The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Video-Based Person Re-Identification via Self Paced Weighting

Wenjun Huang,[2,3] **Chao Liang,**[1,2,3*] **Yi Yu,**[4] **Zheng Wang,**[2,3] **Weijian Ruan,**[2,3] **Ruimin Hu**[1,2,3]
[1]State Key Laboratory of Software Engineering, Wuhan University, China
[2]Collaborative Innovation Center of Geospatial Technology, China
[3]National Engineering Research Center for Multimedia Software, Wuhan University, China
[4]Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan
Email: {2011huang,cliang,wangzwhu,rweij,hrm}@whu.edu.cn yiyu@nii.ac.jp

Abstract

Person re-identification (re-id) is a fundamental technique to associate various person images, captured by different surveillance cameras, to the same person. Compared to the single image based person re-id methods, video-based person re-id has attracted widespread attentions because extra space-time information and more appearance cues that can be used to greatly improve the matching performance. However, most existing video-based person re-id methods equally treat all video frames, ignoring their quality discrepancy caused by object occlusion and motions, which is a common phenomenon in real surveillance scenario. Based on this finding, we propose a novel video-based person re-id method via self paced weighting (SPW). Firstly, we propose a self paced outlier detection method to evaluate the noise degree of video sub sequences. Thereafter, a weighted multi-pair distance metric learning approach is adopted to measure the distance of two person image sequences. Experimental results on two public datasets demonstrate the superiority of the proposed method over current state-of-the-art work.

Introduction

Person re-identification, which aims at identifying a person of interest among different cameras, has become increasingly popular in the research community due to its critical role in many surveillance, security and multimedia applications (Auguste, Martinet, and Tirilly 2015; Loy, Xiang, and Gong 2009; Ye et al. 2016). Currently, major efforts towards this problem focus on still images, in which each person has only one or few images per camera view (Bedagkar-Gala and Shah 2014). Many methods have been developed for accurate person re-identification (Wang et al. 2016b; 2017; Matsukawa et al. 2016; Wang et al. 2016c; Liang et al. 2015). However, the real-world re-id performance is still hindered by limited information extracted from still images, and these methods usually failed to capture complete and robust appearance features in complex scenarios because of ignoring key temporal information among successive image frames (You et al. 2016).

Recently, many research efforts have been devoted into video-based person re-id problem (Wang et al. 2014; Karanam, Li, and Radke 2015a; Liu et al. 2015; Zhu et al. 2016; You et al. 2016; Wang et al. 2016a; Zheng et al. 2016). In this scenario, each person object, in every camera's view field, is described by a long sequence of video images. Hence, more complete and abundant visual information can be utilized for accurate feature representation (Wang et al. 2014; Karanam, Li, and Radke 2015a; Liu et al. 2015; Gao et al. 2016; Liu, Chen, and Wang 2016) and discriminative distance measure (Karanam, Li, and Radke 2015a; Zhu et al. 2016; You et al. 2016; Ye et al. 2017).

However, most video based person re-id methods equally treat all images in each sequence, losing sight of their quality discrepancy caused by object motions and occlusions. Take iLID-VID dataset (Wang et al. 2014) in Fig. 1 as an example, person images are flooded with various object occlusions or background clutters, resulting in highly noisy *unregulated sequences*. On our preliminary comparative experiment conducted on 199-pair unregulated person sequences of iLID-VID dataset, average matching accuracy on original unregulated video sequences is only 7%, 10 percents lower than that obtained on filtered clean video sequences.

Wang *et al.* (Wang et al. 2016a; 2014) first noticed the impact of unregulated sequences, and raised an optical flow based algorithm to detect walking cycles to divide a video sequence into different sub fragments. Then, a ranking model was proposed to select and match aligned video fragment pairs from candidates pool. However, the method is hard to obtain a reliable optical flow estimation under serious noise interference (Ayvaci, Raptis, and Soatto 2010). Furthermore, it exploits all video fragments in candidates pool to select and match fragment pairs without considering the interference of noisy video fragments, resulting in a significant performance degradation.

To address the above problem, we propose a novel video-based person re-id method by self paced weighting (SPW). In our work, two key issues need to be addressed. (1) the algorithm is expected to decompose a video sequence into a series of sub sequences so that images in the same sub sequence have unified state (i.e., either noisy or clean). (2) the algorithm should be able to measure the noisy condition of each sub sequence and make a robust distance metric. For the first issue, we define a Sequence Stability Measure (SSM) to break down automatically unregulated video sequences into multiple fragments with re-

[*]Chao Liang is the corresponding author.

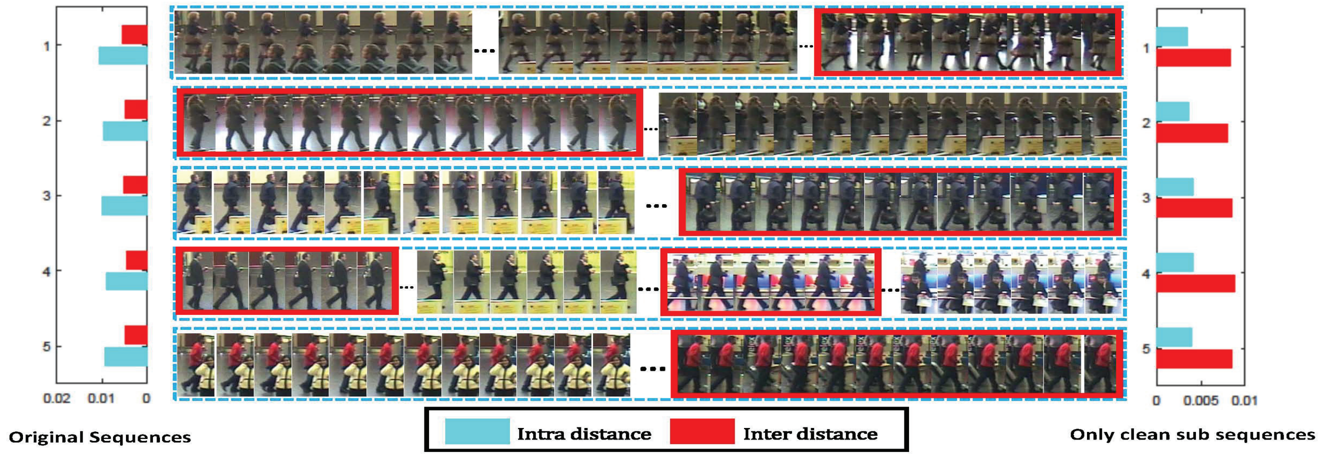


Figure 1: A preliminary comparative experiment which illustrates the negative impact of unregulated sequences. *Intra distance* (the distance between a pair of video sequences of same person captured by two cameras) and the *inter distance* (the minimum distance between a video sequence from one camera and video sequences of other persons from another camera) are computed between video sequence pairs. Five example unregulated video sequences in iLIDS-VID dataset are selected, for each video sequence, we exploit a combined feature representation (You et al. 2016). The *intra distance* is greater than *inter distance* for each video sequence when directly applying unregulated sequences (blue boxes indicate). For comparison, if only considering the clean sub sequences (red boxes indicate), resulting in a nearly perfect result, the *inter distance* is greater than *intra distance*.

spect to the state change of video sequences. For the second issue, motivated by the curriculum learning (CL) idea of self paced learning (Kumar, Packer, and Koller 2010; Lu et al. 2015), in which samples in a curriculum are selected solely in terms of 'easiness'. In this paper, a self paced outlier detection (SPOD) method is proposed to evaluate the noise degree of video sub sequences. Besides, an adaptive weighted multi-pair distance metric learning approach is proposed to jointly measure the distance between video sequences.

To summarize, the contributions of this paper are the followings:

- We find that remarkable performance improvement can be achieved via an self paced weighting (SPW) method.
- We propose a self paced outlier detection method to eliminate noisy sub sequences through noise degree measure.
- We propose a weighted multi-pair distance metric learning approach to measure the distance between video sequences.

Related Work

Feature Representation To extract discriminative features, both image frame level representation and video level representation are adopted. For image frame level representation, most methods use similar features to image-based re-id methods (Karanam, Li, and Radke 2015a; You et al. 2016; Zheng et al. 2016). For video level representation, temporal cues in video is incorporated, which mainly use spatial-temporal descriptors to re-identify pedestrians (Wang et al. 2014; You et al. 2016; Wang et al. 2016a), such as HOG3D (Klaser, Marszałek, and Schmid 2008), the gait energy image (GEI) (Han and Bhanu 2006).

Metric Learning Distance metric learning is also important when matching videos, inter-class distance is much larger for the video-based re-id compared to the image-based re-id because some motion information of different people could be similar. Thus, more stringent constraint is exploited to look for a latent space to maximize the inter-class margin between different persons. Karanam *et al.* (Karanam, Li, and Radke 2015a) proposed to learn a dictionary from all the available images for person that was capable of discriminatively and sparsely encoding features representing different person. Zhu *et al.* (Zhu et al. 2016) proposed a simultaneous intra-video and inter-video distance learning approach to make video representation more compact and to discriminate videos of different identities. You *et al.* (You et al. 2016) proposed a top-push constraint distance learning model which optimized the top-rank matching in video re-id by selecting discriminative features. Wang *et al.* (Wang et al. 2016a) formulated the person re-identification problem as a ranking problem. Ye *et al.* (Ye et al. 2017) proposed a dynamic graph matching (DGM) method to estimate labels for unsupervised video re-id problem.

CNN-based Schemes The main idea of CNN-based methods is to extract useful representations from video images with CNN (or CNN-RNN) models or design an end-to-end deep neural network architecture that simultaneous extracts feature and learns distance metric. In McLaughlin's work, McLaughlin *et al.* (McLaughlin et al. 2016) proposed a recurrent neural network architecture to give an overall appearance feature for the complete sequence. Zhou et al. (Zhou et al. 2017) integrates a temporal attention model within an end-to-end deep neural network architecture. Xu et al. (Xu et al. 2017) proposed a novel deep architecture with jointly attentive spatial-temporal pooling.

