significantly, and it is validated by experiments conducted on three public data sets.

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 formally define the zero-shot re-id problem. In Section IV, we illustrate the cross view consistency and its phenomena. Then, we detail the proposed method with cross-view support consistency and cross-view projection consistency in

Section V. Section VI shows experimental results and finally Section VII concludes this paper.

II. Related Work

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

The feature representation based approaches aim to construct discriminative visual descriptions that can easily separate different persons in various cameras. Gheissari *et al.* [20] used a spatial-temporal segmentation algorithm to generate salient edges and obtained an invariant identity signature by combining normalized color and salient edge histograms. Wang *et al.* [21] studied an appearance model using a co-occurrence matrix to capture the spatial distribution of the appearance relative to each of the object parts. Farenzena *et al.* [22] tried to combine multiple features to describe the appearance image, which was divided into regions by exploiting symmetry and asymmetry perceptual principles. Ma *et al.* [23] developed a representation that relies on the combination of biologically inspired features and covariance descriptors. Layne *et al.* [24] learned a selection and weighting of mid-level semantic attributes to describe people. Kviatkovsky *et al.* [25] found that some aspects of color structure turn out to be invariants under different lighting conditions, and then used shape context descriptors to represent the intra-distribution structure. Zhao *et al.* [26] applied adjacency constrained patch matching to build dense correspondence between image pairs, and assigned salience to each patch in an unsupervised manner. Li *et al.* [27] utilized a unified deep architecture to learn a filter for re-id. To increase the discriminative power of feature representation, feature selection technique is also adopted in the re-id research [28], [29]. However, constructing a robust and discriminative description is extremely challenging if not impossible under realistic conditions [1].

The distance measure based approaches pay attention to seeking out a proper distance measure, which is a crucial aspect [1] in re-id procedure. Among various methods, supervised metric learning algorithms demonstrate an obvious superiority by learning a uniform metric based on the given labeled training data. Hizer *et al.* [31] and Dikmen *et al.* [32] utilized or improved LMNN [30] to learn the optimal metric for person re-id. Zheng *et al.* [1] learned a Mahalanobis distance metric with a probabilistic relative distance comparison (PRDC) method. Kostinger *et al.* [33] used Gaussian distribution to fit pair-wise samples and got a simpler metric function (KISSME), and then Tao *et al.* [13] presented a regularized smoothing KISS metric learning (RS-KISS) by seamlessly integrating smoothing and regularization techniques for robustly estimating the covariance matrices. Mignon *et al.* [34] 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 PCA dimensionality reduction and Local Fisher Discriminant Analysis defined by a training set to do the metric work. Li *et al.* [36] 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.* [14] transformed the metric learning problem to a feature projection matrix learning problem that project image features of one camera to the feature space of other camera. Most of the above methods can be summarized as learning a uniform Mahalanobis-like distance matrix $\mathbf{M}$, and exploiting the uniform learned matrix $\mathbf{M}$ to calculate the distances among testing data and then generate the ranking result [3]. However, it is obviously inappropriate to merely employ a uniform metric for the zero-shot re-id problem, which has been discussed in Section I.

There are still few methods trying to refine the original results generated by the feature-based or distance-based methods. Leng *et al.* [16] utilized the individual and social relationships among images, and proposed bidirectional ranking method without human interaction. Recently, few relevance feedback methods [17]–[19] also have been proposed. However, the effectiveness of these methods is highly depending on the initial results or human feedback. Therefore, this kind of methods is not within the scope of this paper.

III. Zero-Shot Person re-id

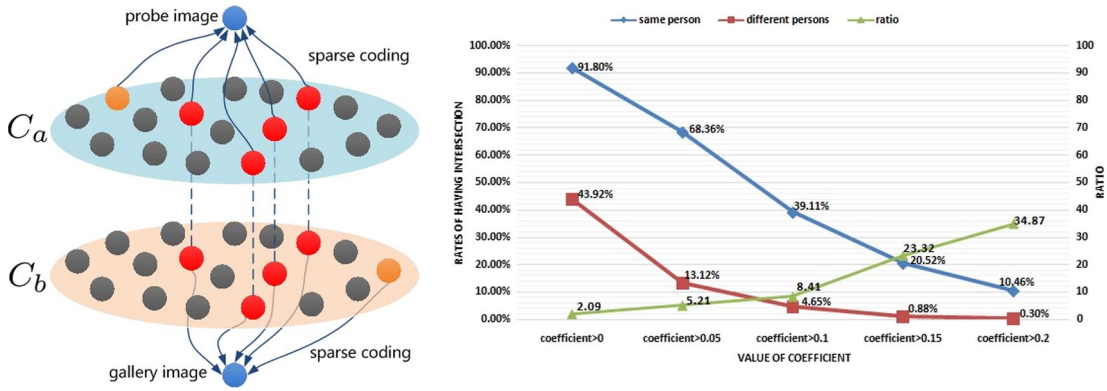
This section gives a brief definition of the zero-shot person re-id problem and the general metric learning approach. We consider a pair of cameras C_a and C_b with non-overlapping field of views. A set of labeled persons $O = \{o_1, o_2, \ldots, o_m\}$ is associated with the two cameras, where m is the number of persons. We denote the representative image of person o_i captured by C_a (or C_b) as x_a^i (or x_b^i), $x_a^i, x_b^i \in R^d$. Let $X_{a,L} = \{x_a^1, \ldots, x_a^i, \ldots, x_a^m\}$, $1 \leq i \leq m$ and $X_{b,L} = \{x_b^1, \ldots, x_b^j, \ldots, x_b^m\}$, $1 \leq j \leq m$ respectively represent the two labeled training sets captured by C_a and C_b, where $i = j$ means the same person o_i.

Let x_a^p stand for a testing probe data from C_a, and $X_{b,U} = \{x_b^{m+1}, \ldots, x_b^q, \ldots, x_b^{m+n}\}$, $m+1 \leq q \leq m+n$ represent the unlabeled test data from C_b. n is the number of testing data in C_b, so the persons for training and testing are different. We denote the probe set in C_a as $X_{a,U}$. Then, for each testing probe data $x_a^p \in X_{a,U}$, the zero-shot re-id task ranks the data in $X_{b,U}$.

A traditional supervised metric learning algorithm learns a discriminative distance function based on $X_{a,L}$ and $X_{b,L}$. Specifically, given a pair of training samples x_a^i and x_b^j, their distance can be defined as a Mahalanobis-like distance $d_{\mathbf{M}}(x_a^i, x_b^j) = (x_a^i - x_b^j)^\top \mathbf{M}(x_a^i - x_b^j)$, where $\mathbf{M}$ is a positive semi-definite matrix for the validity of metric. After learning the uniform metric matrix $\mathbf{M}$, the distance between testing probe data x_a^p and any unlabeled testing data x_b^q in $X_{b,U}$ will be calculated as

$$d_{\mathbf{M}}(x_a^p, x_b^q) = (x_a^p - x_b^q)^\top \mathbf{M}(x_a^p - x_b^q). \tag{1}$$

By performing matrix decomposition on $\mathbf{M}$ with $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$, the above distance can be rewritten as $d_{\mathbf{M}}(x_a^p, x_b^q) = (x_a^p - x_b^q)^\top \mathbf{L}^\top \mathbf{L}(x_a^p - x_b^q) = \|\mathbf{L} \cdot x_a^p - \mathbf{L} \cdot x_b^q\|^2$. With this definition, it is easy to see that the essence of the metric is to seek a projection matrix that transforms original image features into a new feature space where dimensions are not correlated. In this way, it is equivalent to converting two images from camera C_a/C_b to the virtual camera C_v and compute their distance by the Euclidean distance. Hence, the Mahalanobis-like distance



(a) cross-view support consistency

(b) results of cross-view support consistency

Fig. 2. (a) Illustration of cross-view support consistency. The colorful samples in C_a (C_b) are selected by sparse coding method to represent the probe image (the gallery image), acting as the support set. The probe-gallery image pair comes from the same person. Cross-view support consistency means that most of selected images are from same persons as the connected red samples demonstrate. (b) Preliminary experiment results for illustrating the cross-view support consistency phenomenon. The set of 632 image pairs (VIPeR [41]) were randomly split into two subsets, one for the dictionary subset and the other for the evaluation. The rates of having intersection were recorded at different coefficient of sparse coding for the image pairs from the same persons and different persons. The entire evaluation procedure was looped three times for different subset divisions. The dictionary subsets were respectively 500, 400, and 300 image pairs, and evaluation subsets were respectively 132, 232, and 332 image pairs. The results show the average values of three experiments. We also calculated the ratios of the values of the same persons (the blue curve) to that of different persons (the red curve), as the green curve shows.

is applied to the images from two different cameras, while the Euclidean distance is directly exploited for the images from the same camera.

