Person Re-identification by Cross-View Multi-Level Dictionary Learning

Sheng Li, *Member, IEEE,* Ming Shao, *Member, IEEE,* and Yun Fu, *Senior Member, IEEE*

Abstract—Person re-identification plays an important role in many safety-critical applications. Existing works mainly focus on extracting patch-level features or learning distance metrics. However, the representation power of extracted features might be limited, due to the various viewing conditions of pedestrian images in complex real-world scenarios. To improve the representation power of features, we learn discriminative and robust representations via dictionary learning in this paper. First, we propose a Cross-view Dictionary Learning (CDL) model, which is a general solution to the multi-view learning problem. Inspired by the dictionary learning based domain adaptation, CDL learns a pair of dictionaries from two views. In particular, CDL adopts a projective learning strategy, which is more efficient than the l_1 optimization in traditional dictionary learning. Second, we propose a Cross-view Multi-level Dictionary Learning (CMDL) approach based on CDL. CMDL contains dictionary learning models at different representation levels, including image-level, horizontal part-level, and patch-level. The proposed models take advantages of the view-consistency information, and adaptively learn pairs of dictionaries to generate robust and compact representations for pedestrian images. Third, we incorporate a discriminative regularization term to CMDL, and propose a CMDL-Dis approach which learns pairs of discriminative dictionaries in image-level and part-level. We devise efficient optimization algorithms to solve the proposed models. Finally, a fusion strategy is utilized to generate the similarity scores for test images. Experiments on the public VIPeR, CUHK Campus, iLIDS, GRID and PRID450S datasets show that our approach achieves the state-of-the-art performance.

Index Terms—Dictionary learning, cross-view learning, multi-level representation, person re-identification.

1 Introduction

PERSON re-identification is the problem of matching pedestrian images observed from multiple non-overlapping cameras. It saves a lot of human efforts in many safety-critical applications such as video surveillance. In recent years, many research efforts have been focused on developing effective solutions to this problem [1], [2], [3], [4], [5], [6]. The representative person re-identification methods mainly include the distance learning/metric learning methods [7], [8], [9], [10], [11], [12], [13], feature learning methods [2], [14], [15], [16], [17], [18].

The distance learning methods aim to learn distance metrics that are expected to be robust to sample variations. For instance, a logistic metric learning approach with the positive semidefinite constraint is proposed to separate the positive sample pairs from the negative ones [19]. Other effective distance learning methods include the Probabilistic Relative Distance Comparison (PRDC) [8], Keep It Simple and Straightforward Metric Learning (KISSME) [20], etc. The feature learning methods extract discriminative features from pedestrian images, and then perform matching in the feature space. Some effective features include salient features [2], mid-level features [21], salient color features [22], polynomial kernel feature map [17], etc.. The advanced distance learning and feature learning methods have achieved promising performance on person re-identification. However, the representation power of the learned features or metrics might be limited, due to the various viewing conditions of pedestrian images in complex real-world scenarios (e.g., illumination changes and occlusions).

In this paper, we learn discriminative and robust representations via dictionary learning to improve the representation power of features. Our motivations are two-folds. First, dictionary learning is a powerful technique to extract effective and discriminative features from high-dimensional images, and it has shown impressive performance in many vision tasks, such as face recognition [23], which motivates us to design novel dictionary learning methods for person re-identification. Moreover, the success of dictionary learning based domain adaptation inspires us to learn a pair of cross-view dictionaries jointly [24]. The adaptively learned pairs of dictionaries can generate robust representations for pedestrian images. Second, existing works either focus on extracting features from image patches or directly learning global features. However, the complementary information resided in patch-level and image-level are usually ignored. We argue that extracting features from a single level is not sufficient, and it is necessary to design multi-level models, in order to make use of the complementary information.

To address the above problems, we design an effective Cross-view Dictionary Learning (CDL) model. CDL adopts a projective dictionary learning [25] strategy to reformulate the data encoding and reconstruction, and models the connections of samples across different views. The proposed CDL model is a general solution to the multi-view learning problem. Based on CDL, we propose a Cross-view Multi-level Dictionary Learning (CMDL) approach for person re-

- *S. Li is with Adobe Research, San Jose, CA, 95110.*
Email: shengli@ece.neu.edu
- *M. Shao is with the Department of Computer and Information Science, University of Massachusetts Dartmouth, Dartmouth, MA 02747.*
E-mail: mshao@umassd.edu
- *Y. Fu is with the Department of Electrical and Computer Engineering, College of Engineering, and College of Computer and Information Science, Northeastern University, Boston, MA, 02115.*
E-mail: yunfu@ece.neu.edu

Manuscript received March 1, 2016; revised July 13, 2017

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2764893, IEEE Transactions on Pattern Analysis and Machine Intelligence

identification. CMDL learns dictionaries in multiple levels, each level corresponding to image representation with a specific scale, such as different sizes of patches. Notice that most existing methods only consider feature learning in one single level [26]. Specifically, CMDL considers three levels of representation for pedestrian images, including the *image-level*, *horizontal part-level*, and *patch-level*. Three objectives are designed by deploying the CDL model at three different levels. In addition, we add a discriminative regularization function to CMDL in order to enhance the discriminability of the model, and present the CMDL-Dis approach. CMDL-Dis learns discriminative dictionary pairs for image-level and part-level representations.

By far, there are few methods proposed to learn effective representations for the pedestrian images using dictionary learning [26]. The basic assumption in [26] is that each pair of patches in two images shares the same representation coefficients. However, it is not the case in reality, due to the common misalignment problem in person re-identification.

This paper is a substantial extension of our previous work [27]. The major contributions of our work include:

- We propose a general solution named CDL for multi-view dictionary learning, and apply it to person re-identification. CDL extends the projective dictionary learning strategy [25] to multi-view setting, by providing a flexible way to model the relationships across different views.
- We propose to train pairs of dictionaries at different representative levels, and present the CMDL approach. CMDL explicitly models the cross-view interactions in image-level, horizontal part-level and patch-level. To the best of our knowledge, our work is the first attempt to learn representations at three levels for person re-identification.
- We propose the CMDL-Dis approach to train discriminative dictionaries for person re-identification. CMDL-Dis not only enforces the correspondence of the same person across two views, but also minimizes the inter-person correlations.
- We evaluate the performance of our approaches and baselines on five public datasets. Extensive results show that our approach outperforms the state-of-the-art methods.

The rest of the paper is organized as follows. In Section 2, we briefly review the related works and discuss how they differ from our approach. In Section 3, we introduce the model of CDL. Section 4 introduces the CMDL approach, and Section 5 presents the CMDL-Dis approach with solutions. The experimental results and discussions are reported in Section 6. Section 7 is the conclusion.

2 Related Work

In this section, we briefly introduce three research topics that are related to our approach, including person re-identification, dictionary learning, and multi-view learning.

2.1 Person Re-identification

Existing person re-identification methods can be roughly categorized into three groups. The first group is distance and metric learning methods, which focus on learning effective metrics to measure the similarity between two images captured from different camera views [8], [13], [28]. Some recent methods include the kernel based metric learning [12] and relevance metric learning with list-wise similarities [6]. The second group is feature learning method. Extracting expressive features is usually considered as a critical step in many visual learning tasks. By designing various feature extractors, feature learning methods obtain better performance than metric learning. Some effective features include attributes [29], salience features [2], [30], mid-level features [21], salient color features [22], polynomial kernel feature map [17], subject-discriminative features [18]. The third group is deep learning method. Li *et al.* presented a deep filter pairing neural network for person re-identification [4]. Ahmed *et al.* designed an improved deep learning architecture for person re-identification [5]. Zhang *et al.* proposed a deep hashing algorithm and achieved good performance [31]. However, by far these deep learning methods cannot significantly improve the performance on person re-identification, due to the limited size of available training data. In addition, other types of methods have also been introduced to address the person re-identification problem, such as transfer learning methods [32], [33], ranking methods [34], ensemble learning methods [35], post-ranking optimization methods [36], etc.

Although the existing person re-identification methods achieve good performance in certain scenarios, they usually adopt a single level of representation (e.g., patch-level), and the cross-view relationships of pedestrian images haven't been extensively studied. Our CMDL approach explicitly models such relationships in multiple representation levels, and trains discriminative dictionaries to extract expressive features for the re-identification task.

2.2 Dictionary Learning

As a powerful technique for learning expressive basis in sample space, dictionary learning has become an attractive research topic during the past decade [23], [37], [38], [39], [40]. Some popular dictionary learning methods include K-SVD [41], discriminative K-SVD [42], and projective dictionary pair learning [25].

Recently, dictionary learning methods have been applied to person re-identification. Liu *et al.* presented a semi-supervised coupled dictionary learning (SSCDL) method [26] for person re-identification. The major differences between our approach and SSCDL are three-folds. First, SSCDL is a semi-supervised method, while our approach is supervised. Second, SSCDL simply assumes that a pair of patches in two views should have similar codings, which is unreasonable in real scenario due to the misalignment problem. Our approach models the cross-view interactions in image-level, part-level and patch-level, respectively. Third, SSCDL requires solving the l_1 optimization problem that is time-consuming. Our approach adopts an efficient projective dictionary learning strategy. In addition, Jing *et al.* proposed a semi-coupled low-rank discriminant dictionary learning method for super-resolution person re-identification [43]. Our work differs from [43] in that, we employ projective dictionary learning in multi-view settings,

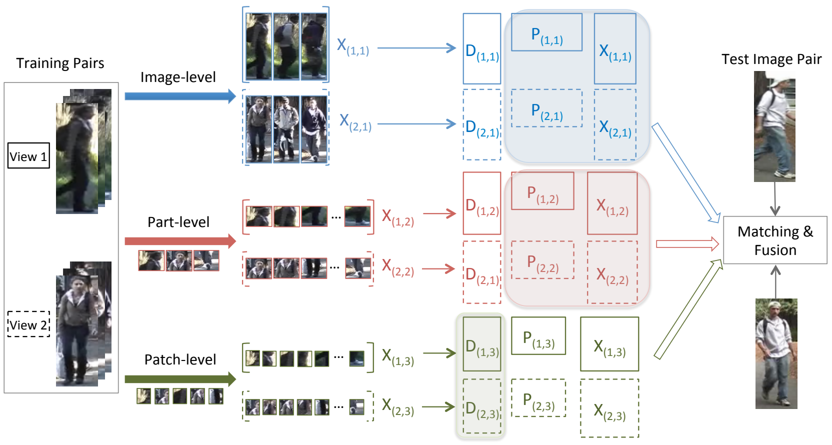


Fig. 1. Framework of our CMDL approach. Solid boxes represent variables related to view 1, while dashed boxes represent variables related to view 2. CMDL learns three pairs of dictionaries at three representation levels, and finally fuses the matching results. The shaded areas indicate view-consistency constraints. In particular, two views share similar codings (e.g., $P_{(1,1)}X_{(1,1)}$, $P_{(2,1)}X_{(2,1)}$) in the image-level and part-level, and share similar dictionary (i.e., $D_{(1,3)}$, $D_{(2,3)}$) in patch-level.

and take advantages of the complementary information in multiple representation levels.

