Learning Heterogeneous Dictionary Pair with Feature Projection Matrix for Pedestrian Video Retrieval via Single Query Image

Xiaoke Zhu,[1,6] **Xiao-Yuan Jing,**[*,1,2] **Fei Wu,**[2] **Yunhong Wang,**[3] **Wangmeng Zuo,**[4] **Wei-Shi Zheng**[5]
[1]State Key Laboratory of Software Engineering, School of Computer, Wuhan University, China
[2]College of Automation, Nanjing University of Posts and Telecommunications, China
[3]School of Computer Science and Engineering, Beihang University, China
[4]School of Computer Science and Technology, Harbin Institute of Technology, China
[5]School of Data and Computer Science, Sun Yat-sen University, China
[6] School of Computer and Information Engineering, Henan University, China. [*]Corresponding author.

Abstract

Person re-identification (re-id) plays an important role in video surveillance and forensics applications. In many cases, person re-id needs to be conducted between image and video clip, e.g., re-identifying a suspect from large quantities of pedestrian videos given a single image of him. We call re-id in this scenario as image to video person re-id (IVPR). In practice, image and video are usually represented with different features, and there usually exist large variations between frames within each video. These factors make matching between image and video become a very challenging task. In this paper, we propose a joint feature *p*rojection matrix and *h*eterogeneous *d*ictionary pair *l*earning (PHDL) approach for IVPR. Specifically, PHDL jointly learns an intra-video projection matrix and a pair of heterogeneous image and video dictionaries. With the learned projection matrix, the influence of variations within each video to the matching can be reduced. With the learned dictionary pair, the heterogeneous image and video features can be transformed into coding coefficients with the same dimension, such that the matching can be conducted using coding coefficients. Furthermore, to ensure that the obtained coding coefficients have favorable discriminability, PHDL designs a point-to-set coefficient discriminant term. Experiments on the public iLIDS-VID and PRID 2011 datasets demonstrate the effectiveness of the proposed approach.

Introduction

Person re-identification (re-id) (Li et al. 2014; Zheng et al. 2015c; Li et al. 2015; Tao et al. 2013; Zhang et al. 2015) has been widely studied in computer vision and pattern recognition communities due to its importance in many safety-critical applications, such as automated video surveillance and forensics. Given an image/video of a person captured from one camera, person re-id is the process of identifying the person from images/videos taken from a different camera (Zheng, Gong, and Xiang 2015; Ma, Yang, and Tao 2014; Zheng et al. 2015b; Su et al. 2015; Shi, Hospedales, and Xiang 2015). According to the scenarios of re-identification, existing person re-id works can be roughly divided into two categories: image-based and video-based person re-id methods. The former focuses on the matching between image and image, and most of existing methods belong to this category (Ma, Yuen, and Li 2013; Liu et al. 2013; Qiu, Ni, and Chellappa 2014; Ahmed, Jones, and Marks 2015; Chen et al. 2015; Liu et al. 2014; Jing et al. 2015; Li, Shao, and Fu 2015; Zheng et al. 2015a; Karanam, Li, and Radke 2015). Different from image-based methods, video-based person re-id methods focus on the matching between video and video (Wang et al. 2014; 2016; Liu et al. 2015; Zhu et al. 2016; McLaughlin, Martinez del Rincon, and Miller 2016; You et al. 2016). In both kinds of methods, the two objects to be matched are homogeneous.

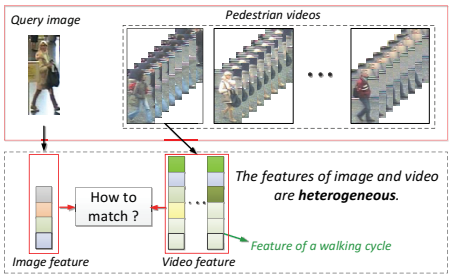


Figure 1: Problem of image to video person re-id.

In many practical cases, person re-id needs to be conducted between image and video. One instance is rapid locating and tracking suspects from masses of city surveillance videos according to an image of the criminal suspect (e.g., Boston marathon bombings event). Another instance is that, an old man who suffers from Alzheimer's disease lost his way in the city, given an image of the old man, the re-id system should retrieve the surveillance video clips containing him. We call re-identification under this scenario as image to video person re-id (IVPR). Figure 1 illustrates the problem of IVPR.

In IVPR, there exist two aspects of difficulties: (1) Image and video are usually represented with different features. In particular, both the visual appearance features and spatial-temporal features can be extracted from a pedestrian video, while only visual appearance features can be extracted from

Figure 2: Image sequences in the iLIDS-VID dataset.

a single image. (2) No matter features are extracted from each frame or each walking cycle, a video can be regarded as a set, and therefore IVPR is actually a point-to-set matching problem. However, there usually exist large variations between different frames or walking cycles within each video, which will make the matching between image and video more tough. Figure 2 shows the intra-video variations.

Motivation

IVPR is an important application in practice, however, it has not been well studied. Existing person re-id methods require that two objects to be matched should be represented with the same kind of feature. Hence, if one tries to apply existing methods to IVPR, the same features should be extracted from image and video. From the first difficulty in IVPR, we can know that only visual appearance features can be extracted from both image and video, which means that the spatial-temporal features contained in video cannot be used by these methods. However, researches in (Wang et al. 2014; 2016; You et al. 2016) have demonstrated the effectiveness of spatial-temporal feature for person re-id, and have also indicated that spatial-temporal feature is complementary to visual appearance features. Therefore, by directly applying these off-the-shelf person re-id methods to IVPR, the useful information contained in video cannot be fully utilized, which will limit their performance. In addition, IVPR is actually a point-to-set matching problem, however, existing methods are not designed for this, and they don't consider the influence of variations within each video to the matching between image and video, which will further hamper their performance.

