An Asymmetric Distance Model for Cross-view Feature Mapping in Person Re-identification

Ying-Cong Chen, Wei-Shi Zheng, Jian-Huang Lai, and Pong C. Yuen

***Abstract*—Person re-identification, which matches person images of the same identify across non-overlapping camera views, becomes an important component for cross-camera-view activity analysis. Most (if not all) person re-identification algorithms are designed based on appearance features. However, appearance features are not stable across non-overlapping camera views under dramatic lighting change, and those algorithms assume that two cross-view images of the same person can be well represented either by exploring robust and invariant features or learning matching distance. Such an assumption ignores the nature that images are captured under different camera views with different camera characteristics and environments, and thus mostly there exists large discrepancy between the extracted features under different views. To solve this problem, we formulate an asymmetric distance model for learning camera-specific projections to transform the unmatched features of each view to a common space where discriminative features across view space are extracted. A cross-view consistency regularization is further introduced to model the correlation between view-specific feature transformations of different camera views, which reflects their nature relations and plays a significant role in avoiding overfitting. A kernel cross-view discriminant component analysis is also presented. Extensive experiments have been conducted to show that asymmetric distance modeling is important for person re-identification, which matches the concerns on cross-disjoint-view matching, reporting superior performance as compared to related distance learning methods on six publically available datasets.**

***Index Terms*—Person re-identification, cross-view matching, visual surveillance**

This research was partly supported by Guangdong Provincial Government of China through the Computational Science Innovative Research Team Program, and partially by Natural Science Foundation Of China (Nos. 61472456, 61522115, 61573387, U1135001), Guangzhou Pearl River Science and Technology Rising Star Project under Grant 2013J2200068, the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265, the GuangDong Program (No. 2015B010105005) and Hong Kong RGC General Research Fund HKBU 12202514. (Corresponding author: Wei-Shi Zheng)

Ying-Cong Chen is with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China.
E-mail: chyingc@mail2.sysu.edu.cn

Wei-Shi Zheng is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, Guangdong Key Laboratory of Information Security Technology, Guangzhou, China, Guangdong Provincial Key Laboratory of Computational Science, Guangzhou, China and Collaborative Innovation Center of High Performance Computing, National University of Defense Technology, Changsha, China.
E-mail: wszheng@ieee.org

Jian-Huang Lai is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, Guangdong Key Laboratory of Information Security Technology, Guangzhou, China. E-mail: stsljh@mail.sysu.edu.cn

Pong C. Yuen is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. E-mail: pcyuen@comp.hkbu.edu.hk
E-mail: pcyuen@comp.hkbu.edu.hk

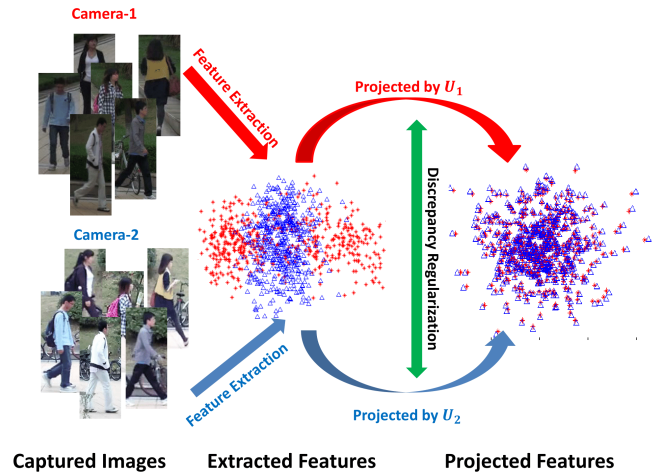


Fig. 1. Illustration of cross-view feature discrepancy problem and our method. These images are selected from the SYSU dataset [1]. After feature extraction, we perform PCA for visualization. It shows that the extracted features are highly divergent, so that the distributions of person images of two views are very distinct and thus the re-identification is extremely difficult. Our method seeks for good view-specific mappings that project the original feature to a common space and make re-identification more reliable. After the feature projection induced by the proposed asymmetric distance model, the person images of two views are more likely to match. To model the correlation nature of different projections, a consistency regularization is imposed to restrict the difference of the projections.

I. Introduction

Nowadays, camera network has been widely deployed in public infrastructure such as airports, railway stations, hospitals for surveillance. Due to economical issue, there are always non-overlapping field between camera views. It then challenges tracking of people and activity prediction over non-overlapping camera networks. Hence it is critical to re-identify a target person when he/she reappears in another camera view. Such a problem is called the *person re-identification*.

However, appearance of a pedestrian would change dramatically across camera views because the environment and camera orientations can be totally different. There are two main feature discrepancy problems: 1) the view-wise discrepancy and 2) the pedestrian-wise discrepancy. The view-wise discrepancy is caused by environment changes such as illumination, the white balance of camera, etc, and the pedestrian-wise discrepancy is caused by pedestrian himself/herself such as those with backpacks or unzipped jackets as well as significant pose changes (see Fig. 2(a) and (b)).

Alleviating the appearance changes across non-overlapping camera views includes 1) seeking discriminative and robust

(a) CUHK01 (b) VIPeR (c) PRID450S (d) SYSU

(e) CAVIAR4REID (f) RAiD(1-3) (g) RAiD(1-4) (h) RAiD(3-4)

Fig. 2. Typical examples of the datasets and sample pairs with view-wise discrepancy or pedestrian-wise discrepancy. Images of the first row of each subfigure were captured by camera a and while images of the second were captured by camera b. Subfigure (a) illustrates the image pairs whose disagreement is more caused by the environmental changes. Subfigure (b) illustrates the images whose disagreement is more caused by the pedestrian himself/herself. Subfigure (c)-(h) illustrates images whose disagreement are caused by both pedestrian himself and environmental changes.

image descriptor [2]–[4], 2) learning reliable distance/subspace models [5]–[9], and 3) preprocessing model such as bright transfer model [10]–[12] and histogram equalization [2], [3]. The first two approaches implicitly assume that one can select a set of features that do not change dramatically. However, appearance could vary dramatically due to indoor/outdoor lighting and pose variations. As such, images of the same person from different camera views will look quite different. Although distance learning methods try to select features robust to those changes, most of these features are extracted based on appearance, especially color features [13] which would be largely affected by illumination or camera characteristics (e.g., white balance). However, the existing methods on using distance learning in person re-identification are all focusing on symmetric modeling, i.e. most of them are based on the following distance form between any two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$:

$$\begin{aligned} d(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \mathbf{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)} \\ &= ||\boldsymbol{U}^T \boldsymbol{x}_i - \boldsymbol{U}^T \boldsymbol{x}_j||_2, \end{aligned} \tag{1}$$

where the positive semidefinite matrix $\mathbf{M}$ is factorized into $\mathbf{M} = \boldsymbol{U}\boldsymbol{U}^T$. [1] The symmetric modeling intrinsically assumes that the same feature transformation is applied to all the camera views, and this ignores feature discrepancy caused by different nature of images captured under different camera views. Since there exists feature discrepancy problem across non-overlapping camera views due to view-wise and pedestrian-wise discrepancy, the conventional unitary projection matrix learning in existing distance/subspace learning methods [5]–[7], [16]–[21] could discard those features with large discrepancy which may be discriminant during the cross-view matching. Sec. III-A will give the details of this analysis.

In this paper, we propose an asymmetric distance model for person re-identification, i.e. we generalize the symmetric form in Eq. (1) and take the view label into account by considering the model based on the following asymmetric form:

$$d(\{\boldsymbol{x}_i^p, p\}, \{\boldsymbol{x}_j^q, q\}) = ||\boldsymbol{U}^{pT} \boldsymbol{x}_i^p - \boldsymbol{U}^{qT} \boldsymbol{x}_j^q||_2, \tag{2}$$

where p and q are the labels of two different camera views and always $\boldsymbol{U}^p \neq \boldsymbol{U}^q$. Essentially speaking, we form the asymmetric learning through learning $\boldsymbol{U}^p$ and $\boldsymbol{U}^q$, which we call the cross-view feature transformation. We hold an assumption that one can seek a latent common space such that the extracted features across different camera views for the same person become more similar and meanwhile for different persons they become more dissimilar. Based on this assumption, we develop a supervised asymmetric distance learning model. We also observe that albeit discrepancy exists across disjoint camera views, there could exist relation between the contents captured by any two camera views, because of the existence of the same person to match and probably similar indoor/outdoor environments. Hence, the discrepancy between feature transformations $\boldsymbol{U}^p$ and $\boldsymbol{U}^q$ should be controlled. To this end, we introduce a cross-view consistency regularization into the cross-view model in order to constrain the difference of view-specific projections, so as to implicitly embed the relation between cross-view images into the distance learning model. Based on the above ideas, we develop a new cross-view matching algorithm for person re-identification, called the Cross-View Discriminant Component Analysis (CVDCA).

In summary, this paper makes the following contributions:

- We propose and develop a new asymmetric distance learning model, called the *cross-view discriminant component analysis* (CVDCA) algorithm to transform the features under different views to a common space for

[1] Conventionally, some works such as [5], [14] directly learn $\mathbf{M}$ under the positive semidefinite constraint, and others like [6], [15] learn $\boldsymbol{U}$, where the learned distance is equivalent to the Euclidean distance of the transformed features. As such, $\boldsymbol{U}$ can be viewed as extracting robust and discriminative transform from the original input space.

person re-identification. The proposed method addresses the feature discrepancy problem by view-specific mappings and models the correlation of different views by a consistency regularization. We also experimentally show that this asymmetric distance model performs much better than the symmetric ones.

- The linear CVDCA is further extended to kernel version and kernelized CVDCA is then proposed.

Extensive experiments have been conducted to demonstrate that the proposed CVDCA and KCVDCA can address the feature discrepancy problem in person re-identification much better.

