Video-Based Person Re-Identification by Simultaneously Learning Intra-Video and Inter-Video Distance Metrics

Xiaoke Zhu[1,3], **Xiao-Yuan Jing**[*,1,2], **Fei Wu**[2,1], **Hui Feng**[1]
[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 and Information Engineering, Henan University, China

Abstract

Video-based person re-identification (re-id) is an important application in practice. However, only a few methods have been presented for this problem. Since large variations exist between different pedestrian videos, as well as within each video, it's challenging to conduct re-identification between pedestrian videos. In this paper, we propose a simultaneous intra-video and inter-video distance learning (SI^{2}DL) approach for video-based person re-id. Specifically, SI^{2}DL simultaneously learns an intra-video distance metric and an inter-video distance metric from the training videos. The intra-video distance metric is to make each video more compact, and the inter-video one is to make that the distance between two truly matching videos is smaller than that between two wrong matching videos. To enhance the discriminability of learned metrics, we design a video relationship model, i.e., video triplet, for SI^{2}DL. Experiments on the public iLIDS-VID and PRID 2011 image sequence datasets show that our approach achieves the state-of-the-art performance.

1 Introduction

The task of person re-identification (re-id) is to match pedestrian images or videos observed from multiple cameras. It has recently drawn much attention in the computer vision and machine learning communities due to its importance in the automated video surveillance and forensics. In recent years, various methods have been presented to tackle this problem. Most of the existing methods focus on the image-based person re-id [Ahmed *et al.*, 2015; Jing *et al.*, 2015; Li *et al.*, 2015a; Liao *et al.*, 2015]. These methods can be further divided into two categories: feature learning and distance learning methods. The former aims to extract distinctive features from pedestrian images, e.g., salience features [Zhao *et al.*, 2013], mid-level features [Zhao *et al.*, 2014], and salient color features [Yang *et al.*, 2014]. The distance learning methods focus on learning effective distance metrics, which can maximize matching accuracy regardless the choice of representation, to measure the similarity between two images. Popular distance learning methods include large margin nearest neighbor (LMNN) [Weinberger and Saul, 2009], the keep it simple and straightforward metric (KISSME) [Kostinger *et al.*, 2012] and relative distance comparison (RDC) [Zheng *et al.*, 2013], and so on.

(a) (b)

Figure 1: Example person sequences in the (a) iLIDS-VID, (b) PRID 2011 datasets. Sequences in the same row are from the same person. Only five frames are shown for each sequence.

Since videos inherently contain more information than independent images, video-based person re-id has attracted much attention. Recently, two methods have been presented for video-based person re-id [Wang *et al.*, 2014; Liu *et al.*, 2015]. Both methods focus on extracting spatial-temporal features to represent each pedestrian video, and then perform video-based person re-id with these features. Specifically, they break down each video to generate several fragments (walking cycles), and extract spatial-temporal feature from each fragment, and then represent each video with a set of the extracted spatial-temporal features. Therefore, re-identification between two videos can be considered as a set to set matching problem. In practice, due to changes in illumination, pose, viewpoint, occlusions, there not only exist severe variations between different pedestrian videos, but also large variations between the frames within each video. Figure 1 shows some demo person sequences in the iLIDS-VID [Wang *et al.*, 2014] and PRID 2011 [Hirzer *et al.*, 2011] datasets. These variations determine that there still exist large variations between spatial-temporal features extracted from different videos (called **inter-video variations**), and between different spatial-temporal features extracted from the same video (called **intra-video variations**). However, both methods don't deal with inter-video and intra-video variations simultaneously, which will directly hamper their performance.

*Corresponding author.

Set-based distance learning is an effective technique to reduce the variations between sets. Recently, a few set-based distance learning methods have been presented, including manifold discriminant analysis (MDA) [Wang and Chen, 2009], set-based discriminative ranking (SBDR) [Wu *et al.*, 2012], covariance discriminative learning (CDL) [Wang *et al.*, 2012], set-to-set distance metric learning (SSDML) [Zhu *et al.*, 2013] and localized multi-kernel metric learning (LMKML) [Lu *et al.*, 2013].

1.1 Motivation and Contribution

Existing distance learning based person re-id methods [Kostinger *et al.*, 2012; Zheng *et al.*, 2013] have demonstrated the effectiveness of distance learning technique for the person re-id task. However, these methods were designed for image-based person re-id, rather than for video-based person re-id particularly. Although a pedestrian video can be regarded as a sample set, existing set-based distance learning methods were not designed to tackle the video-based person re-id. Due to the existence of large intra-video and inter-video variations in the pedestrian video data, it's necessary and meaningful to investigate how to learn the more discriminative distance metric by reducing the influence of these variations.

