Semi-Supervised Coupled Dictionary Learning for Person Re-identification

Xiao Liu[1], Mingli Song[1], Dacheng Tao[2], Xingchen Zhou[1], Chun Chen[1] and Jiajun Bu[1]
[1]*Zhejiang Provincial Key Laboratory of Service Robot, Zhejiang University, China*
{ender_liux, brooksong, zhou_xingchen, chenc, bjj}@zju.edu.cn

[2]*Centre for Quantum Computation and Intelligent Systems*
Faculty of Engineering and Information Technology
University of Technology, Sydney, Australia
dacheng.tao@uts.edu.au

Abstract

The desirability of being able to search for specific persons in surveillance videos captured by different cameras has increasingly motivated interest in the problem of person re-identification, which is a critical yet under-addressed challenge in multi-camera tracking systems. The main difficulty of person re-identification arises from the variations in human appearances from different camera views. In this paper, to bridge the human appearance variations across cameras, two coupled dictionaries that relate to the gallery and probe cameras are jointly learned in the training phase from both labeled and unlabeled images. The labeled training images carry the relationship between features from different cameras, and the abundant unlabeled training images are introduced to exploit the geometry of the marginal distribution for obtaining robust sparse representation. In the testing phase, the feature of each target image from the probe camera is first encoded by the sparse representation and then recovered in the feature space spanned by the images from the gallery camera. The features of the same person from different cameras are similar following the above transformation. Experimental results on publicly available datasets demonstrate the superiority of our method.

1. Introduction

Many multi-camera surveillance-based applications rely on the ability to determine whether a present target has been observed from other cameras. Given a number of images of interest captured by the probe camera, a person re-identification system aims to find all occurrences of these targets from the gallery camera to identify these probe targets. Since the camera views of the surveillance system might be non-overlapped, the system has to rely solely on the visual appearances of the persons most of the time.

The main difficulty of person re-identification arises from the severe changes (e.g. view angle and lighting conditions) from different views that can cause significant variations in appearance, so directly matching the features of person images from different cameras is unreliable due to feature misalignment or even missing features. Early methods such as [21, 22] relies on conventional face recognition technology and has problems dealing with low-quality videos and irregular views. Some studies have investigated seeking a more distinct and reliable low-level feature representation of human appearances, e.g. stripe based rigid blobs [3], spatiotemporal over-segmentation [8], weighted consistent region [17], principle axis-based correspondence [11], and symmetry-driven accumulation [7]. However, it is extremely difficult to compute both distinct and reliable local low-level features under severe changes in different camera views. The same problem arises with feature selection-based methods [10, 20]: if using all features is not reliable, we cannot expect that simply using partial features will lead to substantial performance improvement.

Some other studies [6, 28] have addressed person re-identification as a distance learning problem and show significant improvement in performance. Zheng et al. [28] introduced a Relative Distance Comparison (RDC) model to maximize the probability of a pair of true matches having a smaller distance than a wrongly match pair. The large margin nearest neighbor algorithm is used in [6] to learn the most effective metric to match data from arbitrary cameras. Nevertheless, the previously mentioned distance learning-based methods typically require enormous labeled target pairs which may not be sufficient in practice.

To bridge the human appearance variations across cameras, we present an efficient semi-supervised coupled dictionary learning method for person re-identification in this paper. Our method is inspired by the research on local linear function approximation that a sparse linear combination of

1063-6919/14 $31.00 © 2014 IEEE
DOI 10.1109/CVPR.2014.454

elements from an appropriately chosen dictionary can well-represent the intrinsic structures of features [26]. Based on this observation, it is reasonable to assume that the same person captured from different cameras have the same (but unknown) intrinsic structure, which can be characterized by the jointly learned dictionary in the training phase. Compared to conventional distance learning-based methods, our method requires only a small number of labeled images to carry the relationship between appearance features from different cameras but introduces abundant unlabeled training images to exploit the geometry of the marginal distribution for obtaining better sparse representation. In the testing phase, the feature of each target image from the probe camera is first encoded by the sparse representation and then recovered in the feature space spanned by the images from the gallery camera. The features of the same person from different cameras are similar following the above transformation.

The rest of the paper is organized as follows. Section 2 surveys the related work of person re-identification. Section 3 presents the proposed semi-supervised coupled dictionary learning based on LCC. Section 4 discusses how to apply the semi-supervised coupled dictionary learning for person re-identification. Section 5 shows the experimental results, and Section 6 concludes the paper.

2. Related Work

Distance learning [1, 10, 12, 19, 20, 28] and local feature matching [3, 8, 14, 17, 27] have been widely studied in person re-identification and appearance modeling.

