Camera Compensation Using a Feature Projection Matrix for Person Reidentification

Yimin Wang, Ruimin Hu, *Senior Member, IEEE*, Chao Liang,
Chunjie Zhang, and Qingming Leng

***Abstract*— Matching individuals within a group of spatially nonoverlapping surveillance cameras, also known as person reidentification, has recently attracted a lot of research interest. Current methods mainly focus on feature representation or distance measure, which directly compare person images captured by different cameras. However, it is still a problem because of various surveillance conditions; for example, view switching, lighting variations, and image scaling. Although the brightness transfer function was proposed to address the problem of illumination variation, it could not handle view and scale changes among various cameras. In this paper, we propose a new approach to compensate for the inconsistency of feature distributions of person images captured by different cameras. More precisely, a feature projection matrix (FPM) is learned to project image features of one camera to the feature space of another camera, from which the latent device difference can be effectively eliminated for the person reidentification task. In particular, we formulate the FPM learning as a smooth unconstrained convex optimization problem and use a simple gradient descent algorithm with stochastic samples to accelerate the solving process. Extensive comparative experiments conducted on three standard datasets have shown the promising prospect of the proposed method.**

***Index Terms*—Feature projection matrix, nonoverlapping camera tracking, person reidentification.**

I. Introduction

RECENTLY, more and more nonoverlapping camera networks have been set up for monitoring pedestrian activities over a large public area, such as the airport, metro station and parking lot. To acquire individuals' complete motion trajectories, matching persons across nonoverlapping cameras in a surveillance camera network, also known as person reidentification, is increasingly becoming a hot research topic in the computer vision community [1]–[6]. Because traditional biometrics, such as face and gait, are unreliable or even infeasible in uncontrolled surveillance environment [7], body appearance is exploited for person reidentification [1], [7]–[10] in recent years. However, person reidentification remains an unsolved problem owing to the challenges caused by view change, scale zooming, and illumination variation (see Fig. 1), making different persons appear more alike than the same person in various cameras [3].

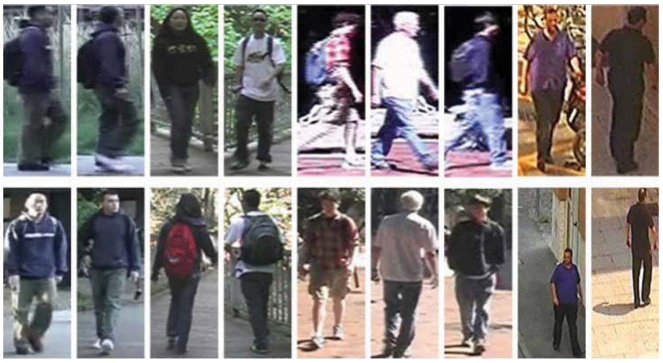


Fig. 1. Examples of appearance changes caused by different views, lighting, and scales from public datasets, VIPeR [11] and 3DPeS [12]. Each column shows two images of the same person taken from two different cameras.

Generally, person reidentification can be regarded as an image retrieval problem [4], that is, given a query person image taken from one camera, the algorithm is expected to search images of the same person captured by other cameras, and generate a final ranking list where top results are more likely of the same person to the query image. The paradigm usually consists of two stages: feature extraction and distance measure. Early research efforts aim to seek a discriminative and robust feature representation which can easily separate different persons in various cameras [8], [13]–[19]. However, designing a set of features that are both distinctive and stable is extremely difficult in itself, let alone under conditions where view changes usually cause significant appearance variations [7].

Recently, more and more researchers change their attentions to the second stage where a proper distance measure is seeked to reflect the identity consistency among persons [1], [2], [4], [7], [9], [20]–[22]. Among various methods, supervised metric learning algorithms demonstrate an obvious

Manuscript received July 10, 2013; revised October 29, 2013 and January 2, 2014; accepted January 25, 2014. Date of publication February 12, 2014; date of current version August 1, 2014. This work was supported in part by the National Natural Science Foundation of China under Grants 61172173, 61170023, 61231015, 61303114, 61303154; in part by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20130141120024; in part by the China Postdoctoral Science Foundation funded project under Grant 2013M530350; in part by the President Fund of UCAS; in part by the Open Project Program of the National Laboratory of Pattern Recognition; and in part by the major Science and Technology Innovation Plan of Hubei Province under Grant 2013AAA020. This paper was recommended by Associate Editor C.-W. Lin.

Y. Wang, R. Hu, C. Liang, and Q. Leng are with the National Engineering Research Center on Multimedia Software, School of Computer, Wuhan University, Wuhan 430072, China (e-mail: fyimin.wang@whu.edu.cn; hrm@whu.edu.cn; cliang@whu.edu.cn; lengqmg@whu.edu.cn).

C. Zhang is with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100044, China (e-mail: cjzhang@jdl.ac.cn).

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

Digital Object Identifier 10.1109/TCSVT.2014.2305519

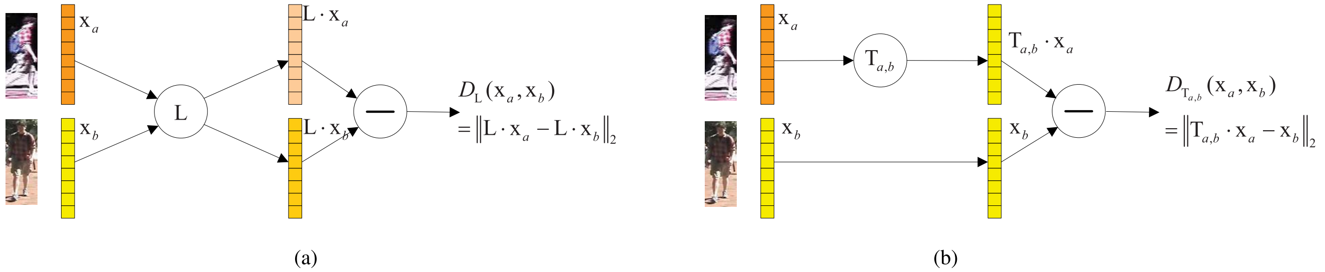

x_a
x_b
L
$L\cdot x_a$
$L\cdot x_b$
$D_L(x_a, x_b) = \|L\cdot x_a - L\cdot x_b\|_2$
(a)
$T_{a,b}$
$T_{a,b}\cdot x_a$
$D_{T_{a,b}}(x_a, x_b) = \|T_{a,b}\cdot x_a - x_b\|_2$
(b)

Fig. 2. Explanation of the proposed method compared to metric learning-based method. (a) Metric learning methods transform original features of images captured by different cameras into a new feature space using the same projection matrix. (b) However, the proposed method compensates the difference between different cameras by projecting feature space from one camera to the other with a FPM.

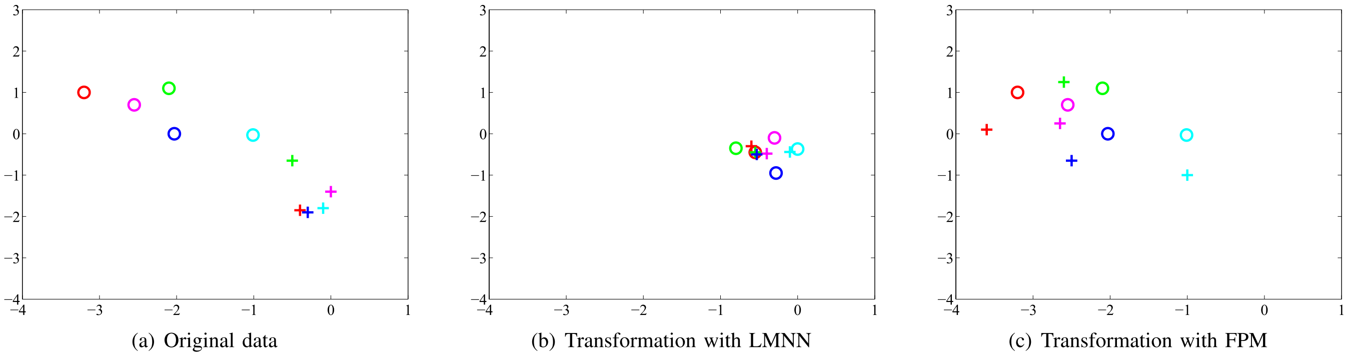


Fig. 3. Comparison of compensating cameras difference. Each point represents a 2-D feature vector of image, where “o” (+) represents samples from camera C_a (C_b); different colors represent different persons. (a) Original feature distributions of images taken from two cameras are inconsistent. (b) LMNN [23] transforms original feature vectors into a new feature space where feature vectors of different cameras are cluttered together. (c) FPM projects the feature vectors from C_a to C_b and keep the feature vector of C_b unchanged, where feature points of C_a are projected to the space nearby the points belonging to the same person and apart from the points of different persons.

advantage in learning a discriminative distance function based on the given training samples. Specifically, given two person image feature $\mathbf{x}_a$ and $\mathbf{x}_b$, their distance can be defined as a Mahalanobis distance $D(\mathbf{x}_a, \mathbf{x}_b) = (\mathbf{x}_a - \mathbf{x}_b)^\top \mathbf{M}(\mathbf{x}_a - \mathbf{x}_b)$, where $\mathbf{M}$ is a positive semidefinite matrix for the validity of metric. Performing eigenvalue decomposition on $\mathbf{M}$ with $\mathbf{M} = \mathbf{L}^\top\mathbf{L}$, the above distance can be rewritten as $D(\mathbf{x}_a, \mathbf{x}_b) = \|\mathbf{L}\cdot(\mathbf{x}_a - \mathbf{x}_b)\|^2 = \|\mathbf{L}\cdot\mathbf{x}_a - \mathbf{L}\cdot\mathbf{x}_b\|^2$. With this definition, it is easy to see that the essence of the metric-based method is to seek a projection matrix that transforms original image features into a new feature space, where feature distance of the same person is smaller than that of different persons. Moreover, for metric learning-based methods, it is noteworthy that the same feature transformation is applied to features of images from different cameras, for example, $\mathbf{x}_a$ and $\mathbf{x}_b$ [see Fig. 2(a)]. Although the differences between different cameras can be partially suppressed by applying the same transformation to different cameras, it is hardly eliminated.