The main contribution of this paper can be summarized as follows.

(1) We propose a novel video-based person re-id approach, namely simultaneous intra-video and inter-video distance learning (SI^{2}DL). To the best of our knowledge, this is the pioneer work to solve video-based person re-id by employing set-based distance learning technique.

(2) We design a new set-based distance learning model, which aims to learn a pair of intra-video and inter-video distance metrics to deal with the intra-video and inter-video variations, respectively. By using the learned intra-video metric, each video becomes more compact. Then by using the learned inter-video metric, the distance between truly matching videos becomes smaller than that between wrong matching videos.

(3) To enhance the discriminability of the learned inter-video distance metric, we design a new video relationship model, i.e., video triplet, which is constituted by a pair of truly matching videos and an "impostor" video.

(4) We evaluate the performance of SI^{2}DL and related methods on the public iLIDS-VID and PRID 2011 pedestrian sequence datasets. Experimental results demonstrate that our approach achieves the state-of-the-art performance.

2 Related Work

In this section, we briefly review two types of works that are related to our approach: (1) Video based person re-identification methods, (2) Set-based distance learning methods.

Video-based Person Re-identification. Video-based person re-identification is an important application in practice. Some early researches related to video-based person re-identification include [Cong *et al.*, 2009; Bedagkar-Gala and Shah, 2011]. Recently, video-based person re-identification methods [Wang *et al.*, 2014] and [Liu *et al.*, 2015] are presented. Both methods focus on extracting spatio-temporal features from videos. Method in [Wang *et al.*, 2014] divides each video into several fragments by employing the flow energy profile signal, and then learns a ranking model with the HOG3D features extracted from these fragments. Method in [Liu *et al.*, 2015] divides each video sequence into small segments corresponding to the action primitives, then each segment is further divided to a series of body-action units, finally, Fisher vectors extracted from all body-action units are concatenated as the representation of the video. **The major differences between our SI^{2}DL approach and methods [Wang *et al.*, 2014; Liu *et al.*, 2015] are two-folds.** Firstly, these methods focus on extracting spatial-temporal features, while SI^{2}DL focuses on learning a pair of distance metrics simultaneously. Secondly, these methods don't deal with the intra-video and inter-video variations simultaneously, while SI^{2}DL copes with them by learning an intra-video distance metric and an inter-video distance metric.

Set-based Distance Learning. Video-based classification is an important problem in several computer vision tasks. Since a video can be considered as an image set, a few set-based distance learning methods have been presented to solve this problem [Huang *et al.*, 2015c]. Method in [Wang and Chen, 2009; Huang *et al.*, 2015b] models each set as a manifold, and learns an embedding space to maximize manifold margin. Methods [Wang *et al.*, 2012] learn a set-to-set distance metric by modeling each set as a covariance matrix. Method [Zhu *et al.*, 2013] extends the point-to-point distance learning to the set-to-set distance learning by modeling each set as a convex hull. **The major differences between our approach and these methods can be summarized into two aspects**. Firstly, these methods are designed for image classification tasks (e.g., face recognition and object categorization), while our approach is designed for video-based person re-id particularly. Secondly, these methods learn a common distance metric to deal with the within-set and between-set variations, while our approach learns a pair of distance metrics to cope with intra-video and inter-video variations.

3 Simultaneous Intra-video and Inter-video Distance Learning (SI^{2}DL)

3.1 Problem Formulation

Denote by $X = [X_1, ..., X_i, ..., X_K]$ a set of p-dimensional training samples from K pedestrian videos, where $X_i \in \mathcal{R}^{p \times n_i}$ is the training sample set corresponding to the i^{th} video, and n_i is the sample number in X_i. Denote by x_{ij} the j^{th} sample in X_i. Since there exist large variations both within each video and between different videos, it's not an easy task to directly match between two videos. Intuitively, reducing the intra-video variation is beneficial to enhancing the inter-video separability. If we can make each video more compact, it will be easier to learn a video-to-video distance metric that has favorable discriminative capability. Therefore, we intend to jointly learn an intra-video distance metric and an inter-video distance metric from the training samples. The intra-video distance metric is to increase the compactness of each video, and the inter-video one is to enhance the separability of videos after intra-video distance learning. The basic idea of our SI^{2}DL approach is illustrated in Figure 2.