Motivated by the above analysis, we intend to design an approach, which can make full use of the heterogeneous features contained in image and video, and simultaneously reduce the influence of intra-video variations to the re-identification, for IVPR.

Contribution

The major contributions of this paper are summarized as the following three points:

(1) We are among the first to investigate the problem of image to video person re-identification (IVPR).

(2) We propose a heterogeneous dictionary pair learning framework, with which heterogeneous features of image and video can be transformed into coding coefficients with the same dimension, such that the matching between image and video can be implemented with the obtained coefficients. To ensure that the obtained coefficients own favorable discriminability, we also design a point-to-set coefficient discriminant term for the framework.

(3) To reduce the influence of intra-video variations to the matching between image and video, we design a video congregating term, which increases the compactness of each video by learning a projection matrix, such that the following matching becomes easier.

We name our approach as joint feature *p*rojection matrix and *h*eterogeneous *d*ictionary pair *l*earning (PHDL). A number of IVPR experiments have been conducted. The experimental results demonstrate the efficacy of our approach.

The Proposed Approach

Problem Formulation

Denote by $\mathbf{X} = [x_1, ..., x_i, ..., x_n]$ the feature set of training images, where $x_i \in \mathbb{R}^p$ is the feature of an image from the i^{th} person, and n is the number of persons. To make full use of the information contained in video, we extract both the visual appearance and spatial-temporal features from each walking cycle of the video. Denote by $\mathbf{Y} = [\mathbf{Y}_1, ..., \mathbf{Y}_i, ..., \mathbf{Y}_n]$ the feature set of n pedestrian videos, where $\mathbf{Y}_i = [y_{i,1}, ..., y_{i,j}, ..., y_{i,n_i}]$ is the feature set corresponding to the i^{th} video, n_i is the number of walking cycles in the i^{th} video, and $y_{i,j} \in \mathbb{R}^q$ is the feature extracted from the j^{th} walking cycle. Here, p and q are the dimensions of image and video features, respectively.

Since features of image and video are heterogeneous (different feature types and dimensions), directly matching between image and video is not an easy task. Dictionary learning (DL) is an effective feature learning technique (Jiang, Lin, and Davis 2013; Lu et al. 2014; Gu et al. 2014). By learning a dictionary, DL methods can represent each sample with a coding coefficient. Inspired by this, we can learn different dictionaries for image and video, such that heterogeneous features of images and videos can be transformed into coding coefficients with the same dimension. In this way, we can directly conduct the re-identification with the coefficients of images and videos. To make the obtained coding coefficients suitable for re-identification, we still need to design a discriminant term, which can ensure that the distance between the coefficients of truly matching image and video should be smaller than that between coefficients of wrong matching image and video.

In practice, there usually exist large variations between frames within each video, as well as between different walking cycles within each video. Figure 2 provides some example image sequences that display the intra-video variations. These variations will lead to the result that the obtained coding coefficients of different walking cycles within each video still contain large variations, which is not conducive to the following matching. Therefore, we should reduce the influence of these variations in the process of dictionary learning. To this end, we can learn a feature projection matrix (FPM) for the video data, under which samples with each video cluster together. Figure 3 illustrates the basic idea of our approach.

Denote by $\mathbf{W} \in \mathbb{R}^{q \times q_1}$ the learned FPM for video data, where q_1 is the dimension of projected video features. De-

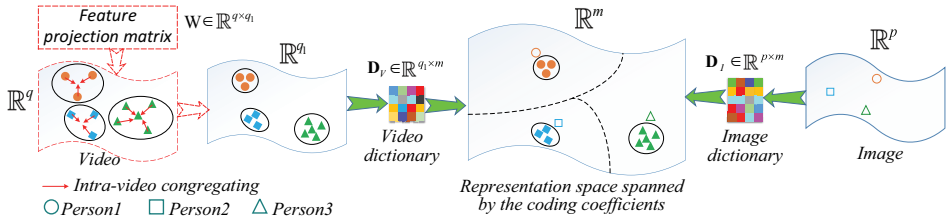


Figure 3: Illustration of the proposed PHDL approach.

note by $\mathbf{D}_I \in \mathbb{R}^{p\times m}$ and $\mathbf{D}_V \in \mathbb{R}^{q_1\times m}$ the learned image and video dictionaries, respectively. Here, m is the number of atoms in $\mathbf{D}_I$ and $\mathbf{D}_V$. Let $\mathbf{A} = [a_1, ..., a_i, ..., a_n]$ represent the coding coefficient matrix of $\mathbf{X}$ over $\mathbf{D}_I$, where a_i is the coefficient of x_i. Let $\mathbf{B}$, $\mathbf{B}_i$, b_{ij} be the coding coefficients of $\mathbf{Y}$, $\mathbf{Y}_i$, $y_{i,j}$ over $\mathbf{D}_V$, respectively. Our objective function is designed as follows:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{D}_I,\mathbf{D}_V} & f_I(\mathbf{D}_I,\mathbf{X},\mathbf{A}) + f_V(\mathbf{W},\mathbf{D}_V,\mathbf{Y},\mathbf{B}) + \\ & \alpha g(\mathbf{W},\mathbf{Y}) + \beta d(\mathbf{A},\mathbf{B}) + \lambda r(\mathbf{W},\mathbf{A},\mathbf{B}) \\ & s.t.\ \|d_{I,i}\|_2^2 \le 1,\ \|d_{V,i}\|_2^2 \le 1,\ \forall i,\end{aligned} \tag{1}$$

where α, β and λ are balancing factors. $d_{I,i}$ ($d_{V,i}$) denotes the i^{th} atom of $\mathbf{D}_I$ ($\mathbf{D}_V$). The constraint is used to restrict the energy of each atom. Details of each term are as follows.

