Image to Video Person Re-identification by Learning Heterogeneous Dictionary Pair with Feature Projection Matrix

Xiaoke Zhu, Xiao-Yuan Jing, Xinge You, Wangmeng Zuo, Shiguang Shan, Wei-Shi Zheng

***Abstract*—Person re-identification plays an important role in video surveillance and forensics applications. In many cases, person re-identification 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 the suspect. We call re-identification in this scenario as image to video person re-identification (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 caused by variations within each video on 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 by using the coding coefficients. Furthermore, to ensure that the obtained coding coefficients own favorable discriminability, PHDL designs a point-to-set coefficient discriminant term. To make better use of the complementary spatial-temporal and visual appearance information contained in pedestrian video data, we further propose a multi-view PHDL approach, which can fuse different video information effectively in the dictionary learning process. Experiments on four publicly available person sequence datasets demonstrate the effectiveness of the proposed approaches.**

***Index Terms*—Person re-identification, Image to video person re-identification, heterogeneous dictionary pair learning, feature projection matrix, multi-view learning.**

I. Introduction

(Corresponding author: Xiao-Yuan Jing.)

Xiaoke Zhu is with the State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China, and also with School of Computer and Information Engineering, Henan University, Kaifeng 475001, China (e-mail: whuzxk@whu.edu.cn).

Xiao-Yuan Jing is with the State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China, and also with the School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: jingxy_2000@126.com).

Xinge You is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, China. (e-mail: youxg@mail.hust.edu.cn).

Wangmeng Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology, China. (e-mail: wmzuo@hit.edu.cn).

Shiguang Shan is with the Key Lab of Intelligent Information Process of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, China. (e-mail: sgshan@ict.ac.cn)

Wei-Shi Zheng is with the School of Data and Computer Science, Sun Yat-sen University, China. (e-mail: wszheng@ieee.org)

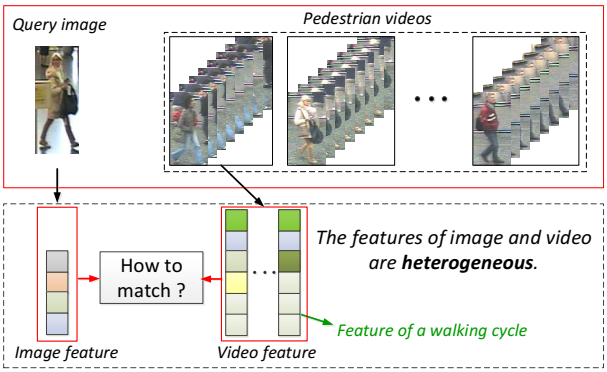


Fig. 1. Problem of image to video person re-identification.

PERSON re-identification [1]–[6] 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-identification is the process of identifying the person from images/videos taken from a different camera [7]–[14]. According to the scenarios of re-identification, existing person re-identification works can be roughly divided into two categories: image-based and video-based person re-identification methods. The former focuses on the matching between image and image, and most of existing methods belong to this category [15]–[24]. Different from image-based methods, video-based person re-identification methods focus on the matching between video and video [25]–[30]. In both kinds of methods, the two objects to be matched are homogeneous.

In many practical cases, person re-identification 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-identification system should retrieve the surveillance video clips that contain him. We call re-identification under this scenario as image to video person re-identification (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,

Fig. 2. Example image sequences in the iLIDS-VID dataset. 20 images are sampled and shown for each sequence.

while only visual appearance features can be extracted from 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 increase the difficulty of matching between image and video. Figure 2 shows the intra-video variations.

A. Motivation

IVPR is an important application in practice, however, it has not been well studied. Existing person re-identification 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 [25], [26], [30] have demonstrated the effectiveness of spatial-temporal feature for person re-identification, and have also indicated that spatial-temporal feature is complementary to visual appearance features. Therefore, by directly applying these off-the-shelf person re-identification 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 on the matching between image and video, which will further hamper their performance.

Motivated by the above analysis, we intend to design an approach for IVPR, which can realize the matching between heterogeneous image and video features, and simultaneously reduce the influence of intra-video variations on the re-identification.

In addition to the above consideration, there is another issue to be solved in the process of dictionary learning, that is, how to make full use of the spatial-temporal and visual appearance information contained in pedestrian video. In practice, the contributions of spatial-temporal information and visual appearance information to the re-identification are usually different, and therefore should be treated differently in the dictionary learning process.

B. Contribution

This is an extended version of our conference paper [31]. The major contributions of this paper are summarized as follows:

- Image to video person re-identification (IVPR) is an important application, but has not been well studied. In this paper, we have systematically investigated the problem of IVPR, and analyzed the difficulties existed in it.
- We have proposed a joint feature *p*rojection matrix and *h*eterogeneous *d*ictionary pair *l*earning (PHDL) approach, 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. The feature projection matrix is used to reduce the intra-video variations, such that the following matching between image and video will become easier. To the best of our knowledge, this is the pioneer work to utilize heterogeneous dictionary pair learning technique to solve the matching between pedestrian image and video.
- We have proposed a multi-view PHDL (MPHDL) approach, which learns different dictionaries and projection matrices for different kinds of video features, and assigns different weights to different features, such that the information contained in video can be exploited more effectively. And we have compared the performances of the proposed PHDL and MPHDL approaches in experiments. Experimental results show that MPHDL can significantly improve the re-identification performance of PHDL.
- We have designed a new model, i.e., image-video triplet (IVT), to describe the relationship between image and video. An IVT consists of an image and two videos (a truly matching video and an "impostor" video), or a video and two images (a truly matching image and an "impostor" image). Based on the designed IVT model, we further design an IVT constraint, which requires that the distance between coefficients of truly matched image and video should be smaller than that between coefficients of wrong matched image and video in each IVT. The designed IVT model and constraint work together to improve the discriminability of learned image to video re-identification model.
- We have conducted extensive experiments on four publicly available person image sequence datasets, including iLIDS-VID, PRID 2011, MARS and HDA+. Experimental results have shown that our proposed approaches can achieve higher matching rates than the competing methods.

II. Related Work

In this section, we briefly review two types of works that are related to our approach: person re-identification and dictionary learning.

A. Person Re-identification

In recent years, many person re-identification methods have been presented [32]–[37]. According to the re-identification

scenarios, existing person re-identification works can be roughly divided into two groups: image-based person re-identification methods and video-based person re-identification methods.

1) Image-based person re-identification: Existing image-based person re-identification methods can be roughly classified into two categories: ***(1)*** Methods based on feature representation [38]–[40]: they focus on seeking a distinct and robust feature representation for matching. For example, a pedestrian image representation is presented in [41], which represents each image with the symmetry-driven accumulation of local features. In [42], a view-independent signature is proposed by mapping local descriptors extracted from RGB-D sensors on a 3D body model. In [38], subject-discriminative feature is presented to relieve the effect of viewpoint changes. In [39], a part-based approach is proposed for representing the appearance of pedestrians. The method presented in [43] designs an efficient descriptor of person appearance, i.e., weighted histograms of overlapping stripes (WHOS). In [44], a clothing context-aware appearance model is presented, which is robust to varying imaging conditions as well as appearance changes. ***(2)*** Methods based on distance learning [45]–[50]: they focus on seeking an optimal distance metric for person re-identification. For example, in [51], Hirzer et al. learned a metric from pairs of samples belonging to different cameras using discriminative Mahalanobis metric learning, which can be efficiently solved after some relaxations. In [52], a distance metric is learned based on equivalence constraints from a statistical inference perspective. In [53], Zheng et al. learned a Mahalanobis distance metric with a probabilistic relative distance comparison (RDC) method. The method presented in [47] utilizes local Fisher Discriminant Analysis (LFDA) to map high dimensional features into a more discriminative low dimensional space. Tao et al. [49] extended the work [52] by integrating smoothing and regularization techniques for robustly estimating covariance matrices.

2) Video-based person re-identification: Video-based person re-identification mainly focuses on matching between pedestrian videos [54]. Recently, a few Video-based person re-identification methods have been presented. The discriminative video fragments selection and ranking (DVR) method [25], [26] selects the most discriminative video fragments for matching by using the HOG3D features and optic flow energy profile. Method STFV3D [27] divides the human images within a walking cycle into a series of body-action units, and then learns Fisher vectors from each unit as the representation of the walking cycle. Top-push distance learning model (TDL) [30] integrates a top-push constraint into the distance learning, such that the matching model is more effective towards selecting more discriminative features to distinguish different persons. In [28], a simultaneous intra-video and inter-video distance learning (SI^{2}DL) approach is developed for video-based person re-identification, which can deal with intra-video and inter-video variations simultaneously.

These methods have relieved the difficulties in person re-identification to an extent. However, as an important practical re-identification scenario, image to video person re-identification has not been well studied in existing methods.