Traditional re-id methods compute the distance $d_{\mathbf{M}}(x_a^p, x_b^q)$ for every pair of test images with the same $\mathbf{M}$, and then obtain the ranking list. In comparison, this paper aims to obtain more suitable $\mathbf{M}$ in distance measure for each specific pair of images.

IV. Cross-View Consistency

This section explains the motivation of and lays a basis for the method proposed in the next section. It addresses two phenomena: images from the same person will have stronger cross-view support and projection consistencies than those from different persons. To facilitate this understanding, we also show some experimental results.

A. Cross-View Support Consistency

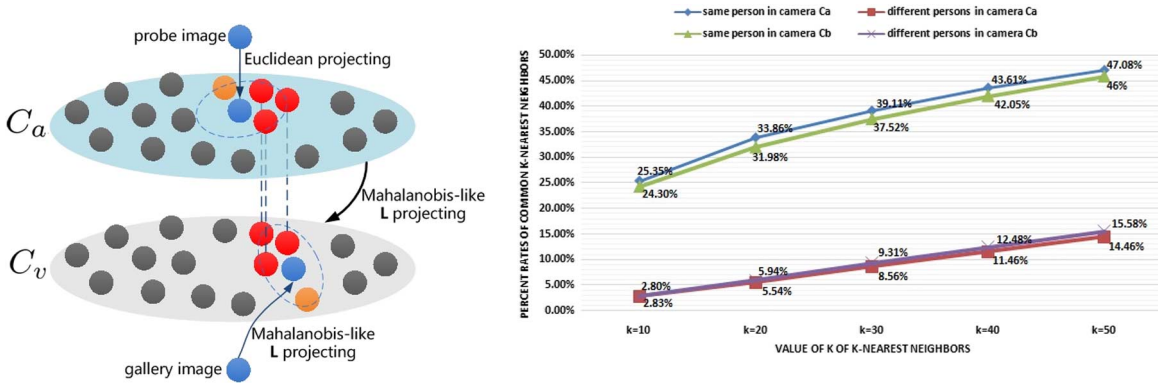
In spite of the existence of salient viewpoint warps and object occlusions, it is reasonable to assume that, from C_a to C_b, the visual appearance of each person will encounter almost the same illumination variation and blur change. For example, assuming that the representations of one person are respectively I_a in C_a and I_b in C_b, and only illumination variation $\mathbf{V}$ and blur change $\mathbf{B}$ exist between two images of the same person from C_a to C_b. Then the transformation is expressed as $I_b = I_a \cdot \mathbf{V} \cdot \mathbf{B}$. If I_a can be represented by three other person images in C_a as $I_a = \omega_1 I_a^1 + \omega_2 I_a^2 + \omega_3 I_a^3$, I_b can be represented as $I_b = \omega_1 I_a^1 \cdot \mathbf{V} \cdot \mathbf{B} + \omega_2 I_a^2 \cdot \mathbf{V} \cdot \mathbf{B} + \omega_3 I_a^3 \cdot \mathbf{V} \cdot \mathbf{B}$. Here, $I_a^1 \cdot \mathbf{V} \cdot \mathbf{B}$, $I_a^2 \cdot \mathbf{V} \cdot \mathbf{B}$ and $I_a^3 \cdot \mathbf{V} \cdot \mathbf{B}$ stand for the transformation form of the selected images I_b^1, I_b^2 and I_b^3 in C_b. Therefore, under this common condition, we can conclude that if the image of one person can be visually represented by a set of images of persons from camera C_a, another image of the same person in camera C_b should also be represented by images of the same set of persons in camera C_b. Considering the noises, occasional warps and occlusions, the rule will not be so precise. However, in a statistical sense, most persons hold proximate consistencies.

We define the support set of an image as the sparsely selected images which together represent the image. The basic idea of cross-view support consistency is that two images of the same person should have similar support set across two camera views [37]–[39]. As Fig. 2(a) shows (which is appropriate to be watched in the color mode), the blue pair of probe-gallery images have support sets respectively, where images that construct a support set of an image are connected to the image by lines. Because the two images are from the same person, they share most of their support sets in common, as marked in the red color.

We made an experiment to explain this phenomenon in re-id. We chose the VIPeR dataset [41], which was collected from two different cameras, and in which each person had a pair of images taken from those two cameras respectively. The dataset considered different influences between two cameras, such as illumination variation, resolution and viewpoint changes. Two personal ID one-to-one corresponding dictionaries were constructed respectively by the images from camera C_a and camera C_b. A sparse coding method [44] was utilized to learn the representation coefficients. The non-zero items, standing for the personal IDs selected, compose the support set. It is evident from Fig. 2(b) that rates of having intersection, the proportion of the number of image pairs holding non-empty intersection of support sets to the total number of image pairs for evaluation, are always much higher for the same person than that for the different persons. It also shows that the ratio of the value of the same persons to that of the different persons will become higher at a higher sparse representation coefficient, which stands for more important support set. So, we can see that the cross-view support consistency is strong for the same person.

B. Cross-View Projection Consistency

The basic idea of cross-view projection consistency is that two images of the same person should have similar neighbor-



(a) cross-view projection consistency

(b) results of cross-view projection consistency

Fig. 3. (a) Illustration of cross-view projection consistency in C_a. The probe image is from C_a, and k-nearest neighbors are obtained by Euclidean distance, as the colorful samples in the neighborhood indicate. The gallery image and images in C_a are projected into virtual camera C_v, and then k-nearest neighbors are obtained. The probe-gallery image pair comes from the same person. Cross-view projection consistency means that most of k-nearest neighbors samples are common as the dark red samples demonstrate. (b) Preliminary experiment results for illustrating the cross-view projection consistency phenomenon. The entire evaluation procedure was looped three times for different subset divisions, as the support consistency experiment did. So, the reference subsets were respectively 500, 400, and 300 image pairs, and evaluation subsets were respectively 132, 232, and 332 image pairs. Common k-nearest neighbors of each evaluation image pairs are obtained as the illustration does. The average percent rates of common k-nearest neighbors were recorded at different values of k for the image pairs from the same persons and different persons respectively in C_a and C_b.

hood across two camera views. The context set of an image is defined as the k-nearest neighbors of the image, and two images of the same person should hold similar context not only in camera C_a, but also in camera C_b.

As defined in Section III, in order to get the common k-nearest neighbors of image pairs in C_a, for images from camera C_a, the distances were calculated directly by Euclidean distance. While, images from camera C_b and the basic reference subset in C_a were both projected into C_v, and the distances were calculated by the learnt metric (KISSME [33]). Fig. 3(a) shows a blue pair of probe-gallery images from different views, and their k-nearest neighborhoods in the same view. The neighborhood of the probe image is directly obtained, while that of the gallery image is obtained by first projecting the gallery image to the view C_v. Because the two images come from the same person, their neighborhoods share most in common, as marked in the red color.

We also made an experiment to explain this phenomenon. Common k-nearest neighbors of each evaluation image pairs are obtained as the illustration does, including the same person pairs and different person pairs. The average percent rates of common k-nearest neighbors were recorded respectively in C_a and C_b. It is evident from Fig. 3(b) that the percentages of common k-nearest neighbors for the same person are always much higher than those for the different persons in both camera views. So, we can also see that the cross-view projection consistency is strong for the same person.