Distance learning-based person re-identification learns the optimal similarity measure between a pair of person images. Porikli [19] proposed a Brightness Transfer Function (BTF) to evaluate the inter-camera radiometric properties. The function is computed for every pair of cameras such that an observed color value in one camera is mapped to the corresponding observation in the other camera. Javed et al. [12] further investigated Porikli's method by showing that all BTFs lie in a low dimensional subspace such that some parameters of BTF are not required for computation. Based on this discovery, BTFs can be estimated efficiently using the Parzen window method. Zheng et al. [28] introduced a Relative Distance Comparison (RDC) model to maximize the probability of a pair of true matches having a smaller distance than a wrongly match pair. This approach avoids treating all features indiscriminately and does not assume the existence of some universally distinctive and reliable features. Pedagadi et al. [18] combined Principle Component Analysis (PCA) and Local Fisher Discriminative Analysis (LFDA) to match the visual appearance features. Using the dimension reduction approach, the high dimensional features can be exploited in an efficient way. Some supervised [10, 20] or unsupervised [1] algorithms have been proposed to select the most relevant features for person re-identification. Gray et al. [10] used the AdaBoost algorithm to find a subset of optimal features for human matching by combining different types of simple features into a single similarity function. Prosser et al. [20] developed a person re-identification system based on RankSVM. In their method, the combinations of local features are learned such that the relative ranking of the matching scores are fitted to the training data.

Local feature matching-based person re-identification matches the carefully designed local features. Bird et al. [3] used stripe based rigid blobs to model the appearances of individuals. The image of a pedestrian is divided into ten equally spaced horizontal strips, and the mean feature vectors of the horizontal strips are learned in a training step. Gheissari et al. [8] proposed a spatiotemporal over-segmentation method for grouping pixels that belong to the same type of fabric, after which they merged connected clusters whenever the distance between clusters was less than the internal variation of each of the individual clusters. The final distance between two individuals was then defined as the sum of the correspondences between these resulting segmentations. Oreifej et al. [17] extracted foreground blobs in aerial images and then assigned a weight to every blob region such that the most consistent regions were given higher weights, since it was more probable that they would lead to the target's identity. Bazzani et al. [2] introduced the histogram plus epitome feature for person re-identification. The proposed feature incorporates both global and local statistical descriptions of human appearances. Zhao et al. [27] used an unsupervised learned salience model for patch matching such that the reliable and discriminative matched patches can be matched for better re-identification performance.

3. Semi-Supervised Coupled Dictionary Learning

In this section, we first briefly review the local coordinate coding method that serves as the basis of the dictionary construction, and then present the semi-supervised coupled dictionary learning.

3.1. LCC Dictionary Learning

Local coordinate coding (LCC) [26] is a high dimensional nonlinear learning method for modeling data distributed on manifolds. LCC approximates a given input point as a weighted linear combination of a few elements called anchor points. The goal of LCC is to discover a good dictionary set of anchor points for better approximation. More specifically, given unlabeled training data $\{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{d\times n}$, the dictionary $D = \{d_1, d_2, \ldots, d_k\} \in \mathbb{R}^{d\times k}$ is learned by minimizing the objective function of the squares reconstruction errors and the

locality penalty. This process is formulated as follows

$$\min_{D\in\mathcal{L},\alpha} \frac{1}{2}||x - D\alpha||^2 + \mu \sum_j \left|\alpha^j\right| \|d_j - x\|^2, \quad (1)$$

where α is the sparse coefficients of x, α^j is the jth component of α and d_j is the jth column of D. $\mathcal{L} = \{D|\|d_i\| \leq 1, i = 1, \ldots, k\}$ is the convex feasible set of D. Given a set of training samples, we need to learn a good dictionary that is adapted to the distribution of the samples. A common approach for this task is to minimize the summed objective functions of all data samples by optimizing D and α

$$\min_{D\in\mathcal{L},\alpha_i} \sum_i \left(\frac{1}{2}\|x_i - D\alpha_i\|^2 + \mu \sum_j \left|\alpha_i^j\right| \|d_j - x_i\|^2\right), \quad (2)$$

where α_i is the coefficients of x_i.

The above objective function is not jointly convex over D and α, which makes it difficult to solve D and α simultaneously. Nevertheless, it is convex over D given fixed α and vice versa. Therefore, we can optimize one variable at a time by fixing the other and alternating between the two variables. Specifically, when D is fixed, the different α_i can be decoupled into individual sparse coding problems, which can be further transformed into a LASSO/LARS problem [23].

Let $\beta = \Lambda\alpha$, where Λ is a diagonal matrix whose elements are $\Lambda_{jj} = \|d_j - x\|^2$ and $\beta = \Lambda\alpha$. If $d_j \neq x$, then Λ^{-1} exists (otherwise, we directly represent x by itself). Therefore, for fixed D and x, optimizing over α can be transformed into optimizing over β as follows

$$\min_{\beta} \frac{1}{2}\|x - D\Lambda^{-1}\beta\|^2 + \mu \left|\beta\right|_1 \quad (3)$$

where $|\beta|_1 = \sum_j \left|\beta^j\right|$ denotes l_1-norm. After solving for β, we obtain $\alpha = \Lambda^{-1}\beta$.