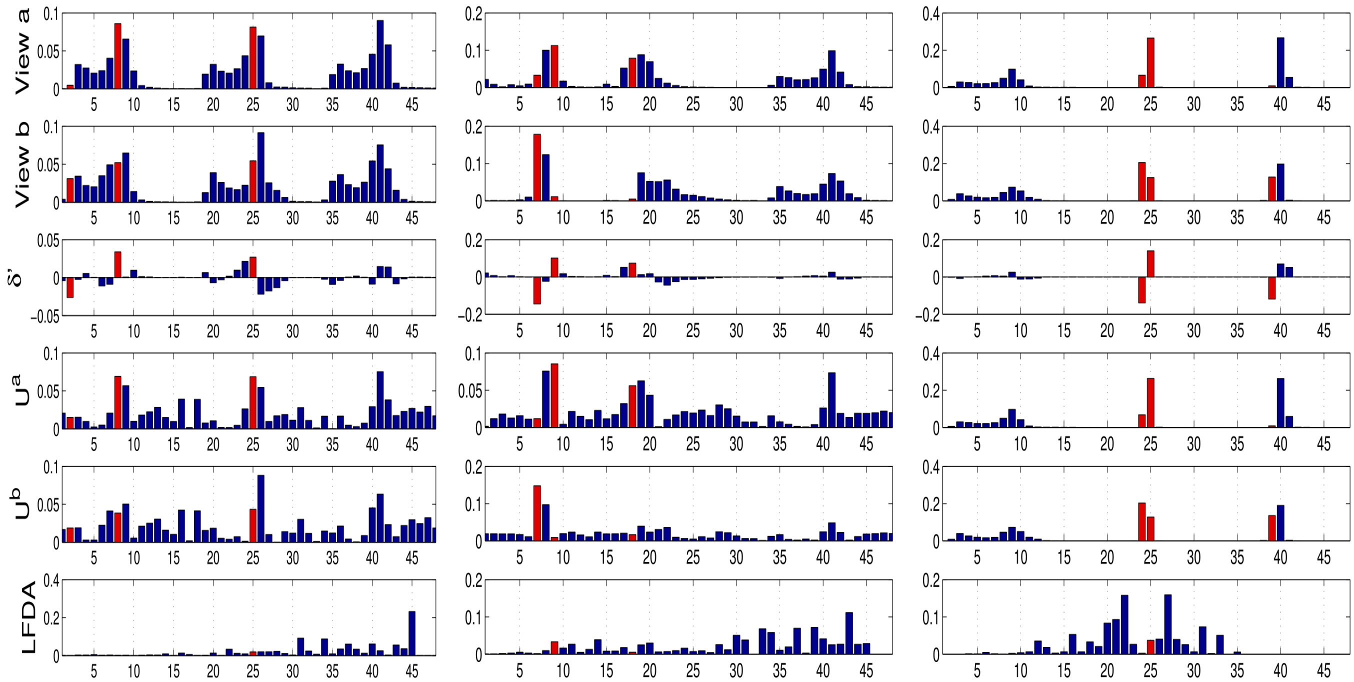
II. Related Work

In order to obtain robust and discriminative representation of pedestrians across different camera views, various methods were proposed to extract color or texture features. Zhao et al. [3] [22] proposed salience-based approaches for person re-identification in which patch matching is employed with adjacency constraint to handle the pose misalignment problem. Later, Zhao et al. [2] proposed mid-level filter which automatically discovers patch clusters was also proposed. However, since color features are used in patch matching, this processing may not be optimal when illumination of different views varies dramatically. Yang et al. [4] proposed a novel salient color name based color descriptor (SCNCD) for person re-identification. However, such a descriptor may be divergent of each view if the lighting of different camera views differs to an extent. Kviakovsky et al. [23] proposed an illumination-invariant color feature based on Log-Chromaticity color space and shape context. However, this method highly depends on high-quality mask, which is usually unavailable in real-world applications. There exist color calibration methods [10]–[12] that aim at learning bright transfer functions (BTFs) to establish a mapping of brightness value between two camera views and thus the gap between them is reduced. However, the cross-camera-view discrepancy is not only caused by lighting. Also, because of incomplete ranges of color value found in the training data, the mapping function may contain many-to-one color correspondence [11], which would cause the loss of useful information. Another approach to deal with the histogram feature mismatch problem is feature warps (FW) [24]. The warp functions are solved by aligning the feature histograms between two camera views, and then they are used as image pair descriptors. Note that this method uses the principles of dynamic time warping to align histograms of each image pair, which implicitly assumes that the divergence of histograms results from histogram shifting.

Due to the difficulty of designing reliable image descriptors across different camera views, some distance/subspace learning methods have been proposed to reduce the variation across views. Zheng et al. [8] formulated person re-identification as a relative distance comparison learning problem by maximizing the probability that relevant samples have smaller distance than the irrelevant ones. Liao et al. [25] proposed a logistic metric learning approach with PSD constraint and asymmetric sample weighting strategy. Li et al. [18] proposed a Locally-Adaptive Decision Function (LADF) to jointly learn the distance matrix and the locally adaptive threshold. Kostinger et al. [17] proposed a simple and effective distance learning called KISSME to conduct hypothesis test on similar/dissimilar pairs. Later, Tao et al. [26] improved KISSME by introducing minimum classification criterion and smoothing technique in order to better estimate the small eigenvalue of the covariance matrix. Liao et al. [27] proposed the Cross-view Quadratic Discriminant Analysis (XQDA) that has similar idea of KISSME but can jointly learn a low-dimension subspace and a metric. Mignon et al. [6] proposed PCCA to learn a projection with sparse pairwise similarity/dissimilarity constraints. Later, Xiong et al. [28] proposed the regularized PCCA (rPCCA) to maximize the inter-class margin and avoid overfitting. Pedagadi et al. [9] applied local fisher discriminant analysis (LFDA) to project the raw features to a discriminative subspace so that the between-class separability is maximized while the multi-modality structure is preserved, and a nonlinear extension using kernel trick of this work was reported in [28]. Paisitkriangkrai et.al [29] proposed a structural learning framework to combine multiple pre-learned distances, which leads to better performance than using an unitary distance measure. All these methods are symmetric-based, and the underlying assumption of the above methods is that features of all camera views have the same properties, while for person re-identification images captured from different camera views could differ notably. Therefore, the unitary projection matrix shared by all views learned by these methods would probably discard the use of divergent features. Recently, sparse reconstruction based classification of face is extended to person re-identification [30], [31]. However, the reconstruction has a underlying assumption that images of the same person should distribute similarly at different camera views, which is not the fact as shown in this work.

Domain adaptation [32]–[34] which can reduce the gap between different distributions seems an alternative solution to cross-view matching. However, those methods cannot be an optimal way to diminish the gap between the two camera views in person re-identification, since they assume the existence of overlapping between training and gallery/testing classes, so that the classifier/metric learned from the training set can be adapted to the gallery/testing one, while for person re-identification there is no overlapping between the training and gallery/testing classes. Note that our work is also different from cross-dataset transferring [35]–[37] since we do not incorporate any source dataset.

There are related works [39]–[44] in person re-identification that can also learn view-specific mappings. An et al. [39], [40] generated a new representation by projecting all samples to the Regularized CCA (rCCA) subspace and constructing the reference descriptors with the reference set. An et al. also proposed robust CCA (ROCCA) [41] to better estimate the data covariance matrices. rCCA and ROCCA are multi-modal learning methods that project heterogeneous features to a common space and thus they are related to our model. However, the person re-identification we discuss in this work is not a multi-modal learning problem and there are important differences between our method and rCCA or ROCCA. Firstly,



(a) RGB (b) HSV (c) YCbCr

Fig. 3. An example of feature discrepancy of RGB, HSV and YCbCr features. The first and the second rows are the distributions of $\frac{1}{n_p^{a,b}}\sum_i \boldsymbol{x}_i^a$ and $\frac{1}{n_p^{a,b}}\sum_j \boldsymbol{x}_j^b$ respectively. The x-axis is the index of bucket and the y-axis is the probability. The third row is the distribution of δ'. The fourth and fifth rows are view-specific mappings of view a and view b trained by our method (will be presented in section III-B). The sixth row is the unitary mapping trained by LFDA [9]. The x-axis corresponds to the buckets in the histogram and the y-axis is the weight of each bucket. The yellow bars indicate those features with large δ'. As shown, the unitary mapping learned by LFDA tends to suppress the weight of the highly divergent features, while our method can utilize those features. This experiment was conducted on PRID450S [38]. Best viewed in color.

rCCA and ROCCA do not control the discrepancy between view-specific feature transformations. Although the feature transformation is specific to each camera view, there should be relation between them, because samples captured from different views are not heterogeneous but related, either from the same identity or from people with similar appearance. As shown in our experiment, this is one of the key factors that makes our model work much better than rCCA. Secondly, rCCA and ROCCA do not consider intra-view modeling which is also useful in our problem. In comparison, our method includes cross-view consistency regularization and intra-view modeling. Besides, we introduce local weighting to the feature transformation processing so as to reduce the impact of extremely different positive sample pairs. Hence our model is more suitable for person re-identification. Liu et al. [42] proposed to learn individual local feature projection for each image sample, which intends to alleviate the influence of configuration variations. In addition, Li et al. [43] proposed to use a gating network to partition the image space of the two camera views into subregions, and some local experts are trained to align the features in the subregions. Some other multi-modality methods like Cross-Modal Metric Learning (CMML) [45] are related to our approach since they also learn view-specific mappings. However, like rCCA and ROCCA, these methods discussed above do not control the discrepancy of inter-view projections or do not incorporate intra-view modeling, which may not be optimal when applying to person re-identification.

III. Approach

A. Feature Discrepancy of Different Camera Views

Let us consider a general case that there are N ($N \geq 2$) cameras with significant feature discrepancy. Let $\boldsymbol{X}^k = [\boldsymbol{x}_1^k, \boldsymbol{x}_2^k, \dots, \boldsymbol{x}_{n^k}^k] \in \mathbb{R}^{d\times n^k}$ denote the feature matrices extracted from the pedestrian images captured by the k-th view, where d is the feature dimension and n^k is the number of samples of the k-th view. The average intra-class variation δ and its lower bound δ' of two specific views (view a and view b) are given by:

$$\begin{aligned}\delta &= \tfrac{1}{n_p^{a,b}} \sum_{i,j\in Ç^{a,b}} |\boldsymbol{x}_i^a - \boldsymbol{x}_j^b| \\ &\geq \tfrac{1}{n_p^{a,b}} \sum_{i,j\in Ç^{a,b}} (\boldsymbol{x}_i^a - \boldsymbol{x}_j^b) = \delta',\end{aligned} \tag{3}$$

where $Ç^{a,b}$ is the set of all positive pairs in view a and view b, and $n_p^{a,b}$ is the cardinality of $Ç^{a,b}$.

Let us consider a single-shot situation, i.e., each pedestrian has only one image for each view with $n^a = n^b$. Then δ' can be rewritten as:

$$\delta' = \frac{1}{n_p^{a,b}} \sum_{i=1}^{n^a} \boldsymbol{x}_i^a - \frac{1}{n_p^{a,b}} \sum_{j=1}^{n^b} \boldsymbol{x}_j^b. \tag{4}$$

Assume that $\boldsymbol{X}^a$ and $\boldsymbol{X}^b$ are histogram features. We draw $\frac{1}{n_p^{a,b}}\sum_i \boldsymbol{x}_i^a$, $\frac{1}{n_p^{a,b}}\sum_j \boldsymbol{x}_j^b$ and δ' in row 1 and row 2 of Fig. 3. We observe that $\frac{1}{n_p^{a,b}}\sum_i \boldsymbol{x}_i^a$ and $\frac{1}{n_p^{a,b}}\sum_j \boldsymbol{x}_j^b$ are not identical, i.e., some features are highly divergent. Those highly different features (the red bars) will generate a high δ'. Note that if $\boldsymbol{X}^a$ and $\boldsymbol{X}^b$ are drawn from identical distribution, $\frac{1}{n_p^{a,b}}\sum_i \boldsymbol{x}_i^a$ and $\frac{1}{n_p^{a,b}}\sum_j \boldsymbol{x}_j^b$ shall be similar. Therefore, we believe that such highly different features between the two views are caused by the unmatched distributions, which will lead to the feature discrepancy problem.

Most supervised subspace/metric learning methods try to reduce the intra-class variation and the lower bound δ' will also be reduced. If a method learns an identical projection or distance matrix for all views, the weights of the divergent features tend to be reduced since those features will cause high intra-class variation. As shown in row 6 of Fig. 3, taking LFDA [9] for example, it learns unitary projection for both views, and thus the weights for highly divergent features are relatively small. However, those features could contain some discriminative information, and deemphasizing them may result in a performance drop. Since using an unitary mapping for all views is not optimal to extract discriminative features, we propose to learn camera-view specific mappings. The camera-view specific mappings are learned so as to transform those features to a common space. As shown in row 3 and row 4 of Fig. 3, by using view-specific mappings, the weights on highly divergent features do not have to be suppressed and more features can be used. By learning view specific transforms, we ultimately formulate an asymmetric distance model called CVDCA for matching person images across disjoint camera views.

