Super-resolution Person Re-identification with Semi-coupled Low-rank Discriminant Dictionary Learning

Xiao-Yuan Jing[1,3], Xiaoke Zhu[1], Fei Wu[1,3], Xinge You[2], Qinglong Liu[1],
Dong Yue[3], Ruimin Hu[4], Baowen Xu[1]

[1]State Key Laboratory of Software Engineering, School of Computer, Wuhan University, China
[2]School of Electronic Information and Communications, Huazhong University of Science and Technology, China
[3]College of Automation, Nanjing University of Posts and Telecommunications, China
[4]National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, China

Abstract

Person re-identification has been widely studied due to its importance in surveillance and forensics applications. In practice, gallery images are high-resolution (HR) while probe images are usually low-resolution (LR) in the identification scenarios with large variation of illumination, weather or quality of cameras. Person re-identification in this kind of scenarios, which we call super-resolution (SR) person re-identification, has not been well studied. In this paper, we propose a semi-coupled low-rank discriminant dictionary learning (SLD^2L) approach for SR person re-identification. For the given training image set which consists of HR gallery and LR probe images, we aim to convert the features of LR images into discriminating HR features. Specifically, our approach learns a pair of HR and LR dictionaries and a mapping from the features of HR gallery images and LR probe images. To ensure that the converted features using the learned dictionaries and mapping have favorable discriminative capability, we design a discriminant term which requires the converted HR features of LR probe images should be close to the features of HR gallery images from the same person, but far away from the features of HR gallery images from different persons. In addition, we apply low-rank regularization in dictionary learning procedure such that the learned dictionaries can well characterize intrinsic feature space of HR and LR images. Experimental results on public datasets demonstrate the effectiveness of SLD^2L.

1. Introduction

Person re-identification is a fundamental task in automated video surveillance and has been widely researched in recent years. Given an image/video of a person taken from one camera, re-identification is the process of identifying the person from images/videos taken from a different camera [2]. Many methods have been presented for person re-identification [34, 33, 13, 15, 17, 21]. These methods can be roughly classified into two categories: ***(1)*** Methods on feature representation [1, 6, 23, 31]: they focus on seeking a distinct and robust feature representation for matching. The Literature [9] seeks a more distinctive representation by exploiting the class information to overcome the large intra-class appearance variations. Literature [24] is presented to solve the feature selection problem. Literature [22] is proposed to learn the most discriminative attribute that characterizes a particular individual, in which saliency detection is utilized to drive automatically the PTZ camera to focus on certain parts of a human body. ***(2)*** Methods on distance learning [10, 14, 19, 35, 36]: they focus on seeking an optimal distance metric for person re-identification. In [10], Hirzer et al. learned a metric from pairs of samples belonging to different cameras using discriminative Mahalanobis metric learning, which can be efficiently solved after some relaxations. In [14], a distance metric is learned based on equivalence constraints from a statistical inference perspective. In [36], Zheng et al. learned a Mahalanobis distance metric with a probabilistic relative distance comparison (RDC) method. Recently, dictionary learning

Figure 1. Super-resolution person re-identification problem.

technique has been introduced into person re-identification. In [17], two coupled dictionaries are learned to bridge the human appearance variations across cameras.

The above methods have relieved some difficulties in person re-identification to some extent. In practice, the captured pedestrian image by cameras usually suffers from low resolution due to insufficient illumination, bad weather, poor quality of camera devices, or even some complex combinations of these factors. Therefore, person re-identification usually should be done between low-resolution and high-resolution pedestrian images. One re-identification scenario is that gallery images are captured with enough illumination at noon while probe images are captured with poor illumination even at night. Another scenario is that gallery images are captured with normal weather while the probe images are captured in rain or snow. Under this kind of scenarios, the captured gallery images are high-resolution (HR) and probe images are low-resolution (LR). We call re-identification under this kind of scenarios the super-resolution (SR) person re-identification. In most cases, person re-identification depends on the visual appearance feature of pedestrian images. However, since low resolution will result in the loss of visual appearance feature, existing person re-identification methods can not well deal with the problem of re-identification between HR and LR pedestrian images. Therefore, it is necessary and meaningful to investigate SR person re-identification. Figure 1 illustrates the SR person re-identification problem.

To improve the quality of LR images, many SR restoration methods [4, 8, 27] have been presented in recent years. Dictionary learning (DL) is an effective feature learning technique in the field of machine learning [11, 30, 37]. Nowadays, some DL based SR restoration methods have been developed, and achieved desirable performance. The coupled dictionary learning model for image super-resolution is proposed in [29, 28] with an assumption that there exist coupled dictionaries of HR and LR images, over which each pair of HR and LR patches have the same sparse representations. The semi-coupled dictionary learning model [25] relaxes the strong assumption of coupled dictionary learning model. And it learns a mapping matrix to capture the relationship of the sparse representations between HR and LR spaces, which brings more flexibility to characterize image structures. However, the above methods are designed for improving human visual perception only, rather than machine perception [32], thus there is no guarantee of identification improvements.

1.1. Motivation

Super-resolution (SR) person re-identification is an important application in practice; however, it has not been well studied. Existing DL based SR restoration methods can improve the quality of LR images by uncovering the relationships between LR and HR images. However, what these methods uncover is the relationship between the features of LR and HR images that are good for human perception, rather than identification.

Motivated by SR restoration works, we intend to uncover the relationship between the features of LR and HR images from the side of person re-identification. The semi-coupled DL technique is introduced to address the SR person re-identification problem. Yet we should address the following two specific problems: (1) Semi-coupled DL is designed for SR restoration, rather than for re-identification. If we directly apply semi-coupled DL for SR person re-identification, the learned dictionary pair and mapping matrix would have no desirable discriminability. (2) Since pedestrian images usually contain noises, the dictionary pair learned by semi-coupled DL directly cannot well characterize the intrinsic feature spaces of LR and HR images.