After solving for α, optimizing over D is a constrained quadratic programming problem as follows

$$\begin{aligned}
&\min_{D\in\mathcal{L}} \sum_i \left(\frac{1}{2}\|x_i - D\alpha_i\|^2 + \mu \sum_j |\alpha_i^j| \|d_j - x_i\|^2\right) \quad (4)\\
&= \min_{D\in\mathcal{L}} \frac{1}{2}\mathrm{tr}\left[D^T D\left(\sum_i \alpha_i\alpha_i^T + 2\mu\Sigma_i\right)\right]\\
&\quad -\mathrm{tr}\left[D^T\left(\sum_i x_i\alpha_i^T + 2\mu x_i\bar{\alpha}_i^T\right)\right]
\end{aligned}$$

where $\bar{\alpha}_i$ is component-wise absolute value of α_i, i.e. $\bar{\alpha}_i^j = \left|\alpha_i^j\right|$ and Σ_i is a diagonal matrix constructed from $\bar{\alpha}_i$.

The above minimization problem can be solved by an iterative updating approach. First, we store two matrices $A = \sum_i \alpha_i\alpha_i^T + 2\mu\Sigma_i$ and $B = \sum_i x_i\alpha_i^T + 2\mu x_i\bar{\alpha}_i^T$ and use block-coordinate descent to find the optimal D. Specifically, we update the jth column d_j^k when other columns are fixed in the kth iteration of dictionary updating. The updating is carried out as follows

$$d_j^{k+1} = \Pi_{\mathcal{L}}\left(d_j^k - \frac{1}{a_{jj}}\left(D^k a_j - b_j\right)\right) \quad (5)$$

where a_j and b_j are the jth columns of matrices A and B respectively, and $\Pi_{\mathcal{L}}$ means projection onto the feasible set $\mathcal{L}$.

In our implementation, the number of dictionary items k is set to 500, and the trade-off coefficient μ is set to 0.15. The number of iterations for the alternating optimization used in dictionary learning is 15.

3.2. Semi-Supervised Coupled LCC Dictionary Learning

In the semi-supervised coupled dictionary learning, we are given two sets of training data $\{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{d\times n}$ and $\{y_1, y_2, \ldots, y_m\} \in \mathbb{R}^{h\times m}$ from two coupled feature spaces, $\mathcal{X}$ and $\mathcal{Y}$. We also know one-to-one correspondences for the top t pairs of points ($t \leq m$ and $t \leq n$). The goal of semi-supervised coupled dictionary learning is to learn two dictionaries D_x and D_y such that the LCC sparse representation $\alpha(x_i)_{i<=t}$ in terms of D_x should be the same as $\alpha(y_i)$ in terms of D_y. The labeled data are used to carry the relationship between $\mathcal{X}$ and $\mathcal{Y}$ while the unlabeled data are introduced to exploit the geometry of the marginal distribution for obtaining robust sparse representations. Putting the above together, we minimize the following objective function

$$\begin{aligned}
&\mathrm{Q}\left(D_x, D_y, \alpha\right) = E_{labeled}\left(D_x, D_y, \alpha^{(s)}\right) + \quad (6)\\
&E_{unlabeled}\left(D_x, \alpha^{(x)}\right) + E_{unlabeled}\left(D_y, \alpha^{(y)}\right)
\end{aligned}$$

where $\alpha^{(s)}$ is the shared coefficients matrix for $\{x_i\}_{i=1\ldots t}$ and $\{y_i\}_{i=1\ldots t}$ and $\alpha^{(x)}$ and $\alpha^{(y)}$ are the coefficients matrices for $\{x_i\}_{i=t+1\ldots n}$ and $\{y_i\}_{i=t+1\ldots m}$, respectively.

$$\begin{aligned}
&E_{labeled}\left(D_x, D_y, \alpha^{(s)}\right) = \quad (7)\\
&\sum_{i=1}^{t}\left(\frac{1}{2}\|x_i - D_x\alpha_i^{(s)}\|^2 + \frac{1}{2}\|y_i - D_y\alpha_i^{(s)}\|^2 + \right.\\
&\left.\mu\sum_j \left|\alpha_i^{(s),j}\right| \|d_j^x - x_i\| + \mu\sum_j \left|\alpha_i^{(s),j}\right| \|d_j^y - y_i\|\right)
\end{aligned}$$

is the labeled term that requires the resulting shared coefficients matrix $\alpha^{(s)}$ that should reconstruct both x_i and y_i,

and

$$E_{unlabeled}\left(D_x, \alpha^{(x)}\right) = \tag{8}$$
$$\sum_{i=t+1}^{n}\left(\frac{1}{2}\|x_i - D\alpha_i^x\|^2 + \mu\sum_j\left|\alpha_i^{(x),j}\right|\|d_j^x - x_i\|^2\right)$$

and

$$E_{unlabeled}\left(D_y, \alpha^{(y)}\right) = \tag{9}$$
$$\sum_{i=t+1}^{m}\left(\frac{1}{2}\|y_i - D\alpha_i^y\|^2 + \mu\sum_j\left|\alpha_i^{(y),j}\right|\|d_j^y - y_i\|^2\right)$$

are the unlabeled terms to guarantee that the sparse representations well reconstruct the unlabeled data.