To provide a further analysis of the discrimination power of symmetric and asymmetric distance, we quantify the power by computing the quotient between the average inter-class distance and the average intra-class distance based on the features generated by CVDCA and LFDA. The quotient is defined as follows:

$$Q = \frac{\sum\limits_{i,j\in\overline{\text{Ç}}^{a,b}} ||\boldsymbol{y}_i^a - \boldsymbol{y}_j^b||^2}{\sum\limits_{i,j\in\text{Ç}^{a,b}} ||\boldsymbol{y}_i^a - \boldsymbol{y}_j^b||^2}, \tag{5}$$

where $\boldsymbol{y}_i^a$ and $\boldsymbol{y}_j^b$ are the projected features, $\overline{\text{Ç}}^{a,b}$ is the set of all negative pairs in view a and view b and $\text{Ç}^{a,b}$ is the set of all positive pairs. Here Q represents the quotient. A larger Q indicates the features can be separated better and thus they are more discriminative. Note that the values of Q of the features extracted by CVDCA are 1.47, 2.27 and 1.37 for RGB, HSV, YCbCr respectively, while those extracted by LFDA are 1.10, 1.07 and 1.15 respectively. Therefore, we claim that by using view-specific mappings, more discriminative features are retained. Also, as shown in our experiments, the proposed method does not dismiss the use of these features and achieves a much better performance than LFDA.

B. Discrepancy Reduction by View-specific Transformations

The asymmetric distance model based person re-identification is formulated by learning feature transformations for each camera view. Let $\boldsymbol{U}^p = [\boldsymbol{u}_1^p, \boldsymbol{u}_2^p, \ldots, \boldsymbol{u}_C^p]$ denote the projection matrices for view p, where $p = 1, 2, \cdots, N$ and C is the dimension of the projected space. We aim at learning $\boldsymbol{U}^p$ that embeds the features $\boldsymbol{X}^p$ into a discriminative common Euclidean space, where the relevant pairs are expected to be with small Euclidean distance and the irrelevant pairs are with large ones.

It is expected that the learned latent common space could model the relations of both cross-view sample pairs and intra-view sample pairs. Hence our model consists of both cross-view modeling and intra-view modeling:

$$f = f_{cross} + \eta f_{intra}, \tag{6}$$

where the cross-view modeling f_{cross} and the intra-view modeling f_{intra} can be formulated as Eq. (7) and Eq. (8), and η is a positive value which controls the weight of intra-view modeling.

$$f_{cross} = \sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\sum_{i=1}^{n^p}\sum_{j=1}^{n^q} \boldsymbol{W}_{ij}^{p,q}||\boldsymbol{U}^{pT}\boldsymbol{x}_i^p - \boldsymbol{U}^{qT}\boldsymbol{x}_j^q||_2^2, \tag{7}$$

$$f_{intra} = \sum_{p=1}^{N}\sum_{i=1}^{n^p}\sum_{j=1}^{n^p} \boldsymbol{W}_{i,j}^{p,p}||\boldsymbol{U}^{pT}\boldsymbol{x}_i^p - \boldsymbol{U}^{pT}\boldsymbol{x}_j^p||_2^2. \tag{8}$$

In the above modeling, $\boldsymbol{W}_{ij}^{p,q}$ is the weight on each pair of samples between view p and view q, and $\boldsymbol{U}^p$ is a projection of view p. We define $\boldsymbol{W}_{ij}^{p,q}$ as:

$$\boldsymbol{W}_{ij}^{p,q} = \begin{cases} \frac{1}{n_{pos}^{p,q}}\boldsymbol{A}_{ij}^{p,q} & if \quad (x_i^p, x_j^q) \in \text{Ç}^{p,q} \\ -\gamma\frac{1}{n_{neg}^{p,q}} & otherwise \end{cases}, \tag{9}$$

where $\boldsymbol{A}_{ij}^{p,q}$ could set as a local weighting term like LFDA [9] or simply set as 1, $n_{pos}^{p,q}$ and $n_{neg}^{p,q}$ are the numbers of positive and negative pairs between view p and q respectively, and γ is a scalar. Since the number of positive pairs is much smaller than the number of negative pairs, we use $\frac{1}{n_{pos}^{p,q}}$ and $\frac{1}{n_{neg}^{p,q}}$ to normalize them, and thus the weight of intra-class modeling and inter-class modeling can be easily modeled by γ. In this way, minimizing the objective function f will reduce the intra-class difference and meanwhile enlarge the inter-class difference. When $p \neq q$, $\boldsymbol{W}^{p,q}$ characterizes the cross-view relationship; when $p = q$, it characterizes the intra-view relationship.

In order to avoid trivial solution, namely $\boldsymbol{U}^k = \boldsymbol{0}$ for $k = 1, 2, \ldots, N$, we additionally incorporate some constraints and formulate an optimization problem as:

$$\begin{aligned} \min_{\boldsymbol{U}^1,\boldsymbol{U}^2,\ldots,\boldsymbol{U}^N} \quad & \sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\sum_{i=1}^{n^p}\sum_{j=1}^{n^q} \boldsymbol{W}_{ij}^{p,q}||\boldsymbol{U}^{pT}\boldsymbol{x}_i^p - \boldsymbol{U}^{qT}\boldsymbol{x}_j^q||_2^2 \\ & + \sum_{p=1}^{N}\sum_{i=1}^{n^p}\sum_{j=1}^{n^p} \boldsymbol{W}_{i,j}^{p,p}||\boldsymbol{U}^{pT}\boldsymbol{x}_i^p - \boldsymbol{U}^{pT}\boldsymbol{x}_j^p||_2^2 \\ s.t. \quad & \boldsymbol{U}^{kT}\boldsymbol{M}^k\boldsymbol{U}^k = \boldsymbol{I}, \quad k = 1, 2, \ldots, N, \end{aligned} \tag{10}$$

where $\boldsymbol{M}^k = \boldsymbol{X}^k\boldsymbol{X}^{kT} + \mu\boldsymbol{I}$ and $\boldsymbol{I}$ denotes the identity matrix which avoids singularity of the covariance matrix.

These constraints ensure the projected features of each view have unit amplitude and thus they are not shrunken to zero.

C. Transformations Constrained by Cross-view Consistency Regularization

Intuitively, if the feature distributions of two views are similar, the learned feature transformations $\boldsymbol{U}^p$ and $\boldsymbol{U}^q$ are also similar; otherwise, the learned $\boldsymbol{U}^p$ and $\boldsymbol{U}^q$ will be different. Since features of corrupted positive pairs are arbitrarily different, e.g. frontal view and dorsal view of a pedestrian wearing a white t-shirt and a black backpack (see Fig. 2), it could make the learned $\boldsymbol{U}^p$ and $\boldsymbol{U}^q$ quite different. These largely different projection basis pairs do not capture the natural property that images from different camera pairs are correlated to an extent, and the performance would drop dramatically when using these projection pairs.

To embed this correlation nature to our model, we propose to penalize those largely different feature transformations. Specifically, the difference of each projection basis pair can be measured by the Bregman discrepancy [46], [47]. Given a strictly convex function $\mathcal{F}:\mathbb{R}^{d\times C}\to\mathbb{R}$, the Bregman discrepancy of a projection pair is given by:

$$d_{\mathcal{F}}(\boldsymbol{U}^p,\boldsymbol{U}^q)=\mathcal{F}(\boldsymbol{U}^p)-\mathcal{F}(\boldsymbol{U}^q)-\nabla\mathcal{F}(\boldsymbol{U}^q)^T(\boldsymbol{U}^p-\boldsymbol{U}^q), \tag{11}$$

where $\nabla\mathcal{F}$ is the derivative of $\mathcal{F}$. For any strictly convex $\mathcal{F}$, $d_{\mathcal{F}}(\boldsymbol{U}^p,\boldsymbol{U}^q)\geq 0$.

The choice of $\mathcal{F}$ is non-trivial to the performance and the computational complexity. If we set $\mathcal{F}(\boldsymbol{x})=\boldsymbol{x}^T\boldsymbol{x}$, the Bregman discrepancy can be simplified to Euclidean distance $||\boldsymbol{U}^p-\boldsymbol{U}^q||_F^2$. As will be shown later, such a regularization term results in an elegant solution and it works empirically well. For all camera pairs, $\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}||\boldsymbol{U}^p-\boldsymbol{U}^q||_F^2$ is added to the objective function Eq. (6). We call this regularization the *cross-view consistency regularization*. In the Appendix, we will explain how this regularization term is related to the prior knowledge of the projection matrices.

Since $\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}||\boldsymbol{U}^p-\boldsymbol{U}^q||_F^2=(N-1)tr(\sum_{k=1}^{N}\boldsymbol{U}^{kT}\boldsymbol{U}^k-2\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\boldsymbol{U}^{pT}\boldsymbol{U}^q)$, where $tr(\cdot)$ denotes the trace operation, we formulate a regularized version of Eq. (10) as:

$$\begin{aligned}&\min_{\boldsymbol{U}^1,\boldsymbol{U}^2,\ldots,\boldsymbol{U}^N}\\&\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\sum_{i=1}^{n^p}\sum_{j=1}^{n^q}\boldsymbol{W}_{ij}^{p,q}||\boldsymbol{U}^{pT}\boldsymbol{x}_i^p-\boldsymbol{U}^{qT}\boldsymbol{x}_j^q||_2^2\\&+\sum_{p=1}^{N}\sum_{i=1}^{n^p}\sum_{j=1}^{n^p}\boldsymbol{W}_{i,j}^{p,p}||\boldsymbol{U}^p\boldsymbol{x}_i^{pT}-\boldsymbol{U}^p\boldsymbol{x}_j^{pT}||_2^2\\&+tr(\lambda\sum_{k=1}^{N}\boldsymbol{U}^{kT}\boldsymbol{U}^k-2\lambda'\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\boldsymbol{U}^{pT}\boldsymbol{U}^q)\\&s.t.\quad \boldsymbol{U}^{kT}\boldsymbol{M}^k\boldsymbol{U}^k=\boldsymbol{I};\quad k=1,2,\ldots,N,\end{aligned} \tag{12}$$

where

$$\lambda=(N-1)\lambda'. \tag{13}$$

This cross-view consistency regularization is important to exploit the intrinsic nature relations between view-specific feature transformations and help alleviate overfitting significantly as evaluated in Sec. IV-D1. We call the above model as *Cross-View Discriminant Component Analysis* (CVDCA).

D. Kernel Extension

The above method learns linear projection matrices for feature transformation and may suffer from the nonlinearity of given data. We further propose the kernel extension to alleviate this problem.

The implicit high dimensional subspace bases of the k-th view could be represented as $\tilde{\boldsymbol{X}}\boldsymbol{\alpha}^k$, where $\tilde{\boldsymbol{X}}$ is the high dimensional column-wise feature matrix of all training data. Therefore, the projected data could be represented as:

$$h^k(\tilde{\boldsymbol{x}}^k)\ =\ \boldsymbol{\alpha}^{kT}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{x}}^k\ =\ \boldsymbol{\alpha}^{kT}\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}^k), \tag{14}$$

where $\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x})=[k(\boldsymbol{X}_1,\boldsymbol{x}),\ldots,k(\boldsymbol{X}_n,\boldsymbol{x})]^T$. $k(\cdot,\cdot)$ is the kernel function and $h^k(\cdot)$ is the projection function of the k-th view and n is the number of training samples.

By substituting Eq. (14) into Eq. (6), we find that the loss function of KCVDCA is similar to the one of CVDCA by replacing $\boldsymbol{U}^p$, $\boldsymbol{U}^q$, $\boldsymbol{x}_i^p$ and $\boldsymbol{x}_j^q$ with $\boldsymbol{\alpha}^p$, $\boldsymbol{\alpha}^q$, $\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}_i^p)$ and $\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}_j^q)$, respectively.