2.3 Multi-View Learning

Multi-view learning has been receiving increasing attention in recent years [44], [45], [46], [47]. One implicit assumption is that either view alone has sufficient information about the samples, but the complexity of learning problems can be reduced by eliminating hypotheses from each view that tend not to agree with each other [48]. The representative multi-view learning methods include co-training [49], manifold co-regularization [50], multi-view subspace learning [51], [52], and multi-view feature learning [53]. The basic idea of these methods is to exploit the *consistency* among multiple views to enhance the learning performance. In this paper, we design a multi-view learning framework through dictionary learning, and apply it to person re-identification.

3 Cross-View Dictionary Learning (CDL)

3.1 Motivation

Dictionary learning aims to learn expressive feature representations for data, and has been widely applied in many visual learning tasks. Given a set of samples $X \in \mathbb{R}^{d \times n}$, traditional dictionary learning methods usually assume that X can be reconstructed by using sparse coefficients $Z \in \mathbb{R}^{m \times n}$ and a dictionary $D \in \mathbb{R}^{d \times m}$:

$$X \approx DZ, \tag{1}$$

where Z is usually constrained by l_1 norm that minimizes the sum of entries in Z.

Existing dictionary learning methods based on Eq. (1) have shown promising performance in many applications like image classification, but there are still some drawbacks. First, solving sparse coefficients Z in Eq. (1) is computationally expensive due to the l_1 norm constraint, which limits the applicability of dictionary learning for large-scale problems. Second, traditional dictionary learning methods mainly focus on single-view data, and therefore they cannot directly handle the multi-view visual data. Nowadays data can be collected from multiple views, and it is of great importance to develop multi-view dictionary learning methods. As we mainly focus on the person re-identification that is essentially a two-view problem, we only consider the two-view setting in this paper.

We aim to tackle the above problems by designing an efficient cross-view dictionary learning (CDL) model. Inspired by the idea of projective dictionary learning [25], we reduce the computational cost of dictionary learning by reformulating the approximation in Eq. (1) as a linear encoding and reconstruction process. Let $P \in \mathbb{R}^{m \times d}$ $(m \ll d)$ denote a low-dimensional projection matrix, Eq. (1) can be reformulated as $X \approx DPX$. Notice that PX denotes the linear encodings of sample set X. Moreover, we will consider the dictionary learning process in different views, and will model the view-consistency in our solution.

3.2 Formulation of CDL

We consider the dictionary learning problem in two-view setting. Let $X_{(1)} \in \mathbb{R}^{d \times n}$ and $X_{(2)} \in \mathbb{R}^{d \times n}$ denote two training sets collected from two different views, respectively. We formulate the encoding and reconstruction in two views as

$$\begin{aligned} X_{(1)} &= D_{(1)}P_{(1)}X_{(1)}, \\ X_{(2)} &= D_{(2)}P_{(2)}X_{(2)}, \end{aligned} \tag{2}$$

where $D_{(1)}$ and $D_{(2)}$ are dictionaries in two views; $P_{(1)}$ and $P_{(2)}$ are the corresponding linear projections in two views. In this way, the codings $P_{(v)}X_{(v)}$ can be obtained analytically, which improves the efficiency of dictionary learning.

Then we can present the objective function of CDL model as

$$\begin{aligned} \min_{\substack{D_{(1)},D_{(2)},\\ P_{(1)},P_{(2)}}} \quad & \|A_{(1)} - D_{(1)}P_{(1)}A_{(1)}\|_{\mathrm{F}}^2 + \|A_{(2)} - D_{(2)}P_{(2)}A_{(2)}\|_{\mathrm{F}}^2 \\ & + \lambda f(D_{(1)}, D_{(2)}, P_{(1)}, P_{(2)}) \\ s.t. \quad & \|d_{1(:,i)}\| \le 1, \ \|d_{2(:,i)}\| \le 1, \ i = 1, \cdots, m, \end{aligned} \tag{3}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2764893, IEEE Transactions on Pattern Analysis and Machine Intelligence

where $f(D_{(1)}, D_{(2)}, P_{(1)}, P_{(2)})$ is a regularization function used to model the view consistency, λ is a trade-off parameter, $\|\cdot\|_\mathrm{F}$ denotes the Frobenius norm, and $d_{1(:,i)}$ and $d_{2(:,i)}$ are the i-th columns in $D_{(1)}$ and $D_{(2)}$, respectively.

The first two terms in Eq. (3) indicate the approximation errors in two views, respectively. CDL aims to minimize the approximation errors in two views jointly. The last term $f(D_{(1)}, D_{(2)}, P_{(1)}, P_{(2)})$ is a regularization function that models the interactions of two views, which can be customized in specific scenarios. We will discuss how to formulate $f(D_{(1)}, D_{(2)}, P_{(1)}, P_{(2)})$ for person re-identification in the next section.

After training, the obtained optimal dictionary pair $\{D_{(1)}, D_{(2)}\}$ can be used to generate new representations (i.e., codings) for test samples.

3.3 Discussions

Design of Regularization Function $f(\cdot)$. The regularization function $f(D_{(1)}, D_{(2)}, P_{(1)}, P_{(2)})$ in Eq. (3) can be customized for specific problems, such as image classification or person re-identification. The function would be able to model the consistency of latent representations (or dictionaries) in two views. In the next section, we will propose to learn dictionaries in three different representation levels for person re-identification, and design regularization functions for each of them.

Comparisons with Related Dictionary Leaning Methods. Compared to the existing multi-view or multi-modal dictionary learning methods, our CDL model provides a more flexible way to learn dictionaries from two data views. The dictionary learning methods proposed in [24], [39] can be considered as special cases of the CDL model.

4 Cross-View Multi-Level Dictionary Learning (CMDL)

In this section, we apply the CDL model to person re-identification, and propose a cross-view multi-level dictionary learning (CMDL) approach. Figure 1 shows the framework.

4.1 Formulation of CMDL

Our goal is to learn robust representations for each pedestrian in different views[1] by virtue of dictionary learning. Person re-identification is a very challenging problem due to the large variations of the same person under different views, such as misalignment, pose and illumination changes. The state-of-the-art person re-identification methods are mainly patch-based models [19], [54], [55], which can partially solve the misalignment problem. The most relevant methods also learn dictionaries in the patch space for person re-identification [26]. However, we notice that learning models only from the patch space is not sufficient. Some useful information, such as the global structure in images, might be discarded when generating local patches. Therefore, we propose to learn dictionaries on multiple levels, and each level corresponds to image representation with a specific scale, such as different sizes of patches.

In particular, CMDL considers three levels of representation for pedestrian images, including the *image-level*, *horizontal part-level*, and *patch-level*. (1) The image-level representation treats the whole image as a single training sample, and each image is vectorized to a d-dimensional vector. (2) In order to achieve horizontal part-level representation, we divide an image (with overlapping) into several parts horizontally, and then vectorize each part accordingly. (3) The patch-level representation is achieved by dividing an image into multiple small patches with overlapping, which is widely used by existing methods and will be detailed in experiments.

The major benefits of adopting three different levels of representation are two-folds. First, three representation levels provide informative cues of pedestrian images at different scales, and capture both local and global characteristics of the pedestrian images, which will be useful for information fusion. Second, three levels of representations can jointly address the issues like misalignment and variations. As patch-level matching is robust to misalignment and pose changes, we aim to extract effective patch-level representations. In addition, the part-level and image-level matching can be helpful when there is severe misalignment across different views, because the part-level/image-level representations of the same person in different views will be very similar. The image-level representation ensures the exact correspondence across views, and the part-level is a trade-off between patch-level and image-level.

We extend the general solution CDL to CMDL, by customizing regularization functions at different representation levels. As we introduce another dimension (i.e., level) in CMDL, we first rewrite (3) and have the unified objective function of CMDL for level l,

$$\begin{aligned} \min_{\substack{D_{(1,l)}, D_{(2,l)}, \\ P_{(1,l)}, P_{(2,l)}}} \quad & \left\|A_{(1,l)} - D_{(1,l)} P_{(1,l)} A_{(1,l)}\right\|_\mathrm{F}^2 \\ & + \left\|A_{(2,l)} - D_{(2,l)} P_{(2,l)} A_{(2,l)}\right\|_\mathrm{F}^2 \\ & + \lambda_l f_l(D_{(1,l)}, D_{(2,l)}, P_{(1,l)}, P_{(2,l)}) \\ s.t. \quad & \left\|d_{1l(:,i)}\right\| \leq 1,\ \left\|d_{2l(:,i)}\right\| \leq 1,\ i = 1, \cdots, m, \end{aligned} \tag{4}$$

where $f_l(D_{(1,l)}, D_{(2,l)}, P_{(1,l)}, P_{(2,l)})$ is a regularization function customized for level l, and λ_l is the corresponding trade-off parameter.

Next, we separately introduce how to design regularization functions for three levels, and present the optimization algorithms accordingly.

4.2 CMDL for Image-Level and Part-Level Representations

Let $X_{(1,1)}$ and $X_{(2,1)}$ denote the training sets of high-dimensional image-level samples in two views, respectively. For the i-th training image in view 1, the dense features of all the patches are concatenated as a high-dimensional vector, which is the i-th column in $X_{(1,1)}$. Clearly, the corresponding columns in $X_{(1,1)}$ and $X_{(2,1)}$ should have similar codings, since they represent the same pedestrian. Hence,

[1] Gallery and probe sets can be considered as two views.

by defining the regularization function $f(\cdot)$ in Eq. (4), we have the following objective

$$\begin{aligned}\min_{\substack{D_{(1,1)},D_{(2,1)},\\ P_{(1,1)},P_{(2,1)}}} \quad & \left\|X_{(1,1)} - D_{(1,1)}P_{(1,1)}X_{(1,1)}\right\|_{\mathrm{F}}^2 \\ & + \left\|X_{(2,1)} - D_{(2,1)}P_{(2,1)}X_{(2,1)}\right\|_{\mathrm{F}}^2 \\ & +\lambda_1 \left\|P_{(1,1)}X_{(1,1)} - P_{(2,1)}X_{(2,1)}\right\|_{\mathrm{F}}^2, \\ s.t. \quad & \|d_{11(:,i)}\| \le 1,\ \|d_{21(:,i)}\| \le 1,\ i = 1,\cdots,m,\end{aligned} \tag{5}$$

where $D_{(1,1)}$ and $D_{(2,1)}$ denote the dictionaries in two views; $P_{(1,1)}$ and $P_{(2,1)}$ denote the linear projection matrices in two views.