1.2. Contribution

The main contributions of our work are summarized as following three points:

(1) We are the first attempt to solve the SR person re-identification problem, and propose a semi-coupled low-rank discriminant dictionary learning (SLD^2L) approach. SLD^2L learns a pair of HR and LR dictionaries and a mapping function from the features of HR and LR training images. With the learned dictionary pair and mapping function, the features of LR images can be converted into discriminating HR features.

(2) To ensure that the converted features using the learned dictionaries and mapping have favorable discriminative capability, we design a discriminant term for semi-coupled dictionary learning. The discriminant term can make the converted HR features of LR probe images be close to the features of HR gallery images from the same person, but far away from the features of HR gallery images from different persons.

(3) To ensure that the learned dictionary pair can well characterize the intrinsic feature spaces of LR and HR images, we introduce the low-rank regularization into semi-coupled DL. So far, the low-rank regularization technique has not been applied to semi-coupled DL in existing works.

2. Brief Review of Related Work

In this section, we briefly review the related coupled or semi-coupled dictionary learning methods including SSCDL [17] and SCDL [25]. Then, discussion of the difference between our approach and related methods is given.

2.1. Semi-Supervised Coupled Dictionary Learning (SSCDL) for Person Re-identification

Assume that $x = \{x_1, x_2, \cdots, x_n\}$ and $y = \{y_1, y_2, \cdots, y_m\}$ are two sets of training data

from two different cameras. To bridge human appearances across cameras, SSCDL [17] employs labeled pairs of persons as well as unlabeled persons from the gallery and probe cameras to jointly learn a coupled dictionary pair. Specifically, SSCDL learns two dictionaries D_x and D_y such that the sparse representation $\alpha(x_i)$ in terms of D_x should be the same as $\alpha(y_i)$ in terms of D_y. The objective function of SSCDL is as follows:

$$Q(D_x,D_y,\alpha)=E_{labeled}(D_x,D_y,\alpha^{(s)})+E_{unlabeled}(D_x,\alpha^{(x)}) + E_{unlabeled}(D_y,\alpha^{(y)})$$

where $\alpha^{(s)}$ is the shared coefficient matrix for labeled image pairs and $\alpha^{(x)}$ and $\alpha^{(y)}$ are the coefficient matrices for unlabeled images from two cameras, respectively.

2.2. Semi-Coupled Dictionary Learning (SCDL)

The SCDL [25] model is designed for image super-resolution and photo-sketch synthesis. With the assumption that there exists a dictionary pair over which the representations of two styles have a stable mapping, SCDL simultaneously learns a pair of dictionaries and a mapping function. The objective function of SCDL is as follows:

$$\min_{D_x,D_y,W} \|X-D_x\Lambda_x\|_F^2+\|Y-D_y\Lambda_y\|_F^2+\gamma\|\Lambda_y-W\Lambda_x\|_F^2 + \lambda_x\|\Lambda_x\|_1+\lambda_y\|\Lambda_y\|_1+\lambda_W\|W\|_F^2 \quad s.t.\ \|d_{x,i}\|_{l_2}\le 1, \|d_{y,i}\|_{l_2}\le 1, \forall i$$

where X and Y represent the training datasets formed by the image patch pairs of two different resolutions (or styles). $\gamma, \lambda_x, \lambda_y, \lambda_W$ are regularization parameters to balance the terms in the objective function. $d_{x,i}$, $d_{y,i}$ are the atoms of D_x and D_y, respectively. Λ_x and Λ_y are the coding coefficients. W is the mapping matrix.

2.3. Comparison with Related Works

Compared with SSCDL: SSCDL is designed for person re-identification. And it is based on a strong assumption that there exists a coupled dictionary pair between two cameras, over which images of the same person from different cameras must have the same sparse representation. While our approach is designed for person re-identification between HR and LR pedestrian images. And our approach aims to learn a pair of HR and LR dictionaries, over which the representations of each pair of HR and LR patches have a mapping.

Compared with SCDL: SCDL is designed for image super-resolution and photo-sketch synthesis, rather than identification. The learned dictionaries and mapping have no favorable discriminative capability. While our approach is designed for SR person re-identification. We design a discriminant term to enhance the discriminative capability of the learned dictionary pair and mappings. In addition, we introduce low-rank matrix recovery to semi-coupled DL to better characterize intrinsic feature spaces of HR and LR images.

3. Semi-coupled Low-rank Discriminant Dictionary Learning ($\mathrm{SLD^2L}$)

In this section, we first describe the problem formulation, and then provide the optimization of the proposed approach.

3.1. Problem Formulation

Assume that C_A is a HR pedestrian image set from camera A and C_B is a LR pedestrian image set from camera B, we aim to learn a pair of HR and LR dictionaries and a mapping function between features of HR and LR images, such that the features of LR images in C_B can be converted into discriminating HR features.

To this end, we firstly generate the LR version of C_A by performing down-sampling and smoothing operations, which has the same resolution as C_B and is denoted by C'_A. By this way, the underlying relationship between HR and LR feature spaces can be revealed. Then we exploit semi-coupled DL to learn a pair of HR and LR dictionaries and a mapping matrix between the corresponding features of C_A and C'_A. Since the dictionaries and mapping matrix learned by semi-coupled DL directly don't have discriminative capability, we require that HR features of images in C_B, which are reconstructed using the learned dictionary pair and mapping matrix, should be close to the features of images from the same person in C_A, but far away from the features of images from different persons in C_A.