Although the form of (6) is complex, it is convex over D given fixed α and vice versa. We minimize (6) by alternating optimization. To solve the dictionaries given fixed coefficients α, we optimize D_x and D_y individually by utilizing (4). Given fixed dictionaries, the coefficients of the unlabeled data α_x and α_y can be directly obtained by exploiting (2) since they are not coupled. For the coupled coefficients α_s, we can concatenate both the descriptors and the corresponding dictionaries to jointly learn the coefficients. The procedure is similar to the coupled learning paradigm, e.g. [25]. Specifically, we group both the data and the dictionaries of the labeled pairs as follows

$$z_i = \left[\begin{array}{c} x_i \\ y_i \end{array}\right]_{i=1\ldots t} \qquad D_s = \left[\begin{array}{c} D_x \\ D_y \end{array}\right], \tag{10}$$

and optimize the following objective function

$$\min_{\alpha^{(s)}}\sum_{i=1}^{t}\left(\frac{1}{2}\|z_i - D_s\alpha_i\|^2 + \mu\sum_j\left|\alpha_i^j\right|\|d_j^s - z_i\|^2\right), \tag{11}$$

which shares the same form as (2) and can be solved by (3).

4. Person Re-identification based on Semi-Supervised Learned Coupled Dictionary

In this paper, our goal is to re-identify each probe image by matching with gallery images. The proposed algorithm handles challenging issues such as resolution and lighting condition changes by the proposed semi-supervised coupled dictionary learning.

Figure 1 illustrates the overall flow of our approach. In the training phase, labeled pairs of images as well as unlabeled images from the gallery and probe cameras are used to jointly learn the coupled LCC dictionaries. In the testing phase, we encode the feature of a given probe image by the LCC sparse representation through the probe dictionary, and then recover the feature through the gallery dictionary. We can finally match the target image with each person in the gallery set.

4.1. Feature Extraction

To extract visual features of the same dimension from different scale images, we divide the images into the same number of local (overlapping) patches. In our implementation, we use 15 rows and 5 columns, leading to 75 patches in an image. In each patch, we extract HSV color histograms quantized into 30 bins, gradient histogram quantized into 9 bins, and LBP [16] histogram quantized into 59 bins, resulting in a 98 dimensional descriptor for each local patch. We then learn the coupled dictionaries for the 98-dimensional features.

4.2. Coupled Dictionary Learning for Person Re-identification

Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ denote the features of persons captured by two different cameras. A person re-identification system aims to find a good matching measure $\mathcal{M} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that matching the features of the same person results in a smaller score. Most of the existing methods try to learn $\mathcal{M}$ from pre-labeled training images; however, even the features from the same person can be very different as a result of resolution and lighting condition changes across cameras, making $\mathcal{M}$ highly nonlinear and requiring enormous numbers of labeled image pairs which may be not sufficient in practice.

We surmise that the condition changes across cameras can be characterized by a transformation function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$. Since the transformation function is heterogeneous, we cannot directly learn $\mathcal{F}$. Nevertheless, we can estimate a homogenous $\hat{\mathcal{F}}$ with a given metric measure $\hat{\mathcal{M}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ on labeled images $\{(x_i, y_i)\}$ by minimizing the summed matching score

$$\hat{\mathcal{F}}_{\hat{\mathcal{M}}} = \underset{F}{\operatorname{argmin}}\sum_i\left(\hat{\mathcal{M}}\left(\mathcal{F}(x_i), y_i\right)\right) + \mu\Omega(\mathcal{F}) \tag{12}$$

where $\sum_i\left(\hat{\mathcal{M}}\left(\mathcal{F}(x_i), y_i\right)\right)$ minimizes the matching score, $\Omega(\mathcal{F})$ is the regularization term, and μ is the trade-off parameter.

Note that $\hat{\mathcal{F}}$ is a nonlinear function that requires an enormous quantity of training images, but as has been shown in [26], an arbitrary nonlinear (β, γ)-Lipschitz[1] smooth function $\mathcal{F}$ can be approximated by a linear function $\sum_j \alpha^j(x)\mathcal{F}(d_j^x)$ with respect to the coding $\alpha(x)$:

$$\left|\mathcal{F}(x) - \sum\nolimits_j \alpha^j(x)\mathcal{F}(d_j)\right| \tag{13}$$
$$\leq \beta\|x - D\alpha(x)\| + \gamma\sum_j\left|\alpha^j(x)\right|\|d_j - D\alpha(x)\|^2,$$

[1] A function $\mathcal{F}(x)$ on $\mathbb{R}^d$ is (β, γ)-Lipschitz smooth with respect to norm $\|\cdot\|$ if $|\mathcal{F}(x') - \mathcal{F}(x) \leq \beta\|x - x'\|$ and $|\mathcal{F}(x') - \mathcal{F}(x) - \nabla\mathcal{F}(x)^T(x' - x)| \leq \gamma\|x - x'\|^2$, where we assume $\beta, \gamma > 0$.