- $f_I(\mathbf{D}_I,\mathbf{X},\mathbf{A}) = \|\mathbf{X}-\mathbf{D}_I\mathbf{A}\|_F^2$ is the image reconstruction fidelity term.
- $f_V(\mathbf{W},\mathbf{D}_V,\mathbf{Y},\mathbf{B}) = \|\mathbf{W}^T\mathbf{Y}-\mathbf{D}_V\mathbf{B}\|_F^2$ is the video reconstruction fidelity term.
- $g(\mathbf{W},\mathbf{Y}) = \sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\|\mathbf{W}^T(y_{i,j}-m_i)\|_2^2$ is the video congregating term, which aims to make each sample move close to the center of video to which it belongs, such that the intra-video variation can be reduced. Here, m_i is the mean vector of $\mathbf{Y}_i$.
- $d(\mathbf{A},\mathbf{B})$ is the point-to-set coefficient discriminant term to ensure that the obtained coding coefficients have good discriminability. Specifically, for each truly matching image-video pair, it requires that the coding coefficient of each sample in the video should move close to that of the image. And for each wrong matching image-video pair, the term requires that the coding coefficient of each sample in the video should be far away from that of the image.

  $$d(\mathbf{A},\mathbf{B}) = \frac{1}{|S|}\sum_{(i,j)\in S} dis(a_i,B_j) - \eta\frac{1}{|Q|}\sum_{(i,j)\in Q} dis(a_i,B_j),$$

  where $dis(a_i,B_j) = \frac{1}{n_j}\sum_{k=1}^{n_j}\|b_{jk}-a_i\|_2^2$, η is a balancing factor, S is the collection of truly matching image-video pairs, and Q represents the collection of wrong matching image-video pairs. Here, $\sharp$ denotes the size of a collection.
- $r(\mathbf{W},\mathbf{A},\mathbf{B}) = \|\mathbf{W}\|_F^2+\|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2$ is the regularization term to regularize FPM and coding coefficients.

The Optimization Algorithm

The objective function in Eq. (1) is not jointly convex to $\mathbf{W},\mathbf{D}_I,\mathbf{D}_V$. However, it is convex w.r.t. each of them if others are fixed. To tackle the energy-minimization in Eq. (1), we separate the objective function into three sub-problems, namely representation coefficient updating, dictionary updating and feature projection matrix updating.

Before solving these three sub-problems, we need to initialize each variable. Specifically, we firstly initialize $\mathbf{W}$ by solving the problem in Eq. (2), which can be easily solved by eigen-decomposition. Then $\mathbf{D}_I$ and $\mathbf{D}_V$ are initialized as random matrices with unit Frobenius norm for each column vector. Finally, we initialize $\mathbf{A}$ and $\mathbf{B}$ by Eq. (3) and Eq. (4), respectively. Both (3) and (4) are ridge regression problems, whose solutions can be analytically derived as $\mathbf{A} = (\mathbf{D}_I^T\mathbf{D}_I+\lambda\mathbf{I})^{-1}\mathbf{D}_I^T\mathbf{X}$ and $\mathbf{B} = (\mathbf{D}_V^T\mathbf{D}_V+\lambda\mathbf{I})^{-1}\mathbf{D}_V^T\mathbf{W}^T\mathbf{Y}$. Here $\mathbf{I}$ is an identity matrix.

$$\min_{\mathbf{W}}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\|\mathbf{W}^T(y_{i,j}-m_i)\|_2^2,\ s.t.\ \mathbf{W}^T\mathbf{W}=\mathbf{I}, \tag{2}$$

$$\min_{\mathbf{A}}\|\mathbf{X}-\mathbf{D}_I\mathbf{A}\|_F^2+\lambda\|\mathbf{A}\|_F^2, \tag{3}$$

$$\min_{\mathbf{B}}\|\mathbf{W}^T\mathbf{Y}-\mathbf{D}_V\mathbf{B}\|_F^2+\lambda\|\mathbf{B}\|_F^2. \tag{4}$$

(1) Update A and B. When $\mathbf{W}$, $\mathbf{D}_I$ and $\mathbf{D}_V$ are fixed, we update $\mathbf{A}$ and $\mathbf{B}$ as follows:

$$\begin{aligned}\min_{a_i}\|x_i-\mathbf{D}_I a_i\|_2^2+\beta\Big(&\frac{1}{|S_{x_i}|}\sum_{(i,j)\in S_{x_i}}\frac{1}{n_j}\sum_{k=1}^{n_j}\|b_{jk}-a_i\|_2^2 \\ &-\eta\frac{1}{|Q_{x_i}|}\sum_{(i,j)\in Q_{x_i}}\frac{1}{n_j}\sum_{k=1}^{n_j}\|b_{jk}-a_i\|_2^2\Big)+\lambda\|a_i\|_2^2,\end{aligned} \tag{5}$$

$$\begin{aligned}\min_{\mathbf{B}_i}\|\mathbf{W}^T\mathbf{Y}_i-\mathbf{D}_V\mathbf{B}_i\|_F^2+\beta\Big(&\frac{1}{|S_{\mathbf{Y}_i}|}\sum_{(j,i)\in S_{\mathbf{Y}_i}} dis(a_j,B_i) \\ &-\eta\frac{1}{|Q_{\mathbf{Y}_i}|}\sum_{(j,i)\in Q_{\mathbf{Y}_i}} dis(a_j,B_i)\Big)+\lambda\|\mathbf{B}_i\|_F^2,\end{aligned} \tag{6}$$

where S_z and Q_z represent the collections of truly matching and wrong matching image-video pairs related to z (x_i or $\mathbf{Y}_i$), respectively.