To conquer the above weakness, we propose a feature projection matrix (FPM) method to directly project feature vectors from one camera to the feature space of the other camera [see Fig. 2(b)], which equals to apply different transformations to features of images from different cameras. Therefore, the difference of feature distributions between two cameras can be more effectively eliminated. The example in Fig. 3 shows the superiority of the proposed approach compared with metric learning-based method, taking the classic large margin nearest neighbors (LMNN) algorithm [23] as an example, where the distribution of five pairs of images of different persons from two cameras is shown for different methods. The proposed approach accurately compensates the differences between different cameras and holds the discriminative ability, which clearly outperforms LMNN.

For learning the FPM, we propose a supervised learning method in which the objective function consists of two terms, that is, consistent term and discriminative term (see Fig. 4). The first acts to project images of the same person close to each other, while the second acts to take images of different persons apart. With the proposed objective function, the FPM learning can be formulated as a smooth unconstrained convex optimization problem, where a simple batch gradient descent algorithm is used on randomly selected samples to efficiently solve the problem without loss of accuracy.

Extensive comparative experiment results have shown the promising prospect of the proposed method by directly compensating the device difference for the person reidentification task.

A similar idea of the transfer function has also been investigated using some early person reidentification methods [24]–[26]. In [24], brightness transfer function (BTF), f_{ab}, was used to compensate different illumination conditions of different cameras. It assumed that the percentage of pixels in an observation x_a with the brightness value less than B_a is equal to the percentage of image points seen in x_b of brightness no more than B_b. More specifically, denoting a person image as I, the count of brightness value B in I as $I(B)$, and the cumulative histogram as $H(I) = \sum I(B)$, their assumption can be formulated as $H_a(B_a) = H_b(B_b)$. Thereafter, f_{ab} can

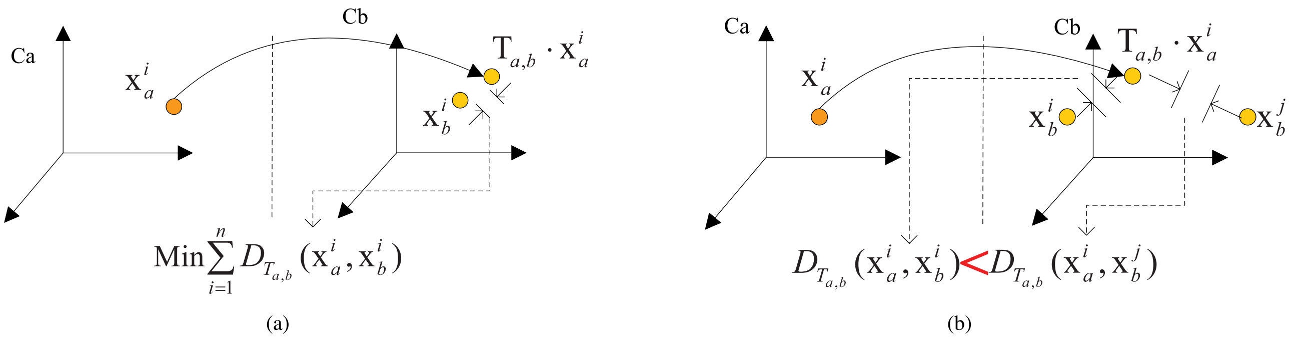

Ca
Cb
x_a^i
$T_{a,b} \cdot x_a^i$
x_b^i
$\mathrm{Min}\sum_{i=1}^{n} D_{T_{a,b}}(x_a^i, x_b^i)$
(a)
Ca
Cb
x_a^i
$T_{a,b} \cdot x_a^i$
x_b^i
x_b^j
$D_{T_{a,b}}(x_a^i, x_b^i) < D_{T_{a,b}}(x_a^i, x_b^j)$
(b)

Fig. 4. Illustration of the proposed two terms on learning the FPM. (a) Consistent term aims to project images close to images of the same person. (b) Discriminative term aims to project images further apart to images of different persons.

be computed by mapping an observed color value in camera C_a to the corresponding observation in camera C_b as

$$f_{ab} = H_b^{-1}(H_a(B_a)) \tag{1}$$

where $H^{-1}(\cdot)$ is the inverted cumulative histogram. In addition to BTF, some extension methods were proposed, such as cumulating multiimages brightness distribution before transformation [25], extending the brightness space to RGB color space [27], and combining multi BTFs with different weight [26].

Compared with BTF and its derivatives [25], [26], the proposed method has three obvious advantages: 1) the BTF is a vector which indicates illumination mapping between cameras, whereas the FPM is a matrix which could handle more complex appearance variation; 2) the BTF is computed under a brightness distribution assumption, which makes BTF to only address the problem caused by illumination changes. On the contrary, the proposed method does not have impractical assumption and hence able to handle various camera differences in principle; and 3) although both methods use the label information, only positive samples corresponding to image pairs of the same person are exploited in computing BTF. In contrast, the learning of FPM not only considers image pairs from the same person, but also those from different persons, making the result more accurate.

The contribution of this paper can be summarized as follows.

1) We propose an FPM method for compensating the camera difference in the person reidentification problem. Compared with BTF, the FPM method compares two images in a common feature space, and hence can handle more complex appearance differences in an implicit way.
2) We formulate the FPM learning problem as a smooth unconstrained convex optimization problem, in which the objective function consists of both consistent term and discriminative term. Moreover, motivated by the idea of stochastic gradient descent (SGD) algorithm [28], we use a simple gradient descent algorithm with a group of randomly selected samples to optimize the objective function, achieving flexible balance between computation cost and accuracy.

The rest of this paper is organized as follows. In Section II, a brief review of related work for person reidentification is given. Then, we detail the proposed FPM method with the objective function and optimization algorithm in Section III. Section IV shows experimental results on three representative datasets and Section V concludes this paper.

II. Related Work

In this section we give a brief review of the related work on person reidentification. Readers who are interested in more detailed reviews are suggested to refer [29] and [30].

Current person reidentification research can be generally categorized into two classes: 1) feature- and 2) distance learning-based methods [7]. The former aims to seek a discriminative and robust feature representation which can easily separate different persons in various cameras. A lot of feature representation methods consisting of low-level visual features, such as color, texture, shape, local features, and their combination, have been developed for person reidentification [8], [13]–[19], [31]. Gheissari *et al.* [13] used a spatial-temporal segmentation algorithm to generate salient edges and obtained an invariant identity signature by combining normalized color and salient edge histograms. In Wang *et al.* [14] 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.* [8] divided the image of person into five regions by exploiting symmetry and asymmetry perceptual principles, and then combined multiple color, texture, and local features to represent the appearance of people, called symmetry-driven accumulation of local feature (SDALF). Cheng *et al.* [18] adopt custom pictorial structure (CPS) to localize the body parts, and extracted and matched descriptors on different parts. Ma *et al.* [19] developed a representation rely on the combination of biologically inspired features and covariance descriptors (BiCov). Rui *et al.* [5] applied adjacency constrained patch matching to build dense correspondence between image pairs. For each patch, they assign salience to it in an unsupervised manner. All these methods focused on designing a robust and distinctive feature representation which is extremely hard if not implausible [3].