V. Data-Driven Distance Metric

This section presents our approach. The framework is first introduced, followed by discussions on cross-view support consistency and cross-view projection consistency. Finally, a data-driven adaption metric is designed to refine the original uniform metric.

A. Framework of the Proposed Method

Based on the general metric learning method, the proposed method obtains adaptive factors to adjust the learnt uniform metric $\mathbf{M}$. For a new image pair x_a^p and x_b^q, our approach exploits adaptive metric $\mathbf{M}_{pq}$ to obtain the data-specific distance

$$d_{\mathbf{M}_{pq}}\left(x_a^p, x_b^q\right) = \left(x_a^p - x_b^q\right)^\top \mathbf{M}_{pq}\left(x_a^p - x_b^q\right). \quad (2)$$

Here, the data-specific metric $\mathbf{M}_{pq}$ is obtained by the uniform metric $\mathbf{M}$ and adaptive factors (Fig. 4), which include the cross-view support adaptive factor f_s and the cross-view projection adaptive factor f_p. Then, the adaptive metric $\mathbf{M}_{pq}$ is obtained by (3) and new distances are generated

$$\mathbf{M}_{pq} = f_s\left(x_a^p, x_b^q\right) \cdot f_p\left(x_a^p, x_b^q\right) \cdot \mathbf{M}. \quad (3)$$

B. Cross-View Support Factor

First, we study the cross-view support consistency. For a probe image x_a^p from C_a, the approach selects some images sparsely in $X_{a,L}$ to encode the image, meanwhile another gallery image x_b^q from C_b is represented sparsely by images in $X_{b,L}$. We generate the dictionary by the training set $X_{a,L}$, denoted as $D_a = [x_a^1 \ldots x_a^i \ldots x_a^m]$, $D_a \in R^{d\times m}$, and then select sparsely from the dictionary to represent x_a^p, as $x_a^p = D_a w_a^{p*}$, where $(\cdot)^*$ stands for the solution and $w_a^{p*} \in R^{m\times 1}$ indicates the selected images and their coefficients [44].[1] If the coefficient of an image is greater than zero, it means that the image is selected as the support data for x_a^p. Here, we use (4) to compute the sparse representation w_a^{p*}

$$w_a^{p*} = \arg\min_{w_a^p \geq 0} \frac{1}{2}\left\|D_a w_a^p - x_a^p\right\|_2^2 + \frac{\rho}{2}\left\|w_a^p\right\|_2^2 + \lambda\left\|w_a^p\right\|_1. \quad (4)$$

[1]The algorithm learns the weights by the SLEP package version 4.1 [48]. The original formulation is $w^* = \arg\min(1/2)\|Dw - x\|_2^2 + (\rho/2)\|w\|_2^2 + \lambda\|w\|_1$, where λ is the l_1-norm regularization parameter, and ρ ($\rho = 0$ by default) is the regularization parameter for the squared l_2-norm.

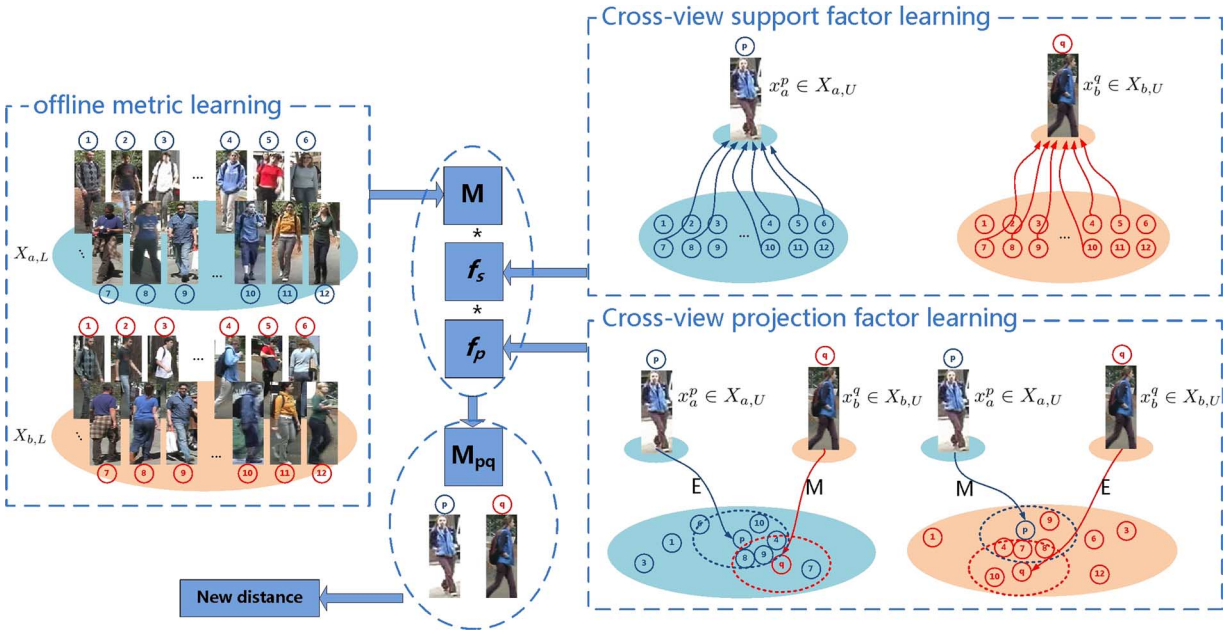


Fig. 4. Framework of the proposed method. In the offline stage, as general metric learning methods do, a uniform metric $\mathbf{M}$ is learned. In the evaluation stage, for a novel image pair x_a^p and x_b^q, through re-exploiting the training set, the cross-view support adaptive factor f_s and the cross-view projection adaptive factor f_p are learned by the associations between training samples and the image pair. (1) Through sparse coding, the algorithm respectively learns the coefficients of representing the image x_a^p and x_b^q. Measuring the consistency of coefficients or judging whether a person is represented by same persons across views, we can obtain the cross-view support adaptive factor f_s. (2) Through projecting the images into both of the camera views/domains, the algorithm respectively learns the context of the image x_a^p and x_b^q, where $\mathbf{E}$ indicates that the distances are obtained by Euclidean distance, and $\mathbf{M}$ indicates that the distances are obtained by the learned metric. Counting the common k-nearest neighbors, we can obtain the cross-view projection adaptive factor f_p. Then, an adaptive metric $\mathbf{M}_{pq}$ can be obtained and a new distance is generated.

In the same way, the dictionary $D_b = [x_b^1 \dots x_b^j \dots x_b^m]$, $D_b \in R^{d\times m}$ is generated by the training set $X_{b,L}$, and x_b^q can be represented sparsely by D_b, as $x_b^q = D_b w_b^{q*}$. We get the sparse representation $w_b^{q*} \in R^{m\times 1}$ for x_b^q by (5)

$$w_b^{q*} = \arg\min_{w_b^q \geq 0} \frac{1}{2}\left\|D_b w_b^q - x_b^q\right\|_2^2 + \frac{\rho}{2}\left\|w_b^q\right\|_2^2 + \lambda\left\|w_b^q\right\|_1. \quad (5)$$

Then, the cross-view support consistency is defined as (6), where $.*$ stands for the element-wise multiplication, and $size(\cdot)$ counts the number of non-zero elements in the vector

$$sc\left(x_a^p, x_b^q\right) = size\left(w_a^{p*} .* w_b^{q*}\right). \quad (6)$$

With the cross-view support consistency, we construct the support adaptive factor. As Fig. 5 illustrates, the cross-view support consistency sc between p and $q1$ is stronger than that between p and $q2$, and after metric adaption, $q1$ is closer to p than $q2$. So, if the support consistency is stronger, the value of support adaptive factor f_s would be smaller, which is denoted by (7)

$$f_s\left(x_a^p, x_b^q\right) = \left[\frac{1}{1 + sc\left(x_a^p, x_b^q\right)}\right]^\alpha = \left[\frac{1}{1 + size\left(w_a^{p*} .* w_b^{q*}\right)}\right]^\alpha, \quad \alpha > 0. \quad (7)$$

The parameter α indicates the contribution of the cross-view support consistency. The greater the value is, the greater the impact of metric adaption derived from the support consistency detection. If α is too large, the origin uniform metric will lose efficacy. While, if $\alpha \to 0$, $\mathbf{M}$ will not change, and the support adaptive factor will have not effect.