Using the reproducing property of the reproduced kernel Hilbert space, $\langle k(\cdot,\boldsymbol{x}),k(\cdot,\boldsymbol{y})\rangle=k(\boldsymbol{x},\boldsymbol{y})$, the regularization terms in the implicit high dimension space can be represented as $\sum_{k=1}^{N}\boldsymbol{\alpha}^{kT}\boldsymbol{K}\boldsymbol{\alpha}^k$ and $-\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\boldsymbol{\alpha}^{pT}\boldsymbol{K}\boldsymbol{\alpha}^q$, where $\boldsymbol{K}$ is the gram matrices defined as $\boldsymbol{K}=[\boldsymbol{k}(\boldsymbol{X},\boldsymbol{X}_1),\boldsymbol{k}(\boldsymbol{X},\boldsymbol{X}_2),\ldots,\boldsymbol{k}(\boldsymbol{X},\boldsymbol{X}_n)]$.

In summary, the optimization problem of KCVDCA is described as below:

$$\begin{aligned}&\min_{\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2,\ldots,\boldsymbol{\alpha}^N}\\&\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\sum_{i=1}^{n^p}\sum_{j=1}^{n^q}\boldsymbol{W}_{ij}^{p,q}||\boldsymbol{\alpha}^{pT}\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}_i^p)-\boldsymbol{\alpha}^{qT}\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}_j^q)||_2^2\\&+\sum_{p=1}^{N}\sum_{i=1}^{n^p}\sum_{j=1}^{n^p}\boldsymbol{W}_{i,j}^{p,p}||\boldsymbol{\alpha}^{pT}\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}_i^p)-\boldsymbol{\alpha}^{pT}\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}_j^p)||_2^2\\&+tr(\lambda\sum_{k=1}^{N}\boldsymbol{\alpha}^{kT}\boldsymbol{K}\boldsymbol{\alpha}^k-2\lambda'\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\boldsymbol{\alpha}^{pT}\boldsymbol{K}\boldsymbol{\alpha}^q)\\&s.t.\quad \boldsymbol{\alpha}^{kT}\boldsymbol{M'}^k\boldsymbol{\alpha}^k=1;\ k=1,2,\ldots,N,\end{aligned} \tag{15}$$

where $\boldsymbol{M'}^k=\boldsymbol{K}^k\boldsymbol{K}^{kT}+\lambda\boldsymbol{K}$ and $\boldsymbol{K}^k=[\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}_1^k),\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}_2^k),\ldots,\boldsymbol{k}(\boldsymbol{X},\boldsymbol{x}_{n_k}^k)]$.

E. A Closed-Form Solution

To show the solution of the objective function, we take the linear case as an example and the kernel case is similar.

The objective function of optimization problem Eq. (12) can be rewritten as:

$$\begin{aligned}f=\ &tr(\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\boldsymbol{U}^{pT}\boldsymbol{H}^{p,q}\boldsymbol{U}^p+\boldsymbol{U}^{qT}\boldsymbol{H}^{q,p}\boldsymbol{U}^q\\&-2\boldsymbol{U}^{pT}\boldsymbol{R}^{p,q}\boldsymbol{U}^q+\lambda\sum_{k=1}^{N}\boldsymbol{U}^{kT}\boldsymbol{U}^k),\end{aligned} \tag{16}$$

where $\boldsymbol{H}^{p,q}=\boldsymbol{X}^p(\boldsymbol{D}^{p,q}+\eta\boldsymbol{D}^{p,p}-\eta\boldsymbol{W}^{p,p})\boldsymbol{X}^{pT}$, $\boldsymbol{D}^{p,q}$ is a diagonal matrix whose diagonal entries are defined as $\boldsymbol{D}_{ii}^{p,q}=\sum_{j=1}^{n}\boldsymbol{W}_{ij}^{p,q}$ and $\boldsymbol{R}^{p,q}=\boldsymbol{X}^p\boldsymbol{W}^{p,q}\boldsymbol{X}^{qT}+\lambda'\boldsymbol{I}$.

The objective function can be further simplified as:

$$f=tr(\boldsymbol{U}^T\boldsymbol{R}\boldsymbol{U}), \tag{17}$$

where $\boldsymbol{U}$ is a row-wise concatenated matrix that consists of projection bases of all N views and is defined as:

$$\boldsymbol{U}=[\boldsymbol{U}^1;\boldsymbol{U}^2;\cdots;\boldsymbol{U}^N]\in\mathbb{R}^{Nd\times C}, \tag{18}$$

and $\boldsymbol{R}$ is defined as:

$$\boldsymbol{R}=\begin{pmatrix} \boldsymbol{G}^1 & -\boldsymbol{R}^{1,2} & \cdots & -\boldsymbol{R}^{1,N} \\ -\boldsymbol{R}^{2,1} & \boldsymbol{G}^2 & \cdots & -\boldsymbol{R}^{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ -\boldsymbol{R}^{N,1} & -\boldsymbol{R}^{N,2} & \cdots & \boldsymbol{G}^N \end{pmatrix}, \tag{19}$$

where $\boldsymbol{G}^k=\sum_{q\neq k}\boldsymbol{H}^{k,q}+\lambda\boldsymbol{I}$

Note that it is reasonable to relax the constraints $\boldsymbol{U}^{kT}\boldsymbol{M}^k\boldsymbol{U}^k=\boldsymbol{I},\ k=1,2,\ldots,N$ to $\sum_{k=1}^N\boldsymbol{U}^{kT}\boldsymbol{M}^k\boldsymbol{U}^k=N\boldsymbol{I}$, since the relaxed version is sufficient to avoid trivial solution. So the optimization problem can be modified as:

$$\begin{array}{ll}\min\limits_{\boldsymbol{U}} & tr(\boldsymbol{U}^T\boldsymbol{R}\boldsymbol{U}) \\ s.t. & \boldsymbol{U}^T\boldsymbol{M}\boldsymbol{U}=N\boldsymbol{I},\end{array} \tag{20}$$

where $\boldsymbol{M}$ is a block diagonal matrix defined as $\boldsymbol{M}=diag(\boldsymbol{M}^1,\boldsymbol{M}^2,\ldots,\boldsymbol{M}^N)$.

The optimization problem Eq. (20) can be solved by computing c eigenvectors corresponding to the smallest eigenvalues of the following generalized eigen-decomposition problem:

$$\boldsymbol{R}\boldsymbol{u}=\nu\boldsymbol{M}\boldsymbol{u}, \tag{21}$$

where ν is the Lagrange multiplier. After getting C eigenvectors $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots\boldsymbol{u}_C$, the c-th transformation basis for the p-th view is: $\boldsymbol{u}_c^p=\frac{\delta_p(\boldsymbol{u}_c)}{||\delta_p(\boldsymbol{u}_c)||_{\boldsymbol{M}}}$ where $\delta_p(\cdot)$ means getting the p-th sub-vector and $||\boldsymbol{v}||_{\boldsymbol{M}}=\sqrt{\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v}}$.

Since the solution of kernel extension Eq. (15) is quite similar to the linear case Eq. (12), we do not present its solution in detail. By replacing $\boldsymbol{\alpha}$ with $\boldsymbol{U}$, $\boldsymbol{M}^k$ with $\boldsymbol{M'}^k$ (see Eq. (15)), $\boldsymbol{R}^{p,q}$ with $\boldsymbol{K}^p\boldsymbol{D}^{p,q}\boldsymbol{K}^{qT}+\lambda'\boldsymbol{K}$ and $\boldsymbol{H}^{p,q}$ with $\boldsymbol{K}^p(\boldsymbol{D}^{p,q}+\eta\boldsymbol{D}^{p,p}-\eta\boldsymbol{W}^{p,p})\boldsymbol{K}^{pT}$, the solution of the kernel extension can be obtained by solving Eq. (21).

F. Properties of the Distance

In this section, we discuss the properties of the proposed asymmetric distance in Eq. (2). Strictly speaking, our asymmetric distance is not a conventional metric, and we prove that it satisfies the non-negativity, symmetry and triangle inequality properties, but not the coincidence property, and it is actually a pseudometric.

1) Non-negativity: Since d is defined as the L_2-norm of a vector, it is naturally equal or larger than 0.

2) Symmetry: Since

$$\begin{aligned} d(\{\boldsymbol{x}_i^p,p\},\{\boldsymbol{x}_j^q,q\}) &= ||\boldsymbol{U}^{pT}\boldsymbol{x}_i^p-\boldsymbol{U}^{qT}\boldsymbol{x}_j^q||_2 \\ &= ||\boldsymbol{U}^{qT}\boldsymbol{x}_j^q-\boldsymbol{U}^{pT}\boldsymbol{x}_i^p||_2 \\ &= d(\{\boldsymbol{x}_j^q,q\},\{\boldsymbol{x}_i^p,p\}), \end{aligned} \tag{22}$$

the distance is symmetric. Note that the reason why we call the distance *asymmetric distance* is that the projection bases are different for different camera views.

3) Triangle Inequality: Note that

$$||\boldsymbol{A}+\boldsymbol{B}||_2\leq||\boldsymbol{B}||_2+||\boldsymbol{A}||_2, \tag{23}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are vectors. By letting $\boldsymbol{A}=\boldsymbol{U}^{pT}\boldsymbol{x}_i^p-\boldsymbol{U}^{qT}\boldsymbol{x}_j^q$, $\boldsymbol{B}=\boldsymbol{U}^{rT}\boldsymbol{x}_k^r-\boldsymbol{U}^{pT}\boldsymbol{x}_i^p$, we obtain

$$\begin{aligned} &||\boldsymbol{U}^{rT}\boldsymbol{x}_k^r-\boldsymbol{U}^{qT}\boldsymbol{x}_j^q||_2\leq \\ &\quad||\boldsymbol{U}^{rT}\boldsymbol{x}_k^r-\boldsymbol{U}^{pT}\boldsymbol{x}_i^p||_2+||\boldsymbol{U}^{pT}\boldsymbol{x}_i^p-\boldsymbol{U}^{qT}\boldsymbol{x}_j^q||_2. \end{aligned} \tag{24}$$

Thus we have

$$\begin{aligned} &d(\{\boldsymbol{x}_k^r,r\},\{\boldsymbol{x}_j^q,q\})\leq \\ &d(\{\boldsymbol{x}_k^r,r\},\{\boldsymbol{x}_i^p,p\})+d(\{\boldsymbol{x}_i^p,p\},\{\boldsymbol{x}_j^q,q\}). \end{aligned} \tag{25}$$

4) Coincidence: It is noted that $d(\{\boldsymbol{x}^p,p\},\{\boldsymbol{x}^q,q\})=0$ holds if and only if $\boldsymbol{U}^{pT}\boldsymbol{x}^p=\boldsymbol{U}^{qT}\boldsymbol{x}^q$, which means $[\boldsymbol{U}^p;-\boldsymbol{U}^q]^T[\boldsymbol{x}^p;\boldsymbol{x}^q]=0$, which could be an underdetermined problem. Therefore there exists infinite number of input $\{\boldsymbol{x}^p,p\},\{\boldsymbol{x}^q,q\}$ that satisfies $d(\{\boldsymbol{x}^p,p\},\{\boldsymbol{x}^q,q\})=0$. That means $d(\{\boldsymbol{x}^p,p\},\{\boldsymbol{x}^q,q\})=0$ does not always imply $\{\boldsymbol{x}^p,p\}=\{\boldsymbol{x}^q,q\}$. However, fortunately one still has $d(\{\boldsymbol{x}^p,p\},\{\boldsymbol{x}^p,p\})=||\boldsymbol{U}^{pT}\boldsymbol{x}^p-\boldsymbol{U}^{pT}\boldsymbol{x}^p||_2=0$. Therefore, the coincidence property does not strictly holds, and our distance is in fact a pseudometric. However, this does not hurt the model for practical use. It is not practical for visual surveillance to have the constraint that $d(\{\boldsymbol{x}^p,p\},\{\boldsymbol{x}^q,q\})=0$ only when $\{\boldsymbol{x}^p,p\}=\{\boldsymbol{x}^q,q\}$. In visual surveillance, it is rare to have the same appearance representation for the same person at different camera views due to the existence of view changes, lighting changes, etc. Hence it is more practical to say two images are from the same person if they are having the same representation in the transformed space (i.e., $\boldsymbol{U}^{pT}\boldsymbol{x}^p=\boldsymbol{U}^{qT}\boldsymbol{x}^q$), while ensuring the optimization that two images of different people have different representations in that space.