In practice, low resolution has different influences on different patches, e.g., patches with pure color suffer little influence, while patches with complex texture suffer more influence. Therefore, learning a common mapping function is not enough to catch all the relationships. Intuitively, we can divide images into patches and group patches into several clusters, and then a pair of HR and LR sub-dictionaries and a more stable mapping function can be learned for each cluster. In this paper we group patches in C'_A and C_B using K-means algorithm [7] according to the similarity of patch features. Then, the patches in C_A are grouped according to clustering results of the corresponding patches in C'_A. We require that each cluster-specific sub-dictionary has good representation ability for the patches from the associated cluster but poor representation ability for other clusters. Denote by D_H^i and D_L^i the HR and LR sub-dictionaries of the i^{th} cluster, respectively. And V_i denotes the mapping of the i^{th} cluster. By separately combining HR and LR sub-dictionaries, we can obtain the structured HR and LR dictionaries, namely $D_H=[D_H^1,D_H^2,\cdots,D_H^c]$ and $D_L=[D_L^1,D_L^2,\cdots,D_L^c]$, where c is the number of clusters.

To ensure that the learned sub-dictionary pairs can well characterize the intrinsic feature spaces of HR and LR images, the noises should be separated from patches in the learning process. Considering that patches from the same cluster are linearly correlated, we can employ low-rank

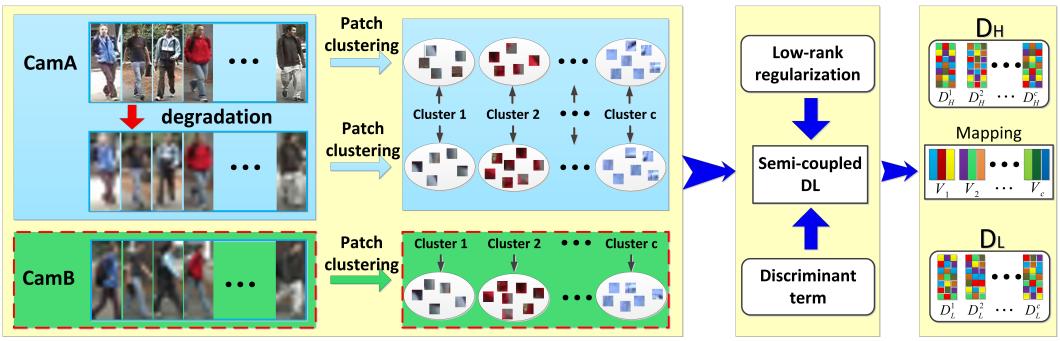

Figure 2. The flowchart of SLD^2L.

matrix recovery technique to separate noises from patches [18, 20, 26]. Figure 2 illustrates the overall flow of SLD^2L.

Let X, X' and Y be the patch sets of C_A, C'_A and C_B, respectively. X_i, X'_i and Y_i separately represent the i^{th} clusters of X, X' and Y, and x_i, x'_i and y_i separately represent the i^{th} patch in X, X' and Y. Denote by a_i, A_i and A the coding coefficients of x_i, X_i and X over D_H, respectively. a'_i, A'_i and A' denote the coding coefficients of x'_i, X'_i and X' over D_L, respectively. b_i, B_i and B separately represent the coding coefficients of y_i, Y_i and Y over D_L. Denote by A_i^j the coding coefficient of X_i over D_H^j. $A_i^{j'}$ and B_i^j are the coding coefficients of X'_i and Y_i over D_L^j, respectively. Let $V = \{V_1, V_2, \cdots, V_c\}$ be the mappings of all clusters.

The objective function of our approach is designed as follows:

$$\min_{D_H, D_L, V} \Phi(D_H, D_L, V, A, A', B) \tag{1}$$

$$s.t.\ X_i = D_H A_i + E_i, X_i = D_H^i A_i^i + E_i,$$
$$X'_i = D_L A'_i + E_j, X'_i = D_L^i A_i^{i'} + E_j,$$
$$Y_i = D_L B_i + E_k, Y_i = D_L^i B_i^i + E_k, i = 1, \cdots, c$$

where

$$\Phi(D_H, D_L, V, A, A', B) = \sum_{i=1}^{c} \{E_{mapping}(V_i, A_i, A'_i) +$$
$$E_{represent}(D_H^i, D_L^i, A_i, A'_i, B_i) + E_{lowrank}(D_H^i, D_L^i) +$$
$$E_{reg}(A_i, A'_i, B_i, V_i, E_i, E_j, E_k)\} + E_{discriminant}(D_H, A, B, V)$$

The constraints mean that the learned dictionaries and sub-dictionaries can well characterize the intrinsic features of training samples. E_i, E_j and E_k represent the separated noises. $E_{mapping}(V_i, A_i, A'_i) = \|A_i - V_i A'_i\|_F^2$ is the mapping fidelity term to represent the mapping error between the coding coefficients of HR and LR image features. $E_{represent}(D_H^i, D_L^i, A_i, A'_i, B_i) = \lambda_1 \sum_{j=1, j\neq i}^{c} \left\|D_H^i A_i^j\right\|_F^2 + \lambda_2 \sum_{j=1, j\neq i}^{c} \left\|D_L^i A_i^{j'}\right\|_F^2 + \lambda_3 \sum_{j=1, j\neq i}^{c} \left\|D_L^i B_i^j\right\|_F^2$ is the sub-dictionary representation capability term to make each sub-dictionary have poor representation ability for other clusters. $E_{lowrank}(D_H^i, D_L^i) = \gamma_1 \left\|D_H^i\right\|_* + \gamma_2 \left\|D_L^i\right\|_*$ is the low-rank regularization term to ensure the learned HR and LR sub-dictionaries being low-rank, where $\|.\|_*$ denotes the nuclear norm of a matrix. $E_{reg}(A_i, A'_i, B_i, V_i, E_i, E_j, E_k) = \|A_i\|_1 + \|A'_i\|_1 + \|B_i\|_1 + \beta_1\|E_i\|_1 + \beta_2\|E_j\|_1 + \beta_3\|E_k\|_1 + \lambda_4\|V_i\|_F^2$ is the regularization term to regularize the coding coefficients and separated noises as well as the mapping matrix. Here, λ_1, λ_2, λ_3, λ_4, γ_1, γ_2, β_1, β_2 and β_3 are balancing factors. $E_{discriminant}(D_H, A, B, V) = \frac{1}{|S|} \sum_{(i,j)\in S} \left\|z_A^i - z_B^j\right\|_2^2 - \frac{1}{|D|} \sum_{(i,j)\in D} \left\|z_A^i - z_B^j\right\|_2^2$ is a discriminant term to ensure that the reconstructed features have good discriminability. S and D are the collections of positive and negative pairs, respectively. $|\cdot|$ represents the size of a collection. $z_A^i = \{D_H a_1^i; D_H a_2^i; \cdots; D_H a_n^i\}$ represents the reconstructed feature of the i^{th} image in C_A, $z_B^i = \{D_H V_{v(1)} b_1^i; D_H V_{v(2)} b_2^i; \cdots; D_H V_{v(n)} b_n^i\}$ represents the converted HR feature of the i^{th} image in C_B. n is the number of patches per image, and $v(i)$ represents the cluster index of i^{th} patch.