B. Dictionary Learning

As a powerful technique for learning expressive bases in sample space, dictionary learning has attracted lots of attention during the past decades [55]–[57]. Some popular DL methods include fisher discrimination dictionary learning (FDDL) [58], label consistent K-SVD [59], projective dictionary pair learning (DPL) [60], semi-coupled dictionary learning (SCDL) [61]. Most recently, some dictionary learning based methods have been proposed for person re-identification [20]–[22], [62], [63]. In [20], a semi-supervised coupled dictionary learning (SSCDL) approach is presented for person re-identification, which learns a pair of dictionaries for two camera views. In [21], Jing et al. proposed a semi-coupled low-rank discriminant dictionary learning (SLD^{2}L) approach for super-resolution person re-identification, which learns a pair of dictionaries from the training HR and LR images. In [22], a cross-view projective dictionary learning (CPDL) approach is presented, which learns effective features for persons across different views. These methods have demonstrated the effectiveness of dictionary learning technique to person re-identification.

The major differences between PHDL and these existing dictionary learning based person re-identification methods are three-fold: (1) These methods are designed for matching between images, while PHDL is designed for matching between image and video. (2) These methods 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) These methods are designed for image-based person re-identification, thus they only can be used to describe the one-to-one relationship between images, while PHDL aims to deal with the one-to-many correspondence between image and video.

III. PHDL

A. Problem Formulation

Let $\mathbf{X} = [x_1, ..., x_i, ..., x_n]$ be the feature set of training images, where $x_i \in \mathbb{R}^p$ represents the feature of an image from the i^{th} person, and n is the number of persons. Let C, where $\mathbf{Y}_i = [y_{i,1}, ..., y_{i,j}, ..., y_{i,n_i}]$ represents 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. To make full use of the information contained in video, both the visual appearance and spatial-temporal features are extracted from each walking cycle of the video. In this paper, we obtain walking cycles from each video according to the following steps: (i) Extract the Flow Energy Profile (FEP) from each image frame of the video. (ii) Transform the original FEP signal into the frequency domain using the discrete Fourier transform, filter out all the frequencies except the dominant one, and obtain the regulated FEP signal E using the inverse discrete Fourier transform on the remaining frequency. (iii) Split the whole video sequence into segments according to the local maxima/minima of E. The local maxima of E correspond to the postures when the person's two legs overlap while at the local minima the two

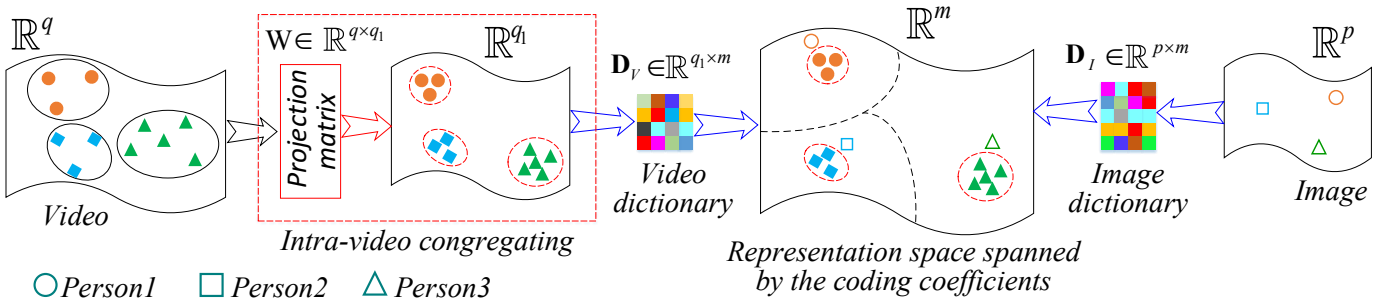


Fig. 3. Illustration of the basic idea of our PHDL approach.

legs are the farthest away. Generally, a full cycle contains two consecutive steps, one step from each leg. However, it is extremely difficult to distinguish between the two steps, hence each step (between two maxima) is treated as a **walking cycle**.

Due to the heterogeneity of image and video features (different feature types and dimensions), it's not an easy task to directly conduct matching between pedestrian image and video. Researches in [59], [60], [64] indicate that dictionary learning (DL) is an effective feature learning technique. With the learned dictionary, each sample can be represented as a coding coefficient. Inspired by this, we intend to learn two heterogeneous dictionaries (an image dictionary and a video dictionary) from image and video data, which have different dimensions but the same atom number. Then, by using the learned image and video dictionaries, heterogeneous image and video features can be transformed into coding coefficients having the same dimension. In this way, the re-identification between image and video can be realized by using their coding coefficients. Since the obtained coding coefficients are used for re-identification, they should own favorable discriminability. To this end, we can design a discriminant term, which requires 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, due to the changes in viewpoint, background and occlusion, there usually exist large variations between different walking cycles within each video. Figure 2 provides some example image sequences that display these intra-video variations. These variations determine 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, the influence of these variations should be reduced in the process of dictionary learning. To this end, we can learn a feature projection matrix (FPM) for the video data, under which features of different walking cycles within each video cluster together. Figure 3 illustrates the basic idea of our PHDL approach.

The notations used in our PHDL approach are shown in Table I. Then, we design the objective function of PHDL 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 in dictionary $\mathbf{D}_I$ and $\mathbf{D}_V$, such that the updating process of $\mathbf{D}_I$ and $\mathbf{D}_V$ will become more stable, which is beneficial to the stability of the optimization process. Details of each term are as follows.

TABLE I
NOTATIONS USED IN OUR PHDL APPROACH.

Notation	Description
$\mathbf{X}$	feature set of training images
$\mathbf{Y}$	feature set of training pedestrian videos
p and q	dimensions of image and video features
x_i	feature of the i^{th} image in $\mathbf{X}$, $x_i \in \mathbb{R}^p$
$\mathbf{Y}_i$	feature set corresponding to the i^{th} video
$y_{i,j}$	feature of the j^{th} walking cycle in $\mathbf{Y}_i$
$\mathbf{W}$	the learned FPM for video data, $\mathbf{W} \in \mathbb{R}^{q\times q_1}$, where q_1 is the dimension of video features after projection
$\mathbf{D}_I$ and $\mathbf{D}_V$	the learned image and video dictionaries, $\mathbf{D}_I \in \mathbb{R}^{p\times m}$ and $\mathbf{D}_V \in \mathbb{R}^{q_1\times m}$, where m is the number of atoms in $\mathbf{D}_I$ and $\mathbf{D}_V$
$\mathbf{A}$ and a_i	coding coefficients of $\mathbf{X}$ and x_i over D_I
$\mathbf{B}$, $\mathbf{B}_i$, b_{ij}	coding coefficients of $\mathbf{Y}$, $\mathbf{Y}_i$, $y_{i,j}$ over $\mathbf{D}_V$

- $f_I(\mathbf{D}_I,\mathbf{X},\mathbf{A}) = \|\mathbf{X}-\mathbf{D}_I\mathbf{A}\|_F^2$ is the reconstruction fidelity term of image data.
- $f_V(\mathbf{W},\mathbf{D}_V,\mathbf{Y},\mathbf{B}) = \|\mathbf{W}^T\mathbf{Y}-\mathbf{D}_V\mathbf{B}\|_F^2$ is the reconstruction fidelity term of video data.
- $g(\mathbf{W},\mathbf{Y}) = \sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\|\mathbf{W}^T(y_{i,j}-m_i)\|_2^2$ is the video congregating term, which aims to make the feature of each walking cycle move towards 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 video feature set $\mathbf{Y}_i$.
- $d(\mathbf{A},\mathbf{B})$ is the point-to-set coefficient discriminant term, which is used to ensure that the obtained coding coefficients own favorable discriminability. Specifically, for each truly matching image-video pair, it requires that the coding coefficient of each walking cycle in the video should move towards that of the image. And for each wrong matching image-video pair, the term requires that the coefficient of each walking cycle in the video should be far away from that of the image. $d(\mathbf{A},\mathbf{B})$ is defined as follows:

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

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 is the collection of wrong matching image-video pairs. Here, $|\cdot|$ represents 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 the coding coefficients.

B. The Optimization Algorithm

Although the objective function in Eq. (1) is not jointly convex to $\{\mathbf{W},\mathbf{D}_I,\mathbf{D}_V\}$, it is convex w.r.t. each of them if others are fixed. To optimize the problem in Eq. (1), we divide the objective function into three sub-problems, including coding coefficient updating, dictionary updating and feature projection matrix updating, and solve these three sub-problems alternatively.

Since the updating of each variable depends on other variables, we need to initialize all variables first. Specifically, $\mathbf{W}$ is initialized by solving the problem in Eq. (2), which can be easily solved by eigen-decomposition.

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

$\mathbf{D}_I$ and $\mathbf{D}_V$ are initialized as random matrices with unit L_2 norm for each column vector. $\mathbf{A}$ and $\mathbf{B}$ are initialized by Eq. (3) and Eq. (4), respectively.