To increase the discriminative power of feature representation, feature selection technique is also adopted in the person reidentification research [9], [20], [32]. Gray and Tao [32] transformed the problem into a classification problem, and used an ensemble of the localized features (ELFs) through AdaBoost algorithm. Prosser *et al.* [9] treated person reidentification problem as

TABLE I

COMPARATIVE RESULTS WITH STATE-OF-THE-ART PERSON REIDENTIFICATION METHODS ON TOP RANKED MATCHING RATE(%)

Methods	Training size = 316					Training size = 200				Training size = 100			
	CMC@1	5	10	25	50	1	5	10	20	1	5	10	20
SDALF [8]	19.9	40	49.4	70.5	84.8	-	-	-	-	-	-	-	-
ELF [32]	8.2	24.2	36.6	58.2	90.9	6.83	19.8	29.8	43.1	4.2	13	20.2	30.7
RankSVM [9]	16.3	38.2	53.7	72	85	10.6	29.7	42.3	58.3	8.9	22.9	32.7	46
PRDC [3]	15.7	38.4	53.9	76	87	12.6	32	44.3	60	9.1	24.2	34.4	48.6
KISSME [1]	19.6	46	62.2	80.7	91.8	13.9	35.9	49.5	64.3	9.9	26.5	38.4	52.5
PCCA [2]	19.3	48.9	64.9	83	**96**	-	-	-	-	9.3	24.9	37.4	52.9
FPM-CON	12.1	35.3	50	71.3	83.8	7.8	23.2	35.7	50.8	0.4	1.7	3.3	6.4
FPM-DIS [33]	15.4	42.2	59.2	81.3	92.1	10.3	31.2	46.3	64.3	5.7	20.1	32.4	48.6
FPM-BOTH	**23.5**	**52.6**	**67.9**	**86.8**	94.1	**17.5**	**43.2**	**57.3**	**72.9**	**13.3**	**33.9**	**46**	**61.6**

[1] FPM_CON only uses the *consistent* term, FPM_DIS only uses the *discriminative* term which is the early version of this work appeared in [33], and FPM_BOTH is the combination of above two terms which is the final version of this work.

a ranking problem and used ensemble RankSVM to learn a subspace where images of the same person should appear in higher positions in the rank list. Schwartz and Davis [20] weighted features according to their discriminative power for each different appearance using a powerful statistical tool called partial least squares (PLS). These methods increase the discriminative power by assigning a higher weight to more discriminative features with a weight vector.

In recent years, a lot of metric learning methods are provided for person reidentification. LMNN [23], which is designed for k-nearest neighbor classification problem, is a popular metric learning method. It introduces two terms. One pulls the same labeled examples closer together, the other pushes examples with different labels further apart. Hirzer *et al.* [22] utilized LMNN metric learning to learning the optimal metric for person reidentification. Dikmen *et al.* [21] improved the LMNN algorithm for person reidentification by exploiting a fixed bound for neighbors. Zheng *et al.* [7] learned a Mahalanobis distance metric with a probabilistic relative distance comparison (PRDC) method. Kostinger *et al.* [1] used Gauss distribution to fit pairwise samples and got a simpler metric function without iterative procedures. As analog to the keep it simple and straightforward (KISS) principle, we named our method KISS metric (KISSME). But, given a small-size training set, the estimation of the covariance matrix is not accurate and result in a poor performance. Li *et al.* [4] presented a regularized smoothing KISS metric learning (RS-KISS) by seamlessly integrating smoothing and regularization techniques for robustly estimating the covariance matrices. Mignon and Jurie [2] introduced a pairwise constrained component analysis (PCCA) to learn distance metric from sparse pairwise similarity or dissimilarity constraints in high-dimensional input space. Sateesh *et al.* [6] exploited a local fisher discriminant analysis, which focuses on local samples, with a regularization term. According to the discussion in the above section, these methods use the same projection matrix to the different cameras.

In addition, an early idea of this paper appeared in [33]. The main improvements of this journal paper include: 1) a new term, consistent term, is introduced to effectively improve the performance of the original FPM model. Relevant evidences can be found in Fig. 9 and Table I; 2) we give more detailed discussion and experimental evaluation on several key parameters of algorithm, including the balance coefficient, μ, the smooth parameter, β, and efficiency of the optimization algorithm (see Section IV-E); and 3) more extensive experiments are conducted on a new dataset, CUHK [34], which is a larger dataset and hence more challenge than those previously used.

III. APPROACH

This section presents our approach. We begin with a brief formulation of the metric learning-based person reidentification problem. Then, the FPM is introduced followed by defining a new feature distance function. Finally, a new objective function consists of consistent and discriminative terms is raised, and meanwhile, a stochastic sampling-based solution method is designed to accelerate the optimization process.

A. Person Reidentification Problem

For the convenience of following discussion, we consider a pair of cameras C_a and C_b with nonoverlapping field of views, and a set of persons $O = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_n\}$ crossing the two cameras. Then, we denote the representing image of person $\mathbf{o}_i$ captured by C_a (or C_b) as x_a^i (or x_b^i), and further let $X_a = \{x_a^1, x_a^2, \ldots, x_a^n\}$ and $X_b = \{x_b^1, x_b^2, \ldots, x_b^n\}$ represent two sets of person images captured by C_a and C_b, respectively. The person reidentification task is that for each instance x_a^i in X_a, the algorithm finds image of the same person, x_b^i, from X_b.

This problem is commonly addressed as visual retrieval problem and consists of two key stages: feature extraction and distance measure [1]. Feature extraction step acts to represent images of persons as feature vectors, whereas distance measure stage acts to define a distance function, such as Euclidean distance, to measure the distance between images. Usually, the instance, x^i is represented by a d-dimensional feature vector, $\mathbf{x}^i \in \mathbf{R}^d$, then the Euclidean distance can be formulated as

$$D\left(\mathbf{x}_a^i, \mathbf{x}_b^j\right) = \left(\mathbf{x}_a^i - \mathbf{x}_b^j\right)^\top \left(\mathbf{x}_a^i - \mathbf{x}_b^j\right) = \left\|\mathbf{x}_a^i - \mathbf{x}_b^j\right\|^2 \tag{2}$$

where $(\cdot)^\top$ is the transpose of a vector or matrix.

Given a query person image $\mathbf{x}_a^i$, after computing distance between it and each instance in gallery set, a ranked list can be achieved, and the algorithm which ranks the correct match on the more top gets better performance.

On this basis, metric learning methods generally learn a Mahalanobis-like distance, that is, $D(\mathbf{x}_a^i, \mathbf{x}_b^j) = (\mathbf{x}_a^i - \mathbf{x}_b^j)^\top \mathbf{M}(\mathbf{x}_a^i - \mathbf{x}_b^j)$, where $\mathbf{M}$ is a positive semidefinite matrix for validity of metric definition. Performing eigenvalue decomposition on $\mathbf{M}$ with $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$, the above distance can be rewritten as

$$\begin{aligned} D(\mathbf{x}_a^i, \mathbf{x}_b^j) &= \left(\mathbf{x}_a^i - \mathbf{x}_b^j\right)^\top \mathbf{M}\left(\mathbf{x}_a^i - \mathbf{x}_b^j\right) && (3)\\ &= \left(\mathbf{x}_a^i - \mathbf{x}_b^j\right)^\top \mathbf{L}^\top \mathbf{L}\left(\mathbf{x}_a^i - \mathbf{x}_b^j\right) && (4)\\ &= \left[\mathbf{L} \cdot \left(\mathbf{x}_a^i - \mathbf{x}_b^j\right)\right]^\top \left[\mathbf{L} \cdot \left(\mathbf{x}_a^i - \mathbf{x}_b^j\right)\right] && (5)\\ &= \left\|\mathbf{L} \cdot \left(\mathbf{x}_a^i - \mathbf{x}_b^j\right)\right\|^2 && (6)\\ &= \left\|\mathbf{L} \cdot \mathbf{x}_a^i - \mathbf{L} \cdot \mathbf{x}_b^j\right\|^2. && (7) \end{aligned}$$

As can be seen from the above derivation, the essence of metric learning is to seek an optimal $\mathbf{M}$ (or $\mathbf{L}$) under the supervised information generally containing two pairwise constraints, that is, similar constraint and dissimilar constraint, which are denoted, respectively, by

$$\begin{aligned} S &= \{(\mathbf{x}^i, \mathbf{x}^j) | \mathbf{x}^i \text{ and } \mathbf{x}^j \text{ belong to the same identity}\}\\ D &= \{(\mathbf{x}^i, \mathbf{x}^j) | \mathbf{x}^i \text{ and } \mathbf{x}^j \text{ belong to different identities}\}. \end{aligned}$$

B. Distance Measure After Compensating With FPM

As shown in (7), metric learning generally applies the same feature transformation $\mathbf{L}$ to the feature vectors (or space) of different cameras, with which the device difference, caused by poses variation, scale zooming, and illumination change between two cameras, can be partially weakened, but hardly eliminated.

To solve the above problem, an FPM, which maps persons from one camera to the other, is introduced to compensate the discrepancy of the surveillance environment. More especially, assume that $\mathbf{T}_{a,b}$ is the FPM from C_a to C_b, then $(\mathbf{x}_a^i)_b = \mathbf{T}_{a,b} \cdot \mathbf{x}_a^i$ is the feature presentation of $\mathbf{x}_a^i$ transferred from C_a to C_b, where $\mathbf{T}_{a,b}$ is a $d \times d$ matrix.