3.2. The Optimization of SLD²L

There is no theoretical guarantee that our objective function in Formula 1 is jointly convex to (D_H, D_L, V); however, it is convex with respect to each of D_H, D_L, V when the others are fixed. To tackle the energy-minimization of our objective function, we divide the objective function in Formula 1 into three sub-problems: (1) updating coding coefficients by fixing D_H, D_L and V; (2) updating D_H and D_L by fixing A, A', B and V; (3) updating V by fixing D_H, D_L and coding coefficients.

(1) Updating the representation coefficients. Here, we update representation coefficients by fixing D_H, D_L and V. First, we should initialize the dictionary pair and the mapping function. The PCA basis is employed to initialize each sub-dictionary. Similar to [25], the mapping function of each cluster is simply initialized as the identity matrix.

The sparse coding coefficients a_i , a'_i and b_i can be cal-

culated as follows:

$$\begin{aligned}&\min_{a_i,e_i}\|a_i\|_1+\beta_1\|e_i\|_1+\left\|a_i-V_{v(i)}a'_i\right\|_F^2+d(a_i)\\&s.t.\ x_i=D_Ha_i+e_i\end{aligned}\tag{2}$$

$$\begin{aligned}&\min_{a'_i,e_j}\|a'_i\|_1+\beta_2\|e_j\|_1+\left\|a_i-V_{v(i)}a'_i\right\|_F^2\\&s.t.\ x'_i=D_La'_i+e_j\end{aligned}\tag{3}$$

$$\begin{aligned}&\min_{b_i,e_k}\|b_i\|_1+\beta_3\|e_k\|_1+d(b_i)\\&s.t.\ y_i=D_Lb_i+e_k\end{aligned}\tag{4}$$

where $d(a_i)$ and $d(b_i)$ represent the discriminant terms associated with a_i and b_i in $E_{discriminant}(D_H,A,B,V)$, respectively.

We first convert Formula 2 to the following equivalent problem:

$$\begin{aligned}&\min_{a_i,e_i}\|Z\|_1+\beta\|J\|_1+l(a_i)\\&s.t.\ x_i=D_Ha_i+e_i,Z=a_i,J=e_i,\end{aligned}\tag{5}$$

where $l(a_i)=\left\|a_i-V_{v(i)}a'_i\right\|_F^2+d(a_i)$. Formula 5 can be addressed by solving the following Augmented Lagrange Multiplier problem [16]:

$$\begin{aligned}&\min_{a_i,e_i}\|Z\|_1+\beta\|J\|_1+l(a_i)+tr[T_1^t(x_i-D_Ha_i-e_i)]\\&+tr[T_2^t(a_i-Z)]+tr[T_3^t(e_i-J)]+\\&\frac{\mu}{2}(\|x_i-D_Ha_i-e_i\|_F^2+\|a_i-Z\|_F^2+\|e_i-J\|_F^2)\end{aligned}\tag{6}$$

where T_1, T_2 and T_3 are Lagrange multipliers and μ is a positive penalty parameter. T_1,T_2,T_3 and μ can be obtained using the similar means as [16, 18]. Formulas 3 and 4 can be solved in the same way as Formula 2.

(2) Updating dictionary pair. Here, we update D_H^i and D_L^i one by one by fixing D_H^j, D_L^j, A, A', B and V, $j\neq i$. If D_H^i is updated, the corresponding coefficients A_i^i for coding X_i should be updated to meet the condition $X_i=D_H^iA_i^i+E_i$. Similarly, $A_i^{i'}$ and B_i^i should also be updated to meet the conditions $X'_i=D_L^iA_i^{i'}+E_j$ and $Y_i=D_L^iB_i^i+E_k$, respectively. So, A_i^i, $A_i^{i'}$ and B_i^i are updated in this step. Let $U_i=[X'_i,Y_i]$, $W_i^i=[A_i^{i'},B_i^i]$, $W_i^j=[A_i^{j'},B_i^j]$, $E=[E_j,E_k]$. D_H^i and D_L^i can be updated as follows:

$$\begin{aligned}&\min_{D_H^i}\left\|A_i^i\right\|_1+\gamma_1\left\|D_H^i\right\|_*+\beta_1\|E_i\|_1+\lambda_1\sum_{j=1,j\neq i}^{c}\left\|D_H^iA_i^j\right\|_F^2+d(D_H^i)\\&s.t.\ X_i=D_H^iA_i^i+E_i,\ \left\|d_H^j\right\|_2^2\leq 1,j=1,2,\cdots,K\end{aligned}\tag{7}$$

$$\begin{aligned}&\min_{D_L^i}\left\|W_i^i\right\|_1+\gamma_2\left\|D_L^i\right\|_*+\beta_2\|E\|_1+\lambda_2\sum_{j=1,j\neq i}^{c}\left\|D_L^iW_i^j\right\|_F^2\\&s.t.\ U_i=D_L^iW_i^i+E,\ \left\|d_L^j\right\|_2^2\leq 1,j=1,2,\cdots,K\end{aligned}\tag{8}$$

where d_H^j and d_L^j are dictionary atoms, K is the number of atoms in each sub-dictionary. $d(D_H^i)$ represents the discriminant term associated with D_H^i in $E_{discriminant}(D_H,A,B,V)$:

$$d(D_H^i)=\frac{1}{|S|}\sum_{(p,q)\in S}\sum_{k=1}^{n}\|f_i(p,q,k)\|^2-\frac{1}{|D|}\sum_{(p,q)\in D}\sum_{k=1}^{n}\|f_i(p,q,k)\|^2$$

where

$$f_i(p,q,k)=D_H^i(a_{p,k}^i-v(k)b_{q,k}^i)+\sum_{j=1,j\neq i}^{c}D_H^j(a_{p,k}^j-v(k)b_{q,k}^j).$$