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

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

Both (3) and (4) are ridge regression problems, whose solutions can be analytically derived as:

$$\mathbf{A}=(\mathbf{D}_I^T\mathbf{D}_I+\lambda\mathbf{I})^{-1}\mathbf{D}_I^T\mathbf{X},$$
$$\mathbf{B}=(\mathbf{D}_V^T\mathbf{D}_V+\lambda\mathbf{I})^{-1}\mathbf{D}_V^T\mathbf{W}^T\mathbf{Y},$$

where $\mathbf{I}$ is an identity matrix.

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

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

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

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 w.r.t. a_i to zero.

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

The solution of (6) can be obtained similarly.

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

where $\mathbf{C}_{j,i}\in\mathbb{R}^{m\times n_i}$ 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 [60]. 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 $\mathbf{P}=\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{i,j}-m_i)(y_{i,j}-m_i)^T$. Algorithm 1 summarizes the optimization process of our PHDL approach.

C. Computational Complexity

In the optimization algorithm of PHDL, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{D}_I$, $\mathbf{D}_V$ and $\mathbf{W}$ are updated alternatively. In each iteration, updating $\mathbf{A}$ costs $O(m^2p+m^3+mpp+n(m^2+mp))$; the time complexity of updating $\mathbf{B}$ is $O(m^2q_1+m^3+mqq_1+N(m^2+mq))$, where N is the total number of samples in $\mathbf{Y}$; updating $\mathbf{D}_I$ costs $Ok(p^2n+pnm+m^2n+m^3+pm^2)$, 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,

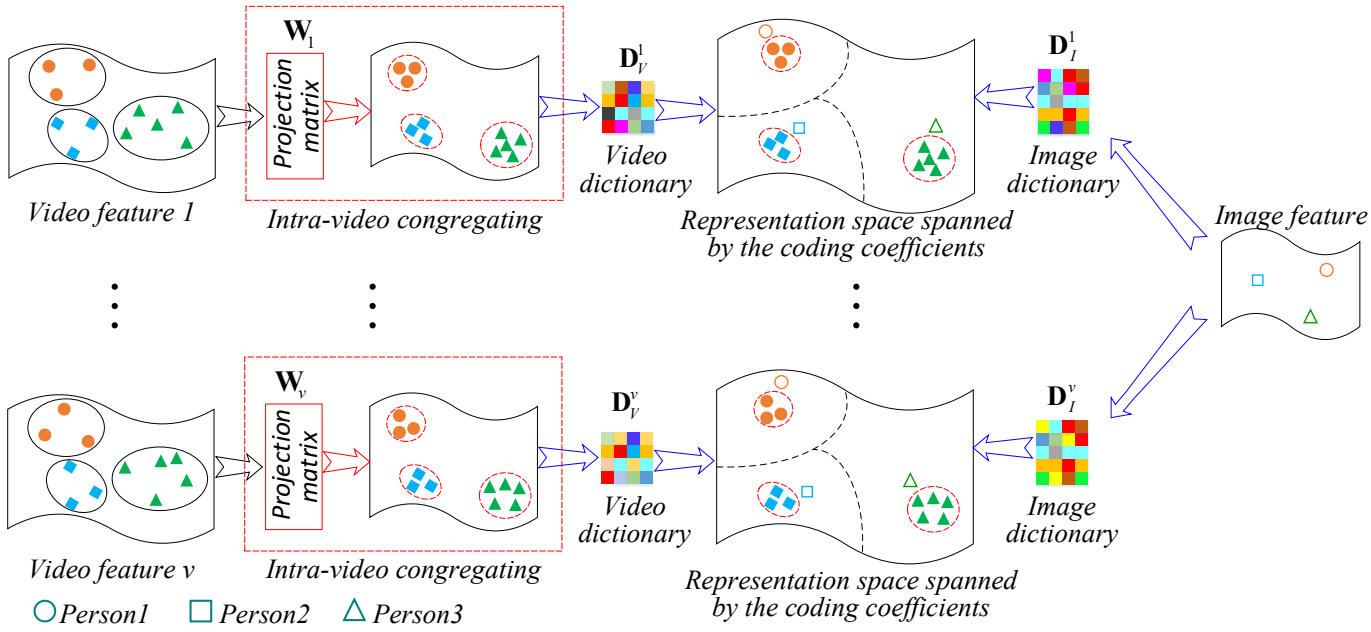


Fig. 4. The basic idea of our MPHDL approach. For the i^{th} $(i = 1, ..., v)$ type of video features, MPHDL learns a pair of heterogeneous video and image dictionaries from this type of video features and the visual appearance features of images.

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

Input: Training image and video sets $\mathbf{X}$ and $\mathbf{Y}$
Output: $\mathbf{D}_I$, $\mathbf{D}_V$ 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 $\mathbf{W}$;

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.

D. 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 subproblem is convex. Figure 5 shows the convergence curves of our algorithm on the iLIDS-VID dataset. We 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.

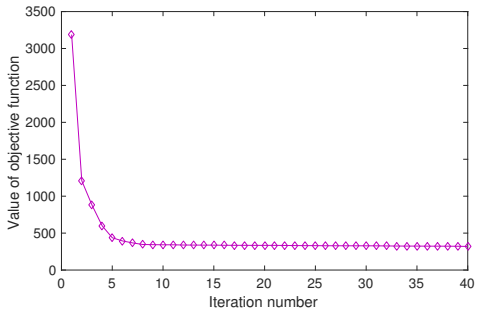


Fig. 5. Convergence curve of PHDL on iLIDS-VID.

IV. Multi-view PHDL

Our PHDL approach aims to seek a common discriminative representation space between the feature space of video and that of image. From the analysis in Section I, we know that both spatial-temporal and visual appearance features can be extracted from pedestrian video, but only visual appearance features can be extracted from a single image. Due to the different characteristics of spatial-temporal and visual appearance features, it's hard to say the most discriminative representation space between the spatial-temporal features of video and visual appearance features of image is the same as that between visual appearance features of video and image. **Therefore, to make better use of the spatial-temporal and visual appearance information contained in pedestrian videos, we should learn different discriminative representation spaces for them.**

In this section, **we propose a multi-view PHDL approach (MPHDL)**. Specifically, for each type of features in videos, MPHDL aims to learn a type-specific discriminative representation space from the visual appearance features of images and

TABLE II
NOTATIONS USED IN OUR MPHDL APPROACH.

Notation	Description
$\mathbf{Y}^v$	the set of the v^{th} type of feature extracted from training pedestrian videos
$\mathbf{Y}_i^v$	feature set corresponding to the i^{th} video in $\mathbf{Y}^v$
$y_{i,j}^v$	feature of the j^{th} walking cycle in $\mathbf{Y}_i^v$
$\mathbf{W}^v$	the learned FPM for the v^{th} type of video features $\mathbf{Y}^v$
$\mathbf{D}_I^v$ and $\mathbf{D}_V^v$	image and video dictionaries learned from the v^{th} type of video features and image features $\mathbf{X}$
$\mathbf{A}^v$ and a_i^v	coding coefficients of $\mathbf{X}$ and x_i over D_I^v
$\mathbf{B}^v$, $\mathbf{B}_i^v$, b_{ij}^v	coding coefficients of $\mathbf{Y}^v$, $\mathbf{Y}_i^v$, $y_{i,j}^v$ over $\mathbf{D}_V^v$

this type of video features. In the testing phase, MPHDL fuses the scores of different video features by assigning different weights to them.

The notations used in our MPHDL approach are shown in Table II. Given $\mathcal{V}$ types of video features, the FPM, image and video dictionaries for the v^{th} (v=1,2,...,$\mathcal{V}$) type of video feature can be obtained as follows:

$$\begin{aligned}\min_{\mathbf{W}^v,\mathbf{D}_I^v,\mathbf{D}_V^v} & f_I(\mathbf{D}_I^v,\mathbf{X},\mathbf{A}^v)+f_V(\mathbf{W}^v,\mathbf{D}_V^v,\mathbf{Y}^v,\mathbf{B}^v)+\\ & \alpha g(\mathbf{W}^v,\mathbf{Y}^v)+\beta d_1(\mathbf{A}^v,\mathbf{B}^v)+\lambda r(\mathbf{W}^v,\mathbf{A}^v,\mathbf{B}^v) \quad (13)\\ & s.t.\ \|d_{I,i}^v\|_2^2\le 1,\ \|d_{V,i}^v\|_2^2\le 1,\ \forall i,\end{aligned}$$

where the definitions of $f_I()$, $f_V()$, $g()$, and $r()$ are the same as those in Eq. (1), and the effect of constraint is also the same as that in Eq. (1). $d_1(\mathbf{A}^v,\mathbf{B}^v)$ is the designed point-to-set coefficient discriminant term. $d_{I,i}^v$ ($d_{V,i}^v$) represents the i^{th} atom in $\mathbf{D}_I^v$ ($\mathbf{D}_V^v$). Figure 4 illustrates the basic idea of our MPHDL approach. We can easily find that: the proposed MPHDL approach can be regarded as a generalization of PHDL, and PHDL can be considered as a special case of MPHDL (when the view number is one).