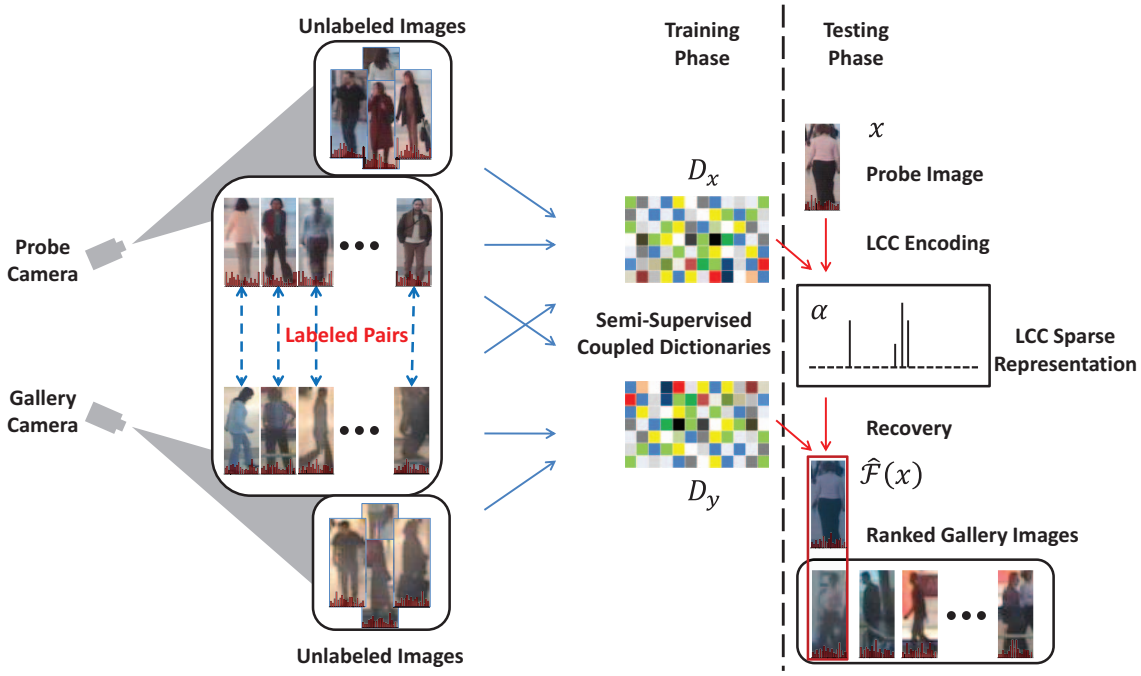

Figure 1. Flowchart of the proposed semi-supervised coupled dictionary learning-based person re-identification. In the training phase, labeled pairs of images as well as unlabeled images from the gallery and probe cameras are used to jointly learn the coupled LCC dictionaries. In the testing phase, we encode the feature of a given probe image by the LCC sparse representation through the probe dictionary, and recover the feature through the gallery dictionary. We can then match the target image with each person in the gallery set.

and the quality of this approximation is bounded by the right hand side, which can be further simplified as

$$\begin{aligned} &\|x - D\alpha(x)\| + \gamma \sum_j \left|\alpha^j(x)\right| \|d_j - D\alpha(x)\|^2 \qquad (14) \\ &\approx \|x - D\alpha(x)\|^2 + \gamma \sum_j \left|\alpha^j(x)\right| \|d_j - x\|^2. \end{aligned}$$

With this approximation, we transform (12) into the following objective function

$$\underset{F}{\operatorname{argmin}} \sum_i \left(\hat{\mathcal{M}} \left(\sum_j \alpha_i^j \mathcal{F}(d_j^x), y_i \right) \right) + \mu\Omega(\mathcal{F}). \quad (15)$$

Let $\hat{\mathcal{M}}$ denote the l_2-norm distance, and then we have

$$\underset{\mathcal{F}(d_j^x)}{\operatorname{argmin}} \sum_i \left\| y_i - \sum_j \alpha_i^j \mathcal{F}(d_j^x) \right\|^2 + \mu\Omega(\mathcal{F}). \qquad (16)$$

Note that when D_x and α are fixed, $\mathcal{F}(d_j^x)$ become constant projection vectors. Therefore, we can concatenate these row vectors into a projection matrix D_y. If we further use the locality penalty defined in (1) as the regularization term, we obtain

$$\underset{D_y}{\operatorname{argmin}} \sum_i \left(\|y_i - D_y\alpha_i\|^2 + \mu \sum_j \left|\alpha_i^j\right| \|d_j - x_i\|^2 \right). \qquad (17)$$

In practice, D_x and α are not fixed but jointly optimized with D_y by minimizing $E_{labeled}\left(D_x, D_y, \alpha^{(s)}\right)$ defined in (7).