The solution of (5) can be easily obtained by setting the derivative with respect to a_i to zero.

$$\begin{aligned}a_i = &(\mathbf{D}_I^T\mathbf{D}_I+(\beta-\beta\eta+\lambda)\mathbf{I})^{-1}(\mathbf{D}_I^T x_i + \\ &\beta\Big(\frac{1}{|S_{x_i}|}\sum_{(i,j)\in S_{x_i}}\frac{1}{n_j}\sum_{k=1}^{n_j}b_{jk}-\eta\frac{1}{|Q_{x_i}|}\sum_{(i,j)\in Q_{x_i}}\frac{1}{n_j}\sum_{k=1}^{n_j}b_{jk}\Big).\end{aligned}$$

The solution of (6) can be obtained similarly.

$$\begin{aligned}\mathbf{B}_i = &(\mathbf{D}_V^T\mathbf{D}_V+(\frac{\beta}{n_i}(1-\eta)+\lambda)\mathbf{I})^{-1}(\mathbf{D}_V^T\mathbf{W}^T\mathbf{Y}_i + \\ &\beta\Big(\frac{1}{|S_{\mathbf{Y}_i}|}\sum_{(j,i)\in S_{\mathbf{Y}_i}}\frac{1}{n_i}\mathbf{C}_{j,i}-\eta\frac{1}{|Q_{\mathbf{Y}_i}|}\sum_{(j,i)\in Q_{\mathbf{Y}_i}}\frac{1}{n_i}\mathbf{C}_{j,i}\Big),\end{aligned}$$

where $\mathbf{C}_{j,i} \in \mathbb{R}^{m*n_i}$ is a matrix with each column vector being a_j.

(2) Update $\mathbf{D}_I$ and $\mathbf{D}_V$. By fixing $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{W}$, we can write the objective functions regarding $\mathbf{D}_I$ or $\mathbf{D}_V$ as follows:

$$\min_{\mathbf{D}_I} \|\mathbf{X}-\mathbf{D}_I\mathbf{A}\|_F^2,\ s.t.\ \|d_{I,i}\|_2^2 \le 1,\ \forall i, \tag{7}$$

$$\min_{\mathbf{D}_V} \|\mathbf{W}^T\mathbf{Y}-\mathbf{D}_V\mathbf{B}\|_F^2,\ s.t.\ \|d_{V,i}\|_2^2 \le 1,\ \forall i, \tag{8}$$

The optimal solutions of $\mathbf{D}_I$ and $\mathbf{D}_V$ can be obtained by using the ADMM algorithm as introduced in (Gu et al. 2014). Specifically, by separately introducing a variable $\mathbf{S}$, (7) and (8) can be rewritten as:

$$\min_{\mathbf{D}_I,\mathbf{S}} \|\mathbf{X}-\mathbf{D}_I\mathbf{A}\|_F^2,\quad s.t.\ \mathbf{D}_I = \mathbf{S},\ \|s_i\|_2^2 \le 1,\ \forall i, \tag{9}$$

$$\min_{\mathbf{D}_V,\mathbf{S}} \|\mathbf{W}^T\mathbf{Y}-\mathbf{D}_V\mathbf{B}\|_F^2,\quad s.t.\ \mathbf{D}_V = \mathbf{S},\ \|s_i\|_2^2 \le 1,\ \forall i, \tag{10}$$

where s_i represents the i^{th} atom in $\mathbf{S}$.

The optimal solution of (9) can be obtained by updating the following three equations iteratively:

$$\begin{cases} \mathbf{D}_I = \min_{\mathbf{D}_I} \|\mathbf{X}-\mathbf{D}_I\mathbf{A}\|_F^2 + \rho\|\mathbf{D}_I - \mathbf{S} + \mathbf{T}\|_F^2 \\ \mathbf{S} = \min_{\mathbf{S}} \rho\|\mathbf{D}_I - \mathbf{S} + \mathbf{T}\|_F^2,\quad s.t.\ \|s_i\|_2^2 \le 1 \\ \mathbf{T} = \mathbf{T} + \mathbf{D}_I - \mathbf{S},\ update\ \rho\ if\ appropriate \end{cases},$$

where the initial values of $\mathbf{S}$ and $\mathbf{T}$ are $\mathbf{D}_I$ and zero matrix, respectively. Problem (10) can be solved in a similar way.

(3) Update $\mathbf{W}$. When $\mathbf{D}_I$, $\mathbf{D}_V$, $\mathbf{A}$ and $\mathbf{B}$ are fixed, the objective function related to $\mathbf{W}$ can be written as follows:

$$\begin{aligned} \min_{\mathbf{W}} \|\mathbf{W}^T\mathbf{Y}-\mathbf{D}_V\mathbf{B}\|_F^2 + \lambda\|\mathbf{W}\|_F^2 + \\ \alpha\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\|\mathbf{W}^T(y_{i,j}-m_i)\|_2^2. \end{aligned} \tag{11}$$

By setting the derivative with respect to $\mathbf{W}$ to zero, the solution of Eq. (11) can be derived as:

$$\mathbf{W} = (\mathbf{Y}\mathbf{Y}^T + \alpha\mathbf{P} + \lambda\mathbf{I})^{-1}\mathbf{Y}\mathbf{B}^T\mathbf{D}_V^T, \tag{12}$$

where $\mathbf{P} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{i,j}-m_i)(y_{i,j}-m_i)^T$. Algorithm 1 summarizes the optimization process of our approach.