Researches in [65]–[67] indicate that "hard" samples (i.e., impostor samples) usually own more discriminative information than other well separable samples, and utilizing these hard samples properly can bring benefits to the performance of person re-identification. Therefore, we intend to introduce the concept of "impostor" into our MPHDL approach. To this end, we design a new image-video relationship model, namely image-video triplet. The image-video triplet has two forms: (i) image-video triplet that is constituted by one image and two videos (one is the truly matching of the image, and the other one is an impostor video of the image); (ii) image-video triplet that consists of one video and two images (one is the truly matching of the video, and the other one is an impostor image of the video). Detailed construction of an image-video triplet can be found in Definition 1. Figure 6 illustrates the designed image-video triplet model. The differences between the designed image-video triplet model and the triplet model used in [67] are two-fold: (1) In the triplet model of [67], each element of the triplet is a pedestrian image. Different from [67], the image-video triplet consists of two types of samples (i.e., pedestrian image and video). (2) The triplet model of [67] has only one form, while the image-video triplet model has two forms (one image versus two videos, one video versus two images).

*Definition 1 (**Image-video Triplet**): Given an image x_i and two videos Y_j and Y_k, where Y_j is the true matching of x_i, while Y_k is the wrong matching of x_i. Let m_j and m_k be the mean vectors of Y_j and Y_k, respectively. If $\|x_i-m_k\|_2^2<\|x_i-m_j\|_2^2$, x_i, Y_j and Y_k constitute an **image-video triplet** (denoted as $<i,j,k>$). It should be noted that the distance between an image and a video is calculated using the visual appearance features. The case of image-video triplet that consists of one video and two images is similar to this.*

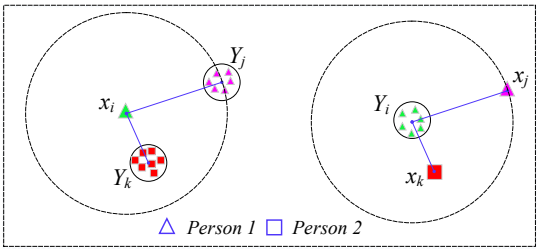


Fig. 6. Illustration of two forms of the designed image-video triplet model.

Based on the designed image-video triplet model, we further design an **image-video triplet constraint**, which requires that the distance between truly matching image and video should be smaller than that between wrong matching image and video in each image-video triplet. Following the idea of image-video triplet constraint, we design the point-to-set coefficient discriminant term $d_1(\mathbf{A}^v,\mathbf{B}^v)$ as follows:

$$\begin{aligned}d_1(\mathbf{A}^v,\mathbf{B}^v)=&\frac{1}{|\mathcal{T}_1|}\sum_{<i,j,k>\in\mathcal{T}_1} dis(a_i^v,\mathbf{B}_j^v)-\rho_1 dis(a_i^v,\mathbf{B}_k^v)\\ &+\frac{1}{|\mathcal{T}_2|}\sum_{<i,j,k>\in\mathcal{T}_2} dis(a_j^v,\mathbf{B}_i^v)-\rho_2 dis(a_k^v,\mathbf{B}_i^v)\end{aligned}$$

Here, $\mathcal{T}_1$ represents the collection of image-video triplets constituted by one image and two videos, and $\mathcal{T}_2$ represents the collection of image-video triplets constituted by one video and two images. $\mathcal{T}_1$ and $\mathcal{T}_2$ are constructed by using the visual appearance features of image and video. ρ_1 and ρ_2 are penalty factors, and can be computed as follows:

$$\begin{aligned}\rho_1 &= exp(-\|x_i-m_k\|_2^2/\|x_i-m_j\|_2^2),\\ \rho_2 &= exp(-\|m_i-x_k\|_2^2/\|m_i-x_j\|_2^2),\end{aligned}$$

where m_i, m_j and m_k are the mean vectors of visual appearance features for the i^{th}, j^{th} and k^{th} videos, respectively.

Problem in Eq. (13) can be solved by using a similar algorithm as Eq. (1). Specifically, when updating $\mathbf{A}^v$ and $\mathbf{B}^v$, Eqs. (5) and (6) should be slightly changed according to the discriminant term $d_1(\mathbf{A}^v,\mathbf{B}^v)$; The other variables $\mathbf{D}_I^v$, $\mathbf{D}_V^v$ and $\mathbf{W}^v$ can be updated by the same equations used in PHDL. Algorithm 2 summarizes the optimization process of MPHDL.

Since we solve Problem in Eq. (13) with a similar algorithm as PHDL, the computational complexity and convergence for solving Eq. (13) are similar to those of PHDL.

V. RE-IDENTIFICATION BETWEEN IMAGE AND VIDEO

Given a probe image and l gallery videos, the steps of re-identifying the probe image in gallery videos with our PHDL and MPHDL approaches are as follows.

Algorithm 2 Multi-view PHDL (MPHDL)

Input: Training image and video sets $\mathbf{X}$ and $\mathbf{Y}^v$, $v = 1, ..., \mathcal{V}$
Output: $\mathbf{D}_I^v$, $\mathbf{D}_V^v$ and $\mathbf{W}^v$

```
for v = 1 : V do
    Initialize D_I^v, D_V^v, W^v, A^v, B^v, α, β and λ
    while not converge do
        Fix W^v, D_I^v and D_V^v, update A^v and B^v with the similar way as (5) and (6), respectively;
        Fix W^v, A^v and B^v, update D_I^v and D_V^v according to (7) and (8), respectively;
        Fix D_I^v, D_V^v, A^v and B^v, update W^v according to (11);
    end while
end for
return D_I^v, D_V^v and W^v, v = 1, ..., V;
```

A. Re-identification With PHDL

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.

B. Re-identification With MPHDL

Let x be the visual appearance feature of a probe image. Let $\mathbf{Z}^v$, $\mathbf{Z}_i^v$ and $z_{i,j}^v$ be the v^{th} type of features of gallery video set, the i^{th} gallery video, the j^{th} walking cycle in the i^{th} gallery video, respectively. Denote by $\mathbf{D}_I^v$, $\mathbf{D}_V^v$ and $\mathbf{W}^v$ the learned dictionary pair and FPM for the v^{th} type of feature. We can re-identify the probe image in gallery video set as follows.

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

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

(3) Compute the distance between the probe image and each gallery video using the obtained representation coefficients of each feature type. The distance between a^v and $\mathbf{G}_i^v$ can be calculated by $d_i^v = \sum_{j=1}^{n_i} \|a^v - g_{ij}^v\|_2^2$.

(4) Fuse the obtained distances of $\mathcal{V}$ types of features as follows.

$$d_i = \sum_{v=1}^{\mathcal{V}} u_v d_i^v \tag{14}$$

where u_v is the weight of the v^{th} type of feature. In experiments, we set the weight of each type of feature by 3-fold cross-validation on each dataset. Detailed setting steps are as follows. Firstly, we first divide the training set into 3 subsets of equal size randomly. Secondly, for each choice of weight, we validate the performance of our model with cross-validation. Specifically, we select data of two subsets as the training data to train our model, and use the remaining one subset as validation data to test the performance of our model. This process is repeated three times, with each subset used exactly once as the validation data. Then, the average validation performance of three times is regarded as the result for current weight choice. Finally, the weight choice that induces the best performance is selected as the final weight used in the testing phase.

(5) Re-identify probe image in gallery videos by using the fused distances. We can obtain the matching result by sorting the obtained distances in ascending order.

VI. Experimental Results

In this section, we conduct extensive experiments on four publicly available person sequence datasets, including iLIDS-VID [25], PRID 2011 [68], MARS [69] and HDA+ [70], to demonstrate the effectiveness of the proposed approaches.

A. Experimental Settings

Baselines. To evaluate the efficacy of the proposed approaches, we compare our approaches with several state-of-the-art person re-identification methods and general point to set based matching methods. The person re-identification methods include **KISSME** [52], **RDC** [53], **ISR** [43], and **XQDA** [71]. The point to set based methods include **PSDML** [72], and **LERM** [73]. For all compared methods, we perform experiments with the source codes provided by the original authors.

Feature Representation. In experiments, we employ the WHOS feature, which is a hybrid visual appearance descriptor proposed in [43], to represent each pedestrian image. For the video, we extract WHOS feature and STFV3D [27], which is a spatial-temporal feature descriptor, from each walking cycle. In experiments, WHOS descriptor is employed for competing methods as the representation of image and video.

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.

Parameter Setting. There are four parameters in our PHDL approach, including α, β, λ, and η. In experiments, we set them by using the 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, $\alpha = 10$, $\beta = 0.6$, $\lambda = 0.01$ and $\eta = 0.1$ for the MARS dataset, $\alpha = 8$, $\beta = 0.7$, $\lambda = 0.014$ and $\eta = 0.12$ for the HDA+ dataset.