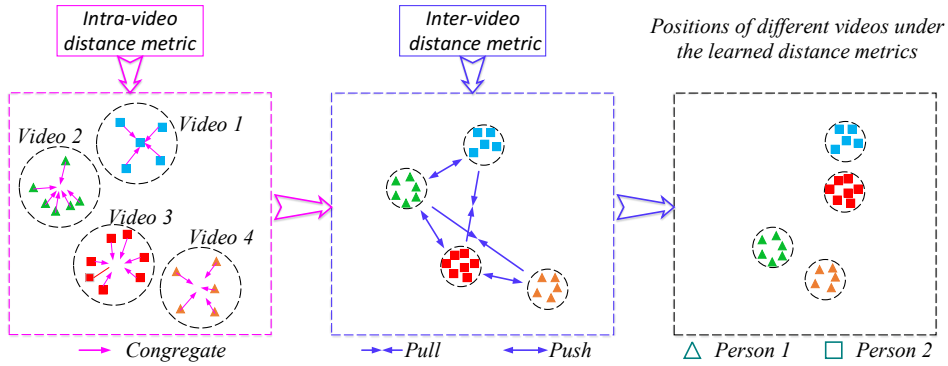


Figure 2: Illustration of our SI^{2}DL approach. Circles containing the same shape represent videos from the same person.

Therefore, we design the framework of SI^{2}DL as follows:

$$J(V,W) = arg\min_{V,W} f(V,X) + \mu g(W,V,X) \quad s.t.\ \|v_i\|_2^2 \le 1,\ \|w_i\|_2^2 \le 1 \tag{1}$$

where $V \in \mathcal{R}^{p\times K_1}$ and $W \in \mathcal{R}^{K_1\times K_2}$ separately represent the intra-video and inter-video distance metrics to be learned, K_1 and K_2 are positive integers. v_i and w_i are the i^{th} column vectors in V and W, respectively. $f(V,X)$ is the intra-video congregating term, which requires that each sample should move close to the center of video to which it belongs. The second term $g(W,V,X)$ is the inter-video discriminant term to ensure that the distance between two truly matching videos is smaller than that between two wrong matching videos. μ is a balancing factor. The constraints are used to restrict the scale of V and W.

There are several models that can be employed to represent a set [Li *et al.*, 2015b]. Considering that the intra-video distance learning results should be able to be directly used for learning the inter-video distance metric, we select the first-order statistics, which shows the averaged position of the sample set in the high dimensional space, to represent each video. For feature set X_i, its first-order statistics, denoted by m_i, can be computed with (2):

$$m_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \tag{2}$$

Therefore, we design $f(V,X)$ as follows:

$$f(V,X) = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left\|V^T(x_{ij}-m_i)\right\|_2^2 \tag{3}$$

where N is the total image number in X, m_i represents the first-order statistics of X_i. $g(W,V,X)$ is designed as follows:

$$g(W,V,X) = \frac{1}{|D|}\sum_{<i,j,k>\in D}\left(\left\|W^TV^T(m_i-m_j)\right\|_2^2 - \rho\left\|W^TV^T(m_i-m_k)\right\|_2^2\right) \tag{4}$$

where D represents the collection of video triplets, with each triplet consisting of a truly matching video pair and one of its "impostor" videos under V. Detailed information about the construction of a video triplet can be found in Definition 1. $|D|$ denotes the number of video triplets in D. $\rho = exp(-\|V^T(m_i-m_k)\|_2^2/\|V^T(m_i-m_j)\|_2^2)$ is a penalty factor. By substituting (3) and (4) into (1), the objective function of our approach can be written as:

$$\min_{V,W}\frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left\|V^T(x_{ij}-m_i)\right\|_2^2+\frac{\mu}{|D|}\sum_{<i,j,k>\in D}\left(\left\|W^TV^T(m_i-m_j)\right\|_2^2 - \rho\left\|W^TV^T(m_i-m_k)\right\|_2^2\right) \quad s.t.\ \|v_i\|_2^2 \le 1,\ \|w_i\|_2^2 \le 1 \tag{5}$$

The learned intra-video distance metric V ensures that all samples in each video move close to the corresponding first-order statistics, such that the first-order statistics can better represent each video, which will facilitate the learning of inter-video distance metric W. Furthermore, W is learned by exploiting the information provided by impostor videos according to the characteristics of pedestrian data, and therefore owns favorable discriminative ability.