IV. Experimental Results

A. Datasets and Settings

1) Datasets: The evaluation of the proposed method is carried out on six challenging datasets: PRID450S[2] [38], VIPeR[3] [48], CUHK01[4] [49], SYSU[5] [1], CAVIAR4REID[6] [50] and RAiD[7] [51]. Significant feature discrepancy can be observed in all six datasets. PRID450S contains 450 image pairs recorded from two different but static surveillance cameras. In this set, masks generated both automatically and manually were provided to define the foreground regions of interest. VIPeR contains 632 pedestrian image pairs captured outdoor with varying viewpoints and illumination conditions. Each image is scaled to 128×48 pixels. CUHK01 contains 971 pedestrians from two disjoint camera views. Each pedestrian has two samples per camera view. SYSU is a large and diverse dataset that contains totally 48,892 images of 502 pedestrians captured by two cameras in a campus environment. One of the camera is positioned around a corner and thus illumination, pose and viewpoint change dramatically. CAVIAR4REID contains 72 pedestrians of which 50 are viewed in disjoint camera views and 22 are not. Totally 1220 images are included in the dataset. RAiD contains 4 camera views with 2 indoor and 2 outdoor. 43 pedestrians are included in the dataset, resulting in 6920 images. Among the 43 pedestrians, 41 of them appeared in all 4 camera pairs. Illumination and pose change greatly across

[2] Available at https://lrs.icg.tugraz.at/download.php

[3] Available at https://vision.soe.ucsc.edu/node/178

[4] Available at http://www.ee.cuhk.edu.hk/∼xgwang/CUHK_identification.html

[5] Available at http://isee.sysu.edu.cn/resource

[6] Available at http://www.lorisbazzani.info/caviar4reid.html

[7] Available at http://www.ee.ucr.edu/∼amitrc/datasets.php

the different camera views. Although there are other datasets publically available, such as iLIDS [52] and ETHZ [53], we do not conduct experiments on them because our method utilize the camera view label information and those datasets do not provide it.

2) Features: To validate the effectiveness of the proposed method, we extracted only low-level color and texture features in the following experiments. Specifically, we equally partitioned each image into 18 non-overlapped horizontal stripes. For each stripe, RGB, HSV, YCbCr, Lab and YIQ color features as well as 16 Gabor texture features were extracted. For each feature channel, a 16-bin histogram was extracted. In order to balance the weight of each type of feature, we normalized all histograms by L_1-norm. All histograms were concatenated together to form a single vector. Since PRID450S provides automatically generated foreground masks, our features were extracted from the foreground; for other datasets, our features were extracted from the whole image.

The extracted feature contains rich information since it was extracted from dense horizon stripe, including various color space and texture features. However, it is sensitive to illumination or viewpoint changes, so features of different views could suffer great discrepancy.

3) Experimental protocol: All datasets were evaluated with the same training protocol: each time half of the pedestrians were selected randomly to form the training set, and the remaining pedestrian images were used to form a testing set. Since there are 22 pedestrians whose images were only captured in a single view in CAVIAR4REID, we did not select them for experiments and only used the rest 50 pedestrians for evaluation. On RAiD, we followed the experimental protocol in [24], i.e., we used camera pairs 1-3, 1-4 and 3-4 for evaluation (denoted as RAiD(1-3), RAiD(1-4) and RAiD(3-4)), and each pedestrian has 10 images for each view in the multi-shot experiment. On SYSU, we randomly pick 3 images of each pedestrian in each view for evaluation.

The performance was evaluated by both single-shot, i.e., only one image image each person was registered in the gallery set, and multi-shot protocol, i.e., at least two images each person were registered. For metric learning methods, when comparing the distance between a probe person and a gallery person in the gallery set, we calculate the average of the learned distance between a probe image and each of the registered image of that gallery person.

The performance was evaluated with both closed-set protocol and open-set protocol. The Cumulative Matching Characteristic (CMC) curve was used for evaluating the closed-set performance. A rank k matching rate in CMC curve indicates the percentage of the probe image with correct matches found in the top k rank against the p gallery images. In practice, a high rank-1 matching rate is critical and the top k matching rank matching rate with a small k value is also important since the top matching images can be verified by human [8].

To simulate the open-set situation, we also randomly discarded 20% of the gallery images, and thus some of the people in the probe set is not known from the gallery set. In order to quantify how well a true target have been verified and how bad a false target have mistakenly passed through the verification, we followed [35] to use the true target rate (TTR) and false target rate (FTR) to evaluate the performance. TTR and FTR are defined as:

$$\begin{aligned} TTR &= \frac{n_{TT}}{n_T}, \\ FTR &= \frac{n_{NT}}{n_N}, \end{aligned} \tag{26}$$

where n_T indicates the number of query target images from target people, n_{TT} indicates the number of query images that are verified as one of the target people, n_{NT} indicates the number of non-target images from non-target people, and n_{NT} indicates the number of query non-target images that are verified as one of the target people.

4) Methods for comparison: We first compared our methods with symmetric distance learning methods including LFDA [9], KLFDA [28], KISSME [17], regularized kernel PCCA (rKPCCA) [28] and RDC [8]. We also evaluated the regularized canonical correlation analysis (rCCA) [54] and cross-modal metric learning (CMML) [45] which provide view-specific mappings. In our comparison, all methods used the same features and thus the performance difference is only due to the different processing on the extracted features.

We also discuss the comparison of our methods with the state-of-the-art methods in section IV-C, which is on the system-level comparison in order to compare to the state-of-the-art.

5) Parameter: In the following experiments, we set both η and γ to 0.1 for both CVDCA and KCVDCA. In order to balance the scale of the objective function and the feature consistency regularization term, λ' was set as 10^{-3} for CVDCA and 0.3 for KCVDCA. All parameters were fixed for all datasets, and we will discuss those parameters in Section IV-D.

For the kernel methods, the chi-square kernel is used in our experiments.

B. Comparison to the distance/subspace learning methods

1) Closed-set Evaluation: We first discuss the closed-set situation where people in the probe set are represented in the gallery set, which is a conventional person re-identification test. Fig. 4 shows the CMC curves on VIPeR, PRID450S, CUHK01, SYSU, CAVIAR4REID and RAiD respectively, and Table I shows the top ranked matching rate on these datasets.

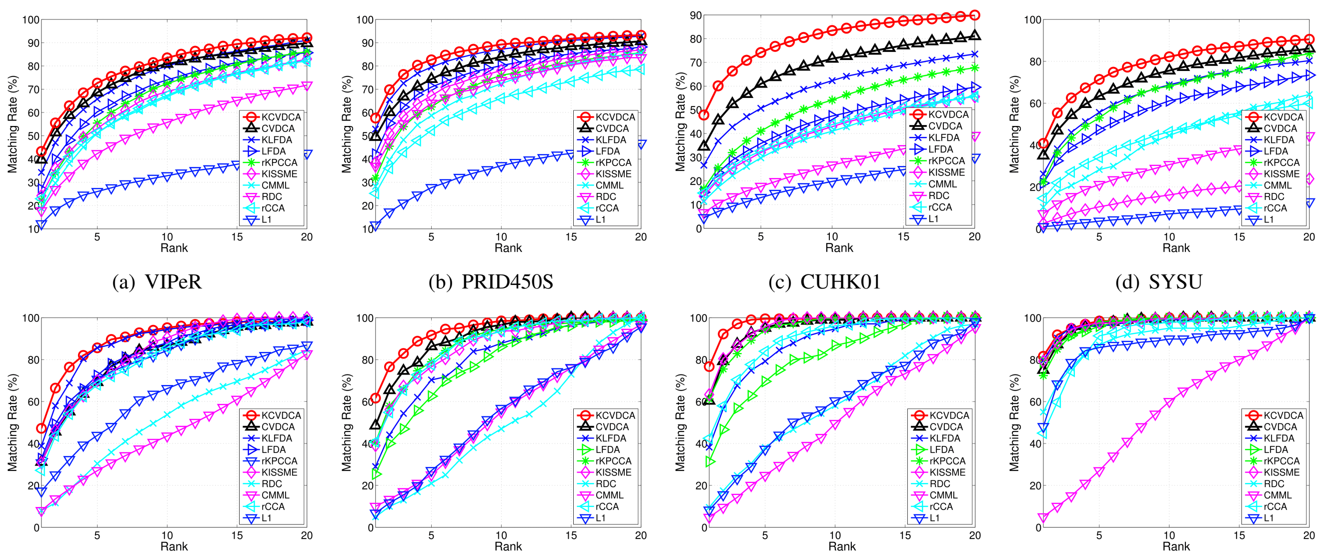
(K)CVDCA vs Baseline: We first compared our methods with the L_1 baseline. Table I shows that L_1 does not perform well in all the six datasets. Note that the features we used consist of low-level color and texture features, which are sensitive to the environment changes across views. Since L_1 is a non-learning distance measure, it is not robust to those changes. Our methods learn asymmetric distance for better measuring the distance of pedestrian images across camera views, and thus they achieve significant improvement.

(K)CVDCA vs Symmetric Distance Learning: Symmetric distance learning methods including (K)LFDA, rKPCCA, KISSME, LMNN and RDC, are most relevant to our methods. The major difference is that symmetric distance learning methods map the original features to a new space with a unitary mapping, while our methods allow different mappings for

TABLE I
TOP RANKED MATCHING RATE (%) ON VIPeR, PRID450S, CUHK01, SYSU, CAVIAR4REID, RAiD(1-3), RAiD(1-4) AND RAiD(3-4).