TABLE III
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+WHOS	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 IV
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+WHOS	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**

In MPHDL, there are three parameters, i.e., α, β and λ. In experiments, α, β and λ are separately set as (1.2, 0.6, 0.01) for iLIDS-VID, (1, 0.7, 0.01) for PRID 2011, (1.2, 0.5, 0.008) for MARS, (1.4, 0.7, 0.012) for HDA+. In the experiments of PHDL, the size of image and video dictionaries is separately set as 120, 180, 420 and 80 for iLIDS-VID, PRID 2011, MARS and HDA+; the number of columns in $\mathbf{W}$ is set as 460, 380, 480 and 320 for iLIDS-VID, PRID 2011, MARS and HDA+, respectively. For MPHDL, the dictionary size is set as 160, 140, 460 and 80 for iLIDS-VID, PRID 2011, MARS and HDA+, respectively; the column number of $\mathbf{W}$ is separately set as 540, 500, 580 and 480 for iLIDS-VID, PRID 2011, MARS and HDA+.

For the experiment on each dataset, we report the rank-k matching rates. For the experiment on each dataset, we report the rank-k matching rates. We repeat each experiment 10 times and report the average results of all methods.

B. Evaluation on the iLIDS-VID Dataset

The **iLIDS-VID** person sequence dataset [25] 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.

Table III shows the detailed rank 1-50 matching rates of all the compared methods. "+WHOS" ("+STFV3D") means that PHDL employs the WHOS (STFV3D) feature to represent the video. "+Both" represents that PHDL uses the concatenation of STFV3D and WHOS features as the representation of video. It can be seen that: (i) PHDL achieves the best matching results; (ii) when both the WHOS 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.

C. Evaluation on the PRID 2011 Dataset

The **PRID 2011** person sequence dataset [68] 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. In experiments, the sequence pairs with less than 20 frames are ignored due to the requirement on the sequence length for extracting walking cycles [27].

Table IV reports 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 8.93% (=33.58%-24.65%). The reasons for the large improvement are two-fold: (1) our PHDL approach takes some targeted measures, which have been provided in the result analysis of Section VI-B, to the difficulties existed in the matching between image and video. (2) There exist large illumination variations between two cameras of PRID 2011, which increases the difficulty of matching between image and video across two cameras for the competing methods. Since our PHDL approach is designed for the matching between heterogeneous image and video features, it can intrinsically handle the differences between two cameras to some extent.

D. Evaluation on the MARS Dataset

MARS dataset [69] contains image sequences of 1261 persons captured by six cameras in the campus scenario, with each person captured by at least two cameras. Each person has 13.2 sequences on average. For evaluation, we randomly select one image per person to form the image set, and randomly select one sequence per person to form the video set, where the image and video of the same person are selected from different cameras, and the length of each sequence is larger than 20 frames. The formed image-video pairs are then randomly split into two sets of equal size, with one for training, and the other for testing.

Table V reports the matching results of all compared methods on the MARS dataset. We can see that PHDL outperforms the compared methods in terms of matching result. In particular, PHDL improves the rank 5 matching rate at least by 2.25% (=45.33%-43.08%). The reason for the better performance of PHDL on MARS is the same as that provided in Section VI-B.

E. Evaluation on the HDA+ Dataset

The HDA+ dataset [70] contains a total of 83 labeled people across 13 indoor cameras in an office environment (Fig.

TABLE V
TOP r RANKED MATCHING RATES (%) ON MARS.

Method	r=1	r=5	r=10	r=20	r=50
RDC	24.95	35.76	46.92	52.58	61.40
KISSME	28.28	41.26	52.41	61.56	68.38
ISR	22.46	33.85	42.69	50.22	59.03
XQDA	29.93	43.08	53.41	60.89	66.88
PSDML	26.78	37.60	49.15	55.77	64.83
LERM	27.52	38.78	51.49	58.63	65.26
PHDL+WHOS	32.65	47.83	58.16	65.24	73.12
PHDL+STFV3D	30.48	45.33	55.52	63.06	72.69
PHDL+Both	**35.72**	**51.49**	**60.88**	**67.28**	**74.35**

TABLE VI
TOP r RANKED MATCHING RATES (%) ON HDA+.

Method	r=1	r=2	r=3	r=4	r=5
RDC	32.14	46.43	51.43	53.57	60.71
KISSME	32.86	52.86	59.29	69.29	72.14
ISR	28.57	35.71	40.00	46.43	51.43
XQDA	39.29	53.57	58.57	70.71	74.29
PSDML	30.00	50.71	56.43	66.43	69.29
LERM	32.14	52.14	57.86	67.14	70.00
PHDL+WHOS	44.29	55.00	60.00	72.14	76.43
PHDL+STFV3D	40.71	53.57	56.43	70.00	73.57
PHDL+Both	**47.14**	**57.14**	**62.86**	**73.57**	**77.14**

5(c,d)). In experiments, we follow the evaluation protocol used in [26]. Specifically, camera pair (19, 40) is selected, where a sufficiently large number of persons appears in both camera views. Camera pair (19, 40) contains pairwise image sequences of 28 different persons at 5fps. Each sequence consists of 15-227 frames, with an average of 88 frames. In experiments, we construct image-video pairs with the strategy introduced in evaluation setting part. For each image-video pair, if the length of video is less than 21 frames, we expand it up to 21 frames by interpolating new frames using duplicates of the temporally-nearest frames.

Table VI provides the matching results of all compared methods on the HDA+ dataset. We can observe that PHDL achieves overall better matching results than the competing methods, which shows the effectiveness of our approach for image to video person re-identification. The reason why our PHDL can achieve better performance on HDA+ is the same as that provided in Section VI-B. From the results on four datasets, we can see that the performance of PHDL is further improved when the concatenation of WHOS and STFV3D features is employed to represent each video, which indicates that there exists the complementary information between the visual appearance and spatial-temporal features of the video.

F. Comparison Between PHDL and MPHDL

To validate that MPHDL can achieve better re-identification performance than PHDL, we conduct comparison experiments on four datasets, including iLIDS-VID, PRID 2011, MARS, and HDA+. For MPHDL, two types of features are used, including STFV3D and WHOS. In the following experiments, MPHDL employs these two types of features by default. For PHDL, STFV3D and WHOS descriptors are concatenated as the representation of video. Table VII reports the matching results of PHDL and MPHDL on four datasets. We can see that MPHDL achieves better results than PHDL. Specifically, the rank 1 matching rate is improved at least by 3.35% (= 39.07%-35.72%, on the MARS dataset). The reason why MPHDL can outperform PHDL is two-fold: (i) MPHDL learns different representation space for different types of video features, such that each learned representation space can own more favorable discriminability; (ii) The contributions of different types of video features to the re-identification are different, and MPHDL assigns different weights to different types of video features in the fusion phase, which is more reasonable. (iii) MPHDL employs the designed image-video triplet constraint to further improve the discriminability of leaned re-identification model.

TABLE VII
TOP r RANKED MATCHING RATES (%) OF PHDL AND MPHDL ON FOUR DATASETS.

iLIDS-VID	r=1	r=5	r=10	r=20	r=50
PHDL+Both	28.15	50.37	65.88	80.35	95.42
MPHDL	32.58	55.76	69.25	83.17	96.36
PRID 2011	r=1	r=5	r=10	r=15	r=20
PHDL+Both	41.92	67.25	85.47	90.04	92.44
MPHDL	46.31	72.62	87.42	91.46	93.26
MARS	r=1	r=5	r=10	r=20	r=50
PHDL+Both	35.72	51.49	60.88	67.28	74.35
MPHDL	39.07	53.84	63.32	69.03	75.66
HDA+	r=1	r=2	r=3	r=4	r=5
PHDL+Both	47.14	57.14	62.86	73.57	77.14
MPHDL	50.71	60.00	66.43	75.00	78.57

VII. DISCUSSION

A. 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 modified versions of PHDL and MPHDL by removing $\mathbf{W}$, and observe the performance of PHDL and MPHDL as well as their modified versions. Here, we name the modified versions of PHDL and MPHDL as PHDL-W and MPHDL-W, respectively. Table VIII reports the top ranked results of PHDL and MPHDL as well as their modified versions on the iLIDS-VID dataset. Here, WHOS feature is employed as the representation of video for PHDL. We can see that without using $\mathbf{W}$, the performance of PHDL and MPHDL decreases, which means that learning FPM is beneficial to improving the discriminability of our image to video re-identification model. More specifically, without using $\mathbf{W}$, the rank 1 matching rate of PHDL and MPHDL are decreased by 2.24% (22.32%-20.08%) and 3.04% (32.58%-29.54%) respectively on the iLIDS-VID dataset. Similar results can be obtained on other datasets.

B. Effect of Image-video Triplet Constraint

The designed image-video triplet constraint is used to further enhance the discriminability of our MPHDL approach, such that the performance of MPHDL can be improved. In

TABLE VIII
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
MPHDL-W	29.54	51.62	66.38	80.74	94.88
MPHDL	32.58	55.76	69.25	83.17	96.36

TABLE IX
RANK 1 RESULTS (%) OF MPHDL AND MPHDL-S ON THE ILIDS-VID DATASET.