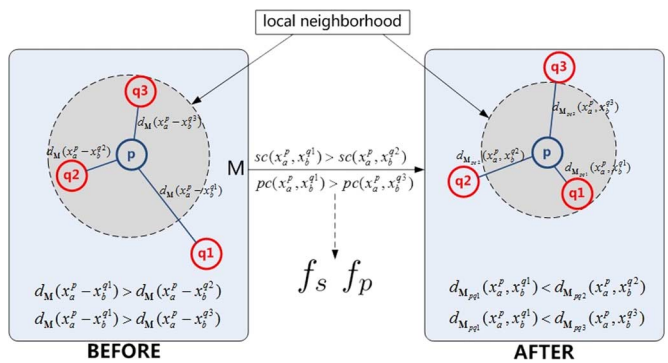


Fig. 5. Example illustrating the improvement by exploiting the support and projection consistency. In the figure, p denotes the probe image in C_a, and $q1$, $q2$ and $q3$ denote three retrieved samples in C_b. $q1$ and p are two images from the same person, $q2$ and $q3$ are from different persons. We can see that although the true target $q1$ is farther away from p than $q2$ and $q3$ with a uniform learnt metric, $q1$ is closer than $q2$ and $q3$ after metric adaptions. In this way, $q1$ can get a higher ranking in the results.

C. Cross-View Projection Factor

Second, we exploit the cross-view projection consistency. It is valid to assume that images coming from the same person should be alike individually as well as socially [16]. In this work, we exploit the property that a probe image x_a^p from C_a and the target image from C_b should hold a similar context with the training sets $X_{a,L}$ and $X_{b,L}$ respectively.

The essence of the metric based method is to seek a projection matrix that transforms original image features into a new feature space. As Section IV-B shows, when we want to obtain the neighbors of x_a^p in $X_{b,L}$, we should project both of the cameras C_a and C_b into the same domain. So the distance between x_a^p and each training data x_b^j is calculated by $d_{\mathbf{M}}(x_a^p, x_b^j) = (x_a^p - x_b^j)^\top \mathbf{M}(x_a^p - x_b^j)$. Meanwhile, every testing data x_b^q and the training set $X_{b,L}$ are in the same camera C_b, and the Euclidean distance $d_{\mathbf{E}}(x_b^q, x_b^j) = (x_b^q - x_b^j)^\top (x_b^q - x_b^j)$ is used.

Through ranking the distances with the training set $X_{b,L}$, the k-nearest neighbors of x_a^p and x_b^q can be acquired respectively. Counting the number of common k-nearest neighbors of x_a^p and x_b^q in $X_{b,L}$, the context similarity is computed. In particular, the set of k-nearest neighbors of x_a^p in $X_{b,L}$ is denoted as $knn(x_a^p|X_{b,L})$, and that of x_b^q is $knn(x_b^q|X_{b,L})$. Then the projection consistency in C_b is defined as

$$pc_b\left(x_a^p, x_b^q\right) = \left|knn\left(x_a^p|X_{b,L}\right) \cap knn\left(x_b^q|X_{b,L}\right)\right| \tag{8}$$

where $|knn(x_a^p|X_{b,L}) \cap knn(x_b^q|X_{b,L})|$ represents the number of common k-nearest neighbors of x_a^p and x_b^q in training data $X_{b,L}$.

In the same way, the distance between x_b^q and every training image x_a^i is computed by $d_{\mathbf{M}}(x_b^q, x_a^i) = (x_b^q - x_a^i)^\top \mathbf{M}(x_b^q - x_a^i)$. So $knn(x_b^q|X_{a,L})$, the k-nearest neighbors of x_b^q in $X_{a,L}$, is generated. The distance between x_a^p and x_a^i is calculated by $d_{\mathbf{E}}(x_a^p, x_a^i) = (x_a^p - x_a^i)^\top (x_a^p - x_a^i)$, and $knn(x_a^p|X_{a,L})$ is generated. Then, the projection consistency in C_a is defined as (9), and the total projection consistency is calculated by (10)

$$pc_a\left(x_a^p, x_b^q\right) = \left|knn\left(x_a^p|X_{a,L}\right) \cap knn\left(x_b^q|X_{a,L}\right)\right| \tag{9}$$

$$pc\left(x_a^p, x_b^q\right) = pc_a\left(x_a^p, x_b^q\right) + pc_b\left(x_a^p, x_b^q\right). \tag{10}$$

As Fig. 5 illustrates, the cross-view projection consistency pc between p and $q1$ is stronger than that between p and $q3$, and after metric adaption, $q1$ is closer to p than $q3$. So, if the projection consistency is stronger, the value of projection adaptive factor f_p would be smaller, which is denoted by (11). The parameter β indicates the contribution of the cross-view projection consistency

$$\begin{aligned} f_p\left(x_a^p, x_b^q\right) &= \left[\frac{1}{1 + pc\left(x_a^p, x_b^q\right)}\right]^\beta \\ &= \left[\frac{1}{1 + pc_a\left(x_a^p, x_b^q\right) + pc_b\left(x_a^p, x_b^q\right)}\right]^\beta, \qquad \beta > 0. \end{aligned} \tag{11}$$

Learning the two factors, we can obtain a new metric specific for each x_a^p and x_b^q by (3). The whole procedure of DDDM is described in Algorithm 1. However, in order to find the suitable value of α and β, the algorithm adopts the idea of cross validation, by dividing the training set into two parts, one for metric learning and the other for obtaining the parameters.

D. Analysis of Complexity

As can be seen from Algorithm 1, the majority of the computation is spent in learning two factors, which consists of mass sparse coding and pair-wise distance computation. A traditional way of this step is directly adjusting the metric of every pair of images, and generating the ranking list based on image distances. As assumed in Section III, the training sets $X_{a,L}$ and $X_{b,L}$ respectively contain m labeled images, and the testing sets $X_{a,U}$ and $X_{b,U}$ respectively contain n labeled images. For each query in $X_{a,U}$, n unlabeled images in $X_{b,U}$ will require n times of sparse coding processes and $n * m$ times of distance measure processes. Therefore, the computation complexity is $O(n^2)$ in sparse coding processes, and $O(n^2 m)$ in distance measure processes. However, we find that gallery images can be obtained before querying the probe image in the practical application. Therefore, our method is revised to two stages as Algorithm 2 describes. In the offline stage, for all the testing images in $X_{b,U}$ sparse representation coefficients are computed with a computation complexity $O(n)$ and distances are computed with a complexity $O(nm)$. In the online stage, it only needs to compute the coefficients and distances of the probe image, whose computation complexity is $O(1)$ and $O(m)$. With the above two-stage implementations, the complexity of the whole algorithm can be greatly reduced and its online part is only proportional to the size of reference set (training samples), which is especially suited to those real-time applications.[2]

Algorithm 1 Algorithm of the data-driven distance metric method.

Input: training sets $X_{a,L}$ and $X_{b,L}$, probe image $x_a^p \in X_{a,U}$ and a gallery set $X_{b,U}$.
Output: a ranking list of $X_{b,U}$ for the probe image x_a^p.
1: learn the uniform metric $\mathbf{M}$ by $X_{a,L}$ and $X_{b,L}$;
2: **for** each $q \in [m+1, m+n]$ **do**
3: get x_b^q from $X_{b,U}$;
4: learn the factor $f_s(x_a^p, x_b^q)$ by (7);
5: learn the factor $f_p(x_a^p, x_b^q)$ by (11);
6: adjust the metric $\mathbf{M}$ to $\mathbf{M}_{pq}$ by (3);
7: compute the distance $d_{\mathbf{M}_{pq}}(x_a^p, x_b^q)$ by (2);
8: **end for**
9: rank the distances from low to high and generate the ranking list.