In addition to the labeled data in our implementation, we also introduce abundant unlabeled data on $\mathcal{X}$ to exploit the geometry of the marginal distribution for obtaining robust linear approximation by minimizing $E_{unlabeled}\left(D_x, \alpha^{(x)}\right)$ and introduce abundant unlabeled data on $\mathcal{Y}$ as the prior of $\hat{\mathcal{F}}$ by minimizing $E_{unlabeled}\left(D_y, \alpha^{(y)}\right)$.

4.3. Appearance Matching

Given a probe image, its feature representation x is transformed to $\hat{\mathcal{F}}(x)$ to avoid resolution and lighting condition changes across different cameras. However, since the feature in a local patch is highly determined by the location of the patch, direct patch-to-patch matching is not robust to pose and view angle changes, so we use a greedy nearest-neighbor patch matching strategy to solve this feature-shift problem.

Our motivation is simple but effective: each patch is

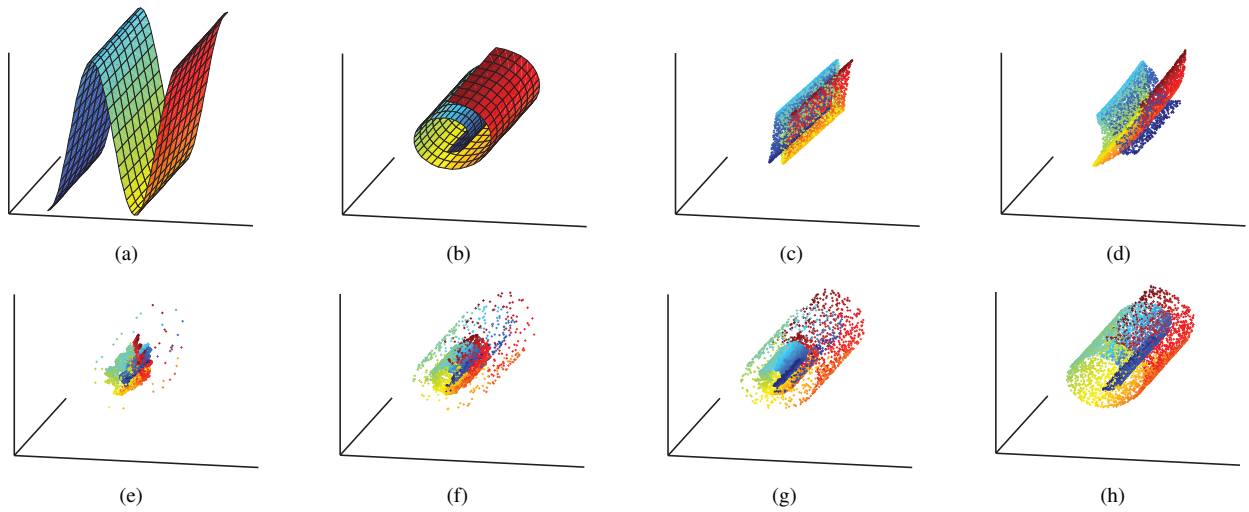

Figure 2. Results of manifold transformation: (a) a wave manifold and (b) a Swissroll manifold that shares the same intrinsic structure, where the pixels with the same color are corresponding matched pairs; (c) result of linear regression, (d) result of polynomial regression, (e) result of coupled dictionary learning with 50 labeled points, (f) result of coupled dictionary learning with 500 labeled points, (g) result of semi-supervised coupled dictionary learning with 50 labeled points and 4950 unlabeled points and (h) result of semi-supervised coupled dictionary learning with 500 labeled points and 4500 unlabeled points.

matched to its nearest neighbor measured by the Euclidean distance within the pre-defined search range. In each iteration, we pick the matched pair of patches with the minimal score and remove the matched patches from the search range in the future matching. The final appearance matching score is the sum of all the patches. A person image is divided into three biometric regions corresponding to head, body and legs, and the search range of a patch is restricted to its own biometric region. According to [7], the partition of biometric regions can be determined by the degree of x-axis asymmetry. However, we find that using a fixed proportion $(5 : 11 : 16)$ works better in practice. It should be mentioned that the nearest-neighbor patch matching is also carried out for the labeled training pairs before learning the coupled dictionaries.

5. Experimental Results

We first validate the proposed semi-supervisd coupled dictionary learning method on a synthetic dataset and then evaluate the performance of our approach on the publicly available VIPeR [9] and CAVIAR4REID [4] datasets. The VIPeR is used for the single-shot evaluation and the CAVIAR4REID is used for the multiple-shot evaluation.

