Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

# Cross-View Projective Dictionary Learning for Person Re-identification

**Sheng Li**
Dept. of ECE
Northeastern University
Boston, MA, USA
shengli@ece.neu.edu

**Ming Shao**
Dept. of ECE
Northeastern University
Boston, MA, USA
mingshao@ece.neu.edu

**Yun Fu**
Dept. of ECE and College of CIS
Northeastern University
Boston, MA, USA
yunfu@ece.neu.edu

## Abstract

Person re-identification plays an important role in many safety-critical applications. Existing works mainly focus on extracting patch-level features or learning distance metrics. However, the representation power of extracted features might be limited, due to the various viewing conditions of pedestrian images in reality. To improve the representation power of features, we learn discriminative and robust representations via dictionary learning in this paper. First, we propose a cross-view projective dictionary learning (CPDL) approach, which learns effective features for persons across different views. CPDL is a general framework for multi-view dictionary learning. Secondly, by utilizing the CPDL framework, we design two objectives to learn low-dimensional representations for each pedestrian in the patch-level and the image-level, respectively. The proposed objectives can capture the intrinsic relationships of different representation coefficients in various settings. We devise efficient optimization algorithms to solve the objectives. Finally, a fusion strategy is utilized to generate the similarity scores. Experiments on the public VIPeR and CUHK Campus datasets show that our approach achieves the state-of-the-art performance.

## 1 Introduction

Person re-identification is the problem of matching pedestrian images observed from multiple non-overlapping cameras. It saves a lot of human efforts in many safety-critical applications such as video surveillance. In recent years, many algorithms have been proposed to tackle this problem [Zheng *et al.*, 2012; Zhao *et al.*, 2013a; Wang *et al.*, 2014; Li *et al.*, 2014b]. These methods can be mainly divided into two categories, including the distance learning/metric learning methods [Weinberger *et al.*, 2005; Zheng *et al.*, 2011; Davis *et al.*, 2007; Mignon and Jurie, 2012; Pedagadi *et al.*, 2013] and feature learning methods [Gray and Tao, 2008; Farenzena *et al.*, 2010; Ma *et al.*, 2012b; Zhao *et al.*, 2013a]. The distance learning methods usually learn distance metrics that are expected to be robust to sample variations. The feature learning methods aim to extract distinctive features from pedestrian images, such as salient features [Zhao *et al.*, 2013a]. However, the representation power of learned features or metrics might be limited, due to the various viewing conditions of pedestrian images in real scenarios.

In this paper, we learn discriminative and robust representations via dictionary learning to improve the representation power of features. Our motivations are two-folds. First, the success of dictionary learning based domain adaptation technique inspires us to learn a pair of cross-view dictionaries for person re-identification [Ni *et al.*, 2013]. The adaptively learned dictionaries can generate robust representations for pedestrian images. Secondly, existing works either focus on extracting features form image patches or directly learning global features, but the complementary information resided in patch-level and image-level are usually ignored.

Based on the motivations above, we propose a cross-view projective dictionary learning (CPDL) approach, which is a general framework for the multi-view dictionary learning problem. We then design two objectives by utilizing the CPDL framework, which learn low-dimensional representations for each person in the patch-level and the image-level, respectively. Different from traditional dictionary learning methods, CPDL adopts the projective learning strategy to avoid solving the $l_1$ optimization problem in training phase. The proposed objectives can capture the intrinsic relationships of different representation coefficients in various settings. We also employ a strategy to fuse the similarity scores estimated in two levels.

By far, there are few methods proposed to learn effective representations for the pedestrian images under different views [Liu *et al.*, 2014]. The basic idea of Liu's method is to learn expressive bases to represent the image patches. It assumes that each pair of patches in two images shares the same representation coefficients. However, it is not the case in reality, due to the common misalignment problem in person re-identification.

The major contributions of this paper are summarized below.

- We propose a general framework, CPDL, for multi-view dictionary learning, and apply it to person re-identification. CPDL adopts the projective dictionary learning strategy, which is more efficient than the traditional dictionary learning methods. We devise efficient optimization algorithms to solve the model.

- We design two objectives using CPDL, which explicitly model the cross-view interactions in different representation levels, including the patch-level and image-level. To the best of our knowledge, this paper is the first attempt to learn representations at different levels for person re-identification.
- We evaluate the performance of CPDL and related methods on the public VIPeR and CUHK Campus datasets. Extensive experimental results show that our approach outperforms the state-of-the-art methods.

## 2 Related Work

There are two types of works that are very related to our approach: (1) person re-identification, (2) dictionary learning.

**Person Re-identification.** In recent years, many algorithms have been proposed for person re-identification. Some traditional methods focus on learning effective metrics to measure the similarity between two images captured from different camera views [Hirzer *et al.*, 2012; Zheng *et al.*, 2011]. Other research works focus on learning expressive features, which usually obtain better performance that the metric learning methods. They suggest that learning effective representations is the key in person re-identification. Some advanced features include attributes [Layne *et al.*, 2012], salience features [Zhao *et al.*, 2013a; 2013b], mid-level features [Zhao *et al.*, 2014], and salient color features [Yang *et al.*, 2014]. Although the existing feature learning methods achieve good performance, the cross-view relationships of pedestrian images haven't been extensively studied. Our CPDL approach explicitly models such relationships in different representation levels, and draws strength from them to enhance the re-identification performance.