dataset	VIPeR				PRID450S				CUHK01				SYSU			
rank	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
KCVDCA	**43.29**	**72.66**	**83.51**	**92.18**	**57.60**	**82.67**	**89.24**	**93.20**	**47.80**	**74.16**	**83.44**	**89.92**	**40.84**	**71.35**	**82.19**	**90.56**
CVDCA	39.72	68.58	80.89	89.78	49.47	74.36	83.96	90.62	34.14	60.95	71.52	81.05	34.98	63.43	75.58	86.10
KLFDA [28]	34.27	65.82	79.94	90.92	52.84	79.51	87.20	92.80	26.62	50.63	62.28	73.50	28.69	58.21	70.96	82.39
LFDA [9]	27.03	60.28	73.99	85.70	42.09	70.93	80.18	88.18	15.22	35.80	47.37	59.52	26.22	55.62	68.80	80.32
rKPCCA [6]	22.28	55.47	72.41	86.04	31.82	62.40	76.00	85.73	16.68	41.00	54.11	67.73	22.31	52.99	68.13	83.67
KISSME [17]	24.21	53.10	68.99	82.97	38.49	67.20	78.09	86.89	13.53	31.99	42.89	55.56	16.85	39.84	54.66	68.96
RDC [8]	17.75	42.34	55.73	71.77	36.89	64.00	73.78	83.56	6.56	17.68	26.69	39.16	7.21	21.16	30.68	44.46
CMML [45]	18.77	51.17	66.77	82.31	28.27	58.71	72.40	85.60	11.45	29.72	40.98	56.03	10.36	28.29	44.62	64.14
rCCA [54]	22.94	51.23	67.44	82.09	25.24	52.31	66.13	78.62	14.90	32.59	43.77	55.53	14.58	34.14	46.37	59.96
L_1	12.15	26.01	32.82	42.47	11.64	27.73	37.16	46.76	4.45	12.97	19.80	29.94	1.00	3.67	7.13	12.91

dataset	CAVIAR4REID				RAiD(1-3)				RAiD(1-4)				RAiD(3-4)			
rank	1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
KCVDCA	**47.20**	**85.60**	**95.20**	**98.40**	**61.64**	**91.69**	**98.50**	**100.00**	**76.69**	**99.50**	99.50	**100.00**	**81.63**	**98.45**	**100.00**	**100.00**
CVDCA	31.20	69.20	86.80	98.00	48.50	86.33	96.57	100.00	60.36	95.12	99.00	100.00	75.05	97.45	99.47	100.00
KLFDA [28]	38.80	85.60	94.40	99.20	29.02	70.36	87.83	100.00	38.21	79.21	94.64	100.00	79.58	96.45	99.00	100.00
LFDA [9]	30.00	67.60	85.20	98.40	25.40	62.57	86.29	99.02	31.38	69.62	86.81	100.00	81.37	98.00	99.50	100.00
rKPCCA [6]	30.77	73.08	80.77	100.00	40.57	79.00	94.62	99.02	61.24	94.07	99.52	100.00	72.37	98.00	100.00	100.00
KISSME [17]	30.77	70.00	90.38	100.00	39.02	76.60	93.71	100.00	63.33	95.64	**100.00**	100.00	79.29	97.50	99.50	100.00
RDC [8]	8.00	30.40	53.76	84.64	5.00	21.00	47.00	100.00	9.52	38.10	58.10	96.83	55.00	87.00	95.00	100.00
CMML [45]	8.00	27.00	43.55	82.74	10.00	25.00	55.00	96.67	4.76	24.76	49.52	95.24	5.00	27.00	60.00	100.00
rCCA [54]	27.20	68.00	85.60	98.00	40.76	77.62	93.60	100.00	42.50	84.00	96.14	100.00	44.95	91.34	97.97	100.00
L_1	17.31	43.85	66.15	86.92	6.76	27.07	56.71	95.62	8.21	37.21	60.33	98.05	48.05	85.76	89.84	100.00



(a) VIPeR (b) PRID450S (c) CUHK01 (d) SYSU

(e) CAVIAR4REID (f) RAiD(1-3) (g) RAiD(1-4) (h) RAiD(3-4)

Fig. 4. CMC curves on VIPeR, PRID450S, CUHK01, SYSU, CAVIAR4REID and RAiD(1-3), RAiD(1-4) and RAiD(3-4). Best viewed in color.

different camera views. As such, symmetric distance learning assumes the features used for distance measurements can be both discriminative and invariant to environment changes, while asymmetric distance learning do not hold such a restricted assumption. Hence our methods weaken the assumption of distance learning previously used in person re-identification, and achieve notable improvement as compared to symmetric distance learning methods. Among those symmetric distance learning methods, KLFDA/LFDA achieves relatively good performance. KCVDCA achieves 9.02%, 4.76%, 21.18%, 12.15%, 9.00%, 32.64%, 38.48% and 2.05% improvement over KLFDA at rank 1 on VIPeR, PRID450S, CUHK01, SYSU, CAVIAR4REID, RAiD(1-3), RAiD(1-4) and RAiD(3-4), and for linear case, the difference of rank-1 between CVDCA and LFDA is 12.69%, 7.38%, 18.94%, 8.76%, 1.20%, 23.10%, 28.98% and -6.32% on those datasets respectively. The improvement is particularly notable on SYSU, CUHK01, RAiD(1-3) and RAiD(1-4). The illumination changes of these datasets are extremely large (see Fig. 2(a),(f) and (g)), and the feature difference of two images is caused more by the lighting change than that by the pedestrian identities. Since our methods learn asymmetric distance, i.e., view-specific mapping is used for each view, the influence of lighting change is suppressed and the distance model for matching is more relevant to pedestrian identities. A latter evaluation in Sec. IV-D1 further shows that the performance of our methods would drop notably when they degrade to symmetric distance models. Also note that in RAiD(3-4), KCVDCA performs only slightly better than KLFDA. It is because in RAiD dataset, camera 3 and 4 are both outdoor cameras, and the environmental condition such as illumination does not change much (see Fig. 2(h)). Thus our methods does not show significant advantage in this setting. For indoor-outdoor setting like RAiD(1-3) and RAiD(1-4), our methods performs much better.

(K)CVDCA vs Multi-Modal Learning: Multi-modal learning methods including CMML and rCCA were also evaluated in our experiments. They are related to our model since they and ours all can learn view-specific feature transformations. However, those methods deal with the problem when features of different views are heterogeneous, while for the person re-identification problem we discuss in this work the features of person images under different camera views are not heterogeneous but related. Nevertheless, these methods do not perform well on person re-identification. In contrast, our methods model the relation between view-specific feature transformations with the feature consistency regularization and thus perform much better.

2) Open-Set Evaluation: In addition to the closed-set performance evaluation, we also report the open-set performance as Fig. 5. Note that in the open-set setting, some identities in the probe set are not known in the gallery set, and our objective is to verify whether a query image comes from the people in the gallery set. The TTR versus FTR curve defined by Eq. (26) was used for evaluation. Clearly, our method also achieves the best performance among all methods in comparison. Specifically, when FTR=10%, the TTR of KCVDCA are 96.21%, 96.33%, 93.93%, 89.05%, 44.52%, 75.32%, 90.55%, and 86.93% on VIPeR, PRID450S, CUHK, SYSU, CAVIAR4REID, RAiD(1-3), RAiD(1-4) and RAiD(3-4) respectively, while for KLFDA they are 90.99%, 89.67%, 80.21%, 73.38%, 31.83%, 38.88%, 47.75% and 85.84% respectively.

C. Comparison to the State-of-the-Art

The proposed method is compared with the state-of-the-art methods using the same evaluation protocols. Table. II-VII show the top matching rate on VIPeR, PRID450S, CUHK01, RAiD and CAVIAR4REID datasets, and Fig. 6 shows the CMC curves. Note that SYSU is a newly released dataset and to the best our our knowledge, there is no supervised method conducted on this dataset and therefore it is not used in this section.

On VIPeR, MLF+LADF [2] combines the result of MLF and LADF. For fair comparison, we trained the proposed KCVDCA method using both our low-level features and high-level texture features [55] used by LADF [18], and then simply summed up the score as MLF+LADF [2] did. Fig. 6(a) shows that our method achieves the best performance on this dataset. We have compared our algorithm with the SCNCD [4], KISSME [17], EIML [19], LMNN [5], LMNN-R [20] and ITML [7] on PRID450S, and compared our algorithm with MLF [2], Ref-Reid [39], ITML [7], LMNN [5], SalMatch [22] and eSDC [3] on CUHK01. Fig. 6(b) and Fig. 6(c) show that our approach outperforms other approaches by a large margin.

On RAiD dataset, we have compared our algorithm with the recently proposed NCR [51], FW [24], WACN [56], LFDA [9], SDALF [57], ICT [58] and ISR [30]. Fig. 6(d)-6(f) and Tab. II show that our method could achieve the state-of-the-art on all the three camera pairs.

On CAVIAR4REID, our approach achieves overall better results. In particular, FW [24] is comparative to our approach as observed from Fig. 6(g) and Fig. 6(h). On the comparison with ISR [30], the proposed method achieves clearly better results. Note that the testing protocol used by ISR in [30] is different from ours, i.e., the gallery and probe images are strictly from different camera views in our setting. The experiment shows that ISR does not perform well when matching person images from disjoint/different camera views probably because the gallery images may not be able to reconstruct the probe one very well.

TABLE III
TOP RANKED MATCHING RATE (%) ON CAVIAR4REID (N = 5).

rank	1	5	10	20
KCVDCA	**36.80**	**78.80**	92.40	99.20
FW [24]	33.20	78.50	**94.10**	**100.00**
LFDA [9]	36.19	66.15	88.56	98.41
ISR [30]	14.40	47.60	69.60	94.40

TABLE IV
TOP RANKED MATCHING RATE (%) ON CAVIAR4REID (N = 10).

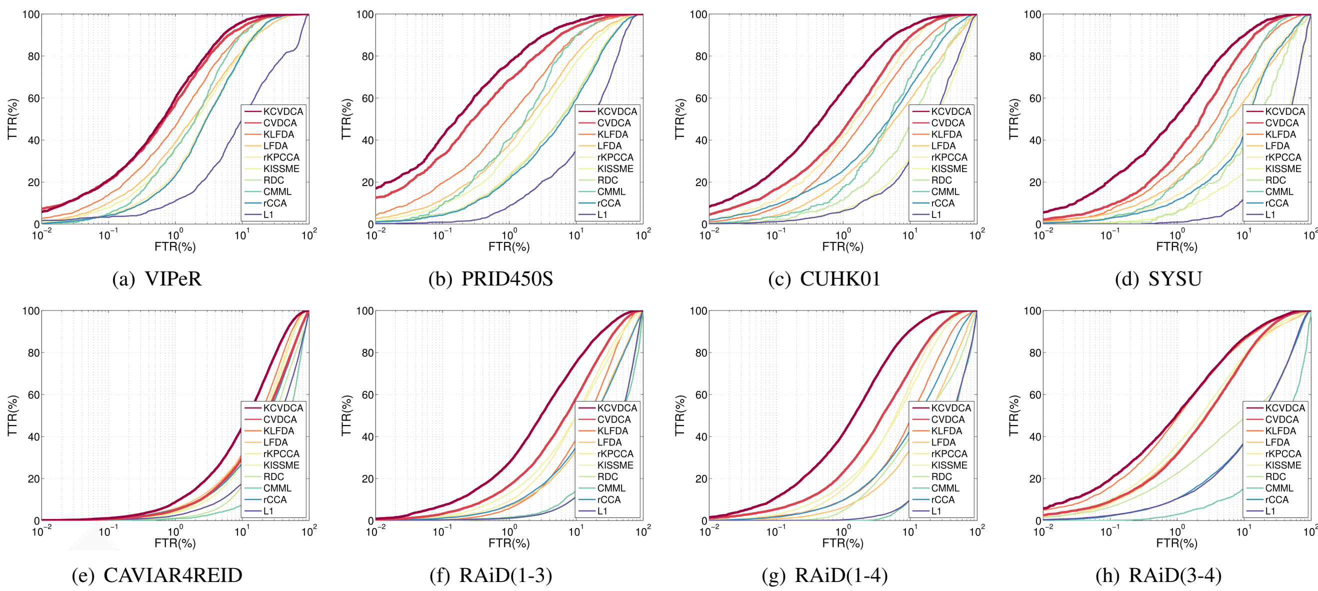
rank	1	5	10	20
KCVDCA	**45.60**	86.00	95.60	99.60
FW [24]	41.90	**86.50**	**96.70**	**100.00**
ICT [58]	26.80	70.40	90.00	99.60
ISR [30]	18.40	50.00	71.20	95.60

TABLE V
TOP RANKED MATCHING RATE (%) ON VIPER COMPARED TO THE STATE-OF-THE-ART.