Algorithm 2: Modified algorithm of the data-driven distance metric method.

Input: training sets $X_{a,L}$ and $X_{b,L}$, probe image $x_a^p \in X_{a,U}$ and a gallery set $X_{b,U}$.
Output: a ranking list of $X_{b,U}$ for the probe image x_a^p.
Offline:
1: learn the uniform metric $\mathbf{M}$ by $X_{a,L}$ and $X_{b,L}$;
2: **for** each $q \in [m+1, m+n]$ **do**
3: get x_b^q from $X_{b,U}$;
4: calculate and save the sparse representation w_b^{q*} by (5);
5: **for** each $i \in [1, m]$ **do**
6: compute the distance $d_{\mathbf{M}}(x_b^q, x_a^i)$ by Mahalanobis-like metric;
7: compute the distance $d_{\mathbf{E}}(x_b^q, x_b^i)$ by Euclidean metric;
8: **end for**

[2]For each probe image, the computation time costs 47 ms for VIPeR [41] data set, 76 ms for CUHK [42] data set, and 17 ms for PRID [40] data set, whose training pairs are respectively 316, 485, and 100. The CPU of the computer includes double core 2.80 GHz processors and 2 GB RAM.

(a) VIPeR dataset (b) CUHK dataset (c) PRID dataset

Fig. 6. Some typical samples of three public datasets. Each column shows two images of the same person from two different cameras. (a) VIPeR dataset. (b) CUHK dataset. (c) PRID dataset.

9: obtain and save the k-nearest neighbors $knn(x_b^q|X_{a,L})$ and $knn(x_b^q|X_{b,L})$;
10: **end for**

Online:
1: **for** each $q \in [m+1, m+n]$ **do**
2: get x_b^q from $X_{b,U}$;
3: calculate w_a^{p*} by (4), and compute the factor $f_s(x_a^p, x_b^q)$ with saved w_b^{q*} by (7);
4: obtain the k-nearest neighbors $knn(x_a^p|X_{a,L})$ and $knn(x_a^p|X_{b,L})$, and compute the factor $f_p(x_a^p, x_b^q)$ by (11);
5: adjust the metric $\mathbf{M}$ to $\mathbf{M}_{pq}$ by (3);
6: compute the distance $d_{\mathbf{M}_{pq}}(x_a^p, x_b^q)$ by (2);
7: **end for**
8: rank the distances from low to high and generate the ranking list.

VI. Experiments

In this section, the proposed method DDDM is evaluated, by comparing with three metric learning methods, standard Mahalanobis distance, complicated learning using LMNN [30] that is a classic metric learning method and the state-of-the-art KISSME [33] that reported a good performance on the VIPeR. It is believed that the proposed method can be proved effective for adapting general metric learning methods, if the three typical methods outperform their original ways. The evaluation is run on three publicly available datasets, the VIPeR dataset [41], the CUHK dataset [42] and the PRID dataset [40]. These datasets cover a wide range of problems faced in the real world person re-id applications, e.g. viewpoint, pose, and lighting changes. They provide two labeled image sets of persons captured by two cameras with non-overlapping fields of views, in which the images of the same person have the same label, while the images of the different persons have different labels.

A. Datasets

The widely used VIPeR dataset [41] contains 1,264 outdoor images obtained from two views of 632 persons. Some example images are shown in Fig. 6(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 with most of the matched image pairs containing a viewpoint change of 90 degrees. Other variations are also considered, such as illumination conditions and the image qualities.

The CUHK dataset is a larger dataset recently released by Wang *et al.* [42] and contains 971 identities from two disjoint camera views. Each identity has two samples per camera view. Some example images are shown in Fig. 6(b). Each identity has two samples in per camera view. Therefore, there are 3,884 images in all. All images are normalized to 160 × 60. Similar to VIPeR, 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. As 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 view for each person as the really used dataset.

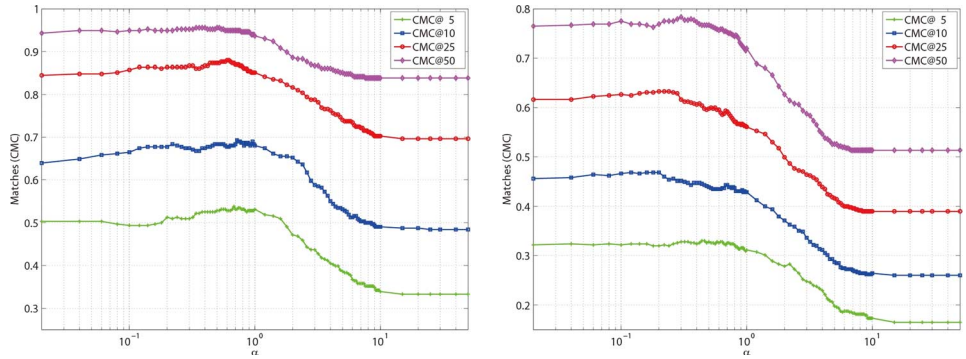
The PRID [40] is a challenge dataset, particularly there is camera characteristics variation. 385 persons images are from one camera and 749 persons images are from the other camera, with 200 common images in both views. All images are normalized to 128 × 48 pixels. Different to VIPeR dataset and CUHK dataset, this dataset has significant and consistent lighting changes [see Fig. 6(c)]. We choose the 200 persons appearing in both views as the used dataset.

B. Feature Representation

A combination feature descriptor consisting of color and texture features is used to represent images of individuals, which follows the model described in [14]. For each image, the color and texture features are extracted from overlapping blocks of size 16 × 16 (16 × 12 for CUHK) and stride of 8 × 8 (8 × 6 for CUHK). The color distribution information is encoded by RGB and HSV histograms, and the texture feature is encoded by LBP descriptors [43]. The bin number of color distribution information is 24, and that of LBP descriptor is 59. All the features are concatenated to a vector. To accelerate the learning process and reduce noise, we conducted principal component analysis (PCA) using covariance [47] to obtain a relatively low dimensional representation as [33], i.e. 50 in this paper unless otherwise specified.

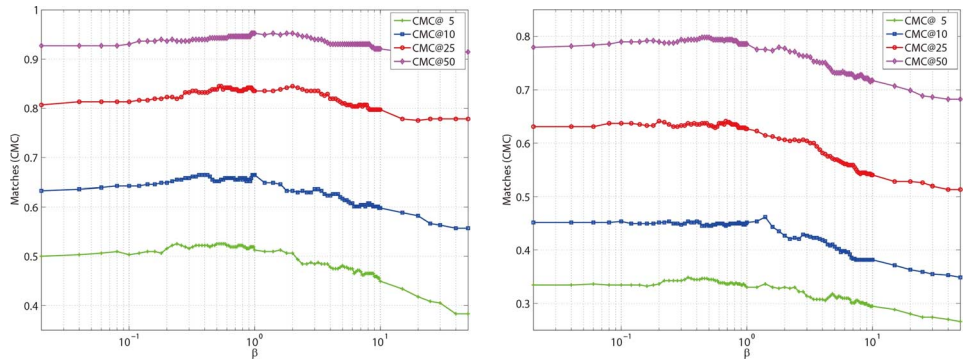
C. Baselines and Settings

To evaluate the effectiveness of the proposed method, we reform the uniform metric methods with DDDM, containing Mahalanobis metric, LMNN and KISSME. Similar to [1], our experiments were designed as follows. First, m (e.g. 316 in VIPeR, 485 in CUHK, 100 in PRID) image pairs were randomly selected as the training set and the rest as the testing set. Second, the test set was divided into a probe set consisting of images taken from Camera C_a, and gallery set made up of images captured by Camera C_b. Finally, each image in the probe set was



(a) On VIPeR dataset (b) On CUHK dataset

Fig. 7. Comparative results of different αs on VIPeR and CUHK datasets. (a) On VIPeR dataset. (b) On CUHK dataset.