$a_{p,k}^i$ and $b_{p,k}^i$ are representation coefficients of the k^{th} patch in the p^{th} and q^{th} images, respectively. n is the number of patches per image.

We convert Formula 7 to the following equivalent problem:

$$\begin{aligned}&\min_{D_H^i,A_i^i,E_i}\|Z\|_1+\gamma_1\|J\|_*+\beta_1\|E_i\|_1+\lambda l(D_H^i)\\&s.t.\ X_i=D_H^iA_i^i+E_i,J=D_H^i,Z=A_i^i,\\&\left\|d_H^j\right\|_2^2\leq 1,j=1,2,\cdots,K\end{aligned}\tag{9}$$

where $l(D_H^i)=\lambda_1\sum_{j=1,j\neq i}^{C}\left\|D_H^iA_i^j\right\|_F^2+d(D_H)$. Formula 9 can be addressed by solving the following Augmented Lagrange Multiplier problem:

$$\begin{aligned}&\min_{D_H^i,A_i^i,E_i}\|Z\|_1+\gamma_1\|J\|_*+\beta\|E_i\|_1+\lambda l(D_H^i)+tr[T_1^t(D_H^i-J)]\\&+tr[T_2^t(A_i^i-Z)]+tr[T_3^t(X_i-D_H^iA_i^i-E_i)]\\&+\frac{\mu}{2}(\left\|D_H^i-J\right\|_F^2+\left\|A_i^i-Z\right\|_F^2+\left\|X_i-D_H^iA_i^i-E_i\right\|_F^2)\end{aligned}\tag{10}$$

where T_1, T_2 and T_3 are Lagrange multipliers and μ is a positive penalty parameter. Formula 8 can be solved in the same way as Formula 7.

(3) Updating mapping function. By fixing D_H^i, D_L^i, A, A', B and V_j, $j\neq i$, V_i can be updated as follows:

$$\min_{V_i}\|A_i-V_iA'_i\|_F^2+\lambda_4\|V_i\|_F^2+d(V_i)\tag{11}$$

where $d(V_i)$ represents the discriminant term associated with V_i in $E_{discriminant}(D_H,A,B,V)$. Formula 11 is a ridge regression problem and the solution can be analytically derived as:

$$V_i=\left(A_x^iA_y^{it}-d'(V_i)\right)(A_y^iA_y^{it}+\lambda_4I)^{-1}$$

where $d'(V_i)$ is the derivative of $d(V_i)$ with respect to V_i. I is an identity matrix. The optimization of our approach is summarized as Algorithm 1.

Algorithm 1 The optimization of SLD^2L

Input: Data X, X' and Y
Output: Dictionaries D_H and D_L, Mapping V
Initialize: Dictionaries D_H and D_L, Mapping V
while $j < m$ (max iteration number) **do**

1. Fix D_H, D_L and V, and update A, A' and B according to Formulas (2), (3) and (4), respectively.

2. Fix A, A', B and V, and update D_H, D_L according to Formulas (7) and (8), respectively.

3. Fix D_H, D_L, A, A' and B, and update V according to Formula (11).

4. Break if the value of $\Phi(D_H, D_L, V, A, A', B)$ in adjacent iterations are close enough.

end while

4. Super-Resolution Person Re-identification with Learned Dictionaries and Mappings

This section elaborates on the SR person re-identification with the learned D_H, D_L and $\{V_1, V_2, \cdots, V_c\}$. Assume that $G = \{g_1, g_2, \cdots, g_m\}$ is a HR gallery set and p is a probe image, the process of re-identifying p in G can be described as follows:

(1) Converting feature of LR probe image into HR feature. Firstly, we divide p into n patches. Denote by y_i the feature of i^{th} patch. We compute the representation coefficient of i^{th} patch over D_L, namely a_i, as:

$$\min_{a_i, e_i} \|a_i\|_1 + \beta \|e_i\|_1 \; s.t. \; y_i = D_L a_i + e_i$$

where e_i denotes the noise. Then, the cluster index j of i^{th} patch can be computed as follows:

$$\min_j \left\| y_i - D_L^j a_i^j - e_i \right\|_F^2, j = 1, 2, \cdots, c$$

With the corresponding mapping of the cluster j and D_H, y_i can be converted into HR feature as: $y_i^H = D_H V_j a_i$. Finally, we concatenate the HR features of n patches as the total HR feature of p.

(2) Computing reconstructed features of gallery images. For the i^{th} gallery image g_i, we divide g_i into n patches. Denote by $\{x_1, x_2, \cdots, x_n\}$ the features of n patches. Then we compute the representation coefficient of each patch over D_H as:

$$\min_{a^i, e_i} \|a_i\|_1 + \beta \|e_i\|_1 s.t. \; x_i = D_H a_i + e_i$$

We take $D_H a_i$ as the new feature of x_i. The feature of g_i can be obtained by concatenating the new features of n patches.

(3) Re-identifying the probe image in gallery images. Firstly, we compute the distance between p and gallery images using the obtained features. Then the nearest neighbor classifier is employed for matching, and the gallery image with the smallest distance is the true match of p.