rank	1	5	10	20
KCVDCA(Fusion)	**47.78**	**76.33**	**86.33**	**94.02**
MLF+LADF [2]	43.39	73.04	84.87	93.70
MLF [2]	29.11	52.34	65.95	79.87
Ref-Reid [40]	33.29	63.54	78.35	88.48
SalMatch [22]	30.16	52.31	65.54	79.15
LADF [18]	29.34	61.04	75.98	88.10
LFDA [9]	24.18	52.00	67.12	82.00
RDC [8]	15.66	38.42	53.86	70.09

D. Discussion

1) Effectiveness of Cross-View consistency regularization: As discussed in III-C, the cross-view consistency regularization is critical to avoid learning arbitrarily different projections for different views. Fig. 7(a) shows the effectiveness of this regularization on PRID450S for example and similar conclusion can be drawn on the others. Note that if the penalty term λ' is infinitely small, the effect of this regularization vanishes and the rank-1 matching rate is less than 10%; as λ'



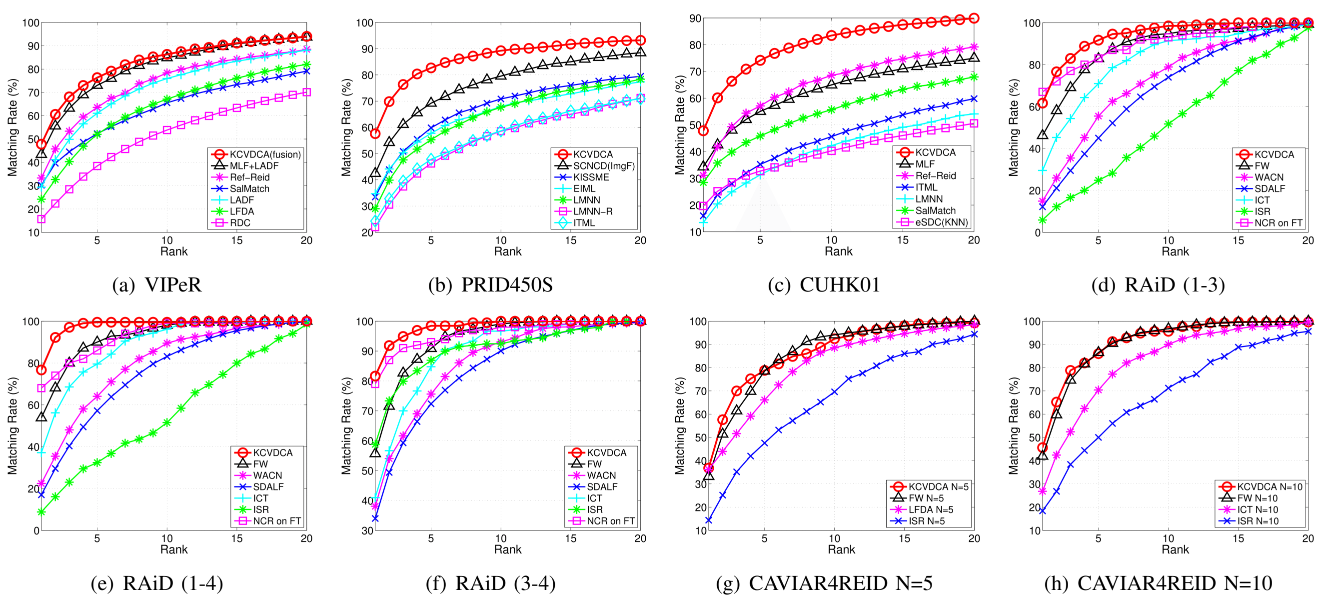
(a) VIPeR (b) PRID450S (c) CUHK01 (d) SYSU

(e) CAVIAR4REID (f) RAiD(1-3) (g) RAiD(1-4) (h) RAiD(3-4)

Fig. 5. TTR-FTR curves on VIPeR, PRID450S, CUHK01, SYSU, CAVIAR4REID, RAiD(1-3), RAiD(1-4) and RAiD(3-4). Better viewed in color.

TABLE II
TOP RANKED MATCHING RATE (%) ON RAID(1-3), RAID(1-4) AND RAID(3-4).

camera pair	pair 1-3				pair 1-4				pair 3-4			
rank	1	5	10	20	1	5	10	20	1	5	10	20
KCVDCA	61.64	**91.69**	**98.50**	**100.00**	**76.69**	**99.50**	**99.50**	**100.00**	**81.63**	**98.45**	**100.00**	**100.00**
NCR on FT [51]	**67.00**	83.00	93.00	100.00	68.00	86.00	99.00	100.00	79.00	93.00	98.00	100.00
FW [24]	46.17	82.86	94.76	99.25	53.81	90.00	98.10	100.00	55.67	90.87	99.12	100.00
WACN [56]	14.89	55.46	78.89	99.28	22.40	64.07	89.48	99.88	38.07	75.62	93.07	99.50
SDALF [57]	12.19	44.99	73.95	99.07	16.99	57.22	83.17	99.60	33.99	72.36	90.07	99.77
ICT [58]	29.52	70.95	91.43	99.05	37.14	79.52	96.19	100.00	40.95	84.76	96.67	100.00
ISR [30]	5.88	24.83	51.71	97.62	8.81	32.40	51.55	98.57	58.79	86.92	92.45	95.25



(a) VIPeR (b) PRID450S (c) CUHK01 (d) RAiD (1-3)

(e) RAiD (1-4) (f) RAiD (3-4) (g) CAVIAR4REID N=5 (h) CAVIAR4REID N=10

Fig. 6. Comparison to the state-of-the-art on VIPeR, PRID450S CUHK01, CAVIAR4REID and RAiD. Specifically, Fig 6(d)-6(f) refer to the experiments conducted on camera pairs 1-3, 1-4 and 3-4 of RAiD respectively. Fig. 6(g) and Fig. 6(h) are multi-shot experiments conducted on CAVIAR4REID with 5 and 10 images per pedestrian respectively. Best viewed in color.

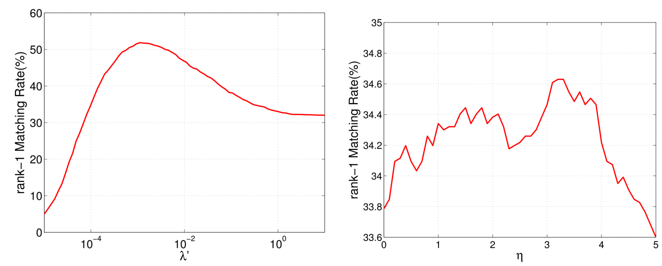
increases, the rank-1 matching rate increases simultaneously until it reaches the maximum which is larger than 50%. If the penalty term is set too large, it then tends to ignore the feature discrepancy across views and thus the performance

TABLE VI
TOP RANKED MATCHING RATE (%) ON PRID450S COMPARED TO THE STATE-OF-THE-ART.

rank	1	5	10	20
KCVDCA	**57.60**	**82.67**	**89.24**	**93.20**
SCNCD(ImgF) [4]	42.44	69.22	79.56	88.44
KISSME [17]	33.47	59.82	70.84	79.47
EIML [19]	34.71	57.73	67.91	77.33
LMNN [5]	28.98	55.29	67.64	78.36
LMNN-R [20]	21.96	46.22	58.53	71.20
ITML [7]	24.27	47.82	58.67	70.89

TABLE VII
TOP RANKED MATCHING RATE (%) ON CUHK01 COMPARED TO THE STATE-OF-THE-ART.

rank	1	5	10	20
KCVDCA	**47.80**	**74.16**	**83.44**	**89.92**
MLF [2]	34.30	55.06	64.96	74.94
Ref-Reid [40]	31.10	57.10	68.55	79.18
ITML [7]	15.98	35.22	45.60	59.81
LMNN [5]	13.45	31.33	42.25	54.11
SalMatch [22]	28.45	45.85	55.67	67.95
eSDC(KNN) [3]	19.67	32.72	40.29	50.58



(a) λ' (b) η

Fig. 7. Parameter analysis of CVDCA. (a) shows the parameter of cross-view consistency regularization; (b) shows the parameter of intra-view modeling.

drops. Note that when this penalty term is infinitely large, the view-specific mappings would be the same and asymmetric distance learning degrades to symmetric distance learning. Hence, the experimental results validate our analysis that a proper cross-view consistency regularization is critical for asymmetric distance learning.

2) Effectiveness of Intra-View Modeling: Our model Eq. (12) consists of both cross-view and intra-view modeling. We argue that the cross-view modeling plays a major role for the person re-identification problem and the intra-view modeling may have relative limited effectiveness to the cross-view matching problem. Fig. 7(b) shows the weight of intra-view modeling η versus the rank-1 matching rate on CUHK01 dataset and similar conclusion can be drawn on the others. Note that when $\eta = 0$, the intra-view modeling part is removed and only the cross-view modeling contributes to the performance, and the rank-1 matching rate is 33.8%. Tuning η does boost the performance and the maximum rank-1 matching rate is 34.6%, which indicates that the intra-view modeling is useful, albeit limited.

Note that the widely used testing protocol for person re-identification is to match pedestrians *across* camera views,

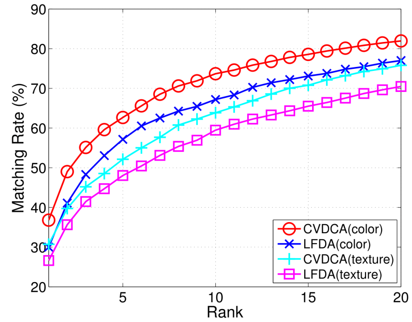


Fig. 8. Performance using different types of features.

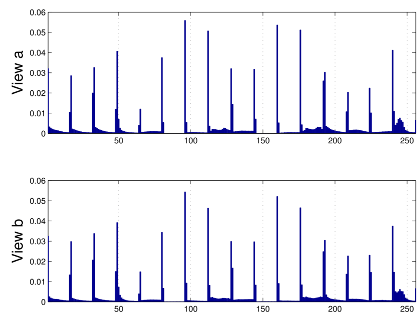


Fig. 9. Illustration of Gabor features of PRID450S. Similar with Fig. 3, the first row is the Gabor feature distributions of $\frac{1}{n_p^{a,b}} \sum_i \boldsymbol{x}_i^a$ and the second row is that of $\frac{1}{n_p^{a,b}} \sum_j \boldsymbol{x}_j^b$.

while the intra-view matching is not the concern. Hence it is reasonable that the cross-view modeling plays more important part in the modeling. The intra-view modeling, to an extent, is related to the matching, and thus incorporating it to the model could help tackling the cross-view matching problem.

3) A Brief Analysis of the Extracted Features: Throughout the experiment section, we have shown that our asymmetric distance performs better than the symmetric ones. Our explanation is that by using different mappings for different camera views, we can extract more discriminative features, even if those features are divergent for each view. On the contrary, using an unitary mapping for all views would have to discard some of the discriminative features if they are divergent.

Taking PRID450S as an example, we give an analysis on the difference of features extracted by asymmetric distance learning method (CVDCA) and symmetric distance learning method (LFDA). We semantically divided the features into color features (including RGB, HSV, YCbCr, Lab and YIQ features) and texture features (the Gabor features). As shown in Fig. 8, color features are more discriminative than texture features, as using color features results in better performance than using texture features. However, as shown in Fig. 3, color features on this dataset are very different across the two camera views, while in Fig. 9 we can see that texture features are much more consistent across the two views.