Then the new distance between $\mathbf{x}_a^i$ and $\mathbf{x}_b^j$ can be defined as

$$D_{\mathbf{T}_{a,b}}\left(\mathbf{x}_a^i, \mathbf{x}_b^j\right) = \left\|\left(\mathbf{x}_a^i\right)_b - \mathbf{x}_b^j\right\|^2 = \left\|\mathbf{T}_{a,b} \cdot \mathbf{x}_a^i - \mathbf{x}_b^j\right\|^2. \quad (8)$$

Compared with (7), where the same projection transformation is applied to both features of images from two cameras, the proposed FPM-based method applies feature projection transformation to only one camera rather than two cameras, which equals to using different transformations to different cameras, and hence is able to eliminate differences in the principle of the device.

C. Objective Function for FPM Learning

Motivated by [23], we formulate the FPM, $\mathbf{T}_{a,b}$, as a smooth unconstrained convex optimization problem, where the objective function consists of two terms. The first term acts to project $\mathbf{x}_a^i$ to the nearby space of $\mathbf{x}_b^i$ [see Fig. 4(a)], through which the inconsistency of two cameras is effectively eliminated. We call it consistent term. The second term acts to project $\mathbf{x}_a^i$ to the space apart from $\mathbf{x}_b^j$, where $i \neq j$ [see Fig. 4(b)], which holds the discriminative ability of the transformed feature space under the FPM, and hence we refer to it as discriminative term.

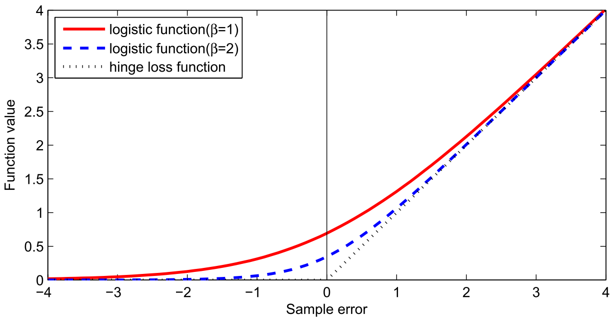


Fig. 5. Explanation for that the logistic loss gives a soft approximation to hinge loss, where the β is larger, the logistic loss is more near to hinge loss.

Specifically, the consistent term can be defined by the sum of feature distance of all similar pairs

$$E_{\text{CON}}(\mathbf{T}_{a,b}) = \sum_{i=1}^{n} D_{\mathbf{T}_{a,b}}\left(\mathbf{x}_a^i, \mathbf{x}_b^i\right). \quad (9)$$

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

Before defining the discriminative term, we first introduce a triple sample $(\mathbf{x}_a^i, \mathbf{x}_b^i, \mathbf{x}_b^j)$, where $\mathbf{x}_a^i$ comes from C_a whereas $\mathbf{x}_b^i$ and $\mathbf{x}_b^j$ from C_b, and $\mathbf{x}_a^i$ and $\mathbf{x}_b^i$ represent the same ith person, whereas $\mathbf{x}_a^i$ and $\mathbf{x}_b^j$ denote different individuals, when $i \neq j$. With this definition, we denote the triple sample set as $S = \{(\mathbf{x}_a^i, \mathbf{x}_b^i, \mathbf{x}_b^j)_k | k = 1, \ldots, s\}$, where s is the size of the set. For each triple sample, the following inequality needs to be satisfied under the FPM:

$$D_{\mathbf{T}_{a,b}}\left(\mathbf{x}_a^i, \mathbf{x}_b^i\right) < D_{\mathbf{T}_{a,b}}\left(\mathbf{x}_a^i, \mathbf{x}_b^j\right). \quad (10)$$

Therefore, we define a error function for one triple sample $(\mathbf{x}_a^i, \mathbf{x}_b^i, \mathbf{x}_b^j)$ as

$$e\left(\mathbf{x}_a^i, \mathbf{x}_b^i, \mathbf{x}_b^j\right) = D_{\mathbf{T}_{a,b}}\left(\mathbf{x}_a^i, \mathbf{x}_b^i\right) - D_{\mathbf{T}_{a,b}}\left(\mathbf{x}_a^i, \mathbf{x}_b^j\right). \quad (11)$$

With this error function, the formulation of the discriminative term can be defined as

$$E_{\text{DIS}}(\mathbf{T}_{a,b}) = \sum_{k=1}^{s} \ell_\beta\left(e\left(\mathbf{x}_a^i, \mathbf{x}_b^i, \mathbf{x}_b^j\right)_k\right) \quad (12)$$

where $\ell_\beta(z) = (1/\beta)\log(1 + e^{\beta z})$ is the generalized logistic loss function. It is easy to see that this term in objective function penalizes triple samples invading the inequality (10).

Different from [23] where the hinge loss function $h(z) = \max(0, z)$ is exploited to guarantee the inequality (10), two reasons are considered for selecting logistic loss function: 1) the hinge loss is not differentiable at zero (see Fig. 5), whereas logistic loss function has derivatives everywhere which makes the solution simpler and 2) the logistic loss gives a soft approximation to hinge loss and more flexible than it. The parameter of logistic loss β is the approximation parameter. As larger as the β is, the logistic is more near to hinge loss, that is, $\lim_{\beta\to\infty} \ell_\beta(z) = h(z)$ (see Fig. 5). The experiments results show that the proposed algorithm achieves better performance under a more suitable β (see Fig. 11).

Finally, we combine $E_{\text{CON}}(\mathbf{T}_{a,b})$ and $E_{\text{DIS}}(\mathbf{T}_{a,b})$ terms with a single objective function for learning FPM[1]

$$E(\mathbf{T}_{a,b}) = (1-\mu)E_{\text{CON}}(\mathbf{T}_{a,b}) + \mu E_{\text{DIS}}(\mathbf{T}_{a,b}) \tag{13}$$

where μ is a balancing factor that can be determined via cross validation.

D. Stochastic Sampling-Based Optimization Algorithm

With the above objective function, that is, (13), the optimal FPM can be learned by solving the following optimization problem:

$$\mathbf{T}^*_{a,b} = \arg\min_{\mathbf{T}_{a,b}} E(\mathbf{T}_{a,b}). \tag{14}$$

It is easy to see that the consistent term is convex. As the logistic loss function is convex, the discriminative term is also convex. Consequently, (14) is a convex optimization problem with respect to $\mathbf{T}_{a,b}$, and can be solved using a simple gradient-descent method.

However, it is time consuming with massive training samples. In particular, the triple sample, that is, $(\mathbf{x}^i_a, \mathbf{x}^i_b, \mathbf{x}^j_b)$, is used in the discriminative term. For easy to discuss, we call the $\mathbf{x}^i_b$ positive instance, and $\mathbf{x}^j_b$ negative instance. Assume there are n persons across two cameras C_a and C_b, for each $\mathbf{x}^i_a$ in C_a, there are one positive instance and $n-1$ negative instances in C_b, and hence the size of the total triple samples is $O(n^2-n)$, which is computationally expensive. In [23], a active set strategy was exploited to improve the efficiency. The active set, which consists of the imposter samples invading a inequality similar to (10), is relatively little. Even so, it is time consuming for using the k closest within-class samples [4] and for recomputing the imposter samples.

Motivated by SGD algorithm [28], we exploit a simple batch gradient descent algorithm with randomly selected samples to accelerate the iteration speed and meanwhile keep the optimization accuracy. Specifically, for each positive examples, we randomly select $m \ll n$ negative samples, with which the size of training set reduces to $O(mn)$ from $O(n^2-n)$.

With the stochastic sampling strategy, a simple gradient-descent method can be exploited to learn the FPM $\mathbf{T}_{a,b}$. The gradient of the objective function is given as

$$\frac{\partial E_{\text{BOTH}}(\mathbf{T}_{a,b})}{\partial \mathbf{T}_{a,b}} = \frac{\partial E_{\text{CON}}(\mathbf{T}_{a,b})}{\partial \mathbf{T}_{a,b}} + \frac{\partial E_{\text{DIS}}(\mathbf{T}_{a,b})}{\partial \mathbf{T}_{a,b}} \tag{15}$$

where

$$\frac{\partial E_{\text{CON}}(\mathbf{T}_{a,b})}{\partial \mathbf{T}_{a,b}} = 2\sum_{i=1}^{n} \left(\mathbf{T}_{a,b}\mathbf{x}^i_a - \mathbf{x}^i_b\right)\left(\mathbf{x}^i_a\right)^\top \tag{16}$$

$$\frac{\partial E_{\text{DIS}}(\mathbf{T}_{a,b})}{\partial \mathbf{T}_{a,b}} = 2\sum_{k=1}^{s} g(e(S_k))\left(\mathbf{x}^j_b - \mathbf{x}^i_b\right)_k\left(\mathbf{x}^i_a\right)^\top_k \tag{17}$$

where $g(x) = (1+e^{-\beta x})^{-1}$ is the derivative of the logistic loss function $\ell_\beta(x)$.