The regularization function defined in Eq. (5) is $\left\|P_{(1,1)}X_{(1,1)} - P_{(2,1)}X_{(2,1)}\right\|_{\mathrm{F}}^2$, indicating that the codings in two views should be as close as possible. In this way, the learned dictionaries $D_{(1,1)}$ and $D_{(2,1)}$ are expected to generate similar codings for the same pedestrian under two views.

Optimization

To facilitate the optimization of Eq. (5), we first add two relaxation variables $A_{(1,1)}$ and $A_{(2,1)}$, and rewrite the objective as

$$\begin{aligned}\min_{\substack{D_{(1,1)},D_{(2,1)},P_{(1,1)},\\ P_{(2,1)},A_{(1,1)},A_{(2,1)}}} \quad & \left\|X_{(1,1)} - D_{(1,1)}A_{(1,1)}\right\|_{\mathrm{F}}^2 \\ & + \left\|X_{(2,1)} - D_{(2,1)}A_{(2,1)}\right\|_{\mathrm{F}}^2 \\ & +\alpha(\left\|P_{(1,1)}X_{(1,1)} - A_{(1,1)}\right\|_{\mathrm{F}}^2 \\ & + \left\|P_{(2,1)}X_{(2,1)} - A_{(2,1)}\right\|_{\mathrm{F}}^2) \\ & +\lambda_1 \left\|A_{(1,1)} - A_{(2,1)}\right\|_{\mathrm{F}}^2, \\ s.t. \quad & \|d_{11(:,i)}\| \le 1, \|d_{21(:,i)}\| \le 1,\end{aligned} \tag{6}$$

where α is a balance parameter.

Although there are many variables in (6), we can alternatively optimize these variables as follows.

1). Fix other variables and update $A_{(1,1)}$ and $A_{(2,1)}$.

By ignoring the irrelevant variables with respect to $A_{(1,1)}$, the objective (6) is reduced to

$$\begin{aligned}\min_{A_{(1,1)}} \quad & J(A_{(1,1)}) = \left\|X_{(1,1)} - D_{(1,1)}A_{(1,1)}\right\|_{\mathrm{F}}^2 \\ & +\alpha \left\|P_{(1,1)}X_{(1,1)} - A_{(1,1)}\right\|_{\mathrm{F}}^2 \\ & +\lambda_1 \left\|A_{(1,1)} - A_{(2,1)}\right\|_{\mathrm{F}}^2.\end{aligned} \tag{7}$$

Setting $\frac{\partial J(A_{(1,1)})}{\partial A_{(1,1)}} = 0$, we get the solution

$$\begin{aligned}A_{(1,1)} = \quad & (D_{(1,1)}^{\mathrm{T}}D_{(1,1)} + (\alpha+\lambda_1)\mathrm{I})^{-1} \\ & (D_{(1,1)}^{\mathrm{T}}X_{(1,1)} + \lambda_1 A_{(2,1)} + \alpha P_{(1,1)}X_{(1,1)}),\end{aligned} \tag{8}$$

where I is an identity matrix.

Similarly, we can obtain solution to $A_{(2,1)}$ as

$$\begin{aligned}A_{(2,1)} = \quad & (D_{(2,1)}^{\mathrm{T}}D_{(2,1)} + (\alpha+\lambda_1)\mathrm{I})^{-1} \\ & (D_{(2,1)}^{\mathrm{T}}X_{(2,1)} + \lambda_1 A_{(1,1)} + \alpha P_{(2,1)}X_{(2,1)}).\end{aligned} \tag{9}$$

2). Fix other variables and update $P_{(1,1)}$ and $P_{(2,1)}$.

The objective function regarding $P_{(1,1)}$ can be written as

$$\min_{P_{(1,1)}} \left\|P_{(1,1)}X_1 - A_{(1,1)}\right\|_{\mathrm{F}}^2. \tag{10}$$

By setting the derivative with respect to $P_{(1,1)}$ to zero, we have the solution $P_{(1,1)} = A_{(1,1)}X_{(1,1)}(X_{(1,1)}X_{(1,1)}^{\mathrm{T}} + \eta\mathrm{I})^{-1}$, where η is a regularization parameter to avoid the singularity problem. Similarly, the solution to $P_{(2,1)}$ is: $P_{(2,1)} = A_{(2,1)}X_{(2,1)}(X_{(2,1)}X_{(2,1)}^{\mathrm{T}} + \eta\mathrm{I})^{-1}$.

3). Fix other variables and update $D_{(1,1)}$ and $D_{(2,1)}$.

By removing the irrelevant terms in (6), we can write the objective function regarding $D_{(1,1)}$ as

$$\begin{aligned}\min_{D_{(1,1)}} \quad & \left\|X_1 - D_{(1,1)}A_{(1,1)}\right\|_{\mathrm{F}}^2 \\ s.t. \quad & \|d_{11(:,i)}\| \le 1,\ i = 1,\cdots,m.\end{aligned} \tag{11}$$

Problem (11) can be effectively solved using an ADMM algorithm as introduced in [25]. We have similar solutions to $D_{(2,1)}$.

The above procedures are repeated until convergence. Finally, we obtain a pair of dictionaries $\{D_{(1,1)}, D_{(2,1)}\}$ that are used to represent high-dimensional image features.

The objective function and optimization algorithm for **part-level representation** are very similar to those of image-level representation. We skip the details due to the space limit.

Discussions

The time complexities of updating variables A, P and D are $O(mdn+m^3+m^2n)$, $O(mdn+d^3+d^2n)$, and $O(t(mdn+m^3+m^2d+d^2m))$, respectively, where t is the number of iterations. In practice, t is a small number as the algorithm converges quickly. The dictionary size m is usually much less than the sample size n and the dimensionality d. Thus, our algorithm is efficient in real-world applications.

The objective in Eq. (6) is a bi-convex problem for variables $\{(D_{v,1}, P_{v,1}), A_{v,1})\}$, $v = 1, 2$. When $D_{v,1}$ and $P_{v,1}$ are fixed, the objective function is convex for $A_{v,1}$. When $A_{v,1}$ is fixed, the objective function is convex for $D_{v,1}$ and $P_{v,1}$. The convergence property of such problems has been extensively studied in [56]. In addition, our optimization algorithm converges quickly in the experiments.

4.3 CMDL for Patch-Level Representation

In addition to modeling the image-level representation in Eq. (5) and part-level representation, we also consider the dictionary learning in patch-level representations. Let $X_{(1,3)}$ and $X_{(2,3)}$ denote the training sets of low-dimensional patch features in two views, respectively. In this case, we cannot simply assume that the codings in two views are close to each other. In reality, the i-th patch in view 1 may not match the i-th patch in view 2 due to the misalignment problem under cross-view settings.

One reasonable assumption for patch-level representation is that the patches in different views could share a similar dictionary. Therefore, the objective function is

$$\begin{aligned}\min_{\substack{D_{(1,3)},D_{(2,3)},\\ P_{(1,3)},P_{(2,3)}}} \quad & \left\|X_{(1,3)} - D_{(1,3)}P_{(1,3)}X_{(1,3)}\right\|_{\mathrm{F}}^2 \\ & + \left\|X_{(2,3)} - D_{(2,3)}P_{(2,3)}X_{(2,3)}\right\|_{\mathrm{F}}^2 \\ & +\lambda_3 \left\|D_{(1,3)} - D_{(2,3)}\right\|_{\mathrm{F}}^2, \\ s.t. \quad & \|d_{13(:,i)}\| \le 1, \|d_{23(:,i)}\| \le 1, i = 1,\cdots,m,\end{aligned} \tag{12}$$

in which the last term emphasizes the similarity of two dictionaries. In this model, we assume that the patch-level dictionaries in two views are very similar to each other. In practice, two different images may still share a lot of similar patches. Thus, it is reasonable to assume that two dictionaries contain similar bases vectors in the image patch space. Another reasonable assumption is that two dictionaries are exactly the same, i.e., $D_{(1,3)} = D_{(2,3)}$. This constraint will reduce the model complexity. And we actually observed very good performance in the experiments by using this simplified model. Without loss of generality, we present the optimization algorithm for the complete model below.

Optimization

To solve the problem Eq. (12), we first reformulate the objective as

$$\begin{aligned} \min_{\substack{D_{(1,3)},D_{(2,3)},P_{(1,3)},\\ P_{(2,3)},A_{(1,3)},A_{(2,3)}}} \quad & \left\|X_{(1,3)} - D_{(1,3)}A_{(1,3)}X_{(1,3)}\right\|_{\mathrm{F}}^2 \\ & + \left\|X_{(2,3)} - D_{(2,3)}A_{(2,3)}X_{(2,3)}\right\|_{\mathrm{F}}^2 \\ & + \beta(\left\|P_{(1,3)}X_{(1,3)} - A_{(1,3)}\right\|_{\mathrm{F}}^2 \\ & + \left\|P_{(2,3)}X_{(2,3)} - A_{(2,3)}\right\|_{\mathrm{F}}^2) \\ & + \lambda_3 \left\|D_{(1,3)} - D_{(2,3)}\right\|_{\mathrm{F}}^2, \\ s.t. \quad & \|d_{13(:,i)}\| \leq 1,\ \|d_{23(:,i)}\| \leq 1 \end{aligned} \tag{13}$$

where β is a balance parameter.

We alternatively update the variables in (13), and obtain the sub-problems (with solutions) as follows.

1). Fix other variables and update $A_{(1,3)}$ and $A_{(2,3)}$.

The sub-problem with respect to $A_{(1,3)}$ is

$$\min_{A_{(1,3)}} \left\|X_{(1,3)} - D_{(1,3)}A_{(1,3)}\right\|_{\mathrm{F}}^2 + \beta \left\|P_{(1,3)}X_{(1,3)} - A_{(1,3)}\right\|_{\mathrm{F}}^2. \tag{14}$$

We set the derivate with respect to $A_{(1,3)}$ to zero, and obtain the solution

$$A_{(1,3)} = (D_{(1,3)}^{\mathrm{T}}D_{(1,3)} + \beta \mathrm{I})^{-1}(D_{(1,3)}^{\mathrm{T}}X_{(1,3)} + \beta P_{(1,3)}X_{(1,3)}). \tag{15}$$

Similarly, the solution to $A_{(2,3)}$ is

$$A_{(2,3)} = (D_{(2,3)}^{\mathrm{T}}D_{(2,3)} + \beta \mathrm{I})^{-1}(D_{(2,3)}^{\mathrm{T}}X_{(2,3)} + \beta P_{(2,3)}X_{(2,3)}). \tag{16}$$

2). Fix other variables and update $P_{(1,3)}$ and $P_{(2,3)}$.

The sub-problem with respect to $P_{(1,3)}$ is

$$\min_{P_{(1,3)}} \quad \left\|P_{(1,3)}X_{(1,3)} - A_{(1,3)}\right\|_{\mathrm{F}}^2. \tag{17}$$

The optimal solution is $P_{(1,3)} = A_{(1,3)}X_{(1,3)}(X_{(1,3)}X_{(1,3)}^{\mathrm{T}} + \eta \mathrm{I})^{-1}$, where η is a small positive value used to avoid the singularity problem. Similarly, the solution to $P_{(1,3)}$ is: $P_{(2,3)} = A_{(2,3)}X_{(2,3)}(X_{(2,3)}X_{(2,3)}^{\mathrm{T}} + \eta \mathrm{I})^{-1}$.

3). Fix other variables and update $D_{(1,3)}$ and $D_{(2,3)}$.

The sub-problem with respect to $D_{(1,3)}$ is

$$\begin{aligned} \min_{D_{(1,3)}} \quad & \left\|X_{(1,3)} - D_{(1,3)}A_{(1,3)}\right\|_{\mathrm{F}}^2 + \lambda_3 \left\|D_{(1,3)} - D_{(2,3)}\right\|_{\mathrm{F}}^2, \\ s.t. \quad & \|d_{13(:,i)}\| \leq 1,\ i = 1, \cdots, m, \end{aligned} \tag{18}$$

We have similar solutions to $A_{(2,3)}$, $P_{(2,3)}$ and $D_{(2,3)}$. The above procedures are repeated until convergence. We finally obtain a pair of optimal dictionaries $\{D_{(1,3)}, D_{(2,3)}\}$ that are used to reconstruct low-dimensional patch features.

4.4 Matching and Fusion

With the learned two pairs of dictionaries, $\{D_{(1,1)}, D_{(2,1)}\}$, $\{D_{(1,2)}, D_{(2,2)}\}$, and $\{D_{(1,3)}, D_{(2,3)}\}$, we can obtain robust representations for the test images in two views by using algorithms like orthogonal matching pursuit. Then we perform the following matching and fusion strategy.

In person re-identification, we need to match a probe image to a set of gallery images. As our approach jointly learns the dictionaries in three different representative levels, we propose a fusion strategy to take full advantages of the robust representations.

Image-level Matching. The image-level matching between the probe image and gallery images is straightforward. As we have already attained the compact representations for each image, we can directly compare the similarity of different images using these new representations. Particularly, the representation coefficients are calculated using the dictionaries $\{D_{(1,1)}, D_{(2,1)}\}$ for each pair of images. We adopt the Cosine function to compute the similarity score $Score_{(1)}(i)$ between the probe image and the i-th gallery image.

Part-level Matching. The part-level matching estimates the similarity of probe and gallery images by evaluating the corresponding horizontal parts of these images. For every horizontal part in each pair of images, we first compute their part-level representation coefficients using the dictionary pair $\{D_{(1,2)}, D_{(2,2)}\}$. Then the coefficients belong to the same probe (or gallery) image are concatenated into a single vector that is the new representation for the image. After that, we can easily use the Cosine function to obtain the similarity score $Score_{(2)}(i)$ between the probe image and the i-th gallery image.

Patch-level Matching. The patch matching methods have been extensively studied in existing works [2], [21]. In order to deal with the misalignment problem, we adopt a constrained patch matching strategy. Particularlly, we search the spatial neighbors of the targeted patch in the gallery images, and calculate the distances between each pair of images. Finally, we can estimate the similarity between a probe image and every gallery image. Instead of comparing the original patches, we match the representation coefficients over the dictionaries $\{D_{(1,3)}, D_{(2,3)}\}$ for each pair of patches. The similarity score $Score_{(3)}(i)$ between the probe image and the i-th gallery image is generated from the similarities between these patches. In detail, we compute the pairwise similarity of patches within horizontal strips, and then compute the sum of patch-wise similarities as the final similarity score.

Fusion. We first normalize each of the three similarity score vectors $Score_{(1)}$, $Score_{(2)}$ and $Score_{(3)}$, and utilize a simple strategy to perform score fusion:

$$Score(i) = Score_{(1)}(i) + \gamma_1 Score_{(2)}(i) + \gamma_2 Score_{(3)}(i), \tag{19}$$

where γ_1 and γ_2 are two user-defined trade-off parameters.

The complete procedures of CMDL is summarized in *Algorithm 1*.

Algorithm 1. CMDL for Person Re-identification

```
Input: Training images in two views $X_1$, $X_2$,
       test images $T_1, T_2$,
       parameters $\lambda_l$, $\alpha$, $\beta$.
Output: Matching scores vector $Score$ .
Training
1: Extract dense features from $X_1, X_2$, and construct
   feature sets $X_{(1,l)}, X_{(2,l)}$, where $l = 1, \cdots, 3$;
2: Learn dictionaries $\{D_{(1,1)}, D_{(2,1)}\}$ from image-level
   features $X_{(1,1)}, X_{(2,1)}$ (Section 4.2);
3: Learn dictionaries $\{D_{(1,2)}, D_{(2,2)}\}$ from horizontal
   part-level features $X_{(1,2)}, X_{(2,2)}$ (Section 4.2);
4: Learn dictionaries $\{D_{(1,3)}, D_{(2,3)}\}$ from patch-level
   features $X_{(1,3)}, X_{(2,3)}$ (Section 4.3);
Testing
5: Extract dense features from $T_1, T_2$, and construct
   feature sets $T_{(1,l)}, T_{(2,l)}$ $l = 1, \cdots, 3$;
6: Encode $T_{(1,1)}, T_{(2,1)}$ using $\{D_{(1,1)}, D_{(2,1)}\}$, and
   perform image-level matching (Section 4.4);
7: Encode $T_{(1,2)}, T_{(2,2)}$ using $\{D_{(1,2)}, D_{(2,2)}\}$, and
   perform part-level matching (Section 4.4);
8: Encode $T_{(1,3)}, T_{(2,3)}$ using $\{D_{(1,3)}, D_{(2,3)}\}$, and
   perform patch-level matching (Section 4.4);
9: Obtain matching score vector $Score$ by fusing
   matching results in three levels using (19).
```

5 CMDL WITH DISCRIMINATIVE REGULARIZATION (CMDL-DIS)

In this section, we improve the CMDL image-level and part-level representations by incorporating a discriminative regularizer. The optimization algorithm is also provided.

5.1 Motivation and Formulation

The CMDL approach described in Section 4 makes use of the correspondence information of samples across two views, and encourages the codings in two views to be as close as possible. Person re-identification is essentially a supervised task, and therefore it is critical to exploit more discriminative information from data, in order to enhance the recognition performance of the model.

Actually, it is inappropriate to incorporate discriminative information into patch-level representations due to the misalignment problem. Therefore, we design a discriminative dictionary learning model for horizontal part-level and image-level representations. In particular, we only show the model details for the image-level one due to the space limit.

Let $X_{(1,1)}$ and $X_{(2,1)}$ denote the training sets of high-dimensional image-level samples in two views, respectively. In CMDL, we assume that the corresponding columns in $X_{(1,1)}$ and $X_{(2,1)}$ should have similar codings, and therefore incorporate a constraint to enforce the similarity between coding matrices $P_{(1,1)}X_{(1,1)}$ and $P_{(2,1)}X_{(2,1)}$. Moreover, we add a discriminative regularization term $f_{dis}(\cdot)$ to further enhance the discriminability of the learned dictionaries. We define $f_{dis} = \mathrm{trace}(X_{(1,1)}^{\mathrm{T}} P_{(1,1)}^{\mathrm{T}} P_{(2,1)} X_{(2,1)} R)$, where R is a prior matrix of size $n \times n$. If there is only one image per person in each view (i.e., the single shot person re-identification), R is defined as

$$R_{ij} = \begin{cases} 0, & \text{if } i = j, \\ \frac{1}{n}, & \text{otherwise.} \end{cases} \tag{20}$$

The motivation of using f_{dis} is to minimize the inter-person correlations in the low-dimensional coding space. Notice that the i-th column in $P_{(2,1)}X_{(2,1)}R$ is the average of $n-1$ columns in $P_{(2,1)}X_{(2,1)}$. Therefore, the i-th diagonal element in matrix $X_{(1,1)}^{\mathrm{T}} P_{(1,1)}^{\mathrm{T}} P_{(2,1)} X_{(2,1)} R$ measures the correlation between the i-th person and all the other $n-1$ persons. In Eq. (20), we assign a constant weight for every off-diagonal element of the prior matrix R. Actually, an adaptive strategy (e.g., assign a few non-zero weights for the nearest neighbors of each sample) may lead to even better performance, but the model complexity will be increased.

By introducing the regularization term f_{dis}, we propose the CMDL with Discriminative regularization (CMDL-Dis). The objective function is

$$\begin{aligned} \min_{\substack{D_{(1,1)}, D_{(2,1)}, \\ P_{(1,1)}, P_{(2,1)}}} \quad & \left\| X_{(1,1)} - D_{(1,1)} P_{(1,1)} X_{(1,1)} \right\|_{\mathrm{F}}^2 \\ & + \left\| X_{(2,1)} - D_{(2,1)} P_{(2,1)} X_{(2,1)} \right\|_{\mathrm{F}}^2 \\ & + \lambda_1 \left\| P_{(1,1)} X_{(1,1)} - P_{(2,1)} X_{(2,1)} \right\|_{\mathrm{F}}^2 , \\ & + \lambda_0 \mathrm{trace}(X_{(1,1)}^{\mathrm{T}} P_{(1,1)}^{\mathrm{T}} P_{(2,1)} X_{(2,1)} R), \\ s.t. \quad & \| d_{11(:,i)} \| \leq 1, \ \| d_{21(:,i)} \| \leq 1, \ i = 1, \cdots, m, \end{aligned} \tag{21}$$

where λ_0 is a balance parameter which controls the contribution of the discriminative regularization term.

5.2 Optimization

By introducing the relaxation variables $A_{(1,1)}$ and $A_{(2,1)}$, we have

$$\begin{aligned} \min_{\substack{D_{(1,1)}, D_{(2,1)}, \\ P_{(1,1)}, P_{(2,1)}}} \quad & \left\| X_{(1,1)} - D_{(1,1)} A_{(1,1)} \right\|_{\mathrm{F}}^2 \\ & + \left\| X_{(2,1)} - D_{(2,1)} A_{(2,1)} \right\|_{\mathrm{F}}^2 \\ & + \alpha (\left\| P_{(1,1)} X_{(1,1)} - A_{(1,1)} \right\|_{\mathrm{F}}^2 \\ & + \left\| P_{(2,1)} X_{(2,1)} - A_{(2,1)} \right\|_{\mathrm{F}}^2) \\ & + \lambda_1 \left\| A_{(1,1)} - A_{(2,1)} \right\|_{\mathrm{F}}^2 \\ & + \lambda_0 \mathrm{trace}(A_{(1,1)}^{\mathrm{T}} A_{(2,1)} R), \\ s.t. \quad & \| d_{11(:,i)} \| \leq 1, \| d_{21(:,i)} \| \leq 1, \end{aligned} \tag{22}$$

Comparing Eq. (22) with Eq. (6), the difference is the new discriminative regularization term, which is only relevant to variables $A_{(1,1)}$ and $A_{(2,1)}$. Therefore, we need to modify the solution to $A_{(1,1)}$ and $A_{(2,1)}$, but we can still follow the optimization procedures described in Section 4.2 to update all the other variables.

We ignore the irrelevant variables with respect to $A_{(1,1)}$, and obtain the sub-problem

$$\begin{aligned} \min_{A_{(1,1)}} \quad & J(A_{(1,1)}) = \left\| X_{(1,1)} - D_{(1,1)} A_{(1,1)} \right\|_{\mathrm{F}}^2 \\ & + \alpha \left\| P_{(1,1)} X_{(1,1)} - A_{(1,1)} \right\|_{\mathrm{F}}^2 \\ & + \lambda_1 \left\| A_{(1,1)} - A_{(2,1)} \right\|_{\mathrm{F}}^2 + \lambda_0 \mathrm{trace}(A_{(1,1)}^{\mathrm{T}} A_{(2,1)} R). \end{aligned} \tag{23}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2764893, IEEE Transactions on Pattern Analysis and Machine Intelligence

Setting $\frac{\partial J(A_{(1,1)})}{\partial A_{(1,1)}} = 0$, we get the solution

$$\begin{aligned} A_{(1,1)} = \ & (D_{(1,1)}^{\mathrm{T}} D_{(1,1)} + (\alpha + \lambda_1)\mathrm{I})^{-1} \\ & (D_{(1,1)}^{\mathrm{T}} X_{(1,1)} + \lambda_1 A_{(2,1)} + \alpha P_{(1,1)} X_{(1,1)} - \lambda_0 A_{(2,1)} R). \end{aligned} \tag{24}$$

Similarly, the solution to $A_{(2,1)}$ is

$$\begin{aligned} A_{(2,1)} = \ & (D_{(2,1)}^{\mathrm{T}} D_{(2,1)} + (\alpha + \lambda_1)\mathrm{I})^{-1} \\ & (D_{(2,1)}^{\mathrm{T}} X_{(2,1)} + \lambda_1 A_{(1,1)} + \alpha P_{(2,1)} X_{(2,1)} - \lambda_0 A_{(1,1)} R^{\mathrm{T}}). \end{aligned} \tag{25}$$

After updating variables $A_{(1,1)}$ and $A_{(2,1)}$, we can follow Eq. (10) and Eq. (11) to update $\{P_{(1,1)}, P_{(2,1)}\}$ and $\{D_{(1,1)}, D_{(2,1)}\}$. We will also adopt similar procedures to learn discriminative dictionaries $\{D_{(1,2)}, D_{(2,2)}\}$ for patch-level representations. Finally, we can perform matching and fusion for person re-identification, as described in Section 4.4.

Comparing Eq. (6) and Eq. (22), we only add a trace function that is related to two variables $A_{(1,1)}^{\mathrm{T}}$ and $A_{(2,1)}^{\mathrm{T}}$. The newly added term is a convex function when updating either $A_{(1,1)}^{\mathrm{T}}$ or $A_{(2,1)}^{\mathrm{T}}$. The optimization algorithm of CMDL-Dis enjoys the convergence property of CMDL.

6 EXPERIMENTS

In this section, we compare our approaches[2] with some related methods on five benchmark datasets that are widely used to evaluate person re-identification algorithms, including VIPeR [57], CUHK01 Campus [21], iLIDS, GRID [58], and PRID 450S [59].

6.1 Settings

Feature Extraction. The pedestrian images in different views are not usually aligned well. Extracting dense features from local patches is a widely used strategy to obtain effective representations, as suggested in [21]. Specifically, the local patches are extracted on a dense grid. The size of each patch is 10×10, and the grid step is 5. Then, for each patch, we extract 32-dimensional color histogram features and 128-dimensional dense SIFT features in each LAB channel. Furthermore, we calculate the color histograms in different sampling scales with the downsampling factors 0.5 and 0.75. All the features of one patch are normalized to unit length. Finally, each patch is represented by a 672-dimensional feature vector. The patches are used as patch-level samples for learning representations as described in Section 4.3, i.e., each patch is regarded as a single sample. Moreover, the patches of a horizontal region of an image are concatenated into a part-level sample for training (used in Section 4.2), and the patches belong to the same image are concatenated into a single vector which is the image-level sample (used in Section 4.2). The dimensions of part-level and image-level feature vectors are very high (usually over 100,000), which vary on different datasets due to different image sizes. To reduce the computational cost, we applied PCA to reduce the dimensions of part-level and image-level features to 200 in the experiments. The machine used in our experiments installs 24 GB RAM and Intel Xeon W3350 CPU. The feature extraction takes about 0.3 seconds to process an image.

Baselines and Evaluation Metrics. We compare our approach with three types of person re-identification methods, which are feature learning methods, metric learning methods, and dictionary learning methods. In addition, the ensemble learning and ranking optimizations methods are also included as baselines.

- *Feature learning methods*: symmetry-driven accumulation of local features (SDALF) [15], local descriptors encoded by Fisher vectors (LDFV) [16], unsupervised salience learning method (eSDC) [30], salience matching method (SalMat) [2], mid-level filters [21], salient color names based color descriptor (SCNCD) [22], local maximal occurrence (LOMO) [13], mirror representations [54], transferring semantic representation [32], eSalMatch [60], and WLC [61].
- *Metric learning methods*: probabilistic relative distance comparison (PRDC) [8], large margin nearest neighbor (LMNN) [7], eBiCov [62], information-theoretic metric learning (ITML) [9], pairwise constrained component analysis (PCCA) [10], KISSME [20], local Fisher discriminant analysis (LF) [11], KLFDA [12], polynomial kernel feature map [17], correspondence structure learning [63], MLAPG [19], and SCSP [64].
- *Dictionary learning and others*: SSCDL [26], cross-view projective dictionary learning (CPDL) [27], multi-task learning [55], metric ensembles [35], and ranking optimization method [65].

We employ the standard cumulated matching characteristics (CMC) curve as our evaluation metric, and report the Rank-k recognition rates.

Parameter Setting. There are several major parameters in our two approaches, including α, β, λ_l, and λ_0. In the experiments, we empirically set these parameters to achieve the best performance. In particular, α and β are set to 2 and 1, respectively. Parameters λ_l ($l = 1, 2, 3$) control the effects of cross-view interactions in different levels. Parameter λ_0 controls the contribution of the discriminative regularization term. In addition, γ_1 and γ_2 used in the fusion strategy are empirically chosen in the range $[0,\ 3]$. The sensitivity of parameters will be discussed in Section 6.7.

6.2 VIPeR Dataset

The VIPeR dataset was collected in an outdoor academic environment [57], which is the most widely used benchmark dataset for evaluating person re-identification algorithms[3]. It contains images of 632 pedestrian pairs under two camera views with different viewpoints. The images in two views have significant variations in pose, viewpoint and illuminations. Figure 2(a) shows some images captured by Camera-1 (first row) and Camera-2 (second row) in the VIPeR dataset. The images are normalized to the size of 128×48 in our experiments.

We follow the evaluation protocol in [14]. In particular, we randomly select 316 pairs of images for training, and the remaining pairs are used for test. Then, two groups of experiments are conducted. First, the images captured by Camera-1 are utilized as probe images, and the images captured by Camera-2 as gallery images. For the probe

[2] Our source code is online available at http://www.sheng-li.org.

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

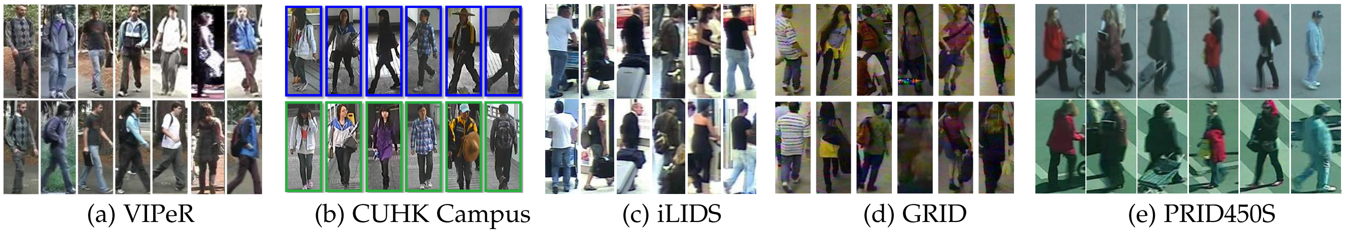

(a) VIPeR (b) CUHK Campus (c) iLIDS (d) GRID (e) PRID450S

Fig. 2. Illustration of images in five benchmark datasets: (a) VIPeR; (b) CUHK01 Campus; (c) iLIDS; (d) GRID and (e) PRID450S.

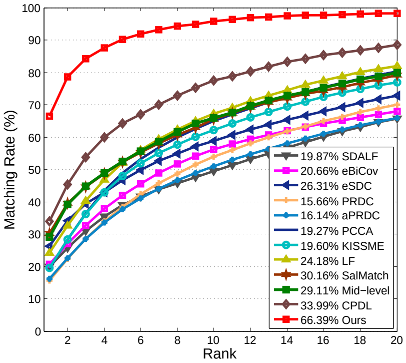


Fig. 3. CMC curves of average matching rates on VIPeR dataset. Rank-1 matching rate is marked before the name of each approach.

TABLE 1
Top ranked matching rates in (%) with 316 persons on VIPeR dataset. r is the number of top ranks.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
PRDC [8]	15.66	38.42	53.86	70.09
PCCA [10]	19.27	48.89	64.91	80.28
SDALF [15]	19.87	38.89	49.37	65.73
eBiCov [62]	20.66	42.00	56.18	68.00
LDFV [16]	22.34	47.00	60.40	71.00
LF [11]	24.11	51.24	67.09	82.01
eSDC [30]	26.31	50.70	62.37	76.36
SalMat [2]	30.16	53.45	65.78	N/A
SSCDL [26]	25.60	53.70	68.10	83.60
Mid-level [21]	29.11	52.50	67.12	80.03
SCNCD [22]	37.80	-	81.20	91.40
Chen et al. [17]	36.80	70.40	83.70	91.70
LOMO+XQDA [13]	40.00	-	80.51	91.08
Shen et al. [63]	34.8	68.7	82.3	91.8
MLAPG [19]	40.73	-	82.34	92.37
MTL-LORAE [55]	42.3	72.2	81.6	89.6
Mirror-KMFA [54]	42.97	75.82	87.28	94.84
TSR [32]	31.1	68.6	82.8	94.9
CPDL [27]	33.99	64.21	77.53	88.58
SCSP [64]	53.54	82.59	91.49	96.65
WLC [61]	51.40	76.40	84.80	-
eSalMatch [60]	44.56	72.10	83.50	-
CMDL (Ours)	62.63	87.97	94.78	97.91
CMDL-Dis (Ours)	**66.39**	**90.25**	**95.85**	**98.29**

images, we match each of them to the gallery set, and obtain the Rank-r rate. The CMC curves are also obtained by using the rates at all ranks. Second, we exchange the training and test sets, and repeat the above procedures. As the raw features for image-level training and part-level training have very high dimensions, we apply PCA to reduce the dimensionality by keeping the 95% energy. We conduct 10 random tests and report the average results. Each random test has two groups of evaluations as described above.

Figure 3 shows the CMC curves of the compared methods. We can observe that our approach achieves higher matching rates in each rank. Table 1 shows the detailed Rank-1, Rank-5, Rank-10, and Rank-20 matching rates of all the compared methods. It shows that the advanced feature learning methods like Mirror-KMFA and metric learning methods like MLAPG obtain much better results than other competitors. Our approaches achieve higher Rank-1/5/10/20 rates than all the other compared methods, which demonstrates the effectiveness of dictionary learning and multi-level representations. Moreover, CMDL-Dis achieves the best results among all competitors, which indicates that discriminative information is critical for person re-identification.

TABLE 2
Comparisons with state-of-the-art ranking and ensemble methods on VIPeR dataset.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
Mid-level+LADF [21]	43.39	73.04	84.87	93.70
Metric Ensembles [35]	45.9	77.5	88.9	95.8
KCCA+DCIA [65]	63.92	78.48	87.50	-
CMDL-Dis (Ours)	**66.4**	**90.3**	**95.9**	**98.3**

As our approaches adopt a fusion strategy, we also compare CMDL-Dis with the state-of-the-art ensemble learning method. Table 2 shows that CMDL-Dis outperforms metric ensembles. In addition, post-ranking optimization methods can usually boost the performance of baseline method significantly. For instance, KCCA+DCIA achieves a Rank-1 rate of 63.9%. Table 2 shows that CMDL-Dis achieves a higher Rank-1 rate of 66.4%, without using any post-ranking optimizations. It demonstrates the importance and great potential of feature learning.

To demonstrate the advantage of cross-view dictionary learning over the single-view dictionary learning for person re-identification, we compare the performance of PDL and CMDL on the image-level representations. The same features are utilized to ensure a fair comparison. The Rank-1/5/10/20 matching rates of CMDL are 21.20%, 56.01%, 70.25%, and 85.13%, while the Rank-1/5/10/20 rates of the original PDL are 8.86%, 19.62%, 31.65%, and 47.47%. Clearly, the strategy of cross-view dictionary learning is very suitable for the application of person re-identification.

6.3 CUHK01 Campus Dataset

The CUHK01 Campus dataset [21] contains pedestrian images of 971 persons in two camera views[4]. It was collected in a campus environment. This dataset shows significant changes of viewpoints. The frontal or back views are captured by Camera-1, while the side views are captured by Camera-2. Figure 2(b) illustrates some images in view 2 (first

[4] http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2764893, IEEE Transactions on Pattern Analysis and Machine Intelligence

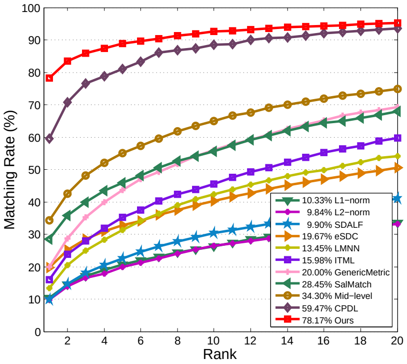


Fig. 4. CMC curves of average matching rates on CUHK01 dataset. Rank-1 matching rate is marked before the name of each approach.

TABLE 3
Top ranked matching rates in (%) on CUHK01 dataset.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
SDALF [15]	9.90	22.57	30.33	41.03
eSDC [30]	19.67	32.71	40.28	50.57
LMNN [7]	13.45	31.33	42.25	54.11
ITML [9]	15.98	35.22	45.60	59.80
SalMat [2]	28.45	45.85	55.67	68.89
Mid-level [21]	34.30	55.06	64.96	73.94
Mirror-KMFA [54]	40.40	64.63	75.34	84.08
XQDA [13]	63.21	83.50	90.04	94.16
MLAPG [19]	64.24	N/A	90.84	94.92
CPDL [27]	59.47	81.26	89.72	93.10
WLC [61]	65.80	81.10	85.90	-
CMDL (Ours)	72.63	87.97	91.78	93.91
CMDL-Dis (Ours)	**78.17**	**88.85**	**92.55**	**95.21**

row) and view 1 (second row). The images are resized to 160×60 in our experiments.

We follow the evaluation protocol in [2]. In particular, we randomly partition the dataset into two parts, 50% for training and 50% for testing, without overlap on person identities. For every person, two images were captured in each view. One image from each person under Camera A are used to build the probe set, and the images (one per person) from Camera B are used to construct the gallery set. We match each image in the probe set to every gallery image, and calculate the correct matched rank and CMC curves. The whole procedure is repeated for 10 times, and the average CMC curves are generated, as shown in Figure 4. Table 3 shows the detailed Rank-1/5/10/20 matching rates of the compared methods. We can observe that both of our approaches obtain much higher matching rates than other methods. The Rank-1 matching rate is improved by almost 13%, compared to the MLAPG method.

6.4 iLIDS Dataset

The iLIDS MCTS dataset is a public video dataset captured at an airport arrival hall under a multi-camera network[5]. From these video sequences, 479 images of 119 pedestrians were extracted to construct a dataset for evaluating the performance of person re-identification algorithms. In particular, the images in this dataset were collected from non-overlapping cameras, which contain considerable illumination changes and occlusions, as shown in Figure 2(c). Each image in this dataset is normalized to the size of 128×64.

TABLE 4
Top ranked matching rates in (%) on iLIDS dataset.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
PCCA [10]	24.1	53.3	69.2	84.8
rPCCA [10]	28.2	56.5	71.8	85.9
LF [11]	32.2	56.0	68.7	81.6
SVMML	20.8	49.1	65.4	81.7
KISSME [20]	28.0	54.2	67.9	81.6
KLFDA [12]	36.9	65.3	78.3	89.4
CPDL [27]	39.5	65.4	78.5	88.2
eSalMatch [60]	44.56	72.10	83.50	-
CMDL (Ours)	45.0	65.6	79.0	88.5
CMDL-Dis (Ours)	**45.1**	**66.7**	**79.2**	**90.1**

For each person, there are multiple images from each view. As we focus on single shot evaluation, we follow the conventional setting [66], and randomly choose one image from each view for each person when constructing the gallery/probe sets. The whole procedure is repeated for 10 times, and we report the average Rank-1/5/10/20 matching rates in Table 4. We compare our approaches with the baselines that report the state-of-the-art results on the iLIDS dataset, including PCCA, rPCCA, LF, SVMML, KISSME, KLFDA, and CPDL. Our approach, CMDL, achieves higher Rank-1 and Rank-5 matching rates than other baselines. Moreover, CMDL-Dis outperforms its competitors in every case, which demonstrates its effectiveness in complex scenarios with illumination changes and occlusions. In addition, we notice that CMDL and CMDL-Dis obtain very similar Rank-1 results on this dataset. Compared with CMDL, CMDL-Dis incorporates a discriminative regularization term, which aims to minimize the correlation between different persons. If different subjects have significant differences on appearance (e.g., VIPeR and CUHK Campus datasets as shown in Figure 2(a-b)), CMDL-Dis can make use of such discriminative information, and achieve better performance than CMDL. However, the iLIDS dataset contains low-quality images, and has less significant variations across subjects, as shown in Figure 2(c). It might be difficult for CMDL-Dis to extract additional discriminative information from this dataset. Moreover, CMDL extends CPDL by introducing the part-level matching. The performance gain by CMDL is limited compared to the results on other datasets such as VIPeR. The major reason is that two camera views in the iLIDS dataset are non-overlapping, which makes it difficult to extract the view-invariant features.

6.5 GRID Dataset

The QMUL underGround Re-IDentification (GRID) dataset[6] contains 250 pairs of pedestrian images [58]. All images are captured from eight disjoint camera views installed in a busy underground station. For each individual, a pair of images was captured from two different camera views. In addition, GRID dataset contains 775 additional images that do not belong to the 250 individuals, and these images can be used to enlarge the gallery set. Figure 2(d) shows six pairs of images, which demonstrate variations of pose, colors, illumination changes. In addition, the resolution of images in GRID dataset is very low, which makes it more challenging to do person re-identification.

[5] https://www.gov.uk/guidance/imagery-library-for-intelligent-detection-systems

[6] http://www.eecs.qmul.ac.uk/~ccloy/downloads_qmul_underground_reid.html

TABLE 5
Top ranked matching rates in (%) with on GRID dataset.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
ELF6+L1-norm	4.40	-	16.24	24.80
ELF6+RankSVM	10.24	24.60	33.28	43.68
ELF6+PRDC [8]	9.68	22.00	32.96	44.32
ELF6+MRank-RankSVM	12.24	27.80	36.32	46.56
ELF6-MRank-PRDC	11.12	26.10	35.76	46.56
ELF6+XQDA [13]	10.48	-	38.64	52.56
LOMO+XQDA [13]	16.56	-	41.84	52.40
MLAPG [19]	16.64	-	41.20	52.96
Chen et. al [17]	16.30	35.80	46.00	57.6
CPDL [27]	21.60	45.85	61.05	65.80
SCSP [64]	24.24	44.56	54.08	65.20
CMDL (Ours)	29.44	53.84	66.40	76.80
CMDL-Dis (Ours)	**30.88**	**56.88**	**67.76**	**78.48**

TABLE 6
Top ranked matching rates in (%) on PRID450S dataset.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
LDML	4	16	23	40
ITML [9]	21	53	67	84
LMNN-R [7]	19	54	66	81
LMNN [7]	24	62	73	87
EIML	29	62	73	86
KISSME [20]	28	65	76	88
SCNCD [22]	42	67	79	88
Shen et al. [63]	44	72	82	90
TSR [32]	43	71	78	86
Mirror-KMFA [54]	**55**	79	88	92
CPDL [27]	48	80	89	96
CMDL (Ours)	51	81	89	95
CMDL-Dis (Ours)	52	**83**	**90**	**96**

In the experiments, we randomly choose 125 pairs of images for training, and use the rest 125 pairs with the additional 775 images for test. The random selection process is repeated for 10 times. Table 5 shows the detailed Rank-1/5/10/20 matching rates of our approaches and the baseline methods. By comparing Table 5 with the results in Table 1 and Table 3, we can observe that the re-identification task on the GRID dataset is more challenging than that on the VIPeR and CUHK01 Campus dataset. There are two major reasons. First, the images in GRID were captured by eight different cameras, while VIPeR and CUHK01 Campus only have two camera views. Second, GRID has a larger test set with 775 additional images, which usually leads to a lower matching rate. We compare the performance of our approaches with the state-of-the-art results reported on the GRID dataset. Table 5 shows the average Rank-1/5/10/20 matching rates. Our approach improves the Rank-1 matching rate by at least 9%. It shows that integrating features extracted from three levels (i.e., image-level, part-level, and patch-level) is an effective strategy to handle the challenging person re-identification problem with multiple viewpoints.

6.6 PRID450S Dataset

The PRID450S dataset[7] is a benchmark dataset for evaluating person re-identification approaches [59], which is based on the PRID 2011 dataset. It contains 450 image pairs captured by two different, static surveillance cameras. PRID450S is also a challenging dataset for person re-identification, due to the partial occlusions and viewpoint changes. We compare the performance of our approaches with the reported state-of-the-art results on this dataset. The baselines include Mirror-KMFA, TSR, SCNCD, etc.. In the experiments, we randomly choose 225 pairs of images for training, and repeat this process 10 times. Table 6 shows the average Rank-1/5/10/20 matching rates of our approaches and the baseline methods. Mirror-KMFA achieves the highest Rank-1 rate among all the compared methods. Both of our approaches, CMDL and CMDL-Dis, outperform Mirror-KMFA in most cases, which demonstrates the effectiveness of cross-view (discriminative) dictionary learning and multi-level representations.

6.7 Discussions

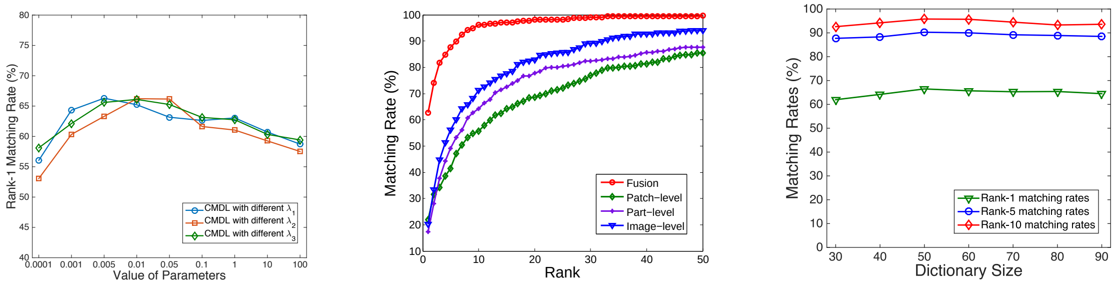
Parameter Sensitivity. Different from existing methods, the proposed CMDL and CMDL-Dis approaches explicitly model the interactions between different views, such as the similarities of codings (in the image-level and part-level) or similarities of dictionaries (in the patch-level). The parameters λ_l $(l = 1, 2, 3)$ control the effects of the cross-view interactions. Figure 5(a) shows the Rank-1 matching rates of our approach with different values of λ_1, λ_2, and λ_3 on the VIPeR dataset. When we observe the effect of one parameter, the values of other two parameters are fixed. Figure 5(a) shows that our approach is not very sensitive to the choice of parameters in the range $[0.005,\ 0.05]$. We set $\lambda_1 = 0.005,\ \lambda_2 = 0.05,\ \lambda_3 = 0.01$. λ_0 used in CMDL-Dis controls the effects of discriminative regularization. CMDL-Dis degrades to CMDL by setting λ_0 to 0. We found that the model is not sensitive to λ_0 in a wide range. Therefore, we set it to 0.1 on all the datasets. Figure 6 shows the sensitivity of γ_1 and γ_2 that are used in similarity score fusion. We can observe that the matching rate is very low by setting both γ_1 and γ_2 to 0. It suggests that the image-level matching and part-level matching play important roles. Moreover, high matching rates are achieved by setting $\gamma_1 = 3$ and $\gamma_2 = 1$. We use the same parameter settings on all the datasets.

Another important factor in our approach is the size of dictionary. We use the same dictionary size in different views. Figure 5(c) shows the Rank-1/5/10 matching rate with different dictionary sizes on the VIPeR dataset. We achieved similar results on the other datasets. Accordingly, the dictionary size is set to 50 on the VIPeR, iLIDS, GRID, and PRID450S datasets, and set to 60 on the CUHK dataset. Also, we note that the matching process in existing feature learning methods (e.g., SalMat or Mid-level filter) is very time-consuming. However, our approach adopts a relatively small dictionary, which leads to compact representations of images, and therefore speeds up the matching process.

Evaluation of Fusion Strategy. To understand the effectiveness of integrating features from multiple representation levels, we report the matching rates of each level in Figure 5(b). It shows the CMC curves of our approach and its three components, i.e., image-level model, part-level model and patch-level model, on the VIPeR dataset. We can observe that the representations in three different levels are complementary to each other, and our approach takes full advantage of the complementary information.

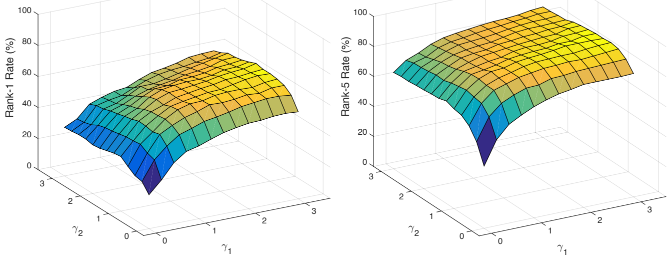
Given a probe image, a similarity score vector can be obtained by computing the similarity between this probe image and all the gallery images. We then rank this similarity score vector. A higher Rank-1 matching rate can be

[7] http://lrs.icg.tugraz.at/download.php



(a) Parameter sensitivity (b) Fusion results (c) Effects of dictionary size

Fig. 5. Experimental analysis on VIPeR dataset. (a) Rank-1 matching rates v.s. different values of parameters; (b) Matching rates of image-level model, patch-level model and the fusion model; (c) Rank-1 matching rates v.s. different dictionary sizes.



(a) Rank-1 matching rate (b) Rank-5 matching rate

Fig. 6. Rank-1 and Rank-5 matching rates v.s. different values of γ_1 and γ_2 for score fusion on VIPeR dataset.

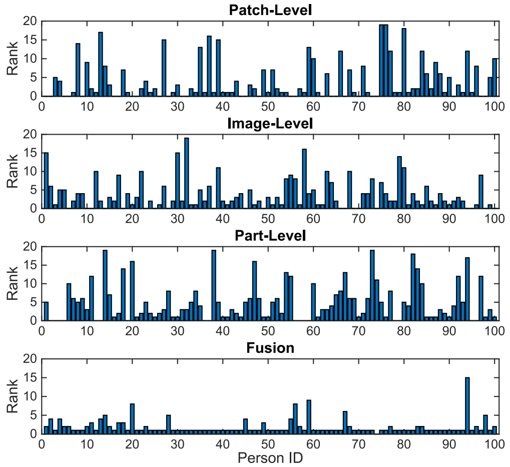


Fig. 7. Ranking of the true gallery samples after matching on VIPeR dataset.

achieved, if more true gallery images are ranked 1st when matching with their corresponding probe images. Figure 7 shows the ranking of true gallery samples after matching on the VIPeR dataset. For the sake of better visualization, we only show the matching of 100 probe images, and only consider the top 20 ranks. In each subfigure, the x-axis indicates 100 probe images, and the y-axis shows the rank of their corresponding true gallery images. The first three subfigures show the ranking results by using similarity scores calculated in different representation levels, and the bottom one shows the results after fusion. Figure 7 explains the effectiveness of our fusion strategy. First, three base models generate quite diverse outputs. It is desirable for fusion, as three representation levels provide with complementary information. Second, the fusion of weak models could become a strong model. For example, the 43th true gallery image is ranked 4th, 3rd, and 2nd in patch-level, image-level, and part-level, respectively. But it was ranked 1st after fusion. Therefore, a high Rank-1 matching rate could be achieved, as shown in Figure 5(b). In addition, we observe that the histogram of part-level is quite different from that of patch-level or image-level, which implies that the part-level representations indeed provide a great amount of complementary information on the VIPeR dataset.

TABLE 7
Top ranked matching rates in (%) on PRID2011 dataset.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
SDALF [15]	5.2	20.7	32	47.9
SalMat [2]	25.8	43.6	52.6	62.0
RPRF [69]	19.3	38.4	51.6	68.1
CLBP+RSVM [67]	34.3	56.0	65.5	77.3
DVDL [67]	40.6	69.7	77.8	85.6
CMDL-ImageLevel (Ours)	41.2	70.5	79.6	88.7
CMDL-ImageLevel-Dis (Ours)	**41.6**	**71.9**	**82.0**	**91.0**

Multi-Shot Experiments. Although we describe multiple representation levels in this paper, the proposed dictionary learning model can be applied to any types of features, under either single-shot setting or multi-shot setting. The experiments described from Section 6.1 to Section 6.5 focus on the single-shot person re-identification problem. In this section, we evaluate the performance of our approach and baselines on multi-shot person re-identification. We follow the feature extraction strategies in [67], and use two multi-shot datasets PRID2011 [68] and iLIDS-VID [3]. The PRID 2011 dataset consists of image sequences for 200 people in two non-overlapping camera views, and the iLIDS-VID dataset consists of image sequences for 300 people in two views. We randomly split the image sequences in each test dataset into equal-sized training and test sets, and report the average results over 10 random training-test splits. As the extracted features are actually image-level representations (i.e., each feature vector corresponds to an image), we only employ the image-level dictionary learning model as introduced in Section 4.2 to train a pair of dictionaries, and perform image-level matching as introduced in Section 4.5. The detailed results are shown in Table 1. It shows that, our method outperforms the baselines in all the cases, which demonstrates the effectiveness of learning view-specific dictionaries. In addition, we can observe that the projective dictionary learning strategy is more effective than the conventional l_1 optimization used in [67].

TABLE 8
Top ranked matching rates in (%) on iLIDS-VID dataset.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
SDALF [15]	6.3	18.8	27.1	37.3
SalMat [2]	10.2	24.8	35.5	52.9
RPRF [69]	14.5	29.8	40.7	58.1
CLBP+RSVM [67]	23.2	44.2	54.1	68.8
DVDL [67]	25.9	48.2	57.3	68.9
CMDL-ImageLevel (Ours)	26.1	52.4	63.3	81.3
CMDL-ImageLevel-Dis (Ours)	**26.7**	**54.67**	**64.7**	**82.7**

7 Conclusions

In this paper, we proposed two cross-view multi-level dictionary learning approaches, CMDL and CMDL-Dis, for person re-identification. Our approaches learned pairs of dictionaries across different views at three representation levels, including the patch-level, part-level, and image-level. The view-consistency was explicitly modeled in each level, and the learned dictionaries can be used to represent probe and gallery images, leading to robust and compact representations. Moreover, a discriminative regularizer was incorporated in CMDL-Dis to improve the discriminability of learned features. Extensive experiments were conducted on five benchmark datasets, including the VIPeR, CUHK01 Campus, iLIDS, GRID and PRID450S datasets. Experimental results showed that our approaches achieved the state-of-the-art performance compared to the related methods, due to the fully exploration of multiple levels of representations and the discriminative regularizer. In our future work, we will apply the proposed cross-view dictionary learning framework to other visual learning tasks, such as multi-view face recognition.

Acknowledgments

This research is supported in part by the NSF IIS award 1651902, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Award W911NF-17-1-0367.

References

[1] W. Zheng, S. Gong, and T. Xiang, "Transfer re-identification: From person to set-based verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2650–2657.

[2] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2528–2535.

[3] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 688–703.

[4] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[5] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[6] J. Chen, Z. Zhang, and Y. Wang, "Relevance metric learning for person re-identification by exploiting listwise similarities," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4741–4755, 2015.

[7] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2005.

[8] W. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 649–656.

[9] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the International Conference on Machine Learning*, 2007, pp. 209–216.

[10] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2666–2672.

[11] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318–3325.

[12] F. Xiong, M. Gou, O. I. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 1–16.

[13] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[14] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proceedings of the European Conference on Computer Vision I*, 2008, pp. 262–275.

[15] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2360–2367.

[16] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Proceedings of the European Conference on Computer Vision Workshops and Demonstration*, 2012, pp. 413–422.

[17] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1565–1573.

[18] Z. Wu, Y. Li, and R. J. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1095–1108, 2015.

[19] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3685–3693.

[20] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295.

[21] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 144–151.

[22] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 536–551.

[23] L. Li, S. Li, and Y. Fu, "Learning low-rank and discriminative dictionary for image classification," *Image and Vision Computing*, vol. 32, no. 10, pp. 814–823, 2014.

[24] J. Ni, Q. Qiu, and R. Chellappa, "Subspace interpolation via dictionary learning for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 692–699.

[25] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Projective dictionary pair learning for pattern classification," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2014, pp. 793–801.

[26] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3550–3557.

[27] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 2155–2161.

[28] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 780–793.

[29] R. Layne, T. M. Hospedales, and S. Gong, "Person re-identification by attributes," in *Proceedings of the British Machine Vision Conference*, 2012, pp. 1–11.

[30] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3586–3593.

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

[32] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4184–4193.

[33] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3567–3574.

[34] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1629–1642, 2014.

[35] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1846–1855.

[36] C. Liu, C. C. Loy, S. Gong, and G. Wang, "POP: person re-identification post-rank optimisation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 441–448.

[37] H. Guo, Z. Jiang, and L. S. Davis, "Discriminative dictionary learning with pairwise constraints," in *Proceedings of the Asian Conference on Computer Vision*. Springer, 2013, pp. 328–342.

[38] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, 2012.

[39] J. Zheng, Z. Jiang, P. J. Phillips, and R. Chellappa, "Cross-view action recognition via a transferable dictionary pair." in *BMVC*, vol. 1, no. 2, 2012, pp. 1–11.

[40] Q. Qiu, V. M. Patel, and R. Chellappa, "Information-theoretic dictionary learning for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2173–2184, 2014.

[41] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[42] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2691–2698.

[43] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 695–704.

[44] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[45] ——, "Multi-view intact space learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2531–2544, 2015.

[46] ——, "Multi-view learning with incomplete views," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5812–5825, 2015.

[47] S. Bhadra, S. Kaski, and J. Rousu, "Multi-view kernel completion," *Machine Learning*, vol. 106, no. 5, pp. 713–739, 2017.

[48] K. Sridharan and S. M. Kakade, "An information theoretic framework for multi-view learning." in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 2008, pp. 403–414.

[49] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. ACM, 1998, pp. 92–100.

[50] V. Sindhwani and D. S. Rosenberg, "An RKHS for multi-view learning and manifold co-regularization," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 976–983.

[51] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3111–3124, 2015.

[52] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188–194, 2016.

[53] R. Memisevic, "On multi-view feature learning," in *Proceedings of the International Conference on Machine Learning*, 2012.

[54] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2015, pp. 3402–3408.

[55] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3739–3747.

[56] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Mathematical Methods of Operations Research*, vol. 66, no. 3, pp. 373–407, 2007.

[57] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETS*, 2007.

[58] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1988–1995.

[59] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*, ser. Advances in Computer Vision and Pattern Recognition, S. Gong, M. Cristani, S. Yan, and C. C. Loy, Eds. London, United Kingdom: Springer, 2014, pp. 247–267.

[60] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 356–370, 2017.

[61] Y. Yang, L. Wen, S. Lyu, and S. Z. Li, "Unsupervised learning of multi-level descriptors for person re-identification." in *AAAI*, 2017, pp. 4306–4312.

[62] B. Ma, Y. Su, and F. Jurie, "BiCov: a novel image representation for person re-identification and face verification," in *Proceedings of the British Machine Vision Conference*, 2012, pp. 1–11.

[63] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[64] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1268–1277.

[65] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel, "Person re-identification ranking optimisation by discriminant context information analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1305–1313.

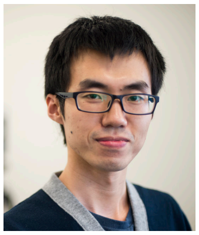
[66] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *BMVC*, 2009.

[67] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4516–4524.

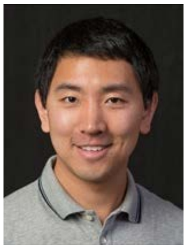
[68] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian Conference on Image Aalysis*. Springer, 2011, pp. 91–102.

[69] Y. Li, Z. Wu, and R. J. Radke, "Multi-shot re-identification with random-projection-based random forests," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 373–380.

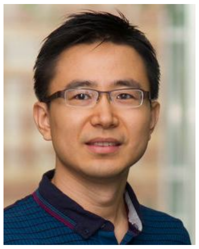
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2764893, IEEE Transactions on Pattern Analysis and Machine Intelligence



Sheng Li (S'11-M'17) received the B.Eng. degree in computer science and engineering and the M.Eng. degree in information security from Nanjing University of Posts and Telecommunications, China, and the Ph.D. degree in computer engineering from the Northeastern University, Boston, MA, in 2010, 2012 and 2017, respectively. He is currently a data scientist at Adobe Research, San Jose, CA. He has published over 50 papers at leading conferences and journals. He received the best paper awards (or nominations) at SDM 2014, IEEE ICME 2014, and IEEE FG 2013. He serves as an Associate Editor of IET Image Processing and Neural Computing and Applications, reviewer for several IEEE Transactions, and program committee member for IJCAI, AAAI, IEEE FG, PAKDD, and DSAA. His research interests include robust machine learning, visual intelligence, and behavior modeling.



Ming Shao (S'11-M'16) received the B.E. degree in computer science, the B.S. degree in applied mathematics, and the M.E. degree in computer science from Beihang University, Beijing, China, in 2006, 2007, and 2010, respectively. He received the Ph.D. degree in computer engineering from Northeastern University, Boston MA, 2016. He is a tenure-track Assistant Professor affiliated with College of Engineering at the University of Massachusetts Dartmouth since 2016 Fall. His current research interests include sparse modeling, low-rank matrix analysis, deep learning, and applied machine learning on social media analytics. He was the recipient of the Presidential Fellowship of State University of New York at Buffalo from 2010 to 2012, and the best paper award winner/candidate of IEEE ICDM 2011 Workshop on Large Scale Visual Analytics, and ICME 2014. He has served as the reviewers for many IEEE Transactions journals including TPAMI, TKDE, TNNLS, TIP, and TMM. He has also served on the program committee for the conferences including AAAI, IJCAI, and FG.



Yun Fu (S'07-M'08-SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. He is an interdisciplinary faculty member affiliated with College of Engineering and the College of Computer and Information Science at Northeastern University since 2012. His research interests are Machine Learning, Computational Intelligence, Big Data Mining, Computer Vision, Pattern Recognition, and Cyber-Physical Systems. He has extensive publications in leading journals, books/book chapters and international conferences/workshops. He serves as associate editor, chairs, PC member and reviewer of many top journals and international conferences/workshops. He received seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; seven Best Paper Awards from IEEE, IAPR, SPIE, SIAM; three major Industrial Research Awards from Google, Samsung, and Adobe, etc. He is currently an Associate Editor of the IEEE Transactions on Neural Networks and Leaning Systems (TNNLS). He is fellow of IAPR, a Lifetime Senior Member of ACM and SPIE, Lifetime Member of AAAI, OSA, and Institute of Mathematical Statistics, member of Global Young Academy (GYA), INNS and Beckman Graduate Fellow during 2007-2008.