**Dictionary Learning.** As a powerful technique for learning expressive bases in sample space, dictionary learning has attracted lots of attention during the past decades [Li *et al.*, 2014a]. Some popular dictionary learning methods include K-SVD [Aharon *et al.*, 2006], discriminative K-SVD [Zhang and Li, 2010], and projective dictionary pair learning [Gu *et al.*, 2014]. Most recently, Liu *et al.* presented a semi-supervised coupled dictionary learning (SSCDL) method [Liu *et al.*, 2014], and applied it to person re-identification. The major differences between our approach and SSCDL are three-folds. First, SSCDL is a semi-supervised method, while our approach is supervised. Secondly, SSCDL simply assumes that a pair of patches in two views should have similar codings, which is unreasonable in real scenario due to the misalignment problem. Our approach models the cross-view interactions in image-level and patch-level, respectively. Thirdly, SSCDL requires solving the $l_1$ optimization problem that is time consuming. Our approach adopts a more efficient learning strategy, i.e., projective dictionary learning.

## 3 A General Framework for Cross-view Projective Dictionary Learning (CPDL)

Traditional dictionary learning methods usually assume that the samples $A \in \mathbb{R}^{d\times n}$ can be reconstructed by sparse coefficients $Z \in \mathbb{R}^{m\times n}$ and a dictionary $D \in \mathbb{R}^{d\times m}$, i.e., $A = DZ$, in which $Z$ is constrained by $l_1$ norm. However, solving sparse coefficients $Z$ often suffers heavy computational costs. Inspired by the projective dictionary learning [Gu *et al.*, 2014], we address this problem by reformulating the DL process as a linear encoding and reconstruction process. Let $P \in \mathbb{R}^{m\times d} (m \ll d)$ denote a low-dimensional projection matrix, we can reconstruct the sample set by $A = DPA$. Note that $PA$ denotes the linear encodings of sample set $A$.

We build a cross-view projective dictionary learning (CPDL) framework in the two-view settings. Let $A_1 \in \mathbb{R}^{d_1\times n}$ and $A_2 \in \mathbb{R}^{d_2\times n}$ denote two training sets that are collected under two different views, respectively. The reconstructions in two views are formulated as

$$A_1 = D_1 P_1 A_1, \qquad A_2 = D_2 P_2 A_2, \tag{1}$$

where $D_1$ (and $D_2$), $P_1$ (and $P_2$) are dictionaries and projections in two views, respectively.

The objective function of CPDL framework is

$$\begin{aligned} \min_{D_1,D_2,P_1,P_2} \quad & \|A_1 - D_1P_1A_1\|_{\mathrm{F}}^2 + \|A_2 - D_2P_2A_2\|_{\mathrm{F}}^2 \\ & +\lambda_1 f(D_1, D_2, P_1, P_2) \\ s.t. \quad & \left\|d_{1(:,i)}\right\| \le 1, \left\|d_{2(:,i)}\right\| \le 1. \end{aligned} \tag{2}$$

where $f(D_1, D_2, P_1, P_2)$ is a regularization function, $\lambda_1$ is a trade-off parameter, and $d_{1(:,i)}$ and $d_{2(:,i)}$ are the $i$-th columns in $D_1$ and $D_2$, respectively.

The first two terms in objective (2) indicate reconstruction errors in two views, respectively. The last term $f(D_1, D_2, P_1, P_2)$ is a regularization function that bridges two views. It can be customized for specific problems, such as multi-view image classification or (cross-view) person re-identification.

Finally, the obtained optimal dictionary pair $\{D_1, D_2\}$ can be used to generate new representations for test samples. Note that, for simplicity, we only formulate two views in this paper, but our model can be extended to the multiple-view case by extending (2).

## 4 CPDL for Person Re-identification

In this section, we first introduce how to extract low-level dense features from the pedestrian images. Then we formulate person re-identification problem using CPDL. Figure 1 shows the training framework of CPDL.

### 4.1 Feature Extraction

The pedestrian images in different camera views are not usually aligned well. Extracting dense features from local patches is a widely used strategy to obtain effective representations, as suggested in [Zhao *et al.*, 2014]. Specifically, the local patches are extracted on a dense grid. The size of each patch is 10×10, and the grid step is 5. Then, for each patch, we extract 32-dimensional color histogram features and 128-dimensional dense SIFT features in each LAB channel. Further, we also calculate the color histograms in different sampling scales with the downsampling factors 0.5 and 0.75. All the features of one patch are normalized with $l_2$ norm. Finally, each patch is represented by a 672-dimensional feature vector.