Algorithm 1 Learning the FPM

```
Input: The training set data: Positive samples with pair form $S_P = \{(\mathbf{x}^i_a, \mathbf{x}^i_b)\}$, Negative Samples with triple form $S_N = \{(\mathbf{x}^i_a, \mathbf{x}^i_b, \mathbf{x}^j_b)'_m\}$
Initialize $\mathbf{L}_0$ as identical matrix;
for $i = 1$ to $MaxIter$ do
    Compute $\nabla E_{\text{BOTH}}(\mathbf{T}_{a,b}) = \frac{\partial E_{\text{BOTH}}(\mathbf{T}_{a,b})}{\partial \mathbf{T}_{a,b}}$ as(15)-(17)
    Choose a proper step $\lambda$ as [35]
    Compute $\mathbf{T}^{i+1}_{a,b} = \mathbf{T}^i_{a,b} - \lambda \nabla E(\mathbf{T}^i_{a,b})$ as(18)
    if converge then
        break;
    end if
end for
Output: The optimal matrix $\mathbf{T}^*_{a,b}$
```

With the gradient, an iterative optimization algorithm can be used to learn the FPM. Starting from an initial identical matrix, which means no projection to instance, the FPM is optimized iteratively with the gradient as

$$\mathbf{T}^{i+1}_{a,b} = \mathbf{T}^i_{a,b} - \lambda \cdot \frac{\partial E\left(\mathbf{T}^i_{a,b}\right)}{\partial \mathbf{T}^i_{a,b}} \tag{18}$$

where $\lambda > 0$ is a step length automatically determined at each gradient update step using a similar strategy in [35]. The iteration of the algorithm is terminated when the update times are greater than the maximum iterative times (1000 in this paper) or the following criterion is met:

$$\mid E_{i+1} - E_i \mid < \varepsilon \tag{19}$$

where ε is a small positive value, 10^{-9}, in this paper. The complete algorithm flow is showed in Algorithm 1.

IV. Experiments

In this section, the proposed approach is validated by comparing with several state-of-the-art person reidentification methods on three publicly available datasets: the VIPeR dataset [11], the CUHK person reidentification dataset [34] and the PRID 2011 dataset (single shot version) [36]. The reasons of selecting these datasets are as follows: 1) these datasets cover a wide range of problems faced in the real world person reidentification applications, for example, viewpoint, pose, and lighting changes and 2) they provide two labeled image sets of persons captured by two cameras with nonoverlapping fields of views, in which images of the same person have the same label, while images of the different persons have different labels.

[1] A similar objective function appeared in LMNN [23]. The difference between the proposed method and LMNN is threefold: 1) LMNN uses the hinge loss function, while we exploit the logistic loss function, making the objective function of our method a smooth convex optimization problem and hence easy to be solved with a gradient descent algorithm; 2) LMNN uses a subgradient descent algorithm to solve the optimization problem, while we use a simple gradient descent algorithm with randomly selected samples to increase speed while maintaining performance; and 3) the last but not the least, LMNN, belonging a classic metric learning method, applies the same feature transformation to both feature vectors coming from different devices, which can hardly eliminate the device difference. In contrast, the FPM is applied to only one camera, aiming to eliminate the device difference in principle.

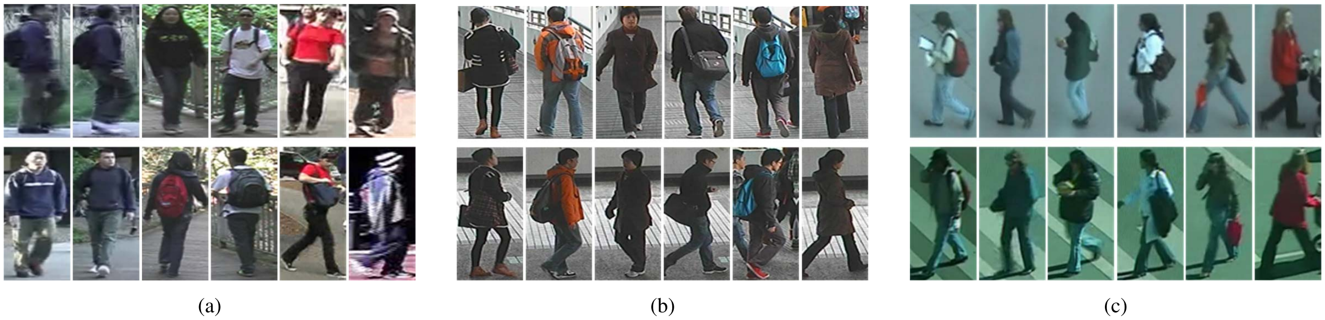


Fig. 6. Some typical samples of three public datasets. Each column shows two images of the same person from two different cameras with significant changes on view point and illumination condition. (a) VIPeR dataset contains significant difference between different views. (b) CUHK is similar to VIPeR, but more challenge as it contains more person pairs. (c) PRID dataset has significant and consistent lighting changes.

A. Datasets

The widely used VIPeR dataset is collected by Gray and Tao [11] and contains 1264 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 using two different cameras, under different viewpoints, pose and light conditions, respectively. All images are normalized to 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°. Other variations are also considered, such as illumination conditions and the image qualities.

CUHK person reidentification dataset is a larger dataset recently proposed in [34] and contains 971 identities from two disjoint camera views. Some example images are shown in Fig. 6(b). Each identity has two samples per camera view. Therefore, there are 3884 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 or 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 views for each people as the really used dataset.

The PRID 2011 dataset [36] consists of person images from two different static surveillance cameras. Camera A contains 385 persons, and camera B contains 749 persons, with 200 of them appearing in both cameras. All images are normalized to 128 × 48 pixels. Different to VIPeR dataset and CUHK dataset, this dataset has significant and consistent lighting changes [Fig. 6(c)]. With this dataset, we mainly evaluate the effectiveness of the proposed approach for different illumination conditions by comparing BTF and CBTF. The 200 image pairs are selected for training and testing.

B. Image Representation

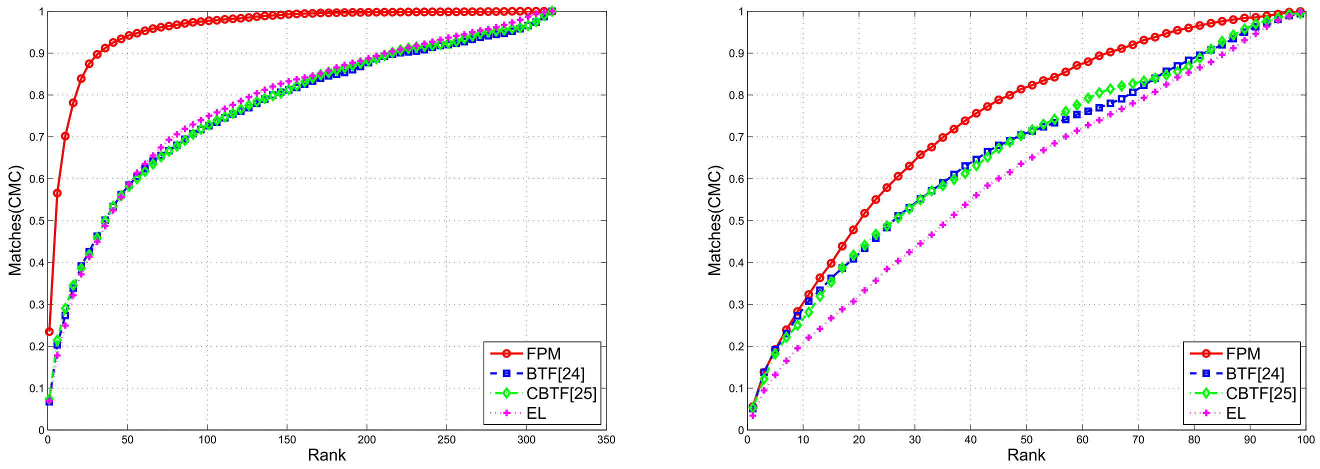
A combination feature descriptor consisting of color and texture features is used to represent images of individuals. Specifically, for each image, the RGB and HSV color histograms and LBP descriptor are extracted from overlapping blocks of size 16×16 (16×12 for CUHK) and stride of 8×8 (8×6 for CUHK), that is, 50% overlap in both directions. RGB and HSV histograms encode the different color distribution information in the RGB and HSV color space, respectively. The uniform rotation-invariant LBP descriptors [37], encoding the texture feature, are extracted in gray-scale images. The bin numbers of RGB and HSV histograms are 24, and the bin number for LBP descriptor is 59. All of the features are then put together to concatenated to a vector. To accelerate the learning process and reduce noise, we conducted principle component analysis (PCA) to obtain a low-dimension representation as [1], that is, 100 in this paper unless otherwise specified. It is worth noting that, the background subtraction technology, such as used SDALF [8], can improve the performance of the algorithm. However, in our experiments, we do not use any background separation technology for a fair comparison, because most of the state-of-the art methods [1], [2], [7], [9], [32], also do not use background modeling technology.