5. Experimental Results

5.1. Compared Methods and Experimental Settings

To evaluate the effectiveness of the proposed approach, we compare SLD^2L with several state-of-the-art person re-identification methods including **SSCDL** [17], **RDC** [36], **RPLM** [10] and **KISSME** [14]. For methods RDC and KISSME, we perform experiment with the code provided by the original authors. For SSCDL and RPLM, the authors don't provide the code, so we re-implement these methods by carefully tuning their parameters.

For each patch, we extract the HSV and LAB histograms and LBP descriptor. All the feature descriptors are concatenated together to represent the patch. For fair comparison, all compared methods use the same data and experimental settings as those of our approach. We repeat each experiment 10 times and compare the average results of all methods in the range of the first 50 ranks on the VIPeR and PRID datasets, and the first 30 ranks on the i-LIDS dataset. To further analyze the advantages of our approach, we compare the matching rates with different down-sampling rates.

Parameter Settings. There are 9 parameters in our approach including $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \gamma_1, \gamma_2, \beta_1, \beta_2, \beta_3$. In experiment, we find that the changes of λ_1, λ_2, λ_3 and λ_4 have little influence on identification results. Thus, we set them as 1 for all the three datasets. $\gamma_1, \gamma_2, \beta_1, \beta_2$ and β_3 are set by using the 5-fold cross validation technique with training data. Specifically, they are set as $\gamma_1 = 1, \gamma_2 = 1.5, \beta_1 = 0.1, \beta_2 = 0.1, \beta_3 = 0.1$ for VIPeR; $\gamma_1 = 1, \gamma_2 = 1, \beta_1 = 0.05, \beta_2 = 0.1, \beta_3 = 0.1$ for i-LIDS; and $\gamma_1 = 1, \gamma_2 = 2, \beta_1 = 0.15, \beta_2 = 0.2, \beta_3 = 0.2$ for PRID. In addition, we experimentally set the number of clusters as 64 (when it reaches 64, the performance of SLD^2L begins to stabilize), image patch size as 8×8 and the number of atoms in each sub-dictionary as 48.

5.2. Evaluation on the VIPeR Dataset

The VIPeR dataset [5] contains 632 persons with each having a pair of images captured from two outdoor cameras. Similar to [12] , down-sampling and smoothing operations are performed on all images from camera B to generate LR images. 632 images from camera A and the generated 632 LR images from camera B form 632 image pairs. All image pairs are randomly split into two sets (316 pairs for each set) with one for training and the other for testing. We take images from camera A in the testing set as the HR gallery image set, and use the LR images from camera B in the testing set to construct the LR probe set.

Figure 3 (a) and Table 1 report the matching results of all compared methods at sampling rate of $1/8$. We can see that the matching results of all competing methods are significantly lower than those provided in the original papers. The reason is that low resolution results in the loss of use-

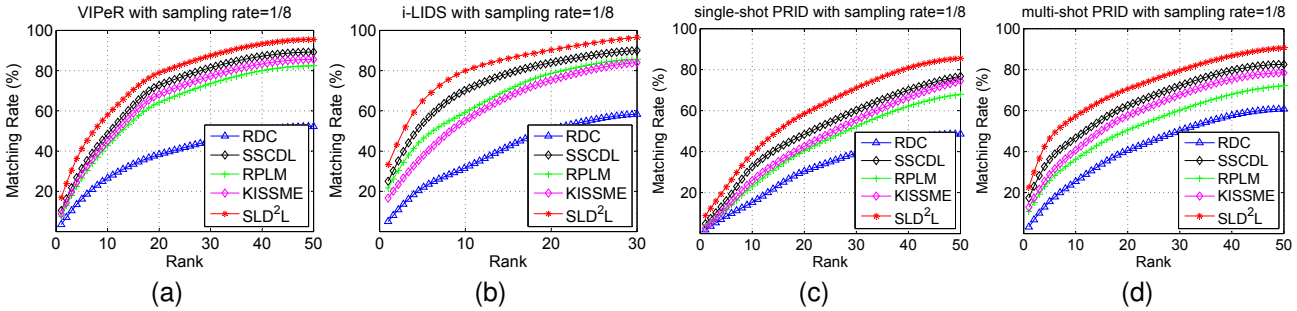

Figure 3. Experimental results on the (a) VIPeR, (b) i-LIDS, (c) single-shot PRID and (d) multi-shot PRID datasets.

Table 1. Top r ranked matching rates (%) on the VIPeR dataset with sampling rate of 1/8.

Method	r=1	r=5	r=10	r=20	r=50
RDC	3.48	16.14	26.58	38.29	52.22
SSCDL	10.44	31.33	48.42	72.78	89.24
RPLM	7.59	26.58	42.72	64.24	82.43
KISSME	8.74	28.58	45.02	68.20	85.62
SLD²L	**16.86**	**41.22**	**58.06**	**79.00**	**95.57**

ful information and these methods cannot work well in this scenario. The experimental results of SLD^2L always outperform these related methods, which demonstrates the effectiveness of our approach for SR person re-identification.

5.3. Evaluation on the i-LIDS Dataset

The i-LIDS dataset [35] consists of 119 persons with a total of 476 shots captured by multiple non-overlapping cameras with an average of four images for each person. We randomly select one image from each person as the HR image set, and select another image from each person and perform down-sampling and smoothing operations to generate LR image set. Thus 119 image pairs are formed. Then, 59 image pairs are randomly selected for training, and the remaining 60 image pairs are used for testing. We further select the HR images in the test set to constitute the gallery set. All the remaining LR images in the test set are used to constitute the LR probe set.

Table 2. Top r ranked matching rates (%) on the i-LIDS dataset with sampling rate of 1/8.