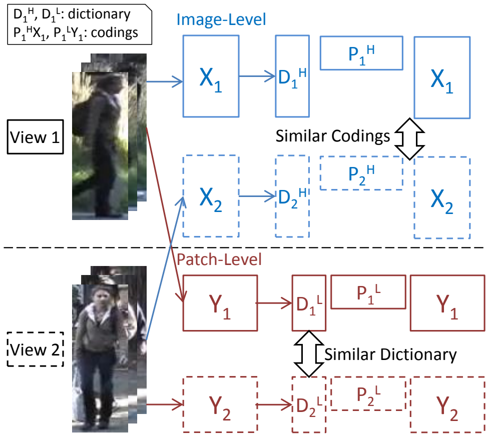


Figure 1: Training framework of CPDL. The solid boxes represent the variables related to view 1, while the dashed boxes represent the variables related to view 2. In the image-level training (blue color), two views share the similar codings (i.e., $P_1^{\mathrm{H}}X_1$, $P_2^{\mathrm{H}}X_2$); in the patch-level training (red color), two views share the similar dictionary (i.e., $D_1^{\mathrm{L}}$, $D_2^{\mathrm{L}}$).

## 4.2 CPDL for Image Representation

Our goal is to learn robust representations for each pedestrian in different camera views by virtue of dictionary learning. It's a challenging problem as the same person under different camera views usually exhibits significant differences in appearance. In this section, we propose to emphasize the feature learning in two levels, *patch level* and *image level*, in order to capture both local and global characteristics from the pedestrian images. Note that most existing methods only consider feature learning in one single level [Liu *et al.*, 2014].

Let $X_1$ and $X_2$ denote the training sets of high-dimensional dense features in two views, respectively. For the $i$-th training image in view 1, the dense features of all the patches are concatenated as a high-dimensional vector [1], which is the $i$-th column in $X_1$. Clearly, the corresponding columns in $X_1$ and $X_2$ should have similar codings, since they represent the same pedestrian. Hence, by defining the regularization function $f(\cdot)$ in (2), we have the following objective

$$\begin{array}{cl} \min\limits_{\substack{D_1^{\mathrm{H}},D_2^{\mathrm{H}},\\ P_1^{\mathrm{H}},P_2^{\mathrm{H}}}} & \left\|X_1 - D_1^{\mathrm{H}}P_1^{\mathrm{H}}X_1\right\|_{\mathrm{F}}^2 + \left\|X_2 - D_2^{\mathrm{H}}P_2^{\mathrm{H}}X_2\right\|_{\mathrm{F}}^2 \\ & +\lambda_1 \left\|P_1^{\mathrm{H}}X_1 - P_2^{\mathrm{H}}X_2\right\|_{\mathrm{F}}^2, \\ s.t. & \|d_{1(:,i)}^{\mathrm{H}}\| \leq 1, \|d_{2(:,i)}^{\mathrm{H}}\| \leq 1, \end{array} \tag{3}$$

where $D_1^{\mathrm{H}}$ (and $D_2^{\mathrm{H}}$), $P_1^{\mathrm{H}}$ (and $P_2^{\mathrm{H}}$) denote the dictionaries and projection matrices in two views, respectively.

The regularization function in (3) is $\left\|P_1^{\mathrm{H}}X_1 - P_2^{\mathrm{H}}X_2\right\|_{\mathrm{F}}^2$, indicating that the codings in two views should be as close as possible. In this way, the learned dictionaries $D_1^{\mathrm{H}}$ and $D_2^{\mathrm{H}}$ are expected to generate similar codings for the same pedestrian under two camera views.

## 4.3 CPDL for Patch Representation

In addition to modeling the image representation in (3), we also consider the dictionary learning in patch-level representations. Let $Y_1$ and $Y_2$ denote the training sets of low-dimensional patch features in two views, respectively. In this case, we cannot simply assume that the codings in two views are close to each other. In reality, the $i$-th patch in view 1 may not match the $i$-th patch in view 2 due to the misalignment problem under cross-view settings. One reasonable assumption is that the patches in different views could share a similar dictionary. Therefore, the objective function is

$$\begin{array}{cl} \min\limits_{D_1^{\mathrm{L}},D_2^{\mathrm{L}},P_1^{\mathrm{L}},P_2^{\mathrm{L}}} & \left\|Y_1 - D_1^{\mathrm{L}}P_1^{\mathrm{L}}Y_1\right\|_{\mathrm{F}}^2 + \left\|Y_2 - D_2^{\mathrm{L}}P_2^{\mathrm{L}}Y_2\right\|_{\mathrm{F}}^2 \\ & +\lambda_2 \left\|D_1^{\mathrm{L}} - D_2^{\mathrm{L}}\right\|_{\mathrm{F}}^2, \\ s.t. & \|d_{1(:,i)}^{\mathrm{L}}\| \leq 1, \|d_{2(:,i)}^{\mathrm{L}}\| \leq 1, \end{array} \tag{4}$$

in which the last term emphasizes the similarity of two dictionaries.

## 4.4 Matching and Fusion

With the learned two pairs of dictionaries, $\{D_1^{\mathrm{L}}, D_2^{\mathrm{L}}\}$ and $\{D_1^{\mathrm{H}}, D_2^{\mathrm{H}}\}$, we can obtain robust representations for the test images in two views, and perform the following matching and fusion strategy.