Computational Complexity

In the designed optimization algorithm, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{D}_I$, $\mathbf{D}_V$ and $\mathbf{W}$ are updated alternatively. In each iteration, the time complexity of updating $\mathbf{A}$ is $O(m^2p+m^3+mpp+n(m^2+mp))$; updating $\mathbf{B}$ takes $O(m^2q_1 + m^3 + mqq_1 + N(m^2 + mq))$, where N is the total number of samples in $\mathbf{Y}$; updating $\mathbf{D}_I$ takes $Ok(p^2n + pnm + m^2n + m^3 + pm^2)$, where k is the iteration number in the ADMM algorithm, and it is usually smaller than 10; similarly, updating $\mathbf{D}_V$ costs $Ok(q_1qN + q_1Nm + m^2N + m^3 + q_1m^2)$; the time complexity of updating $\mathbf{W}$ is $O(q^2N + q^3 + Nmq + qq_1m)$. The dictionary size m is usually much smaller than the sample dimensions p and q, and N may be also large if each video contains a number of walking cycles. Therefore, the major computational burden in the training phase of PHDL is on updating $\mathbf{W}$. Fortunately, the operation that costs $O(q^2N + q^3)$ in Eq. (12), i.e., $(\mathbf{Y}\mathbf{Y}^T + \alpha\mathbf{P} + \lambda\mathbf{I})^{-1}\mathbf{Y}$, will not change in the iteration, and thus can be pre-computed. This greatly accelerates the training process.

Algorithm 1 Joint feature projection matrix and heterogeneous dictionary pair learning (PHDL)

Require: Training image and video sets $\mathbf{X}$ and $\mathbf{Y}$
Ensure: $\mathbf{D}_I$, $\mathbf{D}_V$ and $\mathbf{W}$
1: Initialize $\mathbf{D}_I$, $\mathbf{D}_V$, $\mathbf{W}$, $\mathbf{A}$, $\mathbf{B}$, α, β, λ, and η
2: **while** not converge **do**
3: Fix $\mathbf{W}$, $\mathbf{D}_I$ and $\mathbf{D}_V$, update $\mathbf{A}$ and $\mathbf{B}$ according to (5) and (6), respectively;
4: Fix $\mathbf{W}$, $\mathbf{A}$ and $\mathbf{B}$, update $\mathbf{D}_I$ and $\mathbf{D}_V$ according to (7) and (8), respectively;
5: Fix $\mathbf{D}_I$, $\mathbf{D}_V$, $\mathbf{A}$ and $\mathbf{B}$, update $\mathbf{W}$ according to (11);
6: **end while**
7: **return** $\mathbf{D}_I$, $\mathbf{D}_V$ and $\mathbf{W}$;

Re-identification

Let x be the feature of a probe image, and $\mathbf{Z} = [\mathbf{Z}_1, ..., \mathbf{Z}_i, ..., \mathbf{Z}_l]$ be feature set of l gallery videos, where $\mathbf{Z}_i = [z_{i,1}, ..., z_{i,j}, ...z_{i,n_i}]$ denotes the feature set of the i^{th} gallery video. Here, $z_{i,j}$ is the j^{th} sample in $\mathbf{Z}_i$, n_i is the sample number of $\mathbf{Z}_i$. With the learned dictionary pair and feature projection matrix ($\mathbf{D}_I$, $\mathbf{D}_V$ and $\mathbf{W}$), we can re-identify x in $\mathbf{Z}$ as follows.

(1) Compute the representation coefficient of the probe image x over image dictionary $\mathbf{D}_I$ by solving (3). Denote by a the coefficient of x.

(2) Compute the representation coefficients of gallery videos by solving (4). Denote by $\mathbf{G}$, $\mathbf{G}_i$, g_{ij} the representation coefficients of $\mathbf{Z}$, $\mathbf{Z}_i$ and $z_{i,j}$ over $\mathbf{D}_V$, respectively.

(3) Re-identify x in Z with the obtained coefficients. Firstly, we compute the distance between a and $\mathbf{G}_i$ by $d_i = \sum_{j=1}^{n_i}\|a - g_{ij}\|_2^2$. Then we can obtain the matching result by sorting the obtained distances in ascending order.

Comparison with Existing Dictionary Learning Methods

Dictionary learning (DL) is an effective feature learning technique. Recently, some DL based person re-id methods have been presented, which bridge two different camera views by learning a pair of dictionaries (Liu et al. 2014; Jing et al. 2015; Li, Shao, and Fu 2015). The major differences between PHDL and these methods are three-fold: (1) These methods are designed for matching between images, while PHDL is designed for matching between image and video. (2) They cannot deal with the intra-video variations, while PHDL reduces the influence of intra-video variation by learning a feature projection matrix for video data. (3)They focus on the one-to-one relationship between images from two camera views, while PHDL aims to deal with the one-to-many correspondence between image and video.