Definition 1 (Video Triplet) *Given the intra-video distance metric V and videos X_i, X_j and X_k as well as their corresponding first-order statistics representations m_i, m_j and m_k, where X_j is a true matching of X_i, while X_k is a wrong matching of X_i. If $\|V^T(m_i-m_k)\|_2^2 < \|V^T(m_i-m_j)\|_2^2$, X_k is called an impostor video of X_i under V, then X_i, X_j and X_k constitute a* ***video triplet****, denoted by $<i,j,k>$.*

3.2 The Optimization of SI^{2}DL

The objective function (5) is not jointly convex to (V,W). To update V and W, we introduce two variable matrices A and B, and relax (5) to the following problem:

$$\min_{V,W,A,B}\frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left\|V^T(x_{ij}-m_i)\right\|_2^2+\left\|W^TA\right\|_F^2-\left\|W^TB\right\|_F^2 +\tau_1\left\|V^TM_1-A\right\|_F^2+\tau_2\left\|V^TM_2-B\right\|_F^2 \quad s.t.\ \|v_i\|_2^2 \le 1,\ \|w_i\|_2^2 \le 1 \tag{6}$$

where M_1 and M_2 are matrices with corresponding columns being respectively $\sqrt{\frac{\mu}{|D|}}(m_i-m_j)$ and $\sqrt{\frac{\mu\rho}{|D|}}(m_i-m_k)$, $<$

$i, j, k >\in D$. $\|\cdot\|_F$ represents the Frobenius norm. Then we can solve (6) by updating A, B, V and W iteratively. Detailed steps are as follows.

- **Update A and B by fixing V and W.**

Firstly, we should initialize V and W. Here, V is initialized by solving (7).

$$\min_{V} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left\| V^T (x_{ij} - m_i) \right\|_2^2, \ \ s.t. \ V^T V = I \tag{7}$$

By constructing the Lagrange function and setting the derivative to zero, we get

$$Q_1 V = \gamma V \tag{8}$$

where $Q_1 = \sum_{i=1}^{n} \sum_{j=1}^{n_i} (x_{ij} - m_i)(x_{ij} - m_i)^T$. It is clear that (8) is an eigen-decomposition problem and can be solved easily. We select eigenvectors corresponding to the smallest K_1 eigenvalues as V. After initializing V, we initialize W by solving (9):

$$\begin{aligned} \min_{W} & \sum_{<i,j,k>\in D} \left(\left\| W^T V^T (m_i - m_j) \right\|_2^2 - \rho \left\| W^T V^T (m_i - m_k) \right\|_2^2 \right) \\ & s.t. \ W^T W = I \end{aligned} \tag{9}$$

Similar to (7), this problem can also be solved by eigen-decomposition. Finally, W is set as the eigenvectors corresponding to the smallest K_2 eigenvalues.

When V and W are fixed, A and B can be easily obtained by solving problems (10) and (11), respectively.

$$\min_{A} \ \left\| W^T A \right\|_F^2 + \tau_1 \left\| V^T M_1 - A \right\|_F^2 \tag{10}$$

$$\min_{B} \ - \left\| W^T B \right\|_F^2 + \tau_2 \left\| V^T M_2 - B \right\|_F^2 \tag{11}$$

- **Update V by fixing A, B and W.**

When A, B and W are fixed, the objective function regarding V can be written as

$$\min_{V} \ \ h(V), \ \ s.t. \ \|v_i\|_2^2 \leq 1 \tag{12}$$

where

$$h(V) = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left\| V^T (x_{ij} - m_i) \right\|_2^2 + \tau_1 \left\| V^T M_1 - A \right\|_F^2 + \tau_2 \left\| V^T M_2 - B \right\|_F^2 \tag{13}$$

We optimize (12) with a similar way as [Gu *et al.*, 2014], i.e., introducing a variable S:

$$\min_{V,S} \ \ h(V), \ \ s.t. \ V = S, \ \|s_i\|_2^2 \leq 1 \tag{14}$$

The optimal solution of (14) can be obtained by the ADMM algorithm:

$$\begin{cases} V = \arg\min\limits_{V} \ h(V) + \gamma \|V - S + P\|_F^2 \\ S = \arg\min\limits_{S} \gamma \|V - S + P\|_F^2, \quad s.t. \ \|s_i\|_2^2 \leq 1 \\ P = P + V - S, \ \ update \ \gamma \ if \ appropriate. \end{cases} \tag{15}$$

where the initial value of P is a zero matrix.