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

**Patch-level Matching.** The patch matching methods have been extensively studied in existing works [Zhao *et al.*, 2013a; 2014]. We adopt a similar constrained patch matching strategy. For each patch in the probe image, we can not directly match it to the corresponding patch in gallery images, due to the well-known misalignment problem. Therefore, we search the spatial neighbors of the targeted patch in the gallery images, and calculate the distances between each pairs. Finally, we can estimate the similarity between a probe image and every gallery image. Instead of comparing the original patches, we match the representation coefficients over the dictionaries $\{D_1^{\mathrm{L}}, D_2^{\mathrm{L}}\}$ for each pair of patches. The similarity score $Score_{\mathrm{P}}(i)$ between the probe image and the $i$-th gallery image is generated from the similarities between these patches.

**Image-level Matching.** The image-level matching between the probe image and gallery images is more straightforward, as we have already attained the compact representations for each image. The representation coefficients are calculated using the dictionaries $\{D_1^{\mathrm{H}}, D_2^{\mathrm{H}}\}$ for each pair of patches. We adopt the Gaussian kernel function to compute the similarity score $Score_{\mathrm{I}}(i)$ between the probe image and the $i$-th gallery image.

**Fusion.** We first normalize the similarity score vectors $Score_{\mathrm{P}}$ and $Score_{\mathrm{I}}$, and utilize a simple strategy to perform score fusion:

$$Score(i) = Score_{\mathrm{P}}(i) + \lambda Score_{\mathrm{I}}(i), \tag{5}$$

where $\lambda$ is an user-defined parameter.

[1] As we have high-dimensional image-level features and low-dimensional patch-level features, we use superscripts H and L for the image-level and patch-level variables, respectively.

# 5 Optimization

## 5.1 Solving objective (3)

To facilitate the optimization of (3), we first add two relaxation variables $A_1^{\mathrm{H}}$ and $A_2^{\mathrm{H}}$, and rewrite the objective as

$$\begin{aligned}\min_{\substack{D_1^{\mathrm{H}},D_2^{\mathrm{H}},P_1^{\mathrm{H}},\\ P_2^{\mathrm{H}},A_1^{\mathrm{H}},A_2^{\mathrm{H}}}} \quad & \left\|X_1 - D_1^{\mathrm{H}}A_1^{\mathrm{H}}\right\|_{\mathrm{F}}^2 + \left\|X_2 - D_2^{\mathrm{H}}A_2^{\mathrm{H}}\right\|_{\mathrm{F}}^2 \\ & +\alpha(\left\|P_1^{\mathrm{H}}X_1 - A_1^{\mathrm{H}}\right\|_{\mathrm{F}}^2 + \left\|P_2^{\mathrm{H}}X_2 - A_2^{\mathrm{H}}\right\|_{\mathrm{F}}^2) \\ & +\lambda_1\left\|A_1^{\mathrm{H}} - A_2^{\mathrm{H}}\right\|_{\mathrm{F}}^2, \\ s.t. \quad & \|d^{\mathrm{H}}_{1(:,i)}\| \le 1, \|d^{\mathrm{H}}_{2(:,i)}\| \le 1,\end{aligned} \tag{6}$$

where $\alpha$ is a balance parameter.

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

1). Fix other variables and update $A_1^{\mathrm{H}}$ and $A_2^{\mathrm{H}}$.

By ignoring the irrelevant variables with respect to $A_1^{\mathrm{H}}$, the objective (6) is reduced to

$$\begin{aligned}\min_{A_1^{\mathrm{H}}} \quad & J(A_1^{\mathrm{H}}) = \left\|X_1 - D_1^{\mathrm{H}}A_1^{\mathrm{H}}\right\|_{\mathrm{F}}^2 + \alpha\left\|P_1^{\mathrm{H}}X_1 - A_1^{\mathrm{H}}\right\|_{\mathrm{F}}^2 \\ & +\lambda_1\left\|A_1^{\mathrm{H}} - A_2^{\mathrm{H}}\right\|_{\mathrm{F}}^2.\end{aligned} \tag{7}$$

Setting $\frac{\partial J(A_1^{\mathrm{H}})}{\partial A_1^{\mathrm{H}}} = 0$, we get the solution

$$\begin{aligned}A_1^{\mathrm{H}} = \; & (D_1^{\mathrm{HT}}D_1^{\mathrm{H}} + (\alpha+\lambda_1)\mathrm{I})^{-1} \\ & (D_1^{\mathrm{HT}}X_1 + \lambda_1 A_2^{\mathrm{H}} + \alpha P_1^{\mathrm{H}}X_1),\end{aligned} \tag{8}$$

where I is an identity matrix. We can obtain solution to $A_2^{\mathrm{H}}$ in a very similar way.

2). Fix other variables and update $P_1^{\mathrm{H}}$ and $P_2^{\mathrm{H}}$.

