The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Discriminative Semi-Coupled Projective Dictionary Learning for Low-Resolution Person Re-Identification

Kai Li,[†] Zhengming Ding,[†] Sheng Li,[⋆] Yun Fu[†‡]
[†]Department of Electrical & Computer Engineering, Northeastern University, Boston, USA
[⋆]Adobe Research, USA
[‡]College of Computer & Information Science, Northeastern University, Boston, USA
kaili@ece.neu.edu, allanding@ece.neu.edu, sheli@adobe.com, yunfu@ece.neu.edu

Abstract

Person re-identification (re-ID) is a fundamental task in automated video surveillance. In real-world visual surveillance systems, a person is often captured in quite low resolutions. So we often need to perform low-resolution person re-ID, where images captured by different cameras have great resolution divergences. Existing methods cope problem via some complicated and time-consuming strategies, making them less favorable in practice, and their performances are far from satisfactory. In this paper, we design a novel Discriminative Semi-coupled Projective Dictionary Learning (DSPDL) model to effectively and efficiently solve this problem. Specifically, we propose to jointly learn a pair of dictionaries and a mapping to bridge the gap across low(er) and high(er) resolution person images. Besides, we develop a novel graph regularizer to incorporate positive and negative image pair information in a parameterless fashion. Meanwhile, we adopt the efficient and powerful projective dictionary learning technique to boost the our efficiency. Experiments on three public datasets show the superiority of the proposed method to the state-of-the-art ones.

Introduction

Matching cross-view pedestrians, known as person re-identification (re-ID), is a fundamental task in automated video surveillance. It is a challenging task due to the great variations of human poses, light conditions, and image resolutions, etc. In real-world visual surveillance systems, a person is often captured in low resolutions by cameras, so that it is often required to perform low-resolution person re-ID, where images to be matched have great resolution divergences. For example, it is common that cameras in a visual surveillance system are deployed in different places of a scene, and hence a person could be captured in very different resolutions by different cameras (Ding, Shao, and Fu 2014), if the cameras are located separately with a long distance. In this case, we need to exploit images of great resolution disparities to re-identify persons of interest. Many person re-ID algorithms have been proposed in recent years and proved to be effective to perform person re-ID on cross-view images with similar resolutions. However, their performances are not guaranteed when the images are of great resolution divergences.

There are several attempts targeting for this degenerated scenario. Li *et al.* proposed to jointly learn a shared subspace across different scales and a discriminative distance metric which minimizes a novel heterogeneous class mean discrepancy criterion (Li et al. 2015). Wang *et al.* found that the scale-distance functions between images of the same persons and those of different persons can be distinguished. Based on this observation, they proposed to learn a discriminating surface to separate the two types of scale-distance functions, and used it for re-identifying persons (Wang et al. 2016b). Wang *et al.* built two coupled marginalized denoising auto-encoders to extract effective feature from low and high resolution pedestrian images (Wang, Ding, and Fu 2016). Jing *et al.* proposed a semi-coupled low-rank discriminant dictionary learning (SLD^{2}L) method by dividing images into patches and learning semi-coupled dictionaries for corresponding image patch clusters (Jing et al. 2015). Low-rank constraint was enforced on the dictionaries for all clusters to better characterize intrinsic feature spaces of high and low resolution images. Despite of these sophisticated techniques, the performances of these methods are far from satisfactory, and some of them are too time-consuming.

In this paper, we design a Discriminative Semi-coupled Projective Dictionary Learning (DSPDL) method to cope the resolution difference problem in person re-ID. Figure 1 shows the framework of the proposed method. We adopt the efficient projective dictionary technique, and design a semi-coupled cross-view projective dictionary learning framework. To apply it for person re-ID, we design a novel parameterless graph regularizer which incorporates both intra-person similarity and inter-person dissimilarity in a graph embedding fashion. The main contributions of this paper are summarized as follows:

- We propose a semi-coupled projective dictionary learning framework for matching pedestrian images of great resolution divergences. Our framework adopts the effective projective dictionary learning technique, and jointly learns a mapping function along with the dictionaries. Through the introduction of the mapping function, the stringent correspondence is relaxed between the new codings of cross-view images of the same identity, thereby leaving the dictionaries more flexibility to maximize the feature representation ability.

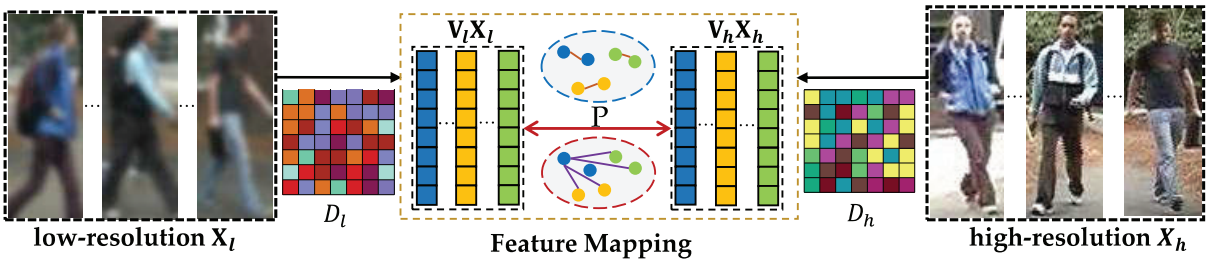


Figure 1: Framework of the proposed method. By utilizing labeled low and high resolution pedestrian images X_l and X_h, we jointly learn a dictionary pair D_l and D_h, and a mapping function P, which relates the new codings of X_l and X_h. V_l and V_h are two projections which map X_l and X_h to their new codings V_lX_l and V_hX_h, respectively. Positive and negative pair information are incorporated in a graph regularization term to guide learning dictionaries of strong discriminability.

- We design a novel parameterless graph regularizer which involves both positive and negative pair knowledge. The incorporation of discriminative information in the regularizer boosts the discriminative ability of the learned dictionaries, thus facilitating our method to distinguish correct person pairs from incorrect ones. The graph regularizer is parameter free, so that our method is supposed to have robust performance with images of great diversity.
- We evaluate our method on three benchmark datasets, i.e., VIPeR, CUHK01, and QMUL-iLIDS by comparing with the state-of-the-art approaches. The results show that our method achieves remarkable improvements.

Related Works

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

Person Re-Identification Person re-ID can generally be classified into three categories: pedestrian description based methods, distance metric learning based methods and deep learning based methods.

Based on the fact that images of the same person in different camera views should be similar in appearance, many pedestrian image description based methods have been proposed for person re-ID (Zhao, Ouyang, and Wang 2013b; Li et al. 2013; Matsukawa et al. 2016). Apart from directly using low-level color and texture features, another good choice is the attribute-based features which can be viewed as mid-level representations. It is believed that attributes are more robust to image transformations compared to low-level descriptors (Liu et al. 2012; Su et al. 2015).

The second category is the distance metric learning based methods. The general idea of metric learning based person re-ID methods is to learn some distance metrics under which the vectors of the same identities are pushed closer while the vectors of different identities are pulled further apart. KISSME (Köstinger et al. 2012) is one of the most acknowledged methods in this category, which decides whether a pair of description vectors is similar or not by formulating it as a likelihood ratio test. The pairwise difference is employed and the difference space is assumed to be a Gaussian distribution with a zero mean. Inspired by KISSME, many metric learning based person re-ID algorithms have been proposed, including LFDA (Xiong et al. 2014), XQDA (Liao et al. 2015), MLAPG (Liao and Li 2015).

The third category is deep learning based methods. Although deep learning has shown extraordinary advantages in many visual learning tasks, the lack of training data becomes the major bottleneck of applying deep learning for person re-ID. Most re-ID datasets provide only two images for each identity such that they are insufficient to train complex deep learning models. For this reason, deep learning based person re-ID methods are often unable to beat the traditional methods (Paisitkriangkrai, Shen, and van den Hengel 2015; Matsukawa et al. 2016). It is also for this reason that many deep learning based person re-ID methods focus on the Siamese model, in which two or more identical subnetworks share parameters during the training process (Yi et al. 2014; Cheng et al. 2016). Some deep learning based methods focus on developing novel loss function (Chen et al. 2017; Wu, Shen, and van den Hengel 2017; Cheng et al. 2016) in training the deep neural networks.

Our method solves the person re-ID problem from the perspective of feature learning. It belongs to the image description based methods, but we generate image descriptions from the perspective of robust feature learning via exploiting dictionary learning.

Dictionary Learning As a powerful feature learning technique, dictionary learning has shown impressive performances in many applications, such as image synthesis and image super-resolution (Wang et al. 2012), classification (Li, Li, and Fu 2014; Ding, Shao, and Fu 2015), etc. K-SVD (Aharon, Elad, and Bruckstein 2006), discriminative K-SVD (Zhang and Li 2010), and projective dictionary pair learning (Gu et al. 2014) are some of the most popular dictionary learning methods. Dictionary learning has also been introduced to person re-ID (Kodirov et al. 2016; Li, Shao, and Fu 2015; Jing et al. 2015; Zhang et al. 2016). The basic idea is to learn some dictionaries under which the images of the same person have similar feature representations. Our method inherits this idea, but we adopt the powerful projective dictionary learning technique to avoid the low efficiency problem of existing methods. Meanwhile, we cope the low-resolution problem by learning a mapping between low and high resolution images along with the dictio-

naries, and define a novel parameterless graph regularizer to incorporate both positive and negative pair knowledge.

Algorithm

In this section, we first introduce our Discriminative Semi-coupled Projective Dictionary Learning (DSPDL) model for matching pedestrian images of great resolution divergence. Next, we present the differences between our method and the existing ones to highlight our novelties. After that, we introduce the details about the optimization of our model, followed by the complexity analysis of our algorithm.

Proposed Model

Denoted by $H = [X_h, T_h] \in \mathbb{R}^{d\times(n+m)}$ and $L = [X_l, T_l] \in \mathbb{R}^{d\times(n+m')}$ two pedestrian image sets of high and low resolutions, respectively. In practice, H and L could be the image collections captured by two cameras, with one being near to a spot of interest and the other being far away from that spot. The goal of (supervised) person re-ID is to utilize training images $X_h \in \mathbb{R}^{d\times n}$ and $X_l \in \mathbb{R}^{d\times n}$ and their corresponding identity label matrices Y_h and Y_l to learn some patterns (classifiers, dictionaries, metrics, etc.) that can be used to re-identify persons in testing image sets $T_h \in \mathbb{R}^{d\times m}$ and $T_l \in \mathbb{R}^{d\times m'}$.

Taking advantage of the good efficiency of the projective dictionary learning, we formulate our semi-coupled projective dictionary learning framework as:

$$\begin{aligned}\min_{D_h,D_l,P,V_h,V_l} \quad & \|X_h - D_hV_hX_h\|_{\mathrm{F}}^2 + \|X_l - D_lV_lX_l\|_{\mathrm{F}}^2 \\ & +\lambda_1\Omega(V_h,X_h,V_l,X_l,P) + \lambda_2\|P\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \left\|d_h^i\right\| \le 1,\ \left\|d_l^i\right\| \le 1, i = 1,\cdots,k,\end{aligned} \tag{1}$$

where D_h (D_l) $\in \mathbb{R}^{d\times k}$ is the dictionary corresponding to X_h (X_l), with d_h^i (d_l^i) being the i-th columns of D_h (D_l). Inherited from projective dictionary learning (Gu et al. 2014), the new codings of input features are analytically projected from the input features, *i.e.*, the new coding of X_h (X_l) under D_h (D_l) is V_hX_h (V_lX_l), which is projected from X_h (X_l) by projection V_h (V_l) $\in \mathbb{R}^{k\times d}$. In this way, we do not need to constrain the new codings of input features to be sparse, thus avoiding to solve an inefficient l_1-norm optimization problem. $\|X_h - D_hV_hX_h\|_{\mathrm{F}}^2$ and $\|X_l - D_lV_lX_l\|_{\mathrm{F}}^2$ are the data fidelity terms for high and low resolution pedestrian image sets, respectively.

$\Omega(V_h, X_h, V_l, X_l, P)$ is the regularizer guiding to learn dictionaries under which the high and low resolution images of the same person will have similar new codings. Intuitively, we can push close the new codings of images of the same person and set the regularizer as $\|V_hX_h - V_lX_l\|_{\mathrm{F}}^2$. However, for cross-view images with great resolution differences, there are significant appearance disparities in images of the same person when their resolution gaps are so large; directly pushing their new codings close could compromise the generalization power of the learned dictionaries. As an improvement, we introduce a mapping function P to bridge the great resolution gaps between pedestrian images from different views. The proposed mapping function introduces much flexibility on the correspondence of the new codings for images of the same persons, thus making it possible to maximize the feature generality power of the dictionaries to be learned. In this way, we formulate our Semi-coupled Projective Dictionary Learning (SPDL) model as:

$$\begin{aligned}\min_{D_h,D_l,V_h,V_l,P} \quad & \|X_h - D_hV_hX_h\|_{\mathrm{F}}^2 + \|X_l - D_lV_lX_l\|_{\mathrm{F}}^2 \\ & +\lambda_1\|V_hX_h - PV_lX_l\|_{\mathrm{F}}^2 + \lambda_2\|P\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \|d_h^i\| \le 1,\ \|d_l^i\| \le 1,\ i = 1,\cdots,k.\end{aligned} \tag{2}$$

The introduction of the mapping function mitigates the resolution divergence between low and high resolution images. However, we find SPDL has a limitation that it constrains only on the positive image pair information of the same identities (via minimizing $\|V_hX_h - PV_lX_l\|_{\mathrm{F}}^2$), but ignores the negative image pair information of different identities. The latter should be beneficial for boosting the discriminative power of the learned dictionaries. We reformulate the regularization term $\Omega(V_h, X_h, V_l, X_l, P)$ using graph embedding technique to incorporate both positive and negative pair information.

We construct two graphs, $i.e.$, intra-person graph G^s and inter-person graph G^d, which can be encoded by affinity matrices W^s and W^d, respectively, for images from both views. Note that for person re-ID, we focus on cross-view discriminative dictionary learning, and consider only edges between cross-view nodes (pedestrians). Therefore, we define

$$W^s = \begin{bmatrix} 0 & W_h^s \\ W_l^s & 0 \end{bmatrix},\ W^d = \begin{bmatrix} 0 & W_h^d \\ W_l^d & 0 \end{bmatrix}, \tag{3}$$

where

$$W_{h,ij}^s = \begin{cases} 1, & \text{if } y_h^i = y_l^j, \\ 0, & \text{otherwise;} \end{cases} \quad W_{h,ij}^d = \begin{cases} \frac{1}{n}, & \text{if } y_h^i \neq y_l^j, \\ 0, & \text{otherwise;} \end{cases} \tag{4}$$

where $y_h^i \in Y_h$ is the label of the i-th person from X_h, and $y_l^j \in Y_l$ is the label of the j-th person from X_l. Let $Z = [V_hX_h, PV_lX_l] = (z_1, z_2, ..., z_n, z_{n+1}, ..., z_{2n})$ be the new codings of high and low resolution pedestrians. Our goal is to maximize intra-person similarity, while minimizing inter-person similarity of person images from both views. Thus, we have the following formulations:

$$\begin{aligned}&\max \textstyle\sum_{i,j} z_i W_{ij}^s z_j^\top = \mathrm{tr}(ZW^sZ^\top) \\ &= \mathrm{tr}\Big([V_hX_h, PV_lX_l]\begin{bmatrix} 0 & W_h^s \\ W_l^s & 0 \end{bmatrix}[V_hX_h, PV_lX_l]^\top\Big) \\ &= \mathrm{tr}\Big((PV_lX_l)W_l^s(V_hX_h)^\top + (V_hX_h)W_h^s(PV_lX_l)^\top\Big)\end{aligned} \tag{5}$$

and

$$\begin{aligned}&\min \textstyle\sum_{i,j} z_i W_{ij}^d z_j^\top = \mathrm{tr}(ZW^dZ^\top) \\ &= \mathrm{tr}\Big([V_hX_h, PV_lX_l]\begin{bmatrix} 0 & W_h^d \\ W_l^d & 0 \end{bmatrix}[V_hX_h, PV_lX_l]^\top\Big) \\ &= \mathrm{tr}\Big((PV_lX_l)W_l^d(V_hX_h)^\top + (V_hX_h)W_h^d(PV_lX_l)^\top\Big),\end{aligned} \tag{6}$$

where tr$(\cdot)$ is the trace operation of a matrix. Combining (5)

and (6), we obtain

$$\begin{aligned}&\Omega(V_h, X_h, V_l, X_l, P)\\ =\ &\mathrm{tr}\Big((PV_lX_l)W_l^d(V_hX_h)^\top + (V_hX_h)W_h^d(PV_lX_l)^\top\Big)\\ -\ &\mathrm{tr}\Big((PV_lX_l)W_l^s(V_hX_h)^\top + (V_hX_h)W_h^s(PV_lX_l)^\top\Big).\end{aligned} \tag{7}$$

This crafted graph regularizer unifies intra-person similarity and inter-person dissimilarity constraints, and meanwhile considers the great resolution gaps between cross-view pedestrian images. No parameter is introduced and thus it is expected that the proposed method will have robust performances on diverse scenarios.

With the graph regularizer, we reach our Discriminative Semi-coupled Projective Dictionary Learning (DSPDL) model as:

$$\begin{aligned}\min_{\substack{D_h,D_l,V_h,\\ V_l,P}}\quad & \|X_h - D_hV_hX_h\|_{\mathrm{F}}^2 + \|X_l - D_lV_lX_l\|_{\mathrm{F}}^2\\ & +\lambda_1\Omega(V_h, X_h, V_l, X_l, P) + \lambda_2\|P\|_{\mathrm{F}}^2\\ \text{s.t.}\quad & \left\|d_h^i\right\| \le 1,\ \left\|d_l^i\right\| \le 1,\ i = 1, \cdots, k,\end{aligned} \tag{8}$$

where $\Omega(V_h, X_h, V_l, X_l, P)$ is defined in (7).

Model Comparison

In this part, we compare our DSPDL model and the three most relevant models to highlight our novelties: SCDL (Wang et al. 2012), CPDL (Li, Shao, and Fu 2015), and SLD^{2}L (Jing et al. 2015).

SCDL is developed for photo-sketch synthesis and image super-resolution. It requires large time consumption to solve the sparse coding problem, while our proposed DSPDL model can be solved efficiently due to the adoption of projective dictionary learning technique. Besides, SCDL is developed to uncover the relationship between different image styles of the same instance, so that it essentially neglects the discriminative information among instances. In contrast, DSPDL is designed for person re-ID, we incorporate discriminative information to learn dictionaries which can help distinguish images of the same identities from those of different identities. CPDL is designed for person re-ID, but it neglects the fact that great image resolution divergences could comprise the generalization ability of the learned dictionaries, when directly pushing close the new codings of images of the same person. Moreover, similar as SCDL, CPDL does not incorporate inter-person dissimilarity to enhance the discriminative power of the dictionaries.

SLD^{2}L is more closely related to our DSPDL: both learn semi-coupled dictionaries that are robust with resolution changes. But DSPDL differs from SLD^{2}L in the following aspects: First, we adopt the more efficient, also more powerful, projective dictionary technique; while SLD^{2}L uses the traditional dictionary learning technique. Second, SLD^{2}L segments images into small patches and clusters the patches into groups, and learns a set of dictionary pairs for all corresponding low and high resolution image patch clusters. Due to the cluster-wise dictionary learning strategy, SLD^{2}L is extremely complicated: It comprises of 15 terms, 9 parameters, and dozens of variables. Solving such a complicated model is definitively a time-consuming task. It is also hard to balance all the terms and tune the parameters to the state that is robust in various scenarios. This is why the parameters for SLD^{2}L are set dataset by dataset in the experiments. Different from SLD^{2}L, our proposed DSPDL learns only one pair of dictionaries using the whole images so that our model is much simpler: we have only 5 terms and 2 parameters. Therefore, our model can be easily and efficiently solved, and promise stable performances on different datasets with fixed parameters. Third, we incorporate positive and negative pair information in a parameterless graph embedding fashion, but SLD^{2}L simply uses the combination of several separated terms, whose weights are hard to balance.

Optimization

To facilitate the optimization of our proposed DSPDL model, we introduce three relaxation variables A_h, B_l, and B_h, and use them to replace V_hX_h, V_lX_l, and PV_lX_l, respectively. In this way, we rewrite our model as:

$$\begin{aligned}\min_{\substack{D_h,D_l,V_h,V_l\\ A_h,B_h,B_l,P}}\quad & \Phi(D_h, D_l, V_h, V_l, A_h, B_h, B_l, P) =\\ & \|X_h - D_hA_h\|_{\mathrm{F}}^2 + \|X_l - D_lB_l\|_{\mathrm{F}}^2\\ & +\lambda_1\Omega(A_h, B_h) + \lambda_2\|P\|_{\mathrm{F}}^2\\ & +\alpha\|A_h - V_hX_h\|_{\mathrm{F}}^2 + \alpha\|B_l - V_lX_l\|_{\mathrm{F}}^2\\ & +\beta\|B_h - PB_l\|_{\mathrm{F}}^2\\ \text{s.t.}\quad & \left\|d_h^i\right\| \le 1,\ \left\|d_l^i\right\| \le 1, i = 1, \cdots, k,\end{aligned} \tag{9}$$

where $\Omega(A_h, B_h) = \mathrm{tr}\Big(B_hW_l^dA_h^\top + A_hW_h^dB_h^\top\Big) - \mathrm{tr}\Big(B_hW_l^sA_h^\top + A_hW_h^sB_h^\top\Big)$, and $\alpha = \beta = 10^{-3}$ are two small penalty parameters.

The variables in (9) can be alternatively optimized by fixing the others when optimizing one of them. The step-by-step optimization procedures are as follows.

(1) Update A_h: By keeping only the terms relevant to A_h, we obtain $\min_{A_h}\Phi(A_h) = \|X_h - D_hA_h\|_{\mathrm{F}}^2 + \lambda_1\Omega(A_h, B_h) + \alpha\|A_h - V_hX_h\|_{\mathrm{F}}^2$. Let the derivative of Φ with respect to A_h be zero, *i.e.*, $\frac{\partial\Phi}{\partial A_h} = 0$, we have the closed-form solution of A_h as

$$\begin{aligned}A_h = (2D_h^\top D_h + 2\alpha\mathrm{I})^{-1}(2D_h^\top X_h + 2\alpha V_hX_h + \lambda_1B_hW_l^s\\ +\lambda_1B_hW_h^{s\top} - \lambda_1B_hW_l^d - \lambda_1B_hW_h^{d\top}).\end{aligned} \tag{10}$$

(2) Update B_h: Ignoring irrelevant terms with respective to B_h, the objective function reduces to $\min_{B_h}\Phi(B_h) = \lambda_1\Omega(A_h, B_h) + \beta\|B_h - PB_l\|_{\mathrm{F}}^2$. Setting $\frac{\partial\Phi}{\partial B_h} = 0$, we have

$$\begin{aligned}B_h = \tfrac{1}{2\beta}(\lambda_1A_hW_l^{s\top} + \lambda_1A_hW_h^s - \lambda_1A_hW_l^{d\top}\\ -\lambda_1A_hW_h^d + 2\beta PB_l).\end{aligned} \tag{11}$$

(3) Update B_l: The objective function regarding to B_l can be written as $\min_{B_l}\Phi(B_l) = \|X_l - D_lB_l\|_{\mathrm{F}}^2 + \alpha\|B_l - V_lX_l\|_{\mathrm{F}}^2 + \beta\|B_h - PB_l\|_{\mathrm{F}}^2$. Setting $\frac{\partial\Phi}{\partial B_l} = 0$, we

have

$$B_l = \begin{array}{l}(D_l^\top D_l + \beta P_l^\top P_l + \alpha \mathrm{I})^{-1} \\ (D_l^\top X_l + \alpha V_l X_l + \beta P^\top B_h).\end{array} \quad (12)$$

(4) Update P: The objective function turns to the following form when keeping the terms relevant only to P: $\min_P \Phi(P) = \lambda_2 \|P\|_{\mathrm{F}}^2 + \beta \|B_h - PB_l\|_{\mathrm{F}}^2$. Let $\frac{\partial \Phi}{\partial P} = 0$, the closed-form solution of P is

$$P = \beta B_h B_l^\top (\beta B_l B_l^\top + \lambda_2 \mathrm{I})^{-1}. \quad (13)$$

(5) Update V_h and V_l: The objective function reduces to $\min_{V_h} \Phi(V_h) = \|A_h - V_h X_h\|_{\mathrm{F}}^2$, when removing all the terms irrelevant to V_h. Setting $\frac{\partial \Phi}{\partial V_h} = 0$, we have

$$V_h = A_h X_h^\top (X_h X_h^\top + \theta \mathrm{I})^{-1}. \quad (14)$$

where $\theta = 10^{-3}$ is a small regularization parameter. We can update V_l in the similar way.

(6) Update D_h and D_l: Keeping the terms relevant only to D_h, the objective function becomes

$$\min_{D_h} \Phi(D_h) = \|X_h - D_h A_h\|_{\mathrm{F}}^2 \quad \text{s.t.} \left\| d_h^i \right\| \leq 1, i = 1, \cdots, k. \quad (15)$$

The famous ADMM algorithm can be employed to effectively solve this problem (Gu et al. 2014). Similar solution to D_l can be obtained.

We repeat above procedures until convergence. Finally, we obtain dictionaries D_h, D_l and mapping function P for matching pedestrian images in testing sets T_h and T_l.

Algorithm 1 summarizes the proposed method.

Algorithm 1: DSPDL for person re-ID.

Input: Training image features X_h and X_l, Testing image features T_h and T_l, and parameters λ_1 and λ_2.

Training stage:
Initialize A_h, B_h, B_l, P, D_h, and D_l.
while not converged **do**
1. Fix other variables and update A_h according to (10);
2. Fix other variables and update B_h according to (11);
3. Fix other variables and update B_l according to (12);
4. Fix other variables and update P according to (13);
5. Fix other variables and update V_h, V_l using (14);
6. Fix other variables and update D_h and D_l using ADMM algorithm.

end while
Test stage:
For each $t_l^i \in T_l$ **do**
For each $t_h^j \in T_h$ **do**
1. calculate new coding f_l^i of t_l^i under D_l;
2. calculate new coding g_h^j of t_h^j under D_h;
3. calculate $f_h^i = P f_l^i$;
4. $S_{ij} = \cos(f_h^i, g_h^j)$.

end for
end for

Output: S.

Time Complexity

In the training phase, $A_h, B_{h/l}, P, V_{h/l}$ and $D_{h/l}$ are updated alternatively. The cost of updating A_h in each iteration is $O(k^3+kdn+kn^2)$, that of updating B_h is $O(kn^2+k^2n)$, that of updating B_l is $O(k^3+kdn+k^2n)$, that of updating P is $O(k^3+k^2n)$, that of updating $V_{h/l}$ is $O(d^3+kdn)$, and that of updating $D_{h/l}$ is $O(\tau(kdn+k^3+k^2d+d^2k))$, where τ is the iteration number in ADMM algorithm for updating $D_{h/l}$.

Experiments

We employed three datasets for performance evaluation: the VIPeR dataset (Gray and Tao 2008), the CUHK01 dataset (Zhao, Ouyang, and Wang 2014), and the QMUL-iLIDS dataset (Zheng, Gong, and Xiang 2009). Following previous works (Jing et al. 2015; Wang, Ding, and Fu 2016), we down-sampled all images from one view with the rate 1/8, and kept images from the other view unchanged to simulate the great resolution differences.

We followed the single-shot protocol and took two images for each person for evaluation (Chen, Zheng, and Lai 2015). Following previous works, all the pedestrian pairs were randomly divided into two equal parts, with one part serving for training and the other for testing. We repeated the random partition procedures for 10 times and calculated the average matching rates. For evaluation metric, we adopt the standard cumulated matching characteristics (CMC) curve.

The following person re-ID methods were employed for comparison: metric learning based methods, LFDA (Xiong et al. 2014), PCCA (Mignon and Jurie 2012), XQDA (Liao et al. 2015), and MLAPG (Liao and Li 2015); feature description or learning based methods, USL (Zhao, Ouyang, and Wang 2013b), PRSM (Zhao, Ouyang, and Wang 2013a), MLF (Zhao, Ouyang, and Wang 2014), and CAE (Wang, Ding, and Fu 2016); deep learning based methods, Deeplist (Wang et al. 2016a); dictionary learning based methods, SLD^{2}L (Jing et al. 2015), SCDL (Wang et al. 2012), CPDL (Li, Shao, and Fu 2015), and SSSVM (Zhang et al. 2016). Please note that SCDL was originally developed for photo synthesis and image super-resolution; we adapted it to person re-ID by feeding it with person re-ID data (the same as ours) and tuning the parameters carefully. Among these methods, SLD^{2}L and CMAE specifically targeted to solve the resolution divergence problem in person re-ID. For fair comparison, whenever possible (*i.e.*, the codes are available and the used features can be replaced), we compared with these methods using the same LOMO features (Liao et al. 2015). Otherwise, we compared with the reported results or the results generated by the defaulted feature extraction methods on the same images.

There are two major parameters in our algorithm: λ_1 and λ_2. λ_2 controls the scale of a variable; thus we set it empirically as a small value $\lambda_2 = 0.01$. λ_1 balances the data fidelity terms and the graph regularizer in the objective function, and we will analyze it later.

Experimental Results

VIPeR. The VIPeR dataset is the most widely used datasets for person re-ID. It contains 632 persons with each having

Table 1: Top r matching rates (%) on the VIPeR dataset. The best/second best results are marked in red/blue.

Methods		$r=1$	$r=5$	$r=10$	$r=20$	$r=30$
Feature learning	USL	14.87	36.08	44.30	56.96	62.66
	PRSM	16.20	34.24	45.06	56.96	63.48
	MLF	16.65	32.91	44.87	57.91	64.87
	CAE	25.95	50.00	64.37	79.75	87.34
Metric learning	LFDA	9.57	28.80	43.33	60.94	71.66
	PCCA	8.55	27.39	41.17	58.68	69.79
	XQDA	23.26	53.86	70.03	84.68	90.98
	MLAPG	24.72	54.91	69.62	83.54	90.25
Dictionary learning	CPDL	19.02	49.97	67.12	81.49	89.11
	SCDL	22.56	54.48	69.59	84.21	90.44
	SLD^{2}L	16.86	41.22	58.06	79.00	N/A
	SSSVM	24.53	53.04	65.54	78.89	85.57
D. L.	Deeplist	25.95	58.23	70.57	83.54	87.97
	SPDL	26.27	58.86	73.73	85.45	91.46
	DSPDL	28.51	61.08	76.11	88.13	92.78

a pair of images. All the images are normalized to 128×48 pixels. There are significant viewpoint changes, pose variations, and illumination differences across the cameras. The synthesized great resolution differences make it even harder to match the images of the same identifies. By randomly dividing the dataset into training and testing parts of equal size, *i.e.*, 316 image pairs for training and the other 316 pairs for testing, and repeating the randomly division procedure for ten times, we obtained the average matching rates.

Table 1 shows the top 1, 5, 10, 20 and 30 matching rates of our proposed SPDL and DSPDL, and the baseline methods on the VIPeR dataset. We can see that our proposed SPDL and DSPDL outperform all the other algorithms for all the ranks, often by large margins. Specifically, compared with the best feature learning based method CMAE (Wang, Ding, and Fu 2016), DSPDL gains about 2.5% and 11% improvements for the rank-1 and rank-5 matching rates, respectively. Compared with the metric learning based methods, DSPDL achieves about 4% and 6% gains in the rank-1 and rank-5 matching rates, respectively, over MLAPG, the best method in this category. Our method is based on dictionary learning, and DSPDL gains the rank-1 and rank-10 matching rate promotions of near 4% and 10.5%, respectively, over the best existing dictionary learning based method SSSVM (Zhang et al. 2016). Recent years have witnessed the overwhelming advantages of deep learning on various research domains, like image classification, object detection, etc., mostly due to the richness of labeled data. However, for person re-ID, existing datasets are usually small, so that deep learning based methods perform worse than traditional methods. This is also reflected in our experiments: our proposed DSPDL beats the recent deep learning based person re-ID method, Deeplist (Wang et al. 2016a) by about 2.5% and 5.5% for the rank-1 and rank-10 matching rates, respectively.

We find that though CMAE and SLD^{2}L are designed specifically for matching pedestrian images of great resolution divergences, they surprisingly perform worse than methods which do not target for this degenerated scenario. For example, although CMAE has a small advantage over

Table 2: Top r matching rates (%) on the CUHK01 dataset. The best/second best results are marked in red/blue.

Methods		$r=1$	$r=5$	$r=10$	$r=20$	$r=30$
Metric learning	LFDA	3.00	11.90	19.89	31.54	40.16
	PCCA	7.07	21.71	31.78	44.77	53.88
	XQDA	18.97	43.58	55.80	68.19	75.72
	MLAPG	19.51	40.41	52.47	65.16	72.74
Dict. learning	CPDL	16.50	37.20	48.46	61.42	68.66
	SCDL	14.05	33.37	44.57	56.40	63.87
	SSSVM	17.02	37.90	48.02	59.71	66.89
	SPDL	20.66	44.84	56.79	69.22	76.89
	DSPDL	21.75	46.50	58.27	69.57	77.24

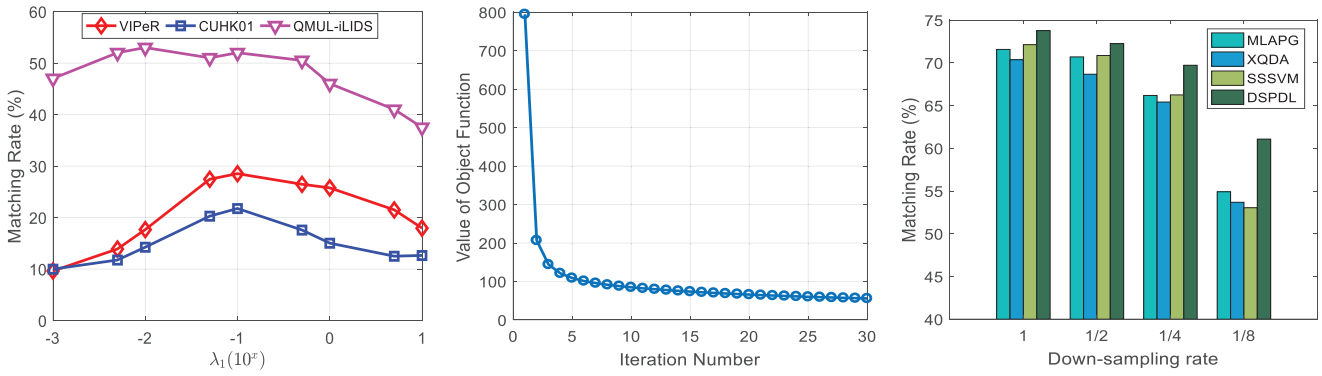
Table 3: Top r matching rates (%) on the QMUL-iLIDS dataset. The best/second best results are marked in red/blue.

Methods		$r=1$	$r=5$	$r=10$	$r=20$
Mtrc. learn.	MLAPG	46.67	77.17	85.67	93.33
	XQDA	50.67	79.17	86.00	93.33
Dictionary learning	SLD^{2}L	33.33	65.00	80.00	90.33
	CPDL	42.33	71.50	83.67	92.83
	SCDL	45.33	79.17	86.50	94.50
	SSSVM	41.33	75.67	88.17	96.67
	SPDL	52.83	79.00	87.50	95.17
	DSPDL	55.17	82.00	90.67	95.67

all the other baseline methods for the rank-1 matching rate, its rank-5 matching rate is much lower than those of XQDA, MLAPG, SCDL and SSSVM. The super-resolution based person re-ID method SLD^{2}L performs even worse. However, our proposed SPDL and DSPDL do perform better than all the baseline methods, and our advantages are significant in many cases. The noticeable advantages of DSPDL over SPDL prove the effectiveness of the proposed discriminative graph regularizer, and incorporating discriminative information among images of different persons helps boost the discriminative power of the learned dictionaries.

CUHK01. The CUHK01 dataset contains pedestrian images of 971 persons in two camera views; each person has two images in both views and we took the first one for use. Images in this dataset are of high resolution, which could be a beneficial factor for person re-ID. By randomly selecting half identities (485 persons) for training and the other half (486 persons) for testing, and repeating the trials for ten times, we obtained the matching rates shown in Table 2. Similar as the observations in the VIPeR dataset, the proposed SPDL and DSPDL beat all the metric learning and dictionary learning based methods, despite that our matching rate promotions are not as significant as we achieve in the VIPeR dataset. The reason could be that the images in this dataset are of higher resolution than those in the VIPeR dataset, so that they are kind of favored by the baseline methods. DSPDL does have some, though not remarkable, matching rate gains over SPDL, which proves that the incorporation of discriminative information does boost the discriminative power of the learned dictionaries to some extent.

QMUL-iLIDS. The QMUL-iLIDS dataset (Berry 2015) consists of 476 images of 119 identities; each person has



(a) (b) (c)

Figure 2: (a) Rank-1 matching rates of the proposed method on the VIPeR, CUHK01, QMUL-iLIDS datasets with different values of λ_1. (b) Convergence curve of the proposed method on the VIPeR dataset. (c) Rank-5 matching rates of several methods on the VIPeR dataset with different down-sampling rates.

four images on average. We randomly selected two images for each person and down-sampled one of them at the rate 1/8, and kept the other unchanged to simulate the resolution difference. Among the 119 images pairs, we randomly selected 59 pairs for training, and the left 60 image pairs for testing. We ran the experiment ten times and calculated the average matching rates. The rank-1, 5, 10, and 20 matching rates are shown in Table 3. We can observe that all the methods achieve high matching rates quickly as the rank increases. This is because the test dataset is small such that for each probe image, there are only 60 images to be queried. We can also observe that our proposed DSPDL beats all the baseline methods by large margins, except a slight inferior to SSSVM for the rank-20 matching rate. SPDL also achieves very competitive results and no baseline method can beat it in the matching rate for all ranks.

Further Analysis

Parameter Analysis. Our proposed DSPDL has an important parameter λ_1, which balances the data fidelity terms and the graph regularization term in the objective function. We varied its value from 10^{-3} to 10, and computed the rank-1 matching rates on all the three experimental datasets. The results are shown in Figure 2(a). We observe that DSPDL reaches the best performances when $\lambda_1 = 0.1$. Thus, λ_1 was set as 0.1 in our method as default.

Convergence Analysis. Figure 2(b) shows the objective function values of DSPDL with the increase of iteration times. We can see that with a small number iterations, our objective function turns to be stable, so that our method has pretty good convergence property.

Impact of Down-Sampling Rate. In all the above experiments, we down-sampled images from one view at the rate $\delta = 1/8$ (images from the other view remained unchanged) to simulate the great resolution differences between cross-view pedestrian images. Here, we further evaluate the performances of methods with different down-sampling rates to show the impacts of different resolution disparities on the matching performances. Figure 2(c) shows the rank-5 matching rates of our proposed DSPDL and several most competitive baseline methods on the VIPeR dataset, with down-sampling rate $\delta = 1$ (*i.e.*, without down-sampling), 1/2, 1/4, and 1/8. We observe that the performances of all methods degenerate as image resolution disparity increases. The performance degenerations are not remarkable with big down-sampling rates (*i.e.*, δ =1/2 and 1/4), but become significant with small one (*i.e.*, $\delta = 1/8$). It is also observed that the proposed DSPDL beats all the compared methods with all down-sampling rates, even in the case that there is no down-sampling (*i.e.*, $\delta = 1$). DSPDL is designed for matching cross-view pedestrian images with great resolution differences, but it owns some advantages over the other methods even in the cases where the resolution differences are not remarkable. We speculate the reason is that the mapping we learned to mitigate the impact of resolution disparities can also reduce the impact of other factors that cause appearance differences, such as different light conditions, human poses, human body occlusions, etc. Meanwhile, we can see that the advantage of DSPDL over the other methods increases as the down-sampling rate decreases. For example, DSPDL beats SSSVM by a margin of 1.5% when $\delta = 1$, but a margin of about 8% when $\delta = 1/8$. This substantiates the advantage of DSPDL on performing person re-ID on images with great resolution divergences.

Conclusion

We presented in this paper a novel algorithm for performing person re-ID on images with great resolution divergences. Our method jointly learns a pair of dictionaries and a mapping function across low and high resolution pedestrian images. The mapping function brings flexibility to the new codings of the images of the same identities for their correspondence, thus leaving more possibility for the dictionaries to maximize the generalization ability. A parameterless graph regularizer is designed to incorporate both positive and negative pair information, so that the discriminative ability of the dictionaries is enhanced. Experimental results on three datasets showed our method outperforms the state-of-the-

art, often by large margins, especially when the resolution disparity among images is large.

Acknowledgment

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

Aharon, M.; Elad, M.; and Bruckstein, A. 2006. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing* 54(11):4311–4322.

Berry, J. M. 2015. *Lobbying for the people: The political behavior of public interest groups*. Princeton University Press.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. *arXiv preprint arXiv:1704.01719*.

Chen, Y.-C.; Zheng, W.-S.; and Lai, J. 2015. Mirror representation for modeling view-specific transform in person re-identification. In *Proc. of IJCAI*.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proc. of CVPR*.

Ding, Z.; Shao, M.; and Fu, Y. 2014. Latent low-rank transfer subspace learning for missing modality recognition. In *Proc. of AAAI*.

Ding, Z.; Shao, M.; and Fu, Y. 2015. Deep low-rank coding for transfer learning. In *Proc. of IJCAI*.

Gray, D., and Tao, H. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. of ECCV*.

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Projective dictionary pair learning for pattern classification. In *Proc. of NIPS*.

Jing, X.-Y.; Zhu, X.; Wu, F.; You, X.; Liu, Q.; Yue, D.; Hu, R.; and Xu, B. 2015. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *Proc. of CVPR*.

Kodirov, E.; Xiang, T.; Fu, Z.; and Gong, S. 2016. Person re-identification by unsupervised l_1 graph learning. In *Proc. of ECCV*.

Köstinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *Proc. of CVPR*.

Li, Z.; Chang, S.; Liang, F.; Huang, T. S.; Cao, L.; and Smith, J. R. 2013. Learning locally-adaptive decision functions for person verification. In *Proc. of CVPR*.

Li, X.; Zheng, W.-S.; Wang, X.; Xiang, T.; and Gong, S. 2015. Multi-scale learning for low-resolution person re-identification. In *Proc. of CVPR*.

Li, L.; Li, S.; and Fu, Y. 2014. Learning low-rank and discriminative dictionary for image classification. *Image and Vision Computing* 32(10):814–823.

Li, S.; Shao, M.; and Fu, Y. 2015. Cross-view projective dictionary learning for person re-identification. In *Proc. of AAAI*.

Liao, S., and Li, S. Z. 2015. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proc. of ICCV*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proc. of CVPR*.

Liu, X.; Song, M.; Zhao, Q.; Tao, D.; Chen, C.; and Bu, J. 2012. Attribute-restricted latent topic model for person re-identification. *Pattern recognition* 45(12):4204–4213.

Matsukawa, T.; Okabe, T.; Suzuki, E.; and Sato, Y. 2016. Hierarchical gaussian descriptor for person re-identification. In *Proc. of CVPR*.

Mignon, A., and Jurie, F. 2012. PCCA: A new approach for distance learning from sparse pairwise constraints. In *Proc. of CVPR*.

Paisitkriangkrai, S.; Shen, C.; and van den Hengel, A. 2015. Learning to rank in person re-identification with metric ensembles. In *Proc. of CVPR*.

Su, C.; Yang, F.; Zhang, S.; Tian, Q.; Davis, L. S.; and Gao, W. 2015. Multi-task learning with low rank attribute embedding for person re-identification. In *Proc. of ICCV*.

Wang, S.; Zhang, L.; Liang, Y.; and Pan, Q. 2012. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *Proc. of CVPR*.

Wang, J.; Wang, Z.; Gao, C.; Sang, N.; and Huang, R. 2016a. Deeplist: Learning deep features with adaptive listwise constraint for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, Z.; Hu, R.; Yu, Y.; Jiang, J.; Liang, C.; and Wang, J. 2016b. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *Proc. of IJCAI*.

Wang, S.; Ding, Z.; and Fu, Y. 2016. Coupled marginalized auto-encoders for cross-domain multi-view learning. In *Proc. of IJCAI*.

Wu, L.; Shen, C.; and van den Hengel, A. 2017. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition* 65:238–250.

Xiong, F.; Gou, M.; Camps, O. I.; and Sznaier, M. 2014. Person re-identification using kernel-based metric learning methods. In *Proc. of ECCV*, 1–16.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Deep metric learning for person re-identification. In *Proc. of ICPR*.

Zhang, Q., and Li, B. 2010. Discriminative k-svd for dictionary learning in face recognition. In *Proc. of CVPR*, 2691–2698.

Zhang, Y.; Li, B.; Lu, H.; Irie, A.; and Ruan, X. 2016. Sample-specific svm learning for person re-identification. In *Proc. of CVPR*.

Zhao, R.; Ouyang, W.; and Wang, X. 2013a. Person re-identification by salience matching. In *Proc. of ICCV*.

Zhao, R.; Ouyang, W.; and Wang, X. 2013b. Unsupervised salience learning for person re-identification. In *Proc. of CVPR*.

Zhao, R.; Ouyang, W.; and Wang, X. 2014. Learning mid-level filters for person re-identification. In *Proc. of CVPR*.

Zheng, W. S.; Gong, S.; and Xiang, T. 2009. Associating groups of people. *Active Range Imaging Dataset for Indoor Surveillance*.