Method	iLIDS-VID	PRID 2011	MARS	HDA+
MPHDL-s	31.27	44.95	37.84	49.29
MPHDL	32.58	46.31	39.07	50.71

this experiment, we aim to evaluate the effect of image-video triplet constraint to the performance of MPHDL. To this end, we generate a modified version of MPHDL that learns FPM, image and video dictionaries for each type of video features by using the same constraint with PHDL, and name the modified version as "MPHDL-s". Then we evaluate the effect of image-video triplet constraint by comparing the performance of MPHDL and MPHDL-s.

Table IX shows the comparison of rank 1 matching rates between MPHDL and MPHDL-s on four datasets. We can see that MPHDL achieves better matching rates than MPHDL-s. In particular, by using the designed image-video triplet constraint, MPHDL can improve the rank 1 match rate at least by 1.23% (39.07%-37.84%, on the MARS dataset), which means that image-video triplet constraint is beneficial to improving the performance of MPHDL.

C. Effect of Different Video Congregating Terms

The video congregating term $g(W, Y)$ is used to reduce the variations within each video. For simplicity, we just require that each walking cycle moves towards the center of the video to which it belongs. In fact, a better (maybe more complex) video congregating term can induce better performance. In this experiment, we aim to evaluate the effects of different video congregating term on our approaches. To this end, we design a new video congregating term. Specifically, the new video congregating term first groups all walking cycles within a video into several clusters, and then makes each walking cycle move towards the center of the cluster to which it belongs. Considering that the orientation of pedestrian includes frontal and side views, the used cluster number is two. Table X shows the performance of our approaches with the original or the new video congregating term on the iLIDS-VID dataset, where "_cluster" means that our approach uses the designed new video congregating term. We can see that our approaches achieve better performance when the new designed cluster-based video congregating term is employed. On the other hand, the cluster-based video congregating term also brings a certain computational burden to MPHDL. In experiments, the average training time of MPHDL_cluster is about 30% longer than that of MPHDL. Therefore, MPHDL_cluster can be used as an enhanced algorithm of the proposed MPHDL approach.

TABLE X
TOP r RANKED MATCHING RATES (%) OF OUR APPROACHES WITH DIFFERENT VIDEO CONGREGATING TERMS ON THE ILIDS-VID DATASET, WHERE PHDL USES THE CONCATENATION OF STFV3D AND WHOS FEATURES AS THE REPRESENTATION OF VIDEO.

Method	r=1	r=5	r=10	r=20	r=50
PHDL	28.15	50.37	65.88	80.35	95.42
PHDL_cluster	28.80	51.73	66.80	81.40	95.53
MPHDL	32.58	55.76	69.25	83.17	96.36
MPHDL_cluster	32.93	56.80	70.33	84.33	96.67

TABLE XI
TOP r RANKED MATCHING RATES (%) OF MPHDL, MPHDL_m AND MPHDL_l ON THE ILIDS-VID DATASET.

Method	r=1	r=5	r=10	r=20	r=50
MPHDL_m	31.42	54.19	67.91	82.03	95.61
MPHDL_l	31.15	53.85	67.64	81.76	95.47
MPHDL	32.58	55.76	69.25	83.17	96.36

D. Effect of Different Penalty Factors

In our MPHDL approach, penalty factors ρ_1 and ρ_2 represent the punishment degrees to the impostor sample in each image-video triplet. In this experiment, we evaluate the effects of different penalty factors to the performance of our MPHDL approach. To this end, we perform MPHDL by employing another two different penalty factors, and compare the performance of different cases. Here, the used penalty factors include: (1) penalty factor based on the distance between image and the median video frame; (2) penalty factor based on the distance between image and the last video frame. We name the manner of MPHDL with the first penalty factor as MPHDL_m, and name the manner of MPHDL with the second penalty factor as MPHDL_l. Table XI reports the top ranked matching rates of MPHDL, MPHDL_m and MPHDL_l on the iLIDS-VID dataset. We can see that MPHDL outperforms its two modified versions. The reason may be that: there usually exist large variations in each video, leading to that the distance between the image and a single video frame cannot well reflect the real position relationship between the image and video, which will further influence the accuracy of the penalty factor based on the distance between image and a single video frame.

E. Can the Competing Methods Work with Heterogeneous Features?

The competing methods (RDC, KISSME, ISR, XQDA, PSDML and LERM) are designed for the matching between samples having the same representation, and thus they cannot work well in the case of matching between heterogeneous image and video features. To validate this, we perform these methods by using (1) WHOS for image and STFV3D for video; (2) WHOS for image and WHOS+STFV3D for video. To make the heterogeneous features can be fed into these methods, we employ the PCA technique to make the image and video features have the same dimension. Table XII and XIII show the top ranked matching rates of all the competing methods under two cases on the iLIDS-VID dataset. We can see that the competing methods achieve poor performance

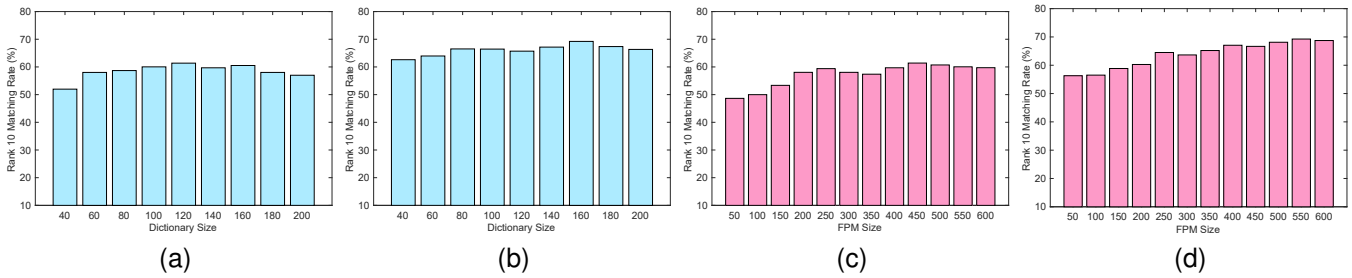


Fig. 7. Rank 10 results of (a) PHDL and (b) MPHDL versus different values of dictionary sizes on iLIDS-VID; Rank 10 results of (c) PHDL and (d) MPHDL versus different values of FPM sizes on iLIDS-VID. For PHDL, WHOS feature is employed as the representation of each walking cycle in the video.

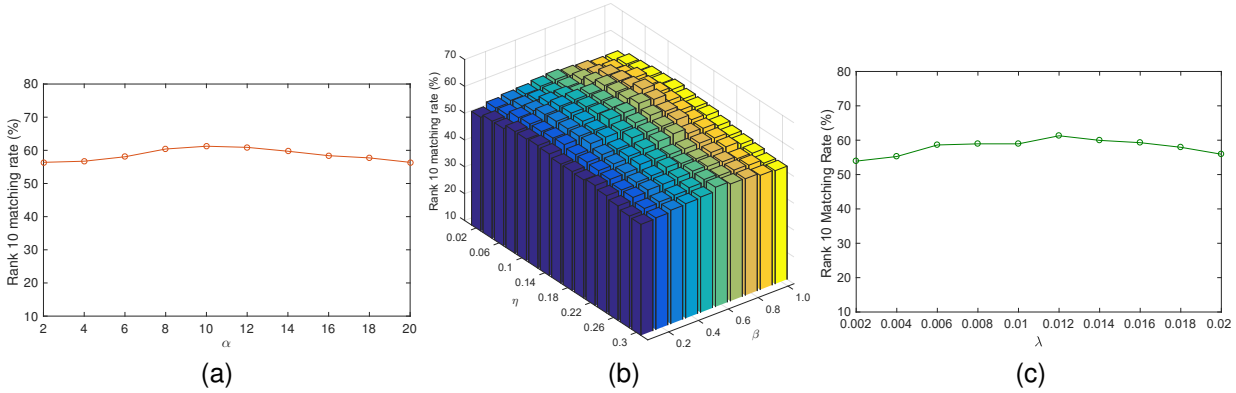


Fig. 8. Rank 10 results of PHDL versus different values of (a) α, (b) β, η (c) λ on iLIDS-VID, where WHOS feature is employed as the representation of each walking cycle in the video.

TABLE XII
TOP r RANKED MATCHING RATES (%) OF THE COMPETING METHODS ON THE ILIDS-VID DATASET, WHERE STFV3D IS EMPLOYED TO REPRESENT EACH VIDEO.

Method	r=1	r=5	r=10	r=20	r=50
RDC	1.49	4.12	8.11	13.51	55.68
KISSME	1.89	6.01	9.59	17.43	58.38
ISR	1.42	3.51	8.38	15.20	56.69
XQDA	1.62	4.59	8.58	15.41	58.58
PSDML	1.76	5.07	9.73	13.99	57.30
LERM	2.09	6.22	10.27	19.59	62.36

TABLE XIII
TOP r RANKED MATCHING RATES (%) OF THE COMPETING METHODS ON THE ILIDS-VID DATASET, WHERE WHOS+STFV3D IS EMPLOYED TO REPRESENT EACH VIDEO.