The objective function regarding $P_1^{\mathrm{H}}$ can be written as

$$\min_{P_1^{\mathrm{H}}} \alpha\left\|P_1^{\mathrm{H}}X_1 - A_1^{\mathrm{H}}\right\|_{\mathrm{F}}^2. \tag{9}$$

By setting the derivative with respect to $P_1^{\mathrm{H}}$ to zero, we have the solution $P_1^{\mathrm{H}} = A_1^{\mathrm{H}}X_1(X_1X_1^{\mathrm{T}} + \gamma\mathrm{I})^{-1}$, where $\gamma$ is a regularization parameter. Similarly, the solution to $P_2^{\mathrm{H}}$ is: $P_2^{\mathrm{H}} = A_2^{\mathrm{H}}X_2(X_2X_2^{\mathrm{T}} + \gamma\mathrm{I})^{-1}$.

3). Fix other variables and update $D_1^{\mathrm{H}}$ and $D_2^{\mathrm{H}}$.

By removing the irrelevant terms in (6), we can write the objective function regarding $D_1^{\mathrm{H}}$ as

$$\min_{D_1^{\mathrm{H}}} \left\|X_1 - D_1^{\mathrm{H}}A_1^{\mathrm{H}}\right\|_{\mathrm{F}}^2 \quad s.t. \; \|d^{\mathrm{H}}_{1(:,i)}\| \le 1. \tag{10}$$

Problem (10) can be effectively solved using ADMM algorithm as introduced in [Gu *et al.*, 2014]. We have similar solutions to $D_2^{\mathrm{H}}$.

The above procedures are repeated until convergence. Finally, we obtain a pair of dictionaries $\{D_1^{\mathrm{H}}, D_2^{\mathrm{H}}\}$ that are used to represent high-dimensional image features.

## 5.2 Solving objective (4)

To solve the problem (4), we first reformulate the objective as

$$\begin{aligned}\min_{\substack{D_1^{\mathrm{L}},D_2^{\mathrm{L}},P_1^{\mathrm{L}},\\ P_2^{\mathrm{L}},A_1^{\mathrm{L}},A_2^{\mathrm{L}}}} \quad & \left\|Y_1 - D_1^{\mathrm{L}}A_1^{\mathrm{L}}\right\|_{\mathrm{F}}^2 + \left\|Y_2 - D_2^{\mathrm{L}}A_2^{\mathrm{L}}Y_2\right\|_{\mathrm{F}}^2 \\ & +\beta(\left\|P_1^{\mathrm{L}}Y_1 - A_1^{\mathrm{L}}\right\|_{\mathrm{F}}^2 + \left\|P_2^{\mathrm{L}}Y_2 - A_2^{\mathrm{L}}\right\|_{\mathrm{F}}^2) \\ & +\lambda_2\left\|D_1^{\mathrm{L}} - D_2^{\mathrm{L}}\right\|_{\mathrm{F}}^2, \\ s.t. \quad & \|d^{\mathrm{L}}_{1(:,i)}\| \le 1, \|d^{\mathrm{L}}_{2(:,i)}\| \le 1,\end{aligned} \tag{11}$$

where $\beta$ is a balance parameter.

***Algorithm 1. CPDL for Person Re-identification***

```
Input: Training images in two views $A_1$, $A_2$,
        test images $T_1, T_2$, parameters $\lambda_1$, $\lambda_2$, $\lambda$, $\alpha$, $\beta$.
Output: Matching results.
Training
  Extract dense features from $A_1, A_2$ (Section 4.1), and
    construct feature sets $X_1, X_2, Y_1, Y_2$;
  Learn dictionaries $\{D_1^{\mathrm{H}}, D_2^{\mathrm{H}}\}$ from image-level
    features $X_1, X_2$ (Section 5.1);
  Learn dictionaries $\{D_1^{\mathrm{L}}, D_2^{\mathrm{L}}\}$ from patch-level
    features $Y_1, Y_2$ (Section 5.2);
Testing
  Extract dense features from $T_1, T_2$ (Section 4.1), and
    construct feature sets $X_{t1}, X_{t2}, Y_{t1}, Y_{t2}$;
  Encode $X_{t1}, X_{t2}$ using $\{D_1^{\mathrm{H}}, D_2^{\mathrm{H}}\}$, and perform
    image-level matching (Section 4.4);
  Encode $Y_{t1}, Y_{t2}$ using $\{D_1^{\mathrm{L}}, D_2^{\mathrm{L}}\}$, and perform
    patch-level matching (Section 4.4);
  Fuse matching results in two-levels using (5).
```

We alternatively update the variables in (11), and obtain the sub-problems (with solutions) as follows

$$\min_{A_1^{\mathrm{L}}} \left\|Y_1 - D_1^{\mathrm{L}}A_1^{\mathrm{L}}\right\|_{\mathrm{F}}^2 + \beta\left\|P_1^{\mathrm{L}}Y_1 - A_1^{\mathrm{L}}\right\|_{\mathrm{F}}^2. \tag{12}$$

The solution to (12) is $A_1^{\mathrm{L}} = (D_1^{\mathrm{LT}}D_1^{\mathrm{L}} + \beta\mathrm{I})^{-1}(D_1^{\mathrm{LT}}Y_1 + \beta P_1^{\mathrm{L}}Y_1)$.