5.1. Synthetic Data

Our first toy experiment is based on a synthetic dataset, where a nonlinear function is learned to transform a manifold from one space to another. As shown in Figure 2-(a) and 2-(b), the wave manifold and the Swissroll manifold share the same intrinsic structure, where the pixels with the same color are corresponding matched pairs. We randomly sampled 5000 points from the wave manifold, and some of them are labeled. The goal is to learn a function that transform the points from the first space to the second one. The transformation is nonlinear and its quality can be evaluated by the root mean square error (RMSE). We can use this toy experiment to validate the performance of the proposed method on nonlinear function learning.

Figure 2-(c) shows the transformation result of a linear regression model learned with 500 labeled points and the RMSE is 10.1. Obviously, it fails to characterize the nonlinear transformation. Figure 2-(d) shows the transformation result of a third order polynomial regression model learned with 500 labeled points, and the RMSE is 7.6. As can be seen, without the prior of an appropriate form, the nonlinear regression fails either. Figure 2-(e) and 2-(f) show the transformation results of coupled dictionary learning with 50 and 500 labeled points. Their RMSEs are 12.5 and 5.4 respectively. The results are getting better than conventional regression methods, but are still not satisfactory. Figure 2-(g) shows the transformation result of semi-supervised coupled dictionary learning with 50 labeled and 4950 unlabeled points and the RMSE is 4.1. Although fewer labeled points are used, it is better than Figure 2-(f) because of considering the distribution of unlabeled points. Figure 2-(h) shows the transformation result of semi-supervised coupled dictionary learning with 500 labeled and 4500 unlabeled points. As can be seen, the result fits the true manifold perfectly and the RMSE is as small as 0.3.

5.2. Single-Shot Person Re-identification Evaluation

We evaluate the performance of our approach for single-shot person re-identification using the publicly available VIPeR dataset [9], which contains 632 pedestrian image pairs. Each pair contains two images of the same individual seen from different camera views under pose changes and varying illumination conditions. Each image has been scaled to 128×48 pixels. In our experiment, the images in camera A are used as the probe images, while the images in camera B are used as the gallery images. The evaluation on this dataset is repeated 10 times, and the average result is reported. In each repetition, the dataset is randomly halved into training data and testing data, and one-third of the training data are labeled while the rest are unlabeled.

We compare our semi-supervised coupled dictionary learning (SSCDL)-based person re-identification with most published results on the VIPeR dataset, including RDC [28], ITML [5], LMNN [24], KISSME [13], ELF [10], eLDFV [15], SDALF [7] and LF [18]. All methods use the same data splitting assignments for training and testing. The performance comparison is graphically depicted by the Cumulated Matching Characteristics (CMC) curves in Figure 3. From the figure, it is clear that our proposed method gives the best result. To show the quantized comparison results more clearly, we also summarize the performance comparison in Table 1. As can be seen, our proposed method achieves 25.6% rank 1 matching rate, which improves the previous best results over 1.5%. The rank 10 matching rate for SSCDL is 68.1% which again outperforms all the other methods. It is worthy to note that SSCDL is a semi-supervised algorithm that only requires one-third of labeled data by comparison with others.

Method	$r = 1$	r=5	r=10	r=20
RDC	15.7	38.4	53.9	70.1
ITML	11.6	31.4	45.8	63.9
LMNN	6.2	19.7	32.6	52.1
KISSME	19.6	48.2	62.2	76.9
ELF	8.1	24.1	36.6	52.1
eLDFA	22.34	47.0	60.0	71.0
SDALF	19.9	38.9	49.4	65.7
LF	24.1	51.2	67.1	82.0
SSCDL	25.6	53.7	68.1	83.6

Table 1. Top ranked matching rates (%) on the VIPeR dataset with 316 testing pairs.

5.3. CAVIAR4REID

CAVIAR4REID [4] is a multi-shot person re-identification dataset which is made by processing 26 sequences captured from two cameras in a shopping center.

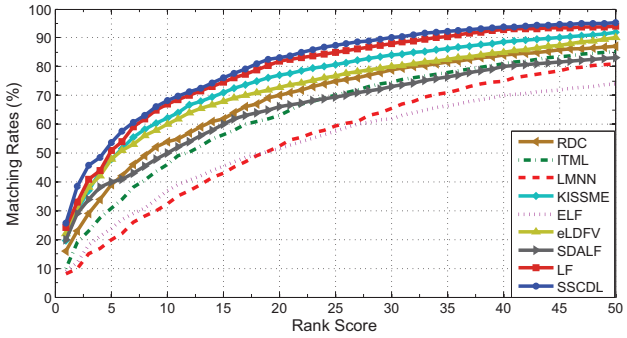

Figure 3. Performance comparison using CMC curves on the VIPeR dataset with 316 testing pairs. A rank r matching rate indicates the percentage that corrected matches in the top r ranks.