(a) On VIPeR dataset (b) On CUHK dataset

Fig. 8. Comparative results of different βs on VIPeR and CUHK datasets. (a) On VIPeR dataset. (b) On CUHK dataset.

matched with all images in the gallery set, and the rank of the real match was recorded. We evaluate the three metric learning methods and their upgraded versions with the proposed DDDM method. The entire evaluation procedure was repeated 10 times. Cumulative Matching Characteristic (CMC) curves [21] were used to calculate the average performance, and the value of CMC@k indicates the percentage of the real match ranked in the top k.

D. Evaluating Parameters of the Proposed Method

In this subsection, we validate the proposed approach under different parameters, including exploiting different parameter α for the contribution of the cross-view support consistency, and different parameter β for the cross-view projection consistency.

Influence of α. We conduct experiments under different α values for further evaluating the effectiveness of cross-view support consistency. To reduce the impact of cross-view projection consistency, we set $\beta = 0$ in this stage. When $\alpha = 0$, it is equal to only exploit the learnt uniform metric; inversely, it only uses the projection consistency to generate ranking result for $\alpha \to \infty$. We change α from 0 to 50, the corresponding results are shown in Fig. 7. It is obvious that 0.5 is a good choice for α on the VIPeR dataset and 0.2 on the CUHK dataset. When $\alpha > 10$, the performance is stable, because the metric is almost entirely depending on the cross-view support factor. The experiment results are consistent with the above claim, that the algorithm achieves good performance under a suitable α.

Influence of β. We also conduct experiments under different β values for further evaluating the effectiveness of cross-view

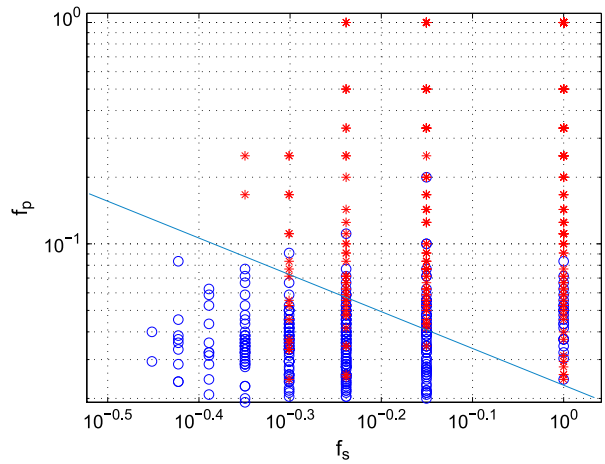


Fig. 9. Distribution of f_s and f_p. A blue "o" point stands for a pair of values of f_s and f_p learnt from two images of the same person, and a red $*$ point stands for that pair learnt from two images of two different person. In the space of $f_s - f_p$, points representing images from the same person and those representing images from different persons can be roughly separated by a blue line, with the former below the line and the latter above the line.

projection consistency. To reduce the impact of cross-view support consistency, we set $\alpha = 0$ in this stage. When $\beta = 0$, it is equal to only exploit the learnt uniform metric; inversely, it only uses the support consistency to generate ranking result for $\beta \to \infty$. We change β from 0 to 50, the corresponding results are shown in Fig. 8. It is obvious that 1 is a good choice for β on the VIPeR dataset and 0.7 on the CUHK dataset. When $\beta > 10$,

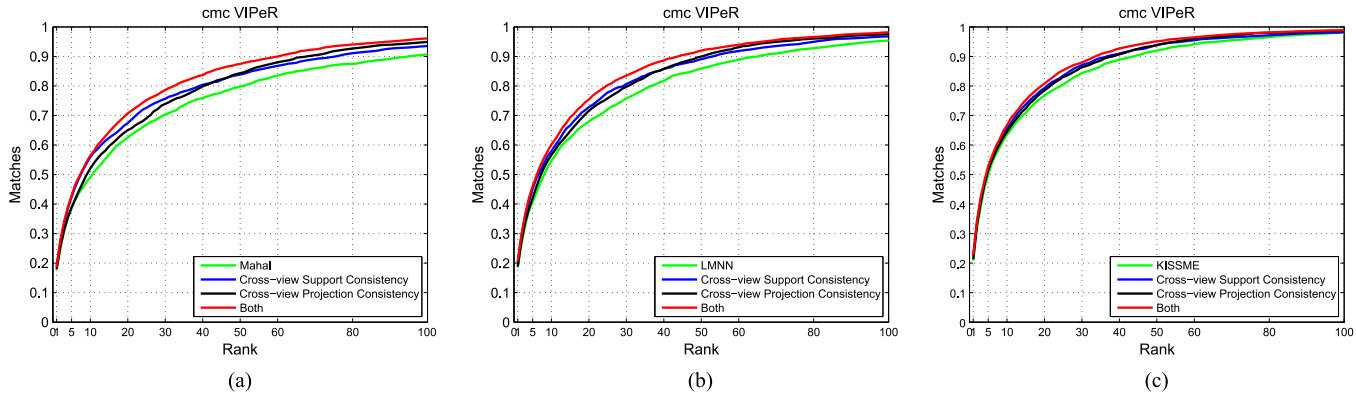


Fig. 10. Comparative results between existing methods and our approach on the VIPeR data set. (a) Comparison with Mahalanobis. (b) Comparison with LMNN [30]. (c) Comparison with KISSME [33].

TABLE I
PERSON RE-ID MATCHING RATES(%) AT DIFFERENT RANKS ON THE VIPER, CUHK, AND PRID DATA SET

	VIPeR				CUHK				PRID			
Method	rank@1	10	25	50	10	20	50	100	5	10	20	30
Mahalanobis	17.9	49.02	66.91	79.81	36.25	47.29	63.44	76.6	31.9	43.2	60.2	69.8
DDDM+Mahalanobis	18.69	56.53	75.23	87.47	41.74	54.25	71.67	85.42	32.4	44.7	62.6	72.8
LMNN [30]	18.54	54.59	72.09	85.89	40.74	52.48	70.57	82.96	29.8	39.7	55.1	67.2
DDDM+LMNN	20.54	60.18	80.28	91.99	45.97	58.39	76.65	88.73	32.5	41.7	57.6	71.5
KISSME [33]	20.78	63.51	80.51	91.96	46.59	59.4	76.66	88.95	31.4	44.8	62.6	75
DDDM+KISSME	22.66	66.71	85.32	95.16	49.72	62.97	80.08	91.48	33.4	46.3	65.1	76.7

the performance is stable, because the metric is almost entirely depending on the cross-view projection factor. The experiment results are consistent with the above claim, that the algorithm achieves good performance under a suitable β.

E. Distribution of f_s and f_p

We plot in Fig. 9 the distributions of two factors on the VIPeR dataset following the above parameters decision. Given 316 person image pairs as the testing set, the proposed method computed 316 pairs of factor f_s and factor f_p for images of the same persons, plotted as the blue points. Then, we randomly selected 316 image pairs of different persons from the testing set, and 316 pairs of factors were also obtained, shown as the red points. We can see from Fig. 9 that most of the testing pairs of the same persons have small factor values, under the divider line illustrated. Small factor values lead to small distances for images of the same person, and this conforms the cross view consistencies.

F. Comparing to Metric Learning Methods

We evaluate the effectiveness of the proposed method by comparing with three metric learning methods, Mahalanobis metric, LMNN and KISSME, on the VIPeR dataset, the CUHK dataset and the PRID dataset, respectively. After the descriptor was generated, Mahalanobis metric, LMNN and KISSME were used to measure the distances between pair of images, respectively. Then, the metrics reformed with cross-view support consistency and cross-view projection consistency were also utilized to measure the distances.

The VIPeR dataset. The obtained results are shown in Fig. 10. Moreover, in Table I we compare the performance of our approach to that of state-of-the-art methods in the range of the first 50 ranks. As can be seen, our approach has 3.5–15.2% improvement compared with Mahalanobis, LMNN and KISSME. Table I shows that our method outperforms all existing methods over the whole range of ranks.[3]

The CUHK dataset. The results are presented in Fig. 11. Our method has evident improvements compared with Mahalanobis, LMNN and KISSME. In Table I we compare the performance of our approach to that of state-of-the-art methods in the range of the first 100 ranks. Our method has 2.8–15.1% ascension with the three compared methods, and exceeds them over the entire range of ranks.