$$\min_{P_1^{\mathrm{L}}} \beta\left\|P_1^{\mathrm{L}}Y_1 - A_1^{\mathrm{L}}\right\|_{\mathrm{F}}^2. \tag{13}$$

The optimal solution is $P_1^{\mathrm{L}} = A_1^{\mathrm{L}}Y_1(Y_1Y_1^{\mathrm{T}} + \gamma\mathrm{I})^{-1}$.

$$\begin{aligned}\min_{D_1^{\mathrm{L}}} \quad & \left\|Y_1 - D_1^{\mathrm{L}}A_1^{\mathrm{L}}\right\|_{\mathrm{F}}^2 + \lambda_2\left\|D_1^{\mathrm{L}} - D_2^{\mathrm{L}}\right\|_{\mathrm{F}}^2, \\ s.t. \quad & \|d^{\mathrm{L}}_{1(:,i)}\| \le 1.\end{aligned} \tag{14}$$

We have similar solutions to $A_2^{\mathrm{L}}$, $P_2^{\mathrm{L}}$ and $D_2^{\mathrm{L}}$. The above procedures are repeated until convergence. We finally obtain a pair of optimal dictionaries $\{D_1^{\mathrm{L}}, D_2^{\mathrm{L}}\}$ that are used to reconstruct low-dimensional patch features.

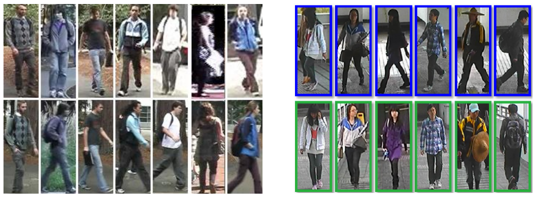
The complete algorithm is summarized in *Algorithm 1*.

# 6 Experiments

In this section, we compare our approach with several related methods on two benchmark datasets, VIPeR [Gray *et al.*, 2007] and CUHK01 Campus [Zhao *et al.*, 2014].

## 6.1 Settings

**Baselines.** We compare our approach with three types of person re-identification methods, which are feature learning methods, metric learning methods and dictionary learning methods. The feature learning methods include symmetry-driven accumulation of local features (SDALF) [Farenzena *et al.*, 2010], local descriptors encoded by Fisher vectors (LDFV) [Ma *et al.*, 2012b], unsupervised salience learning method (eSDC) [Zhao *et al.*, 2013b], salience matching method [Zhao *et al.*, 2013a], and mid-level filters [Zhao *et*



(a) (b)

Figure 2: Illustration of images in (a) VIPeR dataset and (b) CUHK Campus dataset.

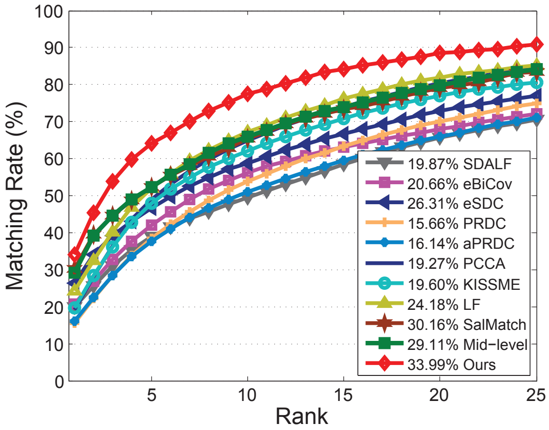


Figure 3: CMC curves of average matching rates on VIPeR dataset. Rank-1 matching rate is marked before the name of each approach.

*al.*, 2014]. The compared metric learning algorithms include probabilistic relative distance comparison (PRDC) [Zheng *et al.*, 2011], large margin nearest neighbor (LMNN) [Weinberger *et al.*, 2005], eBiCov [Ma *et al.*, 2012a], information-theoretic metric learning (ITML) [Davis *et al.*, 2007], pairwise constrained component analysis (PCCA) [Mignon and Jurie, 2012], KISSME [Köstinger *et al.*, 2012], and local Fisher discriminant analysis (LF) [Pedagadi *et al.*, 2013]. We also compare with the dictionary learning method SSCDL [Liu *et al.*, 2014].

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

**Parameter Setting.** There are five parameters in our model, including $\alpha$, $\beta$, $\lambda$, $\lambda_1$ and $\lambda_2$. In the experiments, we empirically set these parameters to achieve the best performance. In particular, $\alpha$ and $\beta$ are set to 2 and 1, respectively. $\lambda$ used in the fusion strategy is chosen in the range $[0\ 1]$. Two parameters $\lambda_1$ and $\lambda_2$ control the effects of cross-view interactions, and we will discuss their settings in the next section.

## 6.2 VIPeR Dataset

The VIPeR dataset was collected in an outdoor academic environment. It contains images of 632 pedestrian pairs under two camera views with different viewpoints. The images in two views have significant variations in pose, viewpoint and illuminations. Figure 2(a) shows some images captured by Camera-1 (first row) and Camera-2 (second row) in the VIPeR dataset. The images are normalized to the size of 128×48 in our experiments.