Method	r=1	r=5	r=10	r=20	r=50
RDC	3.38	9.12	15.07	22.50	60.47
KISSME	4.93	11.28	19.59	27.43	63.72
ISR	3.11	8.51	14.39	21.22	58.72
XQDA	4.59	10.61	20.20	26.42	63.11
PSDML	5.74	11.08	18.72	26.01	62.77
LERM	6.08	12.23	21.28	29.59	66.49

under both heterogeneous cases. A possible reason for this phenomenon is that: different types of features usually have different physical meanings, and cannot be compared directly. Similar effects can be observed on the other datasets.

F. 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 our PHDL and MPHDL approaches. To observe the effect of dictionary size, we conduct experiments by setting different values to it. Figures 7 (a) and (b) plot the rank 10 matching rates of PHDL and MPHDL versus different dictionary sizes on the iLIDS-VID dataset. We can see that PHDL obtains relatively good result when dictionary size is set as 120, and MPHDL obtains good matching rate when dictionary size is set as 160, which means that PHDL and MPHDL are 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 and MPHDL approaches. Figures 7 (c) and (d) show the rank 10 matching rates of PHDL and MPHDL versus different column sizes of $\mathbf{W}$ on the iLIDS-VID dataset. We can see that, PHDL and MPHDL can achieve stable performance when the column size of $\mathbf{W}$ is in the range of [400 600]. Similar effects can be observed on other dataset.

G. Parameter Analysis

In this experiment, we investigate the effects of parameters to the performance of our PHDL and MPHDL approaches. There are four parameters in PHDL, 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.

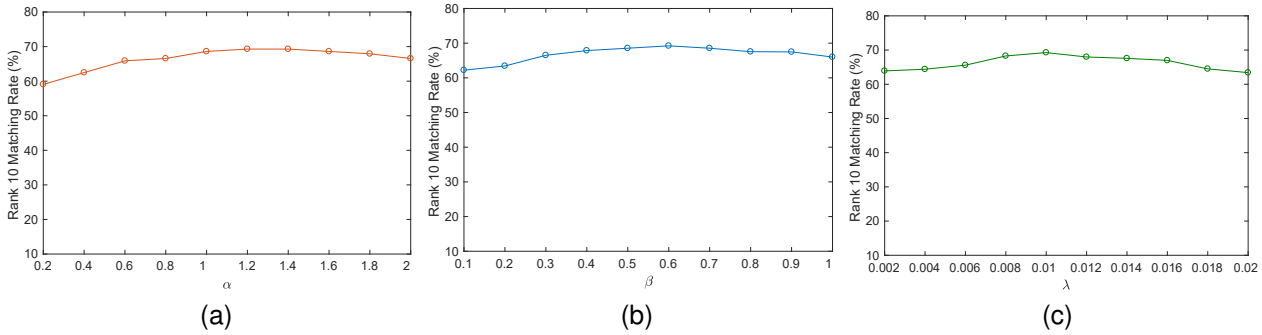


Fig. 9. Rank 10 results of MPHDL versus different values of (a) α, (b) β, (c) λ on the iLIDS-VID dataset.

Parameter η balances the effects of positive and negative image-video pairs. In MPHDL, there are three parameters, including α, β and λ. 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 8 (a)-(c) show the rank 10 matching rates of PHDL 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].

Figure 9 (a)-(c) show the rank 10 matching rates of MPHDL versus different values of α, β and λ on the iLIDS-VID dataset. We can observe that:(1) MPHDL can achieve relatively good performance when α is in the range of [0.8, 2]; (2) MPHDL achieves is insensitive to the choice of β in the range of [0.3, 0.9]. (3) MPHDL can obtain roughly stable performance when λ is in the range of [0.006, 0.016]. Similar effects can be observed on the other three datasets.

H. Results by Using Common Parameter Setting

In this experiment, we evaluate the performance of our approaches by using common parameter setting for all datasets. To this end, we use the median of the optimal parameter values on all datasets (reported in Section VI-A) as the common parameter configuration. Thus, the used common parameter setting for our PHDL approach is α=10, β=0.7, λ=0.011 and η=0.12; the used common parameter setting for our MPHDL approach is α=1.2, β=0.65 and λ=0.01. Table XIV compares the rank 1 matching rates of our PHDL and MPHDL approaches on four datasets using the optimal parameter setting or the common parameter setting, where "_C" means that our approach is performed using the common parameter setting. We can observe that: as compared with PHDL, the rank 1 matching rate of PHDL_C decreases by less than 1% on all datasets except MARS (1.54%=35.72%-34.18%); the performance of MPHDL decreases by 1.23%(=32.58%-31.35%) to 2.14%(=50.71%-48.57%). Therefore, the influence brought by using common parameter setting to our PHDL and MPHDL approaches is limited, and our approaches can still achieve a relatively good performance in this case.

TABLE XIV
TOP *r* RANKED MATCHING RATES (%) OF OUR APPROACHES UNDER OPTIMAL OR COMMON PARAMETER SETTING ON FOUR DATASETS, WHERE PHDL USES THE CONCATENATION OF STFV3D AND WHOS FEATURES AS THE REPRESENTATION OF VIDEO.

Dataset	PHDL	PHDL_C	MPHDL	MPHDL_C
iLIDS-VID	28.15	27.50	32.58	31.35
PRID 2011	41.92	41.11	46.31	44.67
MARS	35.72	34.18	39.07	37.35
HAD+	47.14	46.43	50.71	48.57

TABLE XV
TRAINING TIMES (SECOND) OF OUR PHDL AND MPHDL APPROACHES, WHERE PHDL USES THE CONCATENATION OF STFV3D AND WHOS FEATURES AS THE REPRESENTATION OF VIDEO.

Dataset	PHDL	MPHDL
iLIDS-VID	10.1	19.7
PRID 2011	8.3	14.5
MARS	52.7	96.4
HAD+	2.2	3.9

I. Training Time of Our Approaches

In this section, we report the training times of our approaches. Table XV shows the detailed training time of PHDL and MPHDL on four datasets, where PHDL uses the concatenation of STFV3D and WHOS features to represent each video. We can observe that the training times of our approaches are moderate. In the testing process, the testing time for one query image is less than 0.1 seconds. In practice, the training phase is usually off-line, thus our approaches are suitable for practical use.

VIII. 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 variations within each 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. Furthermore, we propose a multi-view PHDL (MPHDL) approach, which learns different projection matrix and dictionary pair for different

type of video features, and employs the designed image-video triplet constraint. Experimental results on four widely used person sequence datasets demonstrate that: (i) our PHDL can achieve better results than several state-of-the-art methods in the IVPR task. (ii) our MPHDL approach can improve the performance of PHDL.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their constructive comments and suggestions. This work was supported by the National Key Research and Development Program of China under Grant No. 2017YFB0202001, the National Nature Science Foundation of China under Grant Nos. 61671182, 61672208, 61772220, 41571417, U1404618, 61272273, 61572375, 61233011, 91418202, 61472178, 61373038, 61672392, the National Basic Research 973 Program of China under Project No. 2014CB340702, in part by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China Nos. 2015BAK36B00, 2015BAK01B06, in part by the Key Science and Technology of Shenzhen No. CXZZ20150814155434903, in part by the Key Program for International S&T Cooperation Projects of China No. 2016YFE0121200, the Natural Science Foundation of Jiangsu Province under Grant No. BK20170900, the Scientific Research Staring Foundation for Introduced Talents in NJUPT under NUPTSF No. NY217009, the Science and Technology Program in Henan province under Grant Nos. 1721102410064, 172102210186, Research Foundation of Henan University No. 2015YBZR024.

REFERENCES

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

[2] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *ICCV, IEEE Conference on*, 2015, pp. 4678–4686.

[3] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *ICCV, IEEE Conference on*, 2015, pp. 3765–3773.

[4] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4766–4779, 2015.

[5] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *CVPR, IEEE Conference on*, 2016, pp. 1268–1277.

[6] J. García, N. Martinel, A. G. Vicente, I. B. Muñoz, G. L. Foresti, and C. Micheloni, "Discriminant context information analysis for post-ranking person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1650–1665, 2017.

[7] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 591–606, 2015.

[8] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3656–3670, 2014.

[9] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *CVPR, IEEE Conference on*, 2015, pp. 1741–1750.

[10] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *ICCV, IEEE Conference on*, 2015, pp. 3739–3747.

[11] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *CVPR, IEEE Conference on*, 2015, pp. 4184–4193.

[12] Y. Cho and K. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *CVPR, IEEE Conference on*, 2016, pp. 1354–1362.

[13] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific SVM learning for person re-identification," in *CVPR, IEEE Conference on*, 2016, pp. 1278–1287.

[14] N. Martinel, C. Micheloni, and G. L. Foresti, "A pool of multiple person re-identification experts," *Pattern Recognition Letters*, vol. 71, pp. 23–30, 2016.