The PRID dataset. The results are presented in Fig. 12. Our method has evident improvements compared with Mahalanobis, LMNN and KISSME. In Table I we compare the performance of our approach to that of state-of-the-art methods in the range of the first 30 ranks. Our method has 1.6–9.1% ascension with the three compared methods, and exceeds them over the entire range of ranks.

G. Comparing the Relative Importance of Two Consistencies

From Fig. 10, we can find that for the VIPeR dataset, the cross-view support consistency is more effective than the cross-view projection consistency at the top 40 ranks, and the results reverse after the top 40 ranks. Three learning methods on the VIPeR dataset show the similar trend. Therefore, for the VIPeR dataset, the cross-view support consistency is relatively more important at the top ranks, and the cross-view projection consistency will be relatively more important at the rest ranks.

From Fig. 11, we can find that for the CUHK dataset, the cross-view projection consistency improves the results more comparing to the cross-view support consistency at all ranks.

[3]For calculating the projection consistency, we set k-nearest neighbors value $K = 32$ when we evaluate on VIPeR data set, $K = 48$ on CUHK data set, and $K = 10$ on PRID data set. Those values are nearly 10% of the number of each training pairs.

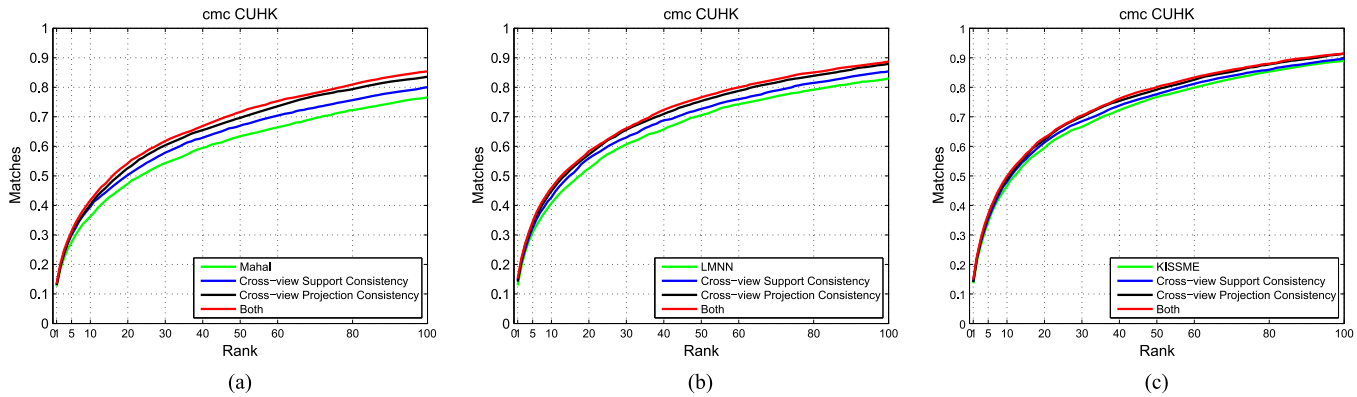


Fig. 11. Comparative results between existing methods and our approach on the CUHK data set. (a) Comparison with Mahalanobis. (b) Comparison with LMNN [30]. (c) Comparison with KISSME [33].

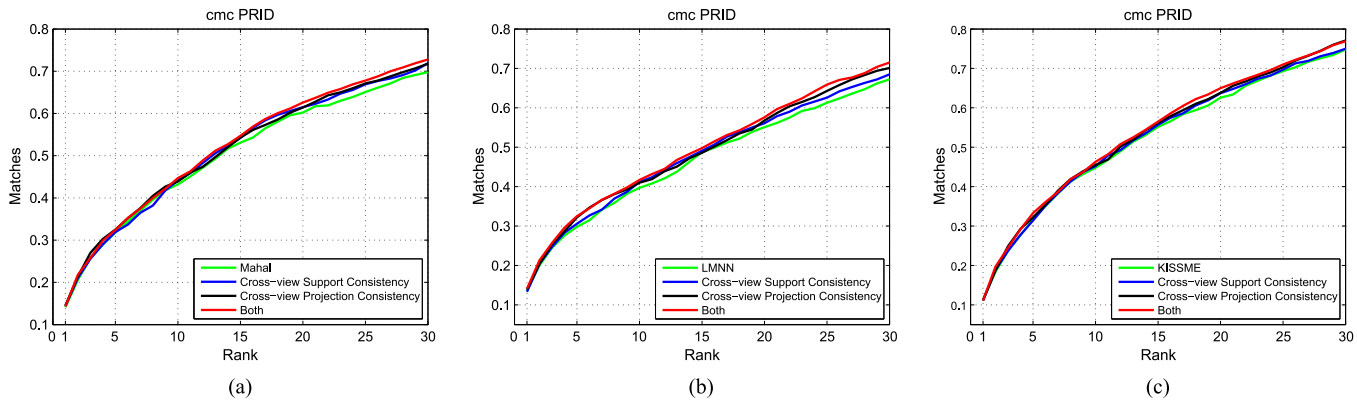


Fig. 12. Comparative results between existing methods and our approach on the PRID data set. (a) Comparison with Mahalanobis. (b) Comparison with LMNN [30]. (c) Comparison with KISSME [33].

So the cross-view projection consistency is more important than the cross-view support consistency for the CUHK dataset.

From Fig. 12, we can find that for the PRID dataset, the improvements are similar for these two consistencies. It is a little difficult to make a conclusion as for which consistency is more important.

In brief, the relative importance of the consistencies depends on the evaluating dataset, not the selected metric learning method.

H. Comparing to the-State-of-the-Art Person Re-id Methods

Table II summarizes the comparing results with the-state-of-the-art re-id methods on the widely used VIPeR dataset with 316 as training set size. For a fair comparison, the results for most of these methods are directly taken from the original public papers. In these methods, SCNCD, PartsSC, eSDC-ocsvm, SDALF and ELF belong to the feature representation based approaches, LMNN, PRDC, PCCA, KISSME, RS-KISSME and LADF are the distance measure based approaches. The Descriptive+Discriminative Model and Bidirectional ranking are two result refinement methods. The results clearly show that our approach improves the KISS and LADF metric learning methods even with complicated SCNCD feature, and also demonstrate that our approach with the SCNCD feature and the KISSME metric gives the best performance in all cases.

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

Method	rank@1	10	25	50
SCNCD [46]	20.7	60.6	79.1	90.4
PartsSC [25]	24	57	73	86
eSDC-ocsvm [26]	26.7	62.4	–	–
SDALF [22]	19.9	49.4	70.5	84.8
ELF [28]	8.2	36.6	58.2	90.9
LMNN	19	58.1	76.9	89.6
PRDC [1]	15.66	53.86	76	87
PCCA [34]	19.27	64.91	83	96
RS-KISSME [13]	24.5	66.6	84.7	93.0
Descriptive+Discriminative Model [40]	19	52	69	80
KISSME + Bidirectional ranking [16]	22	67	85	95
LADF [36]	13.5	56.01	79.64	92.51
DDDM+LADF	13.92	56.22	80.17	93.35
SCNCD+LADF	28.16	78.16	89.56	96.84
DDDM+SCNCD+LADF	28.8	78.48	91.46	97.47
KISSME [33]	19.6	62.2	80.7	91.8
DDDM+KISSME	22.66	66.71	85.32	95.16
SCNCD+KISSME	37.8	81.2	92.7	97.0
DDDM+SCNCD+KISSME	**42.09**	**83.54**	**94.94**	**97.78**

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel and efficient data-driven distance metric (DDDM) method for re-id task. The core idea is: we refine the learnt metric by re-exploiting the training data in the distance measure stage. Applying the cross-view support and projection consistencies, the proposed approach can obtain

a data-specific adaptive metric, and improve the matching result of metric learning methods obviously. Extensive comparative experimental results reported in Section VI shows that our method is effective compared to Mahalanobis, LMNN and KISSME on three challenging public datasets, VIPeR, CUHK and PRID. With the SCNCD feature and the KISSME metric, our method can achieve the best results compared to the state-of-the-art re-id methods.