Table 1: Top ranked matching rates in (%) with 316 persons on VIPeR dataset.

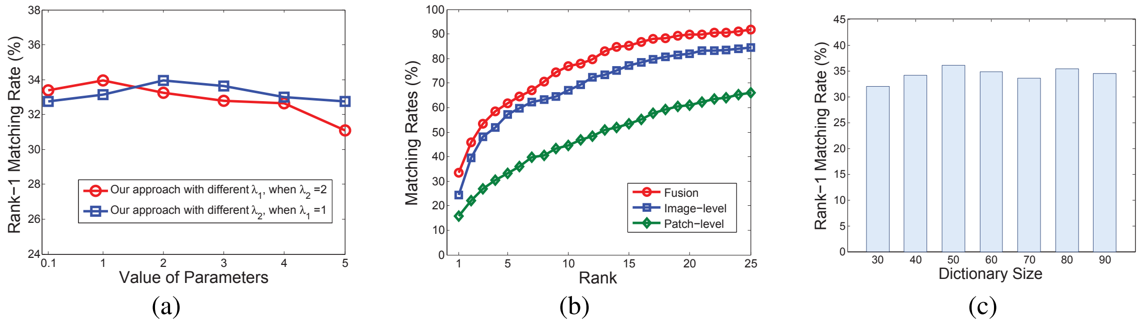
| Method | $r = 1$ | $r = 5$ | $r = 10$ | $r = 20$ |
|---|---|---|---|---|
| PRDC | 15.66 | 38.42 | 53.86 | 70.09 |
| PCCA | 19.27 | 48.89 | 64.91 | 80.28 |
| SDALF | 19.87 | 38.89 | 49.37 | 65.73 |
| eBiCov | 20.66 | 42.00 | 56.18 | 68.00 |
| LDFV | 22.34 | 47.00 | 60.40 | 71.00 |
| LF | 24.11 | 51.24 | 67.09 | 82.01 |
| eSDC | 26.31 | 50.70 | 62.37 | 76.36 |
| SalMat | 30.16 | 53.45 | 65.78 | N/A |
| SSCDL | 25.60 | 53.70 | 68.10 | 83.60 |
| Mid-level | 29.11 | 52.50 | 67.12 | 80.03 |
| Ours | **33.99** | **64.21** | **77.53** | **88.58** |

We follow the evaluation protocol in [Gray and Tao, 2008]. In particular, we randomly select 316 pairs of images for training, and the remaining pairs are used for test. Then, two groups of experiments are conducted. First, the images captured by Camera-1 are utilized as probe images, and the images captured by Camera-2 as gallery images. For the probe images, we match each of them to the gallery set, and obtain the Rank-$k$ rate. The CMC curves are also obtained by using the rates at all ranks. Second, we exchange the training and test sets, and repeat the above procedures. As the raw features for image-level training have very high dimensions, we apply PCA to reduce the dimensionality by keeping the 95% energy. We conduct 10 random tests and report the average results. Each random test has two groups of evaluations as described above.

Figure 3 shows the CMC curves of the compared methods. We can observe that our approach achieves higher matching rates in each rank. Table 1 shows the detailed Rank-1, Rank-5, Rank-10, and Rank-20 matching rates of all the compared methods. It shows that the advanced feature learning methods like salience matching (SalMat) and mid-level filters obtain much better results than metric learning methods. The dictionary learning method SSCDL achieves better Rank-5/10/20 rates than the SalMat and Mid-level methods, which shows the merits of dictionary learning. Our approach achieves the best Rank-1 rate, and significantly improves the Rank-5/10/20 rates, validating the effectiveness of the proposed CPDL framework.

## 6.3 CUHK01 Campus Dataset

The CUHK01 Campus dataset contains pedestrian images of 971 persons in two camera views. It was collected in a campus environment. This dataset shows significant changes of viewpoints. The frontal or back views are captured by Camera-1, while the side views are captured by Camera-2. Figure 2(b) illustrates some images in view 2 (first row) and view 1 (second row). The images are resized to 160×60 in our experiments.



(a) (b) (c)

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

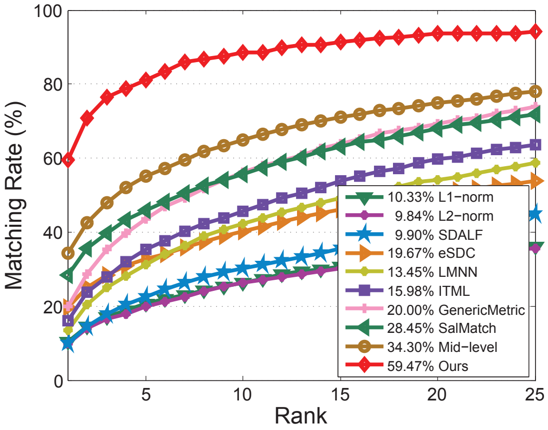


Figure 4: CMC curves of average matching rates on CUHK01 dataset. Rank-1 matching rate is marked before the name of each approach.