- **Update W by fixing A, B and V.**

By fixing A, B and V, the objective function regarding W can be written as

$$\min_{W} \ \left\| W^T A \right\|_F^2 - \left\| W^T B \right\|_F^2 \ \ s.t. \ \|w_i\|_2^2 \leq 1 \tag{16}$$

Similar to (12), problem (16) can also be solved with the ADMM algorithm by introducing a variable S. The proposed SI^{2}DL algorithm is summarized in Algorithm 1.

Algorithm 1 Simultaneous intra-video and inter-video distance learning (SI^{2}DL)

Require: Training sample set X
Ensure: The learned distance metrics V and W
1: Initialize V and W by Eq. (7) and (9), respectively;
2: **while** not converge **do**
3: Fix V and W, update A and B by (10) and (11), respectively;
4: Fix A, B and W, update V according to (15);
5: Fix A, B and V, update W by Eq. (16);
6: **end while**
7: **return** V and W;

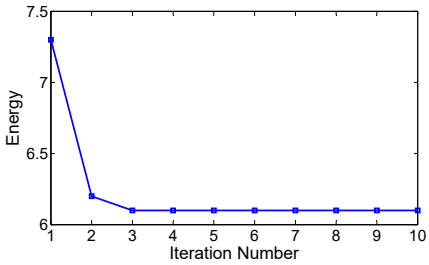


Figure 3: The convergence curve of SI^{2}DL on the PRID 2011 dataset.

3.3 Complexity and Convergence

In the training phase of SI^{2}DL, V and W are firstly initialized, and then V and W are updated alternatively. The time complexity of initializing V and W are respectively $O(Np^2 + p^3)$ and $O(L(K_1 p + {K_1}^2) + {K_1}^3)$, where $L = |D|$. In each iteration, $M_1 {M_1}^T$ and $M_2 {M_2}^T$ as well as Q_1 are not changed, thus the time complexities of updating A, B, V and W are $O({K_1}^2 K_2 + {K_1}^3 + {K_1}^2 p)$, $O({K_1}^2 K_2 + {K_1}^3 + {K_1}^2 p)$, $O(K_1 p^2 + T_1(p^2 K_1 + p^3))$ and $O(K_1 p^2 + {K_1}^2 p + T_2({K_1}^3 + {K_1}^2 K_2))$, respectively, where T_1 and T_2 are the iteration numbers in ADMM algorithm for updating V and W, respectively. We experimentally found that in most cases T_1 and T_2 are less than 10. In practice, the dimension p and N are much smaller than L.

The objective function in (6) is a bi-convex problem for $\{(V, W), (A, B)\}$, i.e., by fixing (A, B) the function is convex for (V, W), and by fixing (V, W) the function is convex for (A, B). The convergence of such a problem has already been intensively studied in [Gorski *et al.*, 2007]. Figure 3 shows the convergence curve of our algorithm on the PRID 2011 dataset. One can see that the energy drops quickly and becomes stable after 3 iterations. In most of our experiments, our algorithm will converge in less than 5 iterations.

3.4 Video-based Person Re-identification with the Learned Distance Metrics

With the learned intra-video and inter-video distance metrics (V, W), we can perform video-based person re-identification easily. Let $Y = [Y_1, ..., Y_i, ..., Y_n]$ be a set of p-dimensional samples from n gallery pedestrian videos, where $Y_i \in \mathcal{R}^{p \times l_i}$ is the sample set corresponding to the i^{th} gallery video, and

l_i is the number of samples in Y_i. Denote by $Z_i \in \mathcal{R}^{p \times n_i}$ the sample set corresponding to the i^{th} probe video, where n_i is the number of training samples in Z_i. Let y_{ij} (z_{ij}) represent the j^{th} sample in Y_i (Z_i). Detailed re-identifying steps between Z_i and Y are as follows.

(1) Calculating the first-order statistics representations of Z_i and each gallery video Y_i under V and W according to Eq. (17).

$$r(Z_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} W^T V^T z_{ij} \tag{17}$$

(2) Computing the distance between the probe video and each gallery video by $\|r(Z_i) - r(Y_k)\|_2^2$, $k = 1, 2, ..., n$.