Method	r=1	r=5	r=10	r=20	r=30
RDC	5.00	21.67	31.67	50.67	58.33
SSCDL	25.00	53.67	70.33	84.00	90.00
RPLM	21.67	46.00	59.33	78.67	85.67
KISSME	16.67	37.33	55.33	75.33	84.00
SLD²L	**33.33**	**65.00**	**80.00**	**90.33**	**96.67**

Figure 3 (b) and Table 2 provide the matching results of all compared methods at sampling rate of $1/8$. It can be seen that SLD^2L constantly achieves the best results. For matching rates at rank 1, SLD^2L improves at least $8.33\%(33.33\% - 25.00\%)$.

5.4. Evaluation on the PRID 2011 Dataset

The PRID 2011 dataset [9] consists of person images recorded from two different cameras. Camera A contains 385 persons and camera B contains 749 persons, with 200 persons appearing in both views. Two scenarios are provided: single-shot and multi-shot. In the single-shot case, there are two images per person (one image per camera view). In the multi-shot case, there are multiple images per person (at least 5 images per camera view). In this paper, we evaluate our approach on both scenarios. In both scenarios, we take images from camera B as HR image set, while images from camera A are used to generate LR image set by performing down-sampling and smoothing operations.

In the single-shot case, the total 200 HR and LR image pairs are randomly divided into a training set of 100 pairs, and a test set with the other 100 pairs. For the test set, we further select the 100 images from camera A to construct the LR probe set, and use all images of camera B except the 100 training samples to construct the gallery set.

Table 3. Top r ranked matching rates (%) on the single-shot PRID dataset with sampling rate of 1/8.

Method	r=1	r=5	r=10	r=20	r=50
RDC	1.80	8.40	15.20	30.40	48.60
SSCDL	4.80	16.00	32.60	48.40	76.80
RPLM	3.90	11.80	23.20	40.40	68.00
KISSME	2.70	12.70	25.90	42.60	74.50
SLD²L	**8.80**	**22.80**	**39.20**	**58.60**	**85.60**

Figure 3 (c) and Table 3 report the results on the single-shot PRID dataset. Compared with RDC, SSCDL, RPLM and KISSME, our approach achieves better results.

In the multi-shot case, we randomly select 5 images per person for each camera view. 200 persons appearing in both views are divided into two sets with equal size, 100 persons for training and the other 100 persons for testing. We further select 500 images from camera B in the test set to constitute the HR gallery set, and the remaining 500 images from camera A are used to construct the LR probe set.

Table 4. Top r ranked matching rates (%) on the multi-shot PRID dataset with sampling rate of 1/8.

Method	r=1	r=5	r=10	r=20	r=50
RDC	3.20	15.60	25.40	40.60	60.80
SSCDL	17.40	36.20	46.70	62.50	82.60
RPLM	10.80	25.90	36.30	50.60	72.20
KISSME	12.80	28.80	40.70	57.60	78.50
SLD^{2}L	**22.60**	**46.60**	**57.40**	**70.70**	**90.80**

Figure 3 (d) and Table 4 provide the matching results of all compared methods at sampling rate of $1/8$. It can be seen that our approach constantly achieves the best results. The results indicate that the proposed approach is also applicable to multi-shot SR person re-identification.

To evaluate the **statistical significance of difference** between SLD^2L and compared methods, we conduct the Mcnemars test [3] on all datasets. The test results show that SLD^2L makes a statistically significant difference in comparison with related methods. Due to the limited space, we do not provide the results in detail.

5.5. Effect of Each Term in the Objective Function

In this experiment, we investigate the effects of $E_{represent}$, $E_{lowrank}$ and $E_{discriminant}$ by performing SLD^2L with/without each term, and evaluate $E_{mapping}$ by performing SLD^2L with each mapping matrix being set as the identity (i.e., coupled DL version of our approach). Figure 4 provides the rank 1 matching results of our approach with/without each term (the coupled DL version for $E_{mapping}$) on the VIPeR dataset with sampling rate of $1/8$. We can see that the matching rates are improved by utilizing these terms, which means that each term in the objective function plays its due role.

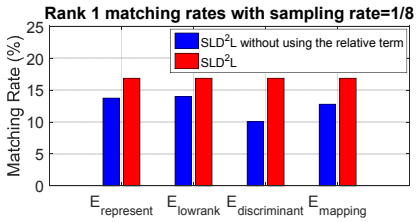

Figure 4. Results of SLD^2L with/without each term (the coupled DL version for $E_{mapping}$) on the VIPeR dataset.

5.6. Computational Cost

In this paper, the computational cost of our approach is proportional to the size of the dictionary, the number of patches. Our approach is ran on a computer with an Intel I7 quad-core 3.4GHZ CPU and 8GB memory. Figure 5 shows the convergence effect of SLD^2L on VIPeR dataset, which indicates that SLD^2L can converge rapidly. The computation time of learning dictionaries and mappings on VIPeR dataset is about 2 hours. However, the testing time for one probe image is less than 0.6 seconds. On other two datasets, our approach can also converge with an acceptable time.

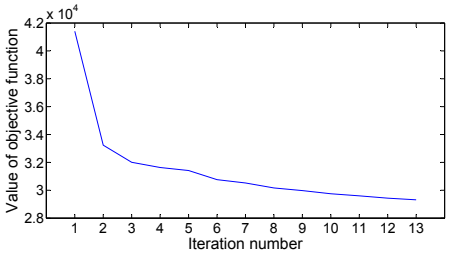

Figure 5. Convergence effect of SLD^2L on VIPeR dataset.

5.7. Impact of Down-sampling Rate

In this experiment, we study the impact of down-sampling rate on each compared method. All compared methods are executed with different down-sampling rates including 1 (i.e., without down-sampling), $1/2, 1/4, 1/8$ and $1/12$. Table 5 gives the corresponding results on the VIPeR dataset. We can see that the larger the down-sampling rate is, the lower the re-identification rate is for every method. At all down-sampling rates, our approach obtains the highest re-identification rate. With the down-sampling rate changing from 1 to $1/12$, the difference between the maximal and minimal matching rates of our approach is 13.27% (i.e., $70.13\% - 56.86\%$) for VIPeR, which shows that our approach is more stable than other methods. Similar effects exist on other two datasets.