We follow the evaluation protocol in [Zhao *et al.*, 2014]. For each person, one image is randomly selected to build the gallery set, and the other one is used to construct the probe set. We map each image in probe set to every gallery image, and calculate the correct matched rank and CMC curves. The whole procedure is repeated for 10 times, and the average CMC curves are generated, as shown in Figure 4. Table 2 shows the detailed Rank-1/5/10/20 matching rates of the compared methods. We can observe that our approach obtains much higher matching rates than other methods. The Rank-1 matching rate is improved by 25.17%, compared to the mid-level filter method.

### 6.4 Discussions

Different from existing methods, the proposed CPDL approach models the interactions between different views, such as the similarities of codings (in the image-level) or dictionaries (in the patch-level). The parameters $\lambda_1$ and $\lambda_2$ control the effects of the cross-view interactions. Figure 5(a) shows the Rank-1 matching rates of our approach with different values of $\lambda_1$ and $\lambda_2$. It shows that our approach is not very sensitive to the choice of parameters in the range $[0\ 5]$. We set $\lambda_1 = 1, \lambda_2 = 2$.

Table 2: Top ranked matching rates in (%) on CUHK01 dataset.

| Method | $r = 1$ | $r = 5$ | $r = 10$ | $r = 20$ |
|---|---|---|---|---|
| SDALF | 9.90 | 22.57 | 30.33 | 41.03 |
| eSDC | 19.67 | 32.71 | 40.28 | 50.57 |
| LMNN | 13.45 | 31.33 | 42.25 | 54.11 |
| ITML | 15.98 | 35.22 | 45.60 | 59.80 |
| SalMat | 28.45 | 45.85 | 55.67 | 68.89 |
| Mid-level | 34.30 | 55.06 | 64.96 | 73.94 |
| Ours | **59.47** | **81.26** | **89.72** | **93.10** |

Figure 5(b) shows the CMC curves of our approach and its two components, i.e., image-level model and patch-level model. We can observe that the representations in image-level and patch-level are complementary to each other, and our approach takes full advantage of the complementary information.

Another important factor in our approach is the size of dictionary. We use the same dictionary size in different views. Figure 5(c) shows the Rank-1 matching rate with different dictionary size. We achieved similar results on the CUHK01 dataset. Accordingly, the dictionary size is set to 50 in our experiments. Also, we note that the matching process in existing feature learning methods (e.g., SalMat or Mid-level filter) is very time consuming. However, our approach adopts a relative small dictionary, which leads to compact representations of images, and therefore speeds up the matching process.

## 7 Conclusions

We proposed a cross-view projective dictionary learning (CPDL) approach for person re-identification in this paper. Our approach learned two pairs of dictionaries across different views in patch-level and image-level, respectively. The learned dictionaries can be used to represent probe and gallery images, leading to robust representations. Experimental results on the public VIPeR and CUHK Campus datasets showed that our approach took full advantages of the complementary information in different views and representation levels, and achieved the state-of-the-art performance compared with the related methods.

## Acknowledgments

This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

## References

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

[Davis *et al.*, 2007] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.

[Farenzena *et al.*, 2010] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010.

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

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

[Gu *et al.*, 2014] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Projective dictionary pair learning for pattern classification. In *NIPS*, 2014.

[Hirzer *et al.*, 2012] Martin Hirzer, Peter M. Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, pages 780–793, 2012.

[Köstinger *et al.*, 2012] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012.

[Layne *et al.*, 2012] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Person re-identification by attributes. In *BMVC*, pages 1–11, 2012.

[Li *et al.*, 2014a] Liangyue Li, Sheng Li, and Yun Fu. Learning low-rank and discriminative dictionary for image classification. *Image Vision Comput.*, 32(10):814–823, 2014.

[Li *et al.*, 2014b] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014.

[Liu *et al.*, 2014] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*, pages 3550–3557, 2014.

[Ma *et al.*, 2012a] Bingpeng Ma, Yu Su, and Frédéric Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, pages 1–11, 2012.

[Ma *et al.*, 2012b] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV Workshops and Demonstration*, pages 413–422, 2012.

[Mignon and Jurie, 2012] Alexis Mignon and Frédéric Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, pages 2666–2672, 2012.

[Ni *et al.*, 2013] Jie Ni, Qiang Qiu, and Rama Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *CVPR*, pages 692–699, 2013.

[Pedagadi *et al.*, 2013] Sateesh Pedagadi, James Orwell, Sergio A. Velastin, and Boghos A. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, pages 3318–3325, 2013.

[Wang *et al.*, 2014] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703, 2014.

[Weinberger *et al.*, 2005] Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005.

[Yang *et al.*, 2014] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li. Salient color names for person re-identification. In *ECCV*, pages 536–551, 2014.

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

[Zhao *et al.*, 2013a] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *ICCV*, pages 2528–2535, 2013.

[Zhao *et al.*, 2013b] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593, 2013.

[Zhao *et al.*, 2014] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *CVPR*, pages 144–151, 2014.

[Zheng *et al.*, 2011] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649–656, 2011.

[Zheng *et al.*, 2012] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Transfer re-identification: From person to set-based verification. In *CVPR*, pages 2650–2657, 2012.