Experimental Results

Datasets

The **iLIDS-VID** person sequence dataset (Wang et al. 2014) consists of 600 image sequences (i.e., video clips) for 300 persons, with each person having one pair of image sequences from two camera views. The length of each image

sequence changes from 22 to 192 frames, with an average of 71. The **PRID 2011** person sequence dataset (Hirzer et al. 2011) consists of image sequences recorded from two disjoint cameras (camera-A and camera-B). Camera-A and camera-B contain 385 and 749 person sequences, respectively. Among them, the first 200 persons appear in both views. Each image sequence has variable length consisting of 5 to 675 image frames, with an average number of 84.

Experimental Settings

Baselines. To evaluate the efficacy of our PHDL approach, we compare PHDL with several state-of-the-art person re-id methods and general point to set based matching methods. The person re-id methods include **KISSME** (Kostinger et al. 2012), **RDC** (Zheng, Gong, and Xiang 2013), **ISR** (Lisanti et al. 2015), and **XQDA** (Liao et al. 2015). The point to set based methods include **PSDML** (Zhu et al. 2013), and **LERM** (Huang et al. 2014). For all compared methods, we perform experiments with the source codes provided by the original authors.

Feature Representation. In experiments, we employ the WHOSE feature, which is a kind of hybrid visual appearance descriptor proposed in (Lisanti et al. 2015), to represent each pedestrian image. For the video, we extract WHOSE feature and STFV3D (Liu et al. 2015), which is a spatial-temporal feature descriptor, from each walking cycle.

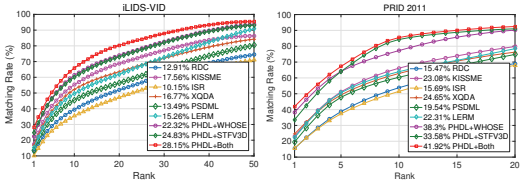
Evaluation Setting. For evaluation, we randomly sample one image from each sequence of the first camera to form the image set, and use the image sequences from the other camera as the video set. Here, the corresponding image and video having the same identity form an image-video pair. Then, all image-video pairs are randomly split into two sets of equal size, with one for training and the other for testing. For the PRID 2011 dataset, the sequence pairs with less than 20 frames are ignored due to the requirement on the sequence length for extracting walking cycles (Liu et al. 2015).

Parameter Setting. There are four parameters in our model, including α, β, λ, and η. In experiments, we set them by using the 5-fold cross validation technique with training data. In particular, they are set as $\alpha = 10$, $\beta = 0.8$, $\lambda = 0.012$ and $\eta = 0.12$ for the iLIDS-VID dataset, $\alpha = 12$, $\beta = 0.7$, $\lambda = 0.01$ and $\eta = 0.14$ for the PRID 2011 dataset. In addition, the size of image and video dictionaries is set as 120 for iLIDS-VID, and 180 for PRID 2011. The number of columns in $\mathbf{W}$ is set as 460 and 380 for iLIDS-VID and PRID 2011, respectively.

We employ the standard cumulated matching characteristics (CMC) curve as our evaluation metric, and report the rank-k matching rates. We repeat each experiment 10 times and report the average results of all methods.

Results and Analysis

In experiments, WHOSE descriptor is employed for competing methods as the representation of image and video. Figure 4 (a) 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 1-50 matching rates of all the compared methods. "+WHOSE" ("+STFV3D", "+Both") means that PHDL employs the



(a) (b)

Figure 4: CMC curves of average matching rates on the (a) iLIDS-VID and (b) PRID 2011 datasets. Rank-1 matching rate is marked before the name of each method.

Table 1: Top r ranked matching rates (%) on iLIDS-VID.

Method	r=1	r=5	r=10	r=20	r=50
RDC	12.91	29.02	39.55	51.94	74.40
KISSME	17.56	41.73	55.28	68.74	86.36
ISR	10.15	25.86	35.39	47.24	71.05
XQDA	16.77	38.58	52.31	63.55	84.30
PSDML	13.49	33.75	45.56	56.33	80.46
LERM	15.26	37.12	49.68	61.95	90.92
PHDL+WHOSE	**22.32**	**46.75**	**61.29**	**73.65**	**93.37**
PHDL+STFV3D	**24.83**	**46.31**	**60.06**	**73.13**	**93.29**
PHDL+Both	**28.15**	**50.37**	**65.88**	**80.35**	**95.42**

Table 2: Top r ranked matching rates (%) on PRID 2011.

Method	r=1	r=5	r=10	r=15	r=20
RDC	15.47	38.75	53.82	62.65	69.02
KISSME	23.08	51.22	66.15	73.91	79.81
ISR	15.69	37.37	51.53	60.47	67.95
XQDA	24.65	49.29	62.83	70.64	76.28
PSDML	19.54	47.81	60.42	67.65	74.83
LERM	22.31	50.66	63.95	71.09	78.47
PHDL+WHOSE	**38.30**	**64.12**	**77.26**	**85.73**	**90.18**
PHDL+STFV3D	**33.58**	**64.04**	**84.27**	**88.76**	**91.01**
PHDL+Both	**41.92**	**67.25**	**85.47**	**90.04**	**92.44**

WHOSE (STFV3D, both the WHOSE and STFV3D) feature to represent the video. It can be seen that: (i) PHDL achieves the best matching results; (ii) when both the WHOSE and STFV3D features are used for matching, the performance of PHDL is significantly improved, which further illustrates the effectiveness of PHDL for IVPR. **The main reasons why our approach can achieve better results are three-fold**: (1) By learning a heterogeneous dictionary pair, PHDL can make full use of the information contained in video. (2) We designed a point-to-set coefficient discriminant term for PHDL, such that the learned dictionary pair has favorable discriminability. (3) PHDL reduces the intra-video variations by learning a feature projection matrix.