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

[16] C. Liu, C. C. Loy, S. Gong, and G. Wang, "Pop: Person re-identification post-rank optimisation," in *ICCV, IEEE Conference on*, 2013, pp. 441–448.

[17] Q. Qiu, J. Ni, and R. Chellappa, "Dictionary-based domain adaptation methods for the re-identification of faces," in *Person Re-Identification*, 2014, pp. 269–285.

[18] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR, IEEE Conference on*, 2015, pp. 3908–3916.

[19] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *CVPR, IEEE Conference on*, 2015, pp. 1565–1573.

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

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

[22] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *IJCAI*, 2015, pp. 2155–2161.

[23] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, J. Bu, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV, IEEE Conference on*, 2015, pp. 1116–1124.

[24] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *ICCV, IEEE Conference on*, 2015, pp. 4516–4524.

[25] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *ECCV*, 2014, pp. 688–703.

[26] ——, "Person re-identification by discriminative selection in video ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pp. 2501–2514, 2016.

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

[28] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," in *IJCAI*, 2016, pp. 3552–3559.

[29] N. McLaughlin, J. M. del Rincón, and P. C. Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR, IEEE Conference on*, 2016, pp. 1325–1334.

[30] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *CVPR, IEEE Conference on*, 2016, pp. 1345–1353.

[31] X. Zhu, X. Jing, F. Wu, Y. Wang, W. Zuo, and W. Zheng, "Learning heterogeneous dictionary pair with feature projection matrix for pedestrian video retrieval via single query image," in *AAAI*, 2017, pp. 4341–4348.

[32] S. Karaman, G. Lisanti, A. D. Bagdanov, and A. D. Bimbo, "Leveraging local neighborhood topology for large scale person re-identification," *Pattern Recognition*, vol. 47, no. 12, pp. 3767–3778, 2014.

[33] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *CVPR, IEEE Conference on*, 2014, pp. 144–151.

[34] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *ECCV*, 2016, pp. 858–877.

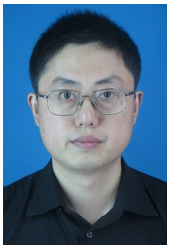
[35] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised l_1 graph learning," in *ECCV*, 2016, pp. 178–195.

[36] J. García, N. Martinel, A. G. Vicente, I. B. Muñoz, G. L. Foresti, and C. Micheloni, "Modeling feature distances by orientation driven classifiers for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 115–129, 2016.

[37] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with block sparse recovery," *Image and Vision Computing*, vol. 60, pp. 75–90, 2017.

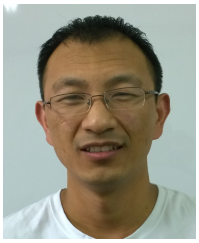
[38] Z. Wu, Y. Li, and R. J. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1095–1108, 2015.
[39] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing attributes assisted person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 869 – 878, 2015.
[40] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR, IEEE Conference on*, 2016, pp. 1363–1372.
[41] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR, IEEE Conference on*, 2010, pp. 2360–2367.
[42] D. Baltieri, R. Vezzani, and R. Cucchiara, "Learning articulated body models for people re-identification," in *Multimedia*, 2013, pp. 557–560.
[43] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1629–1642, 2015.
[44] S. Sunderrajan and B. S. Manjunath, "Context-aware hypergraph modeling for re-identification and summarization," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 51–63, 2016.
[45] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *ICCV, IEEE Conference on*, 2009, pp. 498–505.
[46] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *CVPR, IEEE Conference on*, 2012, pp. 2666–2672.
[47] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR, IEEE Conference on*, 2013, pp. 3318–3325.
[48] W. Li and X. Wang, "Locally aligned feature transforms across views," in *CVPR, IEEE Conference on*, 2013, pp. 3594–3601.
[49] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing kiss metric learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1675–1685, 2013.
[50] J. Chen, Z. Zhang, and Y. Wang, "Relevance metric learning for person re-identification by exploiting listwise similarities," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4741–4755, 2015.
[51] M. Hirzer, P. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *ECCV*, 2012, pp. 780–793.
[52] M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR, IEEE Conference on*, 2012, pp. 2288–2295.
[53] W. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, 2013.
[54] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K. Lam, and Y. Zhong, "Person re-identification by unsupervised video matching," *Pattern Recognition*, vol. 65, pp. 197–210, 2017.
[55] D.-A. Huang and Y.-C. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *ICCV, IEEE Conference on*, 2013, pp. 2496–2503.
[56] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.
[57] Y. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval." in *AAAI*, 2013, pp. 1070–1076.
[58] L. Yang, R. Jin, R. Sukthankar, and Y. Liu, "An efficient algorithm for local distance metric learning," in *AAAI*, 2006, pp. 543–548.
[59] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651–2664, 2013.
[60] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Projective dictionary pair learning for pattern classification," in *NIPS*, 2014, pp. 793–801.
[61] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *ICCV, IEEE Conference on*, 2007, pp. 1–8.
[62] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification," in *BMVC*, 2015, pp. 44.1–44.12.
[63] H. Sheng, B. Zhang, Y. Huang, Y. Zheng, and Z. Xiong, "Discriminative dictionary learning sparse coding for person re-identification," in *IEEE Intelligent Vehicles Symposium, Gotenburg*, 2016, pp. 1338–1343.
[64] J. Lu, G. Wang, W. Deng, and P. Moulin, "Simultaneous feature and dictionary learning for image set based face recognition," in *ECCV*, 2014, pp. 265–280.
[65] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.
[66] M. Hirzer, P. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *AVSS*, 2012, pp. 203–208.
[67] X. Zhu, X. Jing, F. Wu, W. Zheng, R. Hu, C. Xiao, and C. Liang, "Distance learning by treating negative samples differently and exploiting impostors with symmetric triplet constraint for person re-identification," in *ICME, IEEE Conference on*, 2016, pp. 1–6.
[68] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*, 2011, pp. 91–102.
[69] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*, 2016, pp. 868–884.
[70] D. Figueira, M. Taiana, A. Nambiar, J. Nascimento, and A. Bernardino, "The hda+ data set for research on fully automated re-identification systems," in *Workshop of European Conference on Computer Vision*, 2014, pp. 241–255.
[71] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR, IEEE Conference on*, 2015, pp. 2197–2206.
[72] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, "From point to set: Extend the learning of distance metrics," in *ICCV, IEEE Conference on*, 2013, pp. 2664–2671.
[73] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning euclidean-to-riemannian metric for point-to-set classification," in *CVPR, IEEE Conference on*, 2014, pp. 1677–1684.

Xiaoke Zhu received the Ph.D. degree in pattern recognition and intelligence system from the Wuhan University, Wuhan, China, in 2017. He is currently an associate Professor in the School of Computer and Information Engineering, Henan University, China. His current research interests include person re-identification and image classification.



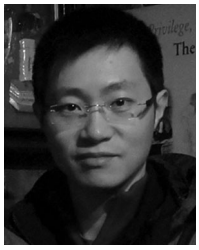
Xiao-Yuan Jing received the Doctoral degree of Pattern Recognition and Intelligent System in the Nanjing University of Science and Technology, 1998. He was a Professor with the Department of Computer, Shenzhen Research Student School, Harbin Institute of Technology, 2005. Now he is a Professor with the State Key Laboratory of Software Engineering, School of Computer, Wuhan University, and with the College of Automation, Nanjing University of Posts and Telecommunications, China.

Xinge You (M'08-SM'10) received the B.S.and M.S.degrees in Mathematics from the Hubei University,Wuhan,China and the Ph.D.degree from the Department of computer Science from the Hong Kong Baptist University, Hong Kong, in 1990, 2000,and 2004, respectively. Currently,he is a Professor at the School of Electronic Information and Communications in Hua zhong University of Science and Technology, China. His current research interests include wavelets and its application, signal and image processing, pattern recognition, machine learning, and computer vision.



Wangmeng Zuo received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007. He is currently a Professor in the School of Computer Science and Technology, Harbin Institute of Technology. His current research interests include image enhancement and restoration, object detection, visual tracking, and image classification. He has published over 60 papers in top-tier academic journals and conferences. He has served as a Tutorial Organizer in ECCV 2016, an Associate Editor of the IET Biometrics and Journal of Electronic Imaging, and the Guest Editor of Neurocomputing, Pattern Recognition, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Neural Networks and Learning Systems.

Shiguang Shan received M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He joined ICT, CAS, in 2002 and has been a Professor since 2010. He is now the Deputy Director of the Key Lab of Intelligent Information Processing of CAS. His research interests cover computer vision, pattern recognition, and machine learning. He especially focuses on face recognition related research topics.



Wei-Shi Zheng received the Ph.D. degree in applied mathematics from Sun Yat-Sen University in 2008. He is an associate professor in Sun Yat-sen University. He has been a postdoctoral researcher on the EU FP7 SAMURAI Project at Queen Mary University of London. His research interests include person/object association and recognition in visual surveillance. He has joined Microsoft Research Asia Young Faculty Visiting Programme.