It includes people walking along, meeting with others, window shopping, entering and existing shops. There are 72 individuals in total: 50 of them appear in both camera views and the remaining 22 persons only appear in one camera view. Multiple images are obtained for each person to maximize the appearance variance over different conditions. The main complexity of the dataset arises from the very severe resolution and lighting changes between the two camera views. We randomly choose 14 of the 50 individuals appearing in two cameras as the labeled training data, and the remaining 36 individuals are used as testing data. The 22 persons appearing in one camera are used as the unlabeled training data. The evaluation is repeated 10 times, and the average result is reported. In each repetition, the labeled training data are again randomly chosen.

We compare our proposed SSCDL method with LF [18] and HPE [2] on this dataset. LF and SSCDL use the same splitting assignment of training and testing. HPE is an unsupervised method that does not require a training step, and we report its published result tested on the 50 individuals. The CMC curves of the three methods are shown in Figure 4. We also summarized the results in Table 2. The proposed SSCDL achieves 49.1% rank 1 matching rate, which outperforms LP significantly.

Method	$r = 1$	r=5	r=10	r=20
HPE	9.7	33.2	55.6	76.3
LF	36.1	51.2	88.6	97.5
SSCDL	49.1	80.2	93.5	97.9

Table 2. Top ranked matching rates (%) on the CAVIAR4REID dataset with 36 testing pairs.

6. Conclusion

In this paper, we propose the semi-supervised coupled dictionary learning method for person re-identification. To

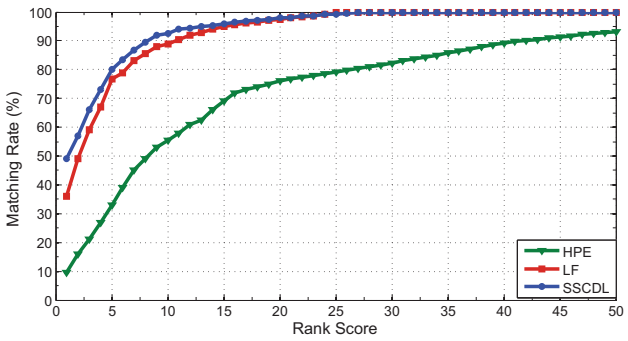

Figure 4. Performance comparison using CMC curves on the CAVIAR4REID dataset with 36 testing pairs. A rank r matching rate indicates the percentage that corrected matches in the top r ranks.

handle the challenge of resolution and lighting condition changes in different cameras, a pair of coupled dictionaries that relate to the probe and gallery cameras are jointly learned from both labeled and unlabeled images. We analyze the principle that semi-supervised coupled dictionary learning is theoretically appropriate for feature transformation across camera views and validate the proposed method in two publicly available datasets.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (61170142), the National Key Technology R&D Program under Grant (2011BAG05B04), the Program of International S&T Cooperation (2013DFG12840), and Australian Research Council Projects (FT-130101457 and DP-140102164). M. Song is the corresponding author.

References

[1] K. Bashir, T. Xiang, and S. Gong. Feature selection on gait energy image for human identification. *Proc. ICASSP*, 2008.

[2] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recogn. Lett.*, pages 893–903, 2012.

[3] N. Bird, O. Masoud, N. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. *IEEE Trans. Intell. Transp. Syst.*, 6:167–177, 2005.

[4] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. *Proc. BMVC*, 2011.

[5] J. Davis, B. Kulis, P. J. ans S. Sra, and I. Dhillon. Information theorietic metric learning. *Proc. ICML*, 2007.

[6] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. *Proc. ACCV*, 2010.

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. *Proc. CVPR*, 2010.

[8] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. *Proc. CVPR*, 2006.

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

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

[11] W. Hu, M. Hu, X. Zhou, J. Lou, T. Tan, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):663–671, 2006.

[12] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Comput. Vis. Image Und.*, 109:146–162, 2008.

[13] M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *Proc. CVPR*, 2012.

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

[15] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. *Proc. ECCV*, 2012.

[16] T. Ojala, M. Pietikainen, and T. Maepaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:971–987, 2002.

[17] O. Oreifej, R. Mehran, and M. Shah. Human identity recognition in aerial images. *Proc. CVPR*, 2010.

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

[19] F. Porikli. Inter-camera color calibration using cross-correlation model function. *Proc. ICIP*, 2003.

[20] B. Prosser, W. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. *Proc. BMVC*, 2010.

[21] J. Sivic, C. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. *Proc. BMVC*, 2006.

[22] Y. Song and T. Leung. Context-aided human recognition - clustering. *Proc. ECCV*, 2006.

[23] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. (Ser.B)*, pages 267–288, 1996.

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

[25] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Trans. Image Process.*, 19(11):2861–2873, 2010.

[26] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. *Proc. NIPS*, 2009.

[27] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. *Proc. CVPR*, 2013.

[28] W. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(3):653–668, 2013.