Table 2 and Figure 4 (b) report the top ranked matching rates on the PRID 2011 dataset. It is observed that our PHDL approach obtains much higher matching rates than other methods. In particular, take the rank-1 matching rate as an example, PHDL improves the average matching rate at least by 12.2% (=36.8%-24.6%).

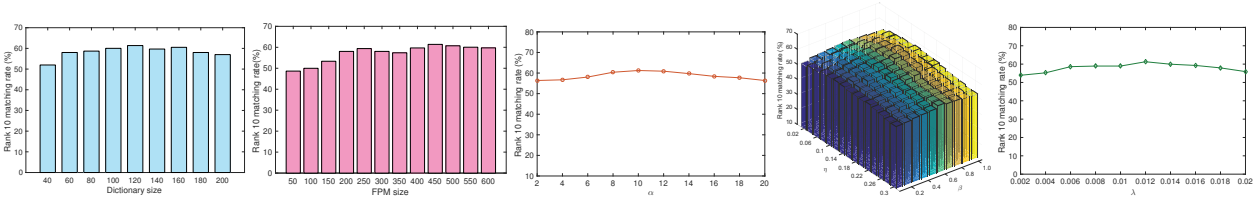


Figure 5: Rank-10 results of PHDL versus different (a) dictionary sizes, (b) FPM sizes, (c) α, (d) β, η (e) λ on iLIDS-VID, where WHOSE feature is employed as the representation of each waling cycle in the video.

Discussion

Effect of Feature Projection Matrix. In PHDL, the feature projection matrix (FPM) $\mathbf{W}$ is used to reduce the intra-video variation, such that the following matching becomes easier. To evaluate the effect of $\mathbf{W}$, we generate a modified version of PHDL by removing $\mathbf{W}$, which is called PHDL-W, and observe its performance. Table 3 reports the top ranked results of PHDL and PHDL-W on the iLIDS-VID dataset. Here, WHOSE feature is employed as the representation of video. We can see that without using $\mathbf{W}$, the performance of PHDL declines, which means that learning FPM is beneficial to improving the discriminability of the coding coefficients. More specifically, without using $\mathbf{W}$, the rank-1 matching rate of PHDL is decreased by 2.24% (22.32%-20.08%) on iLIDS-VID. Similar results can be obtained on PRID 2011.

Effect of Dictionary Size and FPM Size. The size of image and video dictionaries, i.e., the number of atoms in $\mathbf{D}_I$ and $\mathbf{D}_V$, is another important factor in PHDL. To observe the effect of dictionary size, we conduct experiments by setting different values to it. Figure 5 (a) plots the rank-10 matching rates of PHDL versus different dictionary sizes on iLIDS-VID. We can see that PHDL obtains a relatively good result when dictionary size is set as 120, which means that PHDL is able to compute a pair of compact dictionaries.

We also evaluate the effect of FPM size (i.e., the column size of $\mathbf{W}$) to the performance of our PHDL approach. Figure 5 (b) plots the rank-10 matching rates of PHDL versus different column sizes of $\mathbf{W}$ on the iLIDS-VID dataset. We can see that, PHDL can achieve stable performance when the column size of $\mathbf{W}$ is in the range of [400 600]. Similar effects can be observed on the PRID 2011 dataset.

Parameter Analysis. In this experiment, we investigate the effect of parameters of our approach, including α, β, λ and η. α balances the effect of the video congregating term. Parameter β controls the effect of point-to-set coefficient discriminant term. Parameter λ controls the effect of regularization term. Parameter η balances the effects of positive and negative image-video pairs. When some of the parameters are evaluated, the others are fixed as the values given in the section of experimental settings.

We take the experiment on the iLIDS-VID dataset as an example. Figure 5 (c)-(e) shows the rank-10 matching rates of our approach versus different values of α, β, η, and λ on the iLIDS-VID dataset. We can observe that:(1) PHDL is not sensitive to the choice of α in the range of [6, 16]; (2) PHDL achieves the best performance when β and η are set as 0.8 and 0.12, respectively. (3) PHDL can obtain relatively good performance when λ is in the range of [0.006, 0.016]. Similar effects can be observed on the PRID 2011 dataset.

Table 3: Top r ranked matching rates (%) of PHDL and PHDL-W on the iLIDS-VID dataset.

Method	r=1	r=5	r=10	r=20	r=50
PHDL-W	20.08	44.37	58.94	71.46	92.53
PHDL	**22.32**	**46.75**	**61.29**	**73.65**	**93.37**

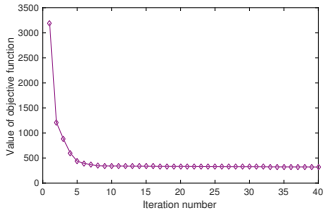


Figure 6: Convergence curve of PHDL on iLIDS-VID.

Convergence Analysis. The proposed optimization algorithm for PHDL is an alternate iterative optimization algorithm. In each iteration, $\{\mathbf{A}, \mathbf{B}\}$, $\{\mathbf{D}_I, \mathbf{D}_V\}$ and $\mathbf{W}$ are updated alternatively, and each sub-problem is convex. In this experiment, we evaluate the performance of PHDL with different numbers of iterations. Figure 6 shows the convergence curves of our algorithm on the iLIDS-VID dataset. One can see that the energy drops quickly and begins to stabilize after 15 iterations. In most of our experiments, our algorithm will converge in less than 20 iterations.