In the list below we outline a number of ideas for future work.

- When the proposed method calculates the factor f_s, each dictionary is generated by all the training images in the corresponding camera. This will bring noise for the sparse coding process, because each person still has unique appearance change across cameras. To this end, we can investigate how to construct a more effective dictionary by introducing some constraints.
- The factor f_p is computed after samples been projected to the same virtual camera, which is depending on the uniform metric $\mathbf{M}$. However, the metric $\mathbf{M}$ is then transformed to a new one, and it means that we might get a more proper f_p with the new metric. This inspires us to study an iterative factor-metric optimization method to obtain a more suitable metric progressively.
- This paper investigates consistencies across two camera views. We believe that there exists more associations among multiple different cameras, and the consistencies across three or more cameras could also be analyzed.

REFERENCES

[1] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 649–656.

[2] N. O'Hare and A. F. Smeaton, "Context-aware person identification in personal photo collections," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 220–228, Feb. 2009.

[3] S. Gong, M. Cristani, S. Yan, and C. Loy, *Person Re-identification*. New York, NY, USA: Springer, 2014.

[4] X. Wang, T. Zhang, Tretter, and Q. Lin, "Personal clothing retrieval on photo collections by color and attributes," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2035–2045, Dec. 2013.

[5] K. W. Chen, C. C. Lai, P. J. Lee, C. S. Chen, and Y. P. Hung, "Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 625–638, Aug. 2011.

[6] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Tracking multiple people under global appearance constraints," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 137–144.

[7] L. L. Presti, S. Sclaroff, and M. L. Cascia, "Path modeling and retrieval in distributed video surveillance databases," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 346–360, Apr. 2012.

[8] J. W. Hsieh, Y. T. Hsu, H. Y. Liao, and C. C. Chen, "Video-based human movement analysis and its application to surveillance systems," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 372–384, Apr. 2008.

[9] F. Chen, C. De Vleeschouwer, and A. Cavallaro, "Resource allocation for personalized video summarization," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 455–469, Feb. 2014.

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

[11] N. A. Fox, R. Gross, J. F. Cohn, and R. B. Reilly, "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 701–714, Jun. 2007.

[12] W. Yin, J. Luo, and C. W. Chen, "Event-based semantic image adaptation for user-centric mobile display devices," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 432–442, Jun. 2011.

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

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

[15] 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.

[16] Q. Leng, R. Hu, and C. Liang, "Bi-directional ranking for person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2013, pp. 1–6.

[17] S. Ali, O. Javed, N. Haering, and T. Kanade, "Interactive retrieval of targets for wide area surveillance," in *Proc. ACM Int. Conf. Multimedia*, 2010.

[18] C. Liu, C. Loy, and S. Gong, "POP: Person re-identification post-rank optimisation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 441–448.

[19] Z. Wang, R. Hu, C. Liang, Q. Leng, and K. Sun, "Region-based interactive ranking optimization for person re-identification," in *Proc. Pacific-Rim Conf. Multimedia*, 2014, pp. 1–10.

[20] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 1528–1535.

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

[22] 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 Recog.*, Jun. 2010, pp. 2360–2367.

[23] 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.*, 2012, p. 11.

[24] B. Layne, T. Hospedales, and S. Gong, "Person re-identification by attributes," in *Proc. Brit. Mach. Vis. Conf.*, 2012, p. 8.

[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] Z. Rui, O. Wanli, and W. Xiaogang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3586–3593.

[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 Recog.*, Jun. 2014, pp. 152–159.

[28] 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.

[29] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person reidentification by support vector ranking," in *Proc. Brit. Mach. Vis. Conf.*, 2010, p. 6.

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

[31] M. Hirzer, C. Beleznai, M. Kstinger, P. M. Roth, and H. Bischof, "Dense appearance modeling and efficient learning of camera transitions for person re-identification," in *Proc. IEEE Int. Conf. Image Process.*, Sep.–Oct. 2012, pp. 1617–1620.

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

[33] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2288–2295.

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

[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 Recog.*, Jun. 2013, pp. 3318–3325.

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

[37] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1268–1281, Aug. 2014.

[38] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Trans. Process.*, vol. 23, no. 10, pp. 4220–4231, Oct. 2014.
[39] J. Jiang, X. Ma, Z. Cai, and R. Hu, "Sparse support regression for image super-resolution," *IEEE Photon. J.*, vol. 7, no. 5, Oct. 2015, Art. ID 6901211.
[40] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person reidentification by descriptive and discriminative classification," in *Proc. Scandinavian Conf. Image Anal.*, 2011, pp. 91–102.
[41] S. B. D. Gray and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveillance*, 2007, vol. 3, no. 5.
[42] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 31–44.
[43] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[44] J. Liu and J. Ye, "Efficient Euclidean projections in linear time," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 657–664.
[45] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. Mitchell, "Zero-shot learning with semantic output codes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1410–1418.
[46] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 536–551.
[47] H. Abdi and L. Williams, "Principal component analysis," *Wiley Interdisciplinary Rev.: Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.
[48] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning With Efficient Projections*. Arizona State Univ., Tempe, AZ, USA, 2009 [Online]. Available: http://www.yelab.net/software/SLEP/

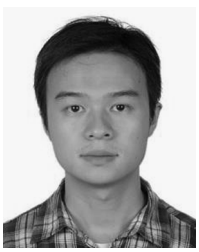
Zheng Wang received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2006 and 2008, respectively, and is currently working toward the Ph.D. degree at the National Engineering Research Center for Multimedia Software (NERCMS), School of Computer, Wuhan University.

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

Mr. Wang was the recipient of the Best Paper Award at the 15th Pacific-Rim Conference on Multimedia (2015).

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 Huazhong University of Science and Technology, Wuhan, China, in 1994.

He is the Dean of School of Computer, Wuhan University, Wuhan, China. He has authored or coauthored two books and over 100 scientific papers. His research interests include audio/video coding and decoding, video surveillance, and multimedia data processing.



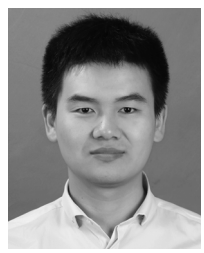
Chao Liang received the Ph.D. degree from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2012.

He is currently an Assistant Professor with the National Engineering Research Center for Multimedia Software (NERCMS), Computer School of Wuhan University, Wuhan, China. He has authored or coauthored over 30 papers, including papers that have appeared in conferences such as CVPR, ACM, and MM, and journals such as the IEEE TRANSACTIONS ON MULTIMEDIA and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His research interests focus on multimedia content analysis and retrieval, computer vision, and pattern recognition.

Prof. Liang was the recipient of the Best Paper Award at the 2014 Pacific-Rim Conference on Multimedia.

Yi Yu received the Ph.D. degree in information and computer science from Nara Women's University, Nara, Japan, in 2009.

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 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 in mathematical sciences from the Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree from the School of Computer, Wuhan University, Wuhan, China, in 2014.

He is currently an Associate Professor with the School of Computer Science, China University of Geosciences, Beijing, China. He has authored or coauthored more than 40 scientific articles, and holds 10 Chinese patents. His research interests include image processing, pattern recognition, hyperspectral remote sensing, and high-resolution remote sensing.

Prof. Jiang was the recipient of the IBM China Excellent Student Scholarship in 2014, the Best Student Paper Runner-Up Award at the 21th International Conference on Multimedia Modelling (2015), and the 2015 ACM Wuhan Doctoral Dissertation Award.

Mang Ye, photograph and biography not available at the time of publication.

Jun Chen received the M.S. degree in instrumentation from Huazhong University of Science and Technology, Wuhan, China, in 1997, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2008.

He is the Deputy Director of the National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University. His research interests include multimedia communications and security emergency information processing.

Qingming Leng, photograph and biography not available at the time of publication.