C. Baselines and Settings

To evaluate the effectiveness of the proposed FPM, we compare with methods based on transfer function, including BTF [24] and its extension cumulative brightness transfer function (CBTF) [25], and metric learning methods, containing Mahalanobis metric (Mahal) [1], LMNN [23] and information theoretical metric learning (ITML) [38], and several representative person reidentification methods, such as ELF [32], RankSVM [9], SDALF [8], PRDC [7], KISSME [1], and PCCA [2]. In addition, Euclidean distance (L2) is used as a baseline in most experiments. As LMNN and ITML are not designed for person reidentification, we use codes of these methods provided by their authors and report their results under the optimal parameter configurations.

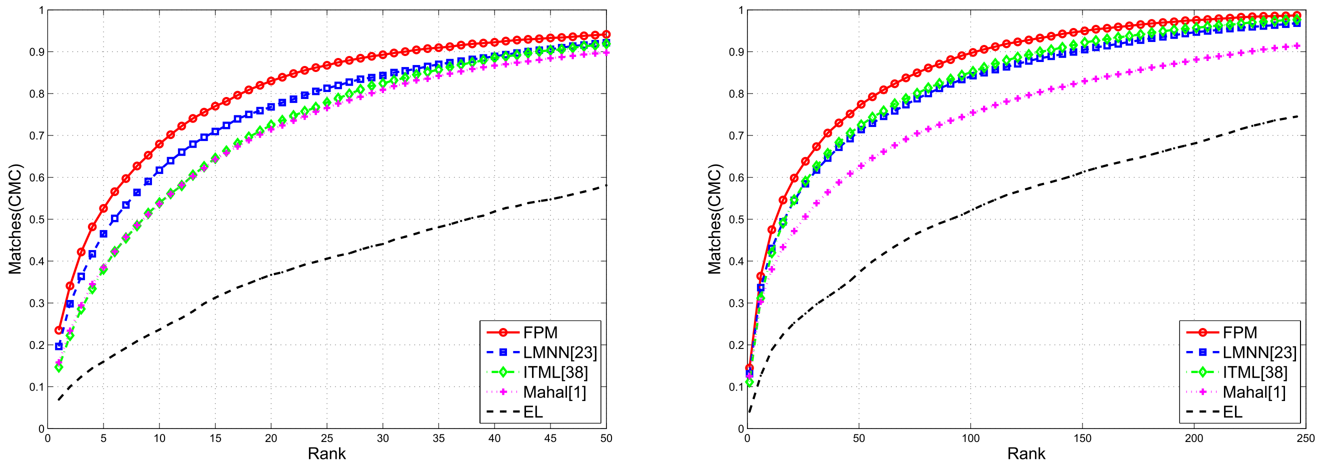
Moreover, we evaluate our approach with different terms. In particular, FPM_CON only uses the consistent term, FPM_DIS only uses the discriminative term, and FPM_BOTH, abbreviated as FPM without confusion, is the combination of above two terms which is the final version of this paper.

Similar to [7], our experiments were designed as follows. Assume that there are two nonoverlapping cameras C_a and C_b and N image pairs $S = \{(x_a^i, x_b^i)_n | n = 1, 2, \ldots, N\}$, in which x_a^i and x_b^i are the images of same person captured



(a) (b)

Fig. 7. Comparative results with BTF [24] and CBTF [25] algorithms on (a) VIPeR and (b) PRID datasets.



(a) (b)

Fig. 8. Comparative results with metric learning algorithms on (a) VIPeR and (b) CUHK datasets.

by C_a and C_b, respectively. First, p (e.g., 316 in VIPeR) image pairs were randomly selected as training set and the rest as testing set. Second, for each training image pair (x_a^i, x_b^i), q different labeled images were selected randomly from C_b to form the triple training sample (x_a^i, x_b^i, x_b^j). Therefore, the count of training samples is $p \times q$. Third, the test set was divided into a probe set consisting of images taken from C_a, and a gallery set made up of images captured by C_b. Finally, each image in the probe set was matched with all images in the gallery set, and the rank of the real match was recorded. For each subexperiment in this paper, the above procedure was repeated 20 times, and the average of the cumulative matching characteristic (CMC) curve, which is suggested in [14], was reported.

The CMC curve is exploited by most papers on the person reidentification problem [1], [2], [4], [7]–[9]. The value of CMC@l indicates the percentage of the real match ranked in the top k. More formally, let $P = \{p_1, \ldots, p_{|P|}\}$ be a probe set, where $|P|$ is the size of P, and $G = \{g_1, \ldots, g_n\}$ be a gallery set. For each probe image $p_i \in P$, all gallery images $g_j \in G$ are ranked based on a defined distance function. The correct match is denoted as g_{p_i}, of which the rank index is denoted as $r(g_{p_i})$. The CMC@l is defined as

$$\text{CMC}_l = \frac{\sum_{i=1}^{|P|} \mathbf{1}(r(g_{p_i}) \leq l)}{|P|} \qquad (20)$$

where $\mathbf{1}(\cdot)$ is the indicator function.

D. Comparing the State-of-the-Art Methods

We first evaluate the performance of our methods by comparing BTF and CBTF, metric learning methods, and the-state-of-the-art person reidentification methods.

1) Comparing BTF and CBTF: We firstly evaluate the effectiveness of the proposed method by comparing with two representative transfer function, BTF and its extension CBTF, on VIPeR dataset and PRID dataset, where the size of training set are 316 and 100, respectively. For BTF and CBTF, the images were first transferred from Camera A to B, then the same feature descriptor (see Section IV-B) was exploited. As shown in Fig. 7, it is obvious that the proposed method leads to a very large performance gain over BTF and CBTF on both datasets. On VIPeR dataset, BTF and CBTF do not improve and even are worse than the Euclidean distance. The reason may be that pose changes are the most significant cause of appearance change, while only a few image pairs have an distinct light changes. On PRID dataset, BTF and CBTF perform almost and better than the Euclidean distance. The results show that FPM still outperforms BTF and CBTF when the appearance changed caused by different illumination conditions. The main reason may be that our approach does not need the assumption of brightness which is not accurate in the practical surveillance condition.

2) Comparing Metric Learning Methods: We also compared FPM with three popular metric learning methods, including Mahal, LMNN, and ITML, on VIPeR dataset and CUHK dataset, where the size of training set are 316 and 485, respectively. Because the public codes of these methods were exploited to learn a optimal distance metric function, we conducted the comparing experiments using the same feature descriptor in Section IV-B, and the same training and testing samples for all methods. The results are shown in Fig. 8. There are two discoveries: a) metric learning-based methods significantly improve the performance for person reidentification, comparing with the widely used standard Euclidean distance. That mainly due to the surprised training samples with which the learned distance function better reflects the characteristics of the data and b) our model outperforms metric learning-based methods on both two datasets. Compensation of the difference of the different cameras with the FPM is the main reason. Thus, our method and metric learning-based methods are not alternative but supplementary. Our method acts to compensate the difference, whereas metric learning acts to seek a distance function.

3) Comparing the-State-of-the-Art Person Reidentification Methods: Table I summarizes the comparing results with the-state-of-the-art person reidentification methods on widely used VIPeR dataset with different sizes of training set, that is, 316, 200, and 100. For a fair comparison, the results for these methods are directly taken from the original public papers. As discussion in Section II, SDALF is of feature-based methods, whereas all other methods exploit a supervised learning. Especially, PRDC, KISSME, and PCCA are of metric learning-based methods. The results clearly show that metric learning methods yield better performance, which is consistent to the discussion in Section I. Our approach gives the best performance in most cases, especially, when training set is small and testing set is large at the same time, for example, when 100 persons are used for training and 532 persons for testing, the improvements of the matching rate for FPM on CMC@1, CMC@5, CMC@10, and CMC@20 are more than 3.4%, 7.4%, 7.6%, and 8.7%, respectively.

E. Evaluating Parameters of the Proposed Method

In this section, we validate the proposed approach under different parameters, including exploiting different terms, using different balance weight μ and smooth approximation parameter β.

1) Different Terms: We evaluate the effectiveness of the two terms on VIPeR dataset and CUHK dataset. As can be seen in Fig. 9 and Table I, FPM_DIS has better performance than FPM_CON, whereas the combination of these two terms achieves the best performance. Especially, the combination improves more significant when the training size is small (see Table I). Given 100 training sample pairs, the combination achieves more than double performance comparing the early version [33] at CMC@1.

2) Influence of μ: We also conduct experiments under different μ values for further evaluating the effectiveness of two terms. When $\mu = 1$, it is equal to only using the discriminative term; inversely, it only uses the consistent term for $\mu = 0$. Hence, we change μ from 0.05 to 0.95, the comparing results are shown in Fig. 10. It is obvious that 0.5 is a good choice for μ on both VIPeR and CUHK datasets. As can be seen in the figure, the performance is optimal and relevant stable when μ between 0.3 and 0.6. Besides, when μ increasing to 1 (or decreasing to 0), the performance descends. So, the combination of two terms outperforms both cases with single term and fixing $\mu = 0.5$, which used in others experiments, is a good choice on both VIPeR dataset and CUHK dataset.