Conclusion

In this paper, we investigate the problem of image to video person re-identification (IVPR) for the first time, and propose a novel approach named PHDL. PHDL can learn a pair of heterogeneous dictionaries as well as a feature projection matrix (FPM) from the training image-video pairs. With the FPM, the variation within video can be reduced. With the dictionary pair, PHDL can realize the matching between heterogeneous image and video features by using their coding coefficients over corresponding dictionaries. Experimental results on two widely used person sequence datasets, i.e., iLIDS-VID and PRID 2011 datasets, demonstrate that our

PHDL can achieve better results than several state-of-the-art methods in the IVPR task.

Acknowledgments

Thanks for the valuable comments of Editor and reviewers. This work was supported by the NSFC (Nos. 61272273, 61671182, 61522115).

References

Ahmed, E.; Jones, M.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *CVPR, IEEE Conference on*, 3908–3916.

Chen, D.; Yuan, Z.; Hua, G.; Zheng, N.; and Wang, J. 2015. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, 1565–1573.

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Projective dictionary pair learning for pattern classification. In *NIPS*, 793–801.

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

Huang, Z.; Wang, R.; Shan, S.; and Chen, X. 2014. Learning euclidean-to-riemannian metric for point-to-set classification. In *CVPR, IEEE Conference on*, 1677–1684.

Jiang, Z.; Lin, Z.; and Davis, L. S. 2013. Label consistent k-svd: Learning a discriminative dictionary for recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35(11):2651–2664.

Jing, X.-Y.; Zhu, X.; Wu, F.; You, X.; Liu, Q.; Yue, D.; Hu, R.; and Xu, B. 2015. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR, IEEE Conference on*, 695–704.

Karanam, S.; Li, Y.; and Radke, R. J. 2015. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV, IEEE Conference on*, 4516–4524.

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

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR, IEEE Conference on*, 152–159.

Li, X.; Zheng, W.-S.; Wang, X.; Xiang, T.; and Gong, S. 2015. Multi-scale learning for low-resolution person re-identification. In *ICCV, IEEE Conference on*, 3765–3773.

Li, S.; Shao, M.; and Fu, Y. 2015. Cross-view projective dictionary learning for person re-identification. In *IJCAI*, 2155–2161.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR, IEEE Conference on*, 2197–2206.

Lisanti, G.; Masi, I.; Bagdanov, A.; and Del Bimbo, A. 2015. Person re-identification by iterative re-weighted sparse ranking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 37(8):1629–1642.

Liu, C.; Loy, C. C.; Gong, S.; and Wang, G. 2013. Pop: Person re-identification post-rank optimisation. In *ICCV*, 441–448.

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

Liu, K.; Ma, B.; Zhang, W.; and Huang, R. 2015. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV, IEEE Conference on*, 3810–3818.

Lu, J.; Wang, G.; Deng, W.; and Moulin, P. 2014. Simultaneous feature and dictionary learning for image set based face recognition. In *ECCV*, 265–280.

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

Ma, A. J.; Yuen, P. C.; and Li, J. 2013. Domain transfer support vector ranking for person re-identification without target camera label information. In *ICCV, IEEE Conference on*, 3567–3574.

McLaughlin, N.; Martinez del Rincon, J.; and Miller, P. 2016. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 1325–1334.

Qiu, Q.; Ni, J.; and Chellappa, R. 2014. Dictionary-based domain adaptation methods for the re-identification of faces. In *Person Re-Identification*. 269–285.

Shi, Z.; Hospedales, T. M.; and Xiang, T. 2015. Transferring a semantic representation for person re-identification and search. In *CVPR, IEEE Conference on*, 4184–4193.

Su, C.; Yang, F.; Zhang, S.; Tian, Q.; Davis, L. S.; and Gao, W. 2015. Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV*, 3739–3747.

Tao, D.; Jin, L.; Wang, Y.; Yuan, Y.; and Li, X. 2013. Person re-identification by regularized smoothing kiss metric learning. *Circuits and Systems for Video Technology, IEEE Transactions on* 23(10):1675–1685.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *ECCV*. 688–703.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2016. Person re-identification by discriminative selection in video ranking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, DOI: 10.1109/TPAMI.2016.2522418*.

You, J.; Wu, A.; Li, X.; and Zheng, W.-S. 2016. Top-push video-based person re-identification. In *CVPR*, 1345–1353.

Zhang, R.; Lin, L.; Zhang, R.; Zuo, W.; and Zhang, L. 2015. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *Image Processing, IEEE Transactions on* 24(12):4766–4779.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Bu, J.; and Tian, Q. 2015a. Scalable person re-identification: A benchmark. In *ICCV, IEEE Conference on*, 1116–1124.

Zheng, L.; Wang, S.; Tian, L.; He, F.; Liu, Z.; and Tian, Q. 2015b. Query-adaptive late fusion for image search and person re-identification. In *CVPR, IEEE Conference on*, 1741–1750.

Zheng, W.-S.; Li, X.; Xiang, T.; Liao, S.; Lai, J.; and Gong, S. 2015c. Partial person re-identification. In *ICCV, IEEE Conference on*, 4678–4686.

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

Zheng, W.-S.; Gong, S.; and Xiang, T. 2015. Towards open-world person re-identification by one-shot group-based verification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 38(3):591–606.

Zhu, P.; Zhang, L.; Zuo, W.; and Zhang, D. 2013. From point to set: Extend the learning of distance metrics. In *ICCV*, 2664–2671.

Zhu, X.; Jing, X.-Y.; Wu, F.; and Feng, H. 2016. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *IJCAI*, 3552–3559.