(3) Sorting the obtained distances, and the gallery video with the smallest distance is the true matching of Z_i.

4 Experimental Results

To evaluate the effectiveness of our approach, we conduct extensive experiments on two publicly available person sequence datasets, including iLIDS-VID [Wang *et al.*, 2014] and PRID 2011 [Hirzer *et al.*, 2011].

4.1 Experimental Settings

Baselines. We compare our approach with the state-of-the-art video-based person re-id methods, including discriminative video fragments selection and ranking (**DVR**) [Wang *et al.*, 2014], and its two enhancements **Salience+DVR** and **MS-Colour&LBP+DVR** [Wang *et al.*, 2014], spatial-temporal fisher vector representation (**STFV3D**) and its enhancement method **STFV3D+KISSME** [Liu *et al.*, 2015].

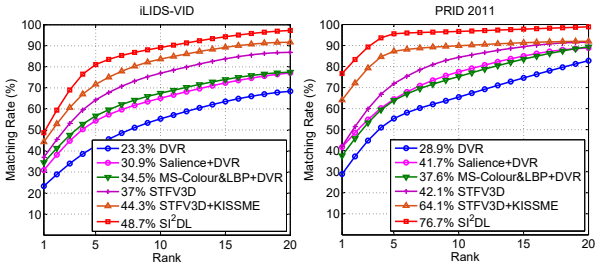
Feature Extraction. In experiments, we employ the effective feature (i.e., STFV3D) provided by the author of [Liu *et al.*, 2015] for both the iLIDS-VID and PRID 2011 datasets, which is the latest pedestrian video feature reported in existing video-base person re-id works. In particular, each video is represented with a sample set, with each sample being a Fisher vector extracted from a waling cycle. The dimensionality of each sample is 2208 and 2512 in the iLIDS-VID and PRID 2011 datasets, respectively.

Parameter Settings. There are three parameters in our SI^{2}DL model, i.e., μ, τ_1 and τ_2. In experiments, we choose these parameters by 5-fold cross-validation on each dataset. With respect to K_1 and K_2, we set them as (2200, 80) for iLIDS-VID, and (2500, 100) for PRID 2011, respectively. The choice of the values of $K1$ and $K2$ will be discussed in Section 4.4.

Evaluation Settings. We follow the evaluation protocol in [Wang *et al.*, 2014] for both iLIDS-VID and PRID 2011 datasets. In particular, we randomly split all sequence pairs into two sets of equal size, with one for training and the other for testing. Then we further select sequences from the first camera in the testing set to form the probe set, and those from the other camera are used as the gallery set. We employ the standard cumulated matching characteristics (CMC) curve as our evaluation metric, and report the Rank-k average matching rates of 10 trials.

4.2 Evaluation on the iLIDS-VID Dataset

The iLIDS-VID dataset [Wang *et al.*, 2014] consists of 600 image sequences for 300 persons, with each person having a pair of image sequences from two camera views. The length of each image sequence ranges from 22 to 192 frames, with an average number of 71. Due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and occlusions, the iLIDS-VID dataset is very challenging. Figure 1 (a) shows some example image sequences in iLIDS-VID. The parameters μ, τ_1 and τ_2 are set as 0.00005, 0.2 and 0.2, respectively.



(a) (b)

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

Figure 4 (a) shows the CMC curves of the compared methods. Table 1 shows the detailed rank-1, rank-5, rank-10, and rank-20 matching rates of all the compared methods. We can observe that our SI^{2}DL approach achieves higher matching rates in each rank, especially the rank-5 matching rate is improved by at least 9.4% (81.1%-71.7%). Compared with the competing methods, the main advantage of SI^{2}DL is that SI^{2}DL learns a pair of distance metrics with favorable discriminability, which can be used for reducing the intra-video and inter-video variations simultaneously.

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

Method	r=1	r=5	r=10	r=20
DVR	23.3	42.4	55.3	68.4
Salience+DVR	30.9	54.4	65.1	77.1
MS-Colour&LBP+DVR	34.5	56.7	67.5	77.5
STFV3D	37.0	64.3	77.0	86.9
STFV3D+KISSME	44.3	71.7	83.7	91.7
SI^{2}DL	**48.7**	**81.1**	**89.2**	**97.3**

4.3 Evaluation on the PRID 2011 Dataset

The PRID 2011 person sequence dataset [Hirzer *et al.*, 2011] consists of image sequences recorded from two disjoint cameras (Cam-A and Cam-B). Cam-A and Cam-B contain 385 and 749 person sequences, respectively. Each sequence contains 5 to 675 image frames, with an average number of 84. Among them, the first 200 persons appear in both views. 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].

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

Method	r=1	r=5	r=10	r=20
DVR	28.9	55.3	65.5	82.8
Salience+DVR	41.7	64.5	77.5	88.8
MS-Colour&LBP+DVR	37.6	63.9	75.3	89.4
STFV3D	42.1	71.9	84.4	91.6
STFV3D+KISSME	64.1	87.3	89.9	92.0
SI^{2}DL	**76.7**	**95.6**	**96.7**	**98.9**

Some example person sequences of the PRID 2011 dataset are shown in Figure 1 (b). The parameters μ, τ_1 and τ_2 are set as 0.00005, 0.1 and 0.1, respectively.

Table 2 and Figure 4 (b) report the top ranked matching rates of compared methods on the PRID 2011 dataset. It is observed that our SI^{2}DL approach obtains much higher matching rates than other methods. In particular, taking the rank-1 matching rate as an example, SI^{2}DL improves the average matching rate at least by 12.6% (=76.7%-64.1%).

4.4 Discussion

Learning a pair of distance metrics or a common one? Our approach learns an intra-video distance metric and an inter-video distance metric to handle the intra-video and inter-video variations, respectively. To evaluate the effectiveness of the manner that we learn the distance metrics, we compare SI^{2}DL with the modified SI^{2}DL version that learns a common distance metric. We call the modified version **SI^{2}DL_c**. Figure 5 reports the rank-1 matching rates of SI^{2}DL and SI^{2}DL_c on the iLIDS-VID and PRID 2011 datasets. We can see that SI^{2}DL significantly outperforms SI^{2}DL_c, which indicates that the manner of learning a common distance metric to catch both the intra-video and inter-video variations will lead to a compromise to the discriminability of the learned distance metric. Therefore, we should learn different distance metrics to handle variations of different levels.

Dimensionality of intra-video and inter-video distance metrics. Since the inter-video distance learning depends on the results of intra-video distance learning, selecting proper dimensionality for the intra-video metric (V), i.e., K_1, is important in the learning of our SI^{2}DL model. Since V is initialized by Eq. (8), which is solved by eigen-decomposition, the value of K_1 should be smaller than the sample dimension, i.e., 2512 for PRID 2011. In experiments, we found that the value of K_1 should be close to the dimensionality of each training sample (p), while K_2 should be small. Figure 6 shows the Rank-1 matching rates of SI^{2}DL with K_1 in the range [2400, 2500] and K_2 in the range [50, 120] on the PRID 2011 dataset. We can see that SI^{2}DL is not very sensitive to the choices of K_1 and K_2 in the observed ranges, and SI^{2}DL achieves the best performance when K_1 and K_2 are separately set as 2500 and 100. Similar performance-stable ranges can be observed on the iLIDS-VID dataset.

Comparison with set-based distance learning methods. In experiments, we also compared SI^{2}DL with three state-of-the-art set-based distance learning methods, including covariance discriminative learning (CDL) [Wang *et al.*, 2012], set-to-set distance metric learning (SSDML) [Zhu *et al.*, 2013], and hybrid euclidean-and-riemannian metric learning (HERML) [Huang *et al.*, 2015a]. These methods used the same features and settings as SI^{2}DL. Experimental results showed that the competing methods obtain rather poor matching rates, so the detailed results of these methods are not provided. The main reasons for this phenomenon can be summarized into three aspects: (i) These methods are designed for image classification tasks, where each class has several image sets (more than two). However, in the setting of video-based person re-id, there is only one video per person for one camera. (ii) They assume that training and testing sets contain the same classes, and conduct the matching between testing samples and training samples. However, in video-based person re-id, people in the training set will not appear in the testing set, and the matching is conducted between two videos from different cameras. (iii) They learn a common distance metric and don't deal with the intra-video variations particularly. All these factors will bring large difficulties to them.

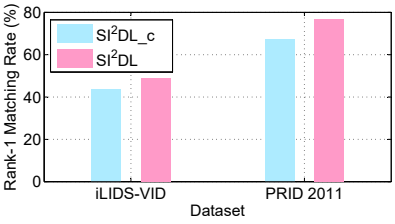


Figure 5: Rank-1 matching rates (%) of SI^{2}DL and SI^{2}DL_c.

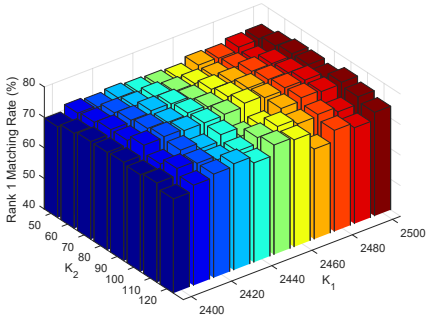


Figure 6: Rank 1 matching rates of SI^{2}DL versus different values of K_1 and K_2 on the PRID 2011 dataset.

5 Conclusion

This paper proposes a novel set-based distance learning approach for video-based person re-identification, which simultaneously learns a pair of intra-video and inter-video distance metrics. The learned intra-video distance metric can make each video more compact, such that the extracted first-order statistics feature can better represent each video. The learned inter-video distance metric can make the distance between truly matching videos smaller than that of wrong matching videos. Experimental results on the public iLIDS-VID and PRID 2011 datasets show that our approach achieves the best matching rates, and also demonstrate the effectiveness of the manner of learning a pair of distance metrics to deal with intra-video and inter-video variations.

Acknowledgments

The authors want to thank the Editor and anonymous reviewers for their constructive comments and suggestions. The work described in this paper was supported by the National Nature Science Foundation of China under Project Nos. 61272273 and the Research Project of NJUPT (XJKY14016).

References

[Ahmed *et al.*, 2015] Ejaz Ahmed, Michael Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916, 2015.

[Bedagkar-Gala and Shah, 2011] Apurva Bedagkar-Gala and Shishir K Shah. Multiple person re-identification using part based spatio-temporal color appearance model. In *ICCV Workshops*, pages 1721–1728, 2011.

[Cong *et al.*, 2009] Dung Nghi Truong Cong, Catherine Achard, Louahdi Khoudour, and Lounis Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In *ICIAP*, pages 179–189. 2009.

[Gorski *et al.*, 2007] Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007.

[Gu *et al.*, 2014] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Projective dictionary pair learning for pattern classification. In *NIPS*, pages 793–801, 2014.

[Hirzer *et al.*, 2011] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. 2011.

[Huang *et al.*, 2015a] Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Hybrid euclidean-and-riemannian metric learning for image set classification. In *ACCV*, pages 562–577. 2015.

[Huang *et al.*, 2015b] Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *CVPR*, pages 140–149, 2015.

[Huang *et al.*, 2015c] Zhiwu Huang, Ruiping Wang, Shiguang Shan, Xianqiu Li, and Xilin Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *ICML*, 2015.

[Jing *et al.*, 2015] Xiao-Yuan Jing, Xiaoke Zhu, Fei Wu, Xinge You, Qinglong Liu, Dong Yue, Ruimin Hu, and Baowen Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, pages 695–704, 2015.

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

[Li *et al.*, 2015a] Sheng Li, Ming Shao, and Yun Fu. Cross-view projective dictionary learning for person re-identification. In *IJCAI*, pages 2155–2161, 2015.

[Li *et al.*, 2015b] Yan Li, Ruiping Wang, Zhiwu Huang, Shiguang Shan, and Xilin Chen. Face video retrieval with image query via hashing across euclidean space and riemannian manifold. In *CVPR*, pages 4758–4767, 2015.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, June 2015.

[Liu *et al.*, 2015] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*, pages 3810–3818, 2015.

[Lu *et al.*, 2013] Jiwen Lu, Gang Wang, and Philippe Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, pages 329–336, 2013.

[Wang and Chen, 2009] Ruiping Wang and Xilin Chen. Manifold discriminant analysis. In *CVPR*, pages 429–436, 2009.

[Wang *et al.*, 2012] Ruiping Wang, Huimin Guo, Larry S Davis, and Qionghai Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503, 2012.

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

[Weinberger and Saul, 2009] K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[Wu *et al.*, 2012] Yang Wu, Michihiko Minoh, Masayuki Mukunoki, and Shihong Lao. Set based discriminative ranking for recognition. In *ECCV*, pages 497–510. 2012.

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

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

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

[Zheng *et al.*, 2013] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(3):653–668, 2013.

[Zhu *et al.*, 2013] Pengfei Zhu, Lei Zhang, Wangmeng Zuo, and Dejing Zhang. From point to set: Extend the learning of distance metrics. In *ICCV*, pages 2664–2671, 2013.