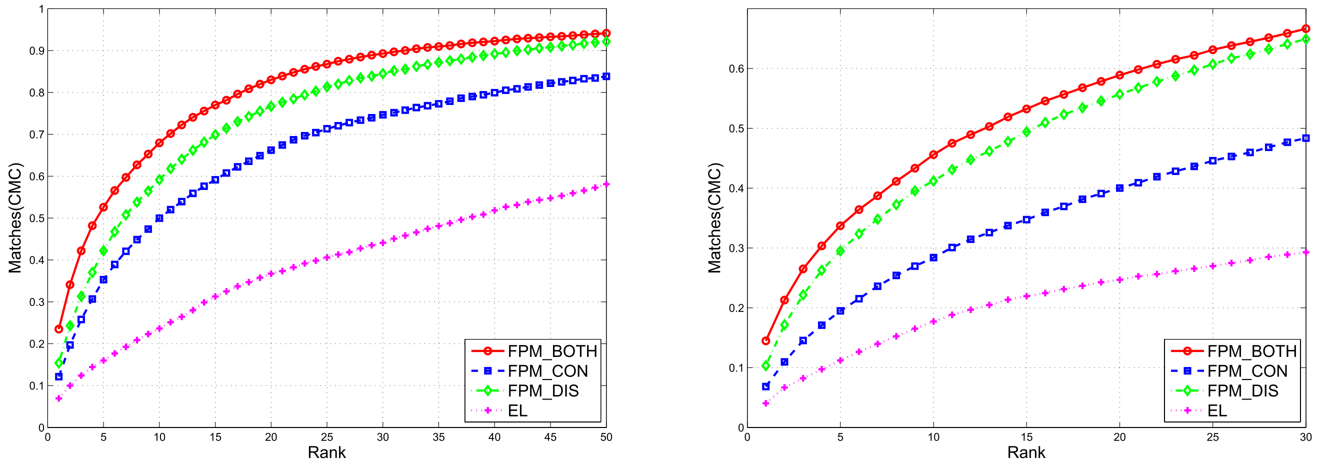
3) Influence of β: Moreover, we use logistic loss function rather than hinge loss function and claim that logistic loss gives a soft approximation to hinge loss and more flexible than it. We change β from 0.001 to 10, the comparing results are shown in Fig. 11. It is obvious that the optimal value of β is about 0.01 for both VIPeR dataset and CUHK dataset. When $\beta > 0.1$, the performance is stable, that is because the logistic loss is equal to hinge loss function. The experiment results are consistent to the discussion on Section III-C, that the algorithm achieves better performance under a suitable β.

F. Evaluating the Efficiency of the Optimization Algorithm

We further validate the claim that the designed optimization algorithm can greatly reduces computation cost and maintains effectiveness at the same time. Under the procedure in Section IV-C, we vary the randomly sampling number and record average CMC and elapsed time. When the random number reach the max value, that is, selecting all samples, the designed optimization algorithm degenerates into the original gradient descent algorithm.

1) Effectiveness of Randomly Sampling: Firstly, we evaluate the effect of the number of stochastic sampling negative sample. The randomly sampling number is changed from 1 to 50, and the average results of 20 times are reported in Fig. 12. We can find that a relatively small value, 10 on VIPeR dataset and 30 on CUHK dataset, has achieved a stable and good effectiveness.

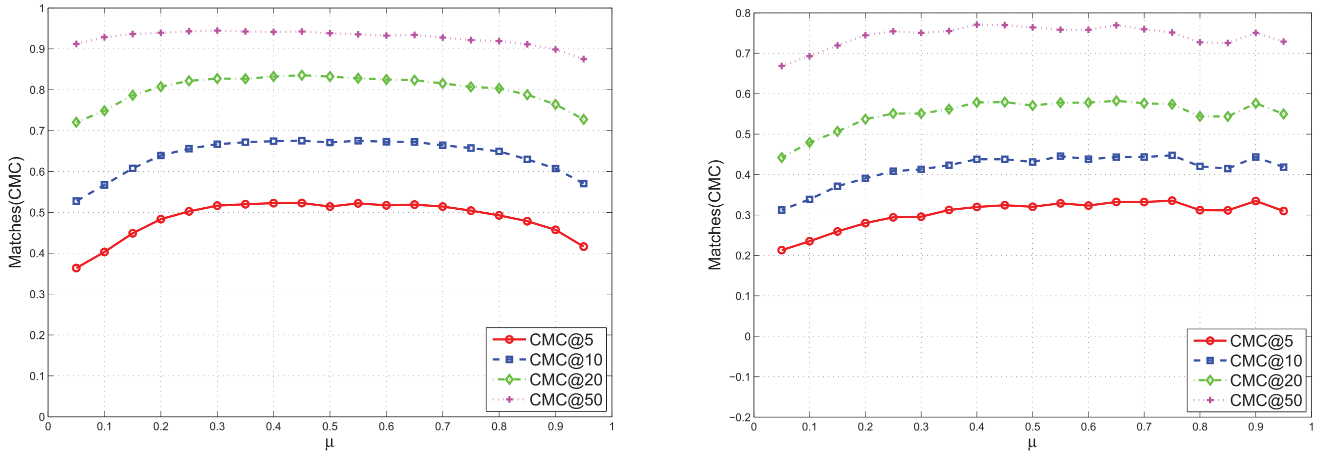
2) Efficiency of Randomly Sampling: Table II summarizes the average and variance of computation times of 20 random experiments with varying numbers of negative samples from



(a)

(b)

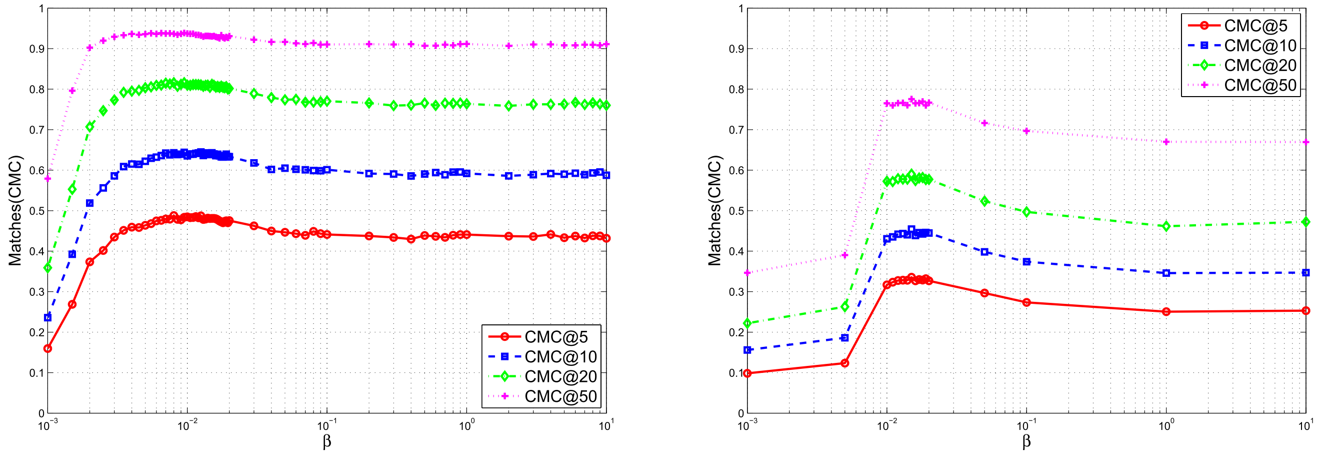
Fig. 9. Comparative results of different constrains on (a) VIPeR and (b) CUHK datasets. FPM_CON only uses the consistent term, FPM_DIS only uses the discriminative term which is the early version of this paper appeared in [33], and FPM_BOTH is the combination of above two terms which is the final version of this paper.



(a)

(b)

Fig. 10. Comparative results of different μs on (a) VIPeR and (b) CUHK datasets.



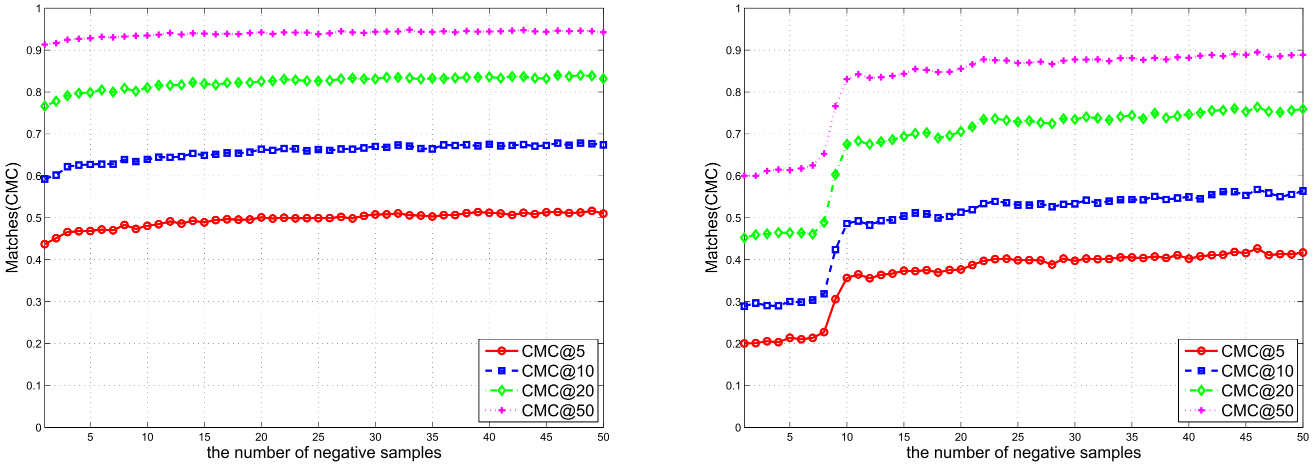
(a)

(b)

Fig. 11. Comparative results of different βs on (a) VIPeR and (b) CUHK datasets.

5 to 45 on VIPeR dataset. From the table, we can observe that the compute time decreases rapidly with the decrease in the randomly sampling number.

With Fig. 12 and Table II, we can see that a relatively small value, such as 10, is good enough for both effectiveness and efficiency, which outperforms LMNN and ITML and is faster



(a) (b)

Fig. 12. Comparative results of different negative number on (a) VIPeR and (b) CUHK datasets.

TABLE II

Comparative Results of Computing Time With Varying Numbers of Negative Samples on VIPeR Dataset

Methods	5	10	15	20	25	30	35	40	45	ITML	LMNN
Times(s) (±variance)	0.78 (±0.23)	1.45 (±0.39)	2.93 (±0.78)	4.3 (±1.26)	6.15 (±1.9)	6.39 (±1.48)	7.61 (±1.89)	9.4 (±2.32)	11.16 (±2.45)	5.9 (±1.45)	18.32 (±0.58)

than them simultaneously, although LMNN exploits a active set technology the ITML uses only one negative sample.

V. Conclusion

In this paper, we propose an FPM method to address the person reidentification. To matching the person images taken from two nonoverlapping cameras, an optimal FPM, which is used to transfer the person images from one camera to the other, is learned by solving a smooth unconstrained convex optimization problem whose objective function consist two terms, consistent term and discriminative term. The first term acts to project images of the same person close together, while the second term acts to make images of different individuals widely separated. Extensive comparative experimental results described in Section IV show that our method is both effective and robust compared with BTF, CBTF, some popular metric learning methods, such as, LMNN and ITML, and several the-state-of-the-art person reidentification methods on three challenging public datasets, VIPeR, CUHK, and PRID.

Although FPM can be used to compensate cameras difference, there are some constrains of using this concept in the real application. First, the FPM is learned by a supervised learning method, which needs a large amount of manual labels, and hence it is usually labor-intensive in the city-level camera network. Therefore, some novel learning technologies collaborating both labeled and unlabeled samples, such as semisupervised or unsupervised learning methods, can be considered. Second, the cameras' difference varies dynamically. Therefore, how to adaptively update the FPM to deal with the dynamical changing environment is an important problem. Motivated by the popularity of transfer learning, an adaptive updating scheme based on a basic FPM would be a promising research direction. Third, in the practical surveillance application, training data is usually obtained in a sequential manner rather than a batch mode, making the online learning of FPM of great importance.

References

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

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

[3] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

[4] 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.

[5] Z. Rui, O. Wanli, and W. Xiaogang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. CVPR*, May 2013, pp. 1–8.

[6] P. Sateesh, O. James, V. Sergio, and B. Boghos, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 3318–3325.

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

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

[9] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. BMVC*, vol. 1. Sep. 2010, no. 3, pp. 1–5.

[10] M. Hirzer, P. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. 12th ECCV*, Oct. 2012, pp. 780–793.

[11] S. B. D. Gray and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop PETS*, Sep. 2007, pp. 1–7.

[12] D. Baltieri, R. Vezzani, and R. Cucchiara, "3DPeS: 3D people dataset for surveillance and forensics," in *Proc. Joint ACM Workshop Human Gesture Behavior Understand.*, Dec. 2011, pp. 59–64.

[13] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Conf. CVPR*, vol. 2. Oct. 2006, pp. 1528–1535.

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

[15] C. Madden, E. Cheng, and M. Piccardi, “Tracking people across disjoint camera views by an illumination-tolerant appearance representation,” *Mach. Vis. Appl.*, vol. 18, nos. 3–4, pp. 233–247, Aug. 2007.

[16] G. Lian, J. Lai, and W.-S. Zheng, “Spatial–temporal consistent labeling of tracked pedestrians across non-overlapping camera views,” *Pattern Recognit.*, vol. 44, no. 5, pp. 1121–1136, May 2011.

[17] G. Lian, J.-H. Lai, C. Y. Suen, and P. Chen, “Matching of tracked pedestrians across disjoint camera views using CI-DLBP,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 1087–1099, Jul. 2012.

[18] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, “Custom pictorial structures for re-identification,” in *Proc. BMVC*, vol. 2. Sep. 2011, no. 5, pp. 1–6.

[19] B. Ma, Y. Su, and F. Jurie, “BiCov: A novel image representation for person re-identification and face verification,” in *Proc. BMVC*, 2012, pp. 1–11.

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

[21] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja, “Pedestrian recognition with a learned metric,” in *Proc. ACCV*, Nov. 2010, pp. 501–512.

[22] M. Hirzer, C. Beleznai, M. Kostinger, P. M. Roth, and H. Bischof, “Dense appearance modeling and efficient learning of camera transitions for person re-identification,” in *Proc. IEEE ICIP*, Oct. 2012, pp. 1617–1620.

[23] 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, Feb. 2009.

[24] K. Shafique and K. Shafique, “Appearance modeling for tracking in multiple non-overlapping cameras,” in *Proc. IEEE Conf. CVPR*, vol. 2. Jun. 2005, pp. 26–33.

[25] B. Prosser, S. Gong, and T. Xiang, “Multi-camera matching using bidirectional cumulative brightness transfer functions,” in *Proc. BMVC*, vol. 8. 2008, p. 164.

[26] A. Datta, L. M. G. Brown, R. Feris, and S. Pankanti, “Appearance modeling for person re-identification using weighted brightness transfer functions,” in *Proc. 21st ICPR*, Nov. 2012, pp. 2367–2370.

[27] A. Gilbert and R. Bowden, “Incremental, scalable tracking of objects inter camera,” *Comput. Vis. Image Understand.*, vol. 111, no. 1, pp. 43–58, Jul. 2008.

[28] T. Zhang, “Solving large scale linear prediction problems using stochastic gradient descent algorithms,” in *Proc. 21st ICML*, 2004, p. 116.

[29] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, “Appearance-based person re-identification in camera networks: Problem overview and current approaches,” *J. Ambient Intell. Humanized Comput.*, vol. 2, no. 2, pp. 127–151, Jun. 2011.

[30] X. Wang, “Intelligent multi-camera video surveillance: A review,” *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3–19, Jan. 2013.

[31] 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.

[32] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in *Proc. 10th ECCV*, Oct. 2008, pp. 262–275.

[33] W. Yimin, H. Ruimin, L. Chao, Z. Chunjie, and L. Qingming, “Camera compensation using feature projection matrix for person re-identification,” in *Proc. IEEE ICME*, Jul. 2013, pp. 1–6.

[34] W. Li, R. Zhao, and X. Wang, “Human reidentification with transferred metric learning,” in *Proc. 11th ACCV*, Nov. 2012, pp. 31–44.

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

[36] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, “Person re-identification by descriptive and discriminative classification,” in *Proc. 17th SCIA*, May 2011, pp. 91–102.

[37] 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.

[38] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in *Proc. ICML*, Feb. 2007, pp. 209–216.

Yimin Wang received the B.S. degree from the School of Computer, Wuhan University, Wuhan, China, in 2008, where he is currently working toward the Ph.D. degree from the National Engineering Research Center for Multimedia Software.

His research interests include computer vision and machine learning.

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

He is the Director with National Engineering Research Center for Multimedia Software, with the Key Laboratory of Multimedia Network Communication Engineering, Wuhan University, Wuhan. He is an Executive Chairman with the Audio Section, Audio Video Coding Standard Workgroup of China, Beijing, China. He has published two books and more than 100 scientific papers. His research interests include audio and video coding and decoding, video surveillance, and multimedia data processing.

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

He is a Post-Doctoral Teacher with National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China. He has published more than 20 papers, including premier conferences and journals. His research interests include multimedia content analysis and retrieval, computer vision, and pattern recognition.

Chunjie Zhang received the B.E. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2006 and 2011, respectively.

He was an Engineer with Henan Electric Power Research Institute from 2011 to 2012. He is currently a Post-Doctoral with the School of Computer and Control, University of Chinese Academy of Sciences, Beijing. His research interests include image processing, machine learning, cross media content analysis, pattern recognition, and computer vision.

Qingming Leng received the B.S. degree in life science from Nanchang University, Nanchang, China, and the M.S. degree from International School of Software, Wuhan University, Wuhan, China, in 2007 and 2009, respectively. He is currently working toward the Ph.D. degree from National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University.

His research interests include computer vision, machine learning, and person reidentification.