We trained CVDCA and LFDA with the color+texture features and obtained the projection bases $\boldsymbol{U}^a$ and $\boldsymbol{U}^b$ (for LFDA, $\boldsymbol{U}^a = \boldsymbol{U}^b$). Then we calculated the energy for each

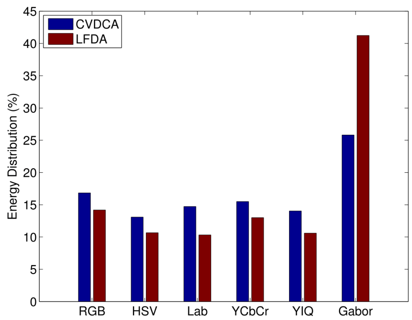


Fig. 10. Energy distribution of different types of features. Energy is defined as Eq. (27). We compare the percentage of energy of different feature types between CVDCA and LFDA.

type of features as follows:

$$E(f) = \sum_{k} \sum_{j \in B_f} (\boldsymbol{U}^a(k,j)^2 + \boldsymbol{U}^b(k,j)^2), \tag{27}$$

where $f \in \{RGB, HSV, YCbCr, Lab, YIQ, Gabor\}$ indicates the set of feature types, B_f is the set of indices of feature type f, and $\boldsymbol{U}^a(k,j)$ is the k-th column and j-th row of matrix $\boldsymbol{U}^a$. Fig. 10 shows the energy distribution of different feature types of CVDCA and LFDA. It shows that CVDCA allots more energy to the color features, while LFDA allots more to the texture features. Recall that color features are more discriminative but divergent, while texture features are less discriminative but more divergent. Therefore, we conclude that more color features, which is shown to be more discriminative, are preserved by our approach.

V. Conclusion

In this paper, we address the feature discrepancy problem across non-overlapping camera views for person re-identification. A cross-view discriminant component analysis method which forms an asymmetric distance model for matching person images between disjoint camera views by learning view-specific mappings is proposed to overcome this problem. To model the correlation nature of feature transformations of different views, a cross-view consistency regularization is introduced in our model. Experimental results demonstrate that 1) asymmetric distance model performs notably better than symmetric ones; 2) the influence of feature discrepancy can be effectively alleviated by view specific modeling.

In this work, we use Euclidean distance as the measure of the discrepancy of different mappings in the cross-view consistency regularization, which implicitly assumes Gaussian distribution for the projection matrices. However, how to relax such an assumption remains a future issue to investigate. In our future works, we would like to investigate other examples of the Bregman distance, which could work better for more general distributions in the exponential families.

Acknowledgments

The authors are most grateful for the constructive advice on the revision of the manuscript from the anonymous reviewers.

References

[1] C.-C. Guo, S.-Z. Chen, J.-H. Lai, X.-J. Hu, and S.-C. Shi, "Multi-shot person re-identification with automatic ambiguity inference and removal," in *ICPR*, 2014.

[2] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identfiation," in *CVPR*, 2014.

[3] ——, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.

[4] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *ECCV*, 2014.

[5] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *NIPS*, 2006.

[6] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012.

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *ICML*, 2007.

[8] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE TPAMI*, vol. 35, no. 3, pp. 653–668, 2013.

[9] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013.

[10] F. Porikli, "Inter-camera color calibration by correlation model function," in *ICIP*, 2003.

[11] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions." in *BMVC*, 2008.

[12] C. Siebler, K. Bernardin, and R. Stiefelhagen, "Adaptive color transformation for person re-identification in camera networks," in *ICDSC*, 2010.

[13] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: what features are important?" in *ECCV Workshops*, 2012.

[14] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *NIPS*, 2002, pp. 505–512.

[15] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. Salakhutdinov, "Neighbourhood components analysis," in *NIPS*, 2004, pp. 513–520.

[16] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *JMLR*, vol. 8, pp. 1027–1061, 2007.

[17] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.

[18] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *CVPR*, 2013.

[19] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *AVSS*, 2012.

[20] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *ACCV*, 2011.

[21] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE TPAMI*, 2007.

[22] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *ICCV*, 2013.

[23] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE TPAMI*, vol. 35, no. 7, pp. 1622–1634, 2013.

[24] N. Martinel, A. Das, C. Micheloni, and A. Roy-Chowdhury, "Re-identification in the function space of feature warps," *PAMI*, 2015.

[25] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification."

[26] D. Tao, L. Jin, Y. Wang, and X. Li, "Person reidentification by minimum classification error-based kiss metric learning," *IEEE Transactions on Cybernetics*, 2015.

[27] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[28] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *ECCV*, 2014.

[29] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," 2015.

[30] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *PAMI*, 2014.

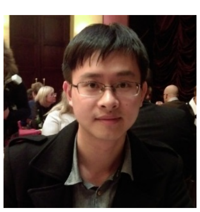
[31] D.-N. T. Cong, C. Achard, and L. Khoudour, "People re-identification by classification of silhouettes based on sparse representation," in *IPTA*. IEEE, 2010.

[32] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE TNN*, 2011.

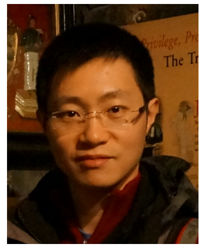
[33] B. Geng, D. Tao, and C. Xu, "Daml: Domain adaptation metric learning," *IEEE TIP*, 2011.

[34] M. Xiao and Y. Guo, "Feature space independent semi-supervised domain adaptation via kernel matching," *IEEE TPAMI*, 2014.

[35] W.-S. Zheng, S. Gong, and T. Xiang, “Transfer re-identification: From person to set-based verification,” *CVPR*, 2012.
[36] A. J. Ma, P. C. Yuen, and J. Li, “Domain transfer support vector ranking for person re-identification without target camera label information,” in *ICCV*, 2013.
[37] D. T. Lianyang Ma, Xiaokang Yang, “Person re-identification over camera networks using multi-task distance metric learning,” *IEEE TIP*, 2014.
[38] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof, “Mahalanobis distance learning for person re-identification,” in *Person Re-Identification*, S. Gong, M. Cristani, S. Yan, and C. C. Loy, Eds. Springer, 2014, pp. 247–267.
[39] L. An, M. Kafai, S. Yang, and B. Bhanu, “Reference-based person re-identification,” in *AVSS*, 2013.
[40] ——, “Person re-identification with reference descriptor,” *IEEE TCSVT*.
[41] L. An, S. Yang, and B. Bhanu, “Person re-identification by robust canonical correlation analysis,” *Signal Processing Letters, IEEE*, vol. 22, no. 8, pp. 1103–1107, 2015.
[42] K. Liu, Z. Zhao, and A. Cai, “Datum-adaptive local metric learning for person re-identification,” *SPL, IEEE*, 2015.
[43] W. Li and X. Wang, “Locally aligned feature transforms across views,” in *CVPR*, 2013.
[44] Y.-C. Chen, W.-S. Zheng, and J. Lai, “Mirror representation for modeling view-specific transform in person re-identification,” in *Proc. IJCAI*, 2015, pp. 3402–3408.
[45] A. Mignon and F. Jurie, “Cmml: a new metric learning approach for cross modal matching,” in *ACCV*, 2012.
[46] J. B. HH Bauschke, “Joint and separate convexity of the bregman distance,” *Inherently parallel algorithms in feasibility and optimization and their applications*, 2001.
[47] S. Si, D. Tao, and B. Geng, “Bregman divergence-based regularization for transfer subspace learning,” *IEEE TKDE*, 2010.
[48] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” in *VS-PETS Workshop*. Citeseer, 2007.
[49] W. Li, R. Zhao, and X. Wang, “Human reidentification with transferred metric learning,” in *ACCV*, 2012.
[50] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, “Custom pictorial structures for re-identification.” in *BMVC*. Citeseer, 2011.
[51] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, “Consistent re-identification in a camera network,” in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 330–345.
[52] W.-S. Zheng, S. Gong, and T. Xiang, “Associating groups of people,” in *BMVC*, 2009.
[53] W. Schwartz and L. Davis, “Learning Discriminative Appearance-Based Models Using Partial Least Squares,” in *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009.
[54] T. De Bie and B. De Moor, “On the regularization of canonical correlation analysis,” *Int. Sympos. ICA and BSS*, pp. 785–790, 2003.
[55] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang, “Hierarchical gaussianization for image classification,” in *ICCV*, 2009.
[56] N. Martinel and C. Micheloni, “Re-identify people in wide area camera network,” in *CVPR Workshop*. IEEE, 2012.
[57] L. Bazzani, M. Cristani, and V. Murino, “Symmetry-driven accumulation of local features for human characterization and re-identification,” *CVIU*, 2013.
[58] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, “Learning implicit transfer for person re-identification,” in *ECCV Workshop*. Springer, 2012.



Ying-Cong Chen received his B.Sc degree from Sun Yat-sen University in 2013. Now he is a Master student in Sun Yat-sen University. His research interest includes computer vision and machine learning.
Homepage: http://isee.sysu.edu.cn/∼chenyingcong



Wei-Shi Zheng received the PhD degree in applied mathematics from Sun Yat-sen University in 2008. He is an associate professor in Sun Yat-sen University. He has been a postdoctoral researcher on the EU FP7 SAMURAI Project at Queen Mary University of London. He has now published more than 70 papers, including more than 40 publications in main journals (TPAMI,TNN,TIP,TSMC-B,PR) and top conferences (ICCV, CVPR, IJCAI, AAAI). He has joined the organisation of four tutorial presentations in ACCV 2012, ICPR 2012, ICCV 2013 and CVPR 2015 along with other colleagues. His research interests include person/object association and activity understanding in visual surveillance. He is a recipient of excellent young scientists fund of the national natural science foundation of China. He has joined Microsoft Research Asia Young Faculty Visiting Programme.
Homepage: http://isee.sysu.edu.cn/∼zhwshi

Jian-Huang Lai is Professor of School of Information Science and Technology in Sun Yat-sen university. He received his M.Sc. degree in applied mathematics in 1989 and his Ph.D. in mathematics in 1999 from SUN YAT-SEN University, China.. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, wavelet and its applications. He has published over 100 scientific papers in the international journals and conferences on image processing and pattern recognitione.g. IEEE TPAMI , IEEE TNN, IEEE TIP, IEEE TSMC (Part B), Pattern Recognition, ICCV, CVPR and ICDM.

Pong-Chi Yuen received his B.Sc. degree in Electronic Engineering with First Class Honours in 1989 from City Polytechnic of Hong Kong, and his Ph.D. degree in Electrical and Electronic Engineering in 1993 from The University of Hong Kong. He joined the Hong Kong Baptist University in 1993 and, currently is a Professor and Head of the Department of Computer Science. Dr. Yuen was a recipient of the University Fellowship to visit The University of Sydney in 1996. In 1998, Dr. Yuen spent a 6-month sabbatical leave in The University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland at college park. From June 2005 to January 2006, he was a visiting professor in GRAVIR laboratory (GRAphics, VIsion and Robotics) of INRIA Rhone Alpes, France. Dr. Yuen was the director of Croucher Advanced Study Institute (ASI) on biometric authentication in 2004 and the director of Croucher ASI on Biometric Security and Privacy in 2007.

Dr. Yuen has been actively involved in many international conferences as an organizing committee and/or technical program committee member. He was the track co-chair of International Conference on Pattern Recognition (ICPR) 2006 and the program co-chair of IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS) 2012. Currently, Dr. Yuen is an Editorial Board Member of Pattern Recognition and Associate Editor of IEEE Transactions on Information Forensics and Security, and SPIE Journal of Electronic Imaging. He is also serving as a Hong Kong Research Grant Council Engineering Panel Member.

Dr. Yuen’s current research interests include video surveillance, human face recognition, biometric security and privacy.