Table 5. Rank 10 results (%) with different down-sampling rates on the VIPeR dataset.

Methods	1	1/2	1/4	1/8	1/12
RDC	54.37	30.35	28.16	26.58	21.62
SSCDL	68.16	56.58	52.37	48.42	44.78
RPLM	64.31	49.99	46.54	42.72	35.52
KISSME	62.28	51.90	48.42	45.02	40.84
SLD^{2}L	70.13	60.22	59.13	58.06	56.86

6. Conclusion

We propose the SLD^2L approach for the SR person re-identification problem. It can jointly learn a dictionary pair and a mapping function from HR gallery images and LR probe images. With the designed discriminant term, the learned dictionary pair and mapping have favorable discriminative capability. By applying the designed low-rank regularization on sub-dictionaries, the influence of noises contained in patches can be effectively reduced. With the learned dictionary pair and mapping, features of LR images can be converted into discriminating HR features.

Experimental results on three public datasets demonstrate the effectiveness of the proposed approach for super-resolution person re-identification problem.

Acknowledgement

The work described in this paper was supported by the National Nature Science Foundation of China under Project Nos. 61272273, 61233011, 61231015, 61172173, 61272203, 61374055, the National High Technology Research and Development Program of China (863 Program) under Project Nos. 2015AA016306, 2013AA014602, the Internet of Things Development Funding Project of Ministry of industry in 2013 (No. 25), and the Research Project of NJUPT (XJKY14016).

References

[1] S. Bak, E. Corvee, F. Brémond, and M. Thonnat. Person re-identification using haar-based and dcd-based signature. In *Advanced Video and Signal Based Surveillance (AVSS), IEEE Conference on*, pages 1–8, 2010.

[2] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014.

[3] B. A. Draper, W. S. Yambor, and J. R. Beveridge. Analyzing pca-based face recognition algorithms: Eigenvector selection and distance measures. *Empirical Evaluation Methods in Computer Vision, Singapore*, pages 1–15, 2002.

[4] X. Gao, K. Zhang, D. Tao, and X. Li. Image super-resolution with sparse neighbor embedding. *Image Processing, IEEE Transactions on*, 21(7):3194–3205, 2012.

[5] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Performance Evaluation of Tracking and Surveillance, IEEE workshop on*, 2007.

[6] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. 2008.

[7] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108, 1979.

[8] L. He, H. Qi, and R. Zaretzki. Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution. In *CVPR, IEEE Conference on*, pages 345–352, 2013.

[9] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. 2011.

[10] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, pages 780–793. 2012.

[11] D.-A. Huang and Y.-C. F. Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *ICCV, IEEE Conference on*, pages 2496–2503, 2013.

[12] H. Huang and H. He. Super-resolution method for face recognition using nonlinear mappings on coherent features. *Neural Networks, IEEE Transactions on*, 22(1):121–130, 2011.

[13] A. Ilyas, M. Scuturici, and S. Miguet. Inter-camera color calibration for object re-identification and tracking. In *Soft Computing and Pattern Recognition (SoCPaR), IEEE Conference on*, pages 188–193, 2010.

[14] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR, IEEE Conference on*, pages 2288–2295, 2012.

[15] P. Kumar and K. Dogancay. Analysis of brightness transfer function for matching targets across networked cameras. In *Digital Image Computing Techniques and Applications (DICTA), IEEE Conference on*, pages 250–255, 2011.

[16] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[17] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR, IEEE Conference on*, pages 3550–3557, 2014.

[18] L. Ma, C. Wang, B. Xiao, and W. Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In *CVPR, IEEE Conference on*, pages 2586–2593, 2012.

[19] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *Image Processing, IEEE Transactions on*, 23(8):3656–3670, 2014.

[20] H. Peng, B. Li, R. Ji, W. Hu, W. Xiong, and C. Lang. Salient object detection via low-rank and structured sparse matrix decomposition. In *AAAI*, pages 796–802, 2013.

[21] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *BMVC*, volume 8, pages 164–1, 2008.

[22] P. Salvagnini, L. Bazzani, M. Cristani, and V. Murino. Person re-identification with a ptz camera: An introductory study. In *ICIP*, pages 3552–3556, 2013.

[23] W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *Computer Graphics and Image Processing (SIBGRAPI), Brazilian Symposium on*, pages 322–329, 2009.

[24] S. Tahir and A. Cavallaro. Cost-effective features for reidentification in camera networks. *Circuits and Systems for Video Technology, IEEE Transactions on*, 24(8):1362–1374, 2014.

[25] S. Wang, L. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *CVPR, IEEE Conference on*, pages 2216–2223, 2012.

[26] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, pages 2080–2088, 2009.

[27] J. Yang, Z. Lin, and S. Cohen. Fast image super-resolution based on in-place example regression. In *CVPR, IEEE Conference on*, pages 1059–1066, 2013.

[28] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *Image Processing, IEEE Transactions on*, 21(8):3467–3478, 2012.

[29] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *Image Processing, IEEE Transactions on*, 19(11):2861–2873, 2010.

[30] M. Yang, D. Zhang, and X. Feng. Fisher discrimination dictionary learning for sparse representation. In *ICCV, IEEE Conference on*, pages 543–550, 2011.

[31] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *ECCV*, pages 536–551. 2014.

[32] H. Zhang, J. Yang, Y. Zhang, N. M. Nasrabadi, and T. S. Huang. Close the loop: Joint blind image restoration and recognition with sparse representation prior. In *ICCV, IEEE Conference on*, pages 770–777, 2011.

[33] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV, IEEE Conference on*, pages 2528–2535, 2013.

[34] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR, IEEE Conference on*, pages 3586–3593, 2013.

[35] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, pages 23.1–23.11, 2009.

[36] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(3):653–668, 2013.

[37] Y. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, pages 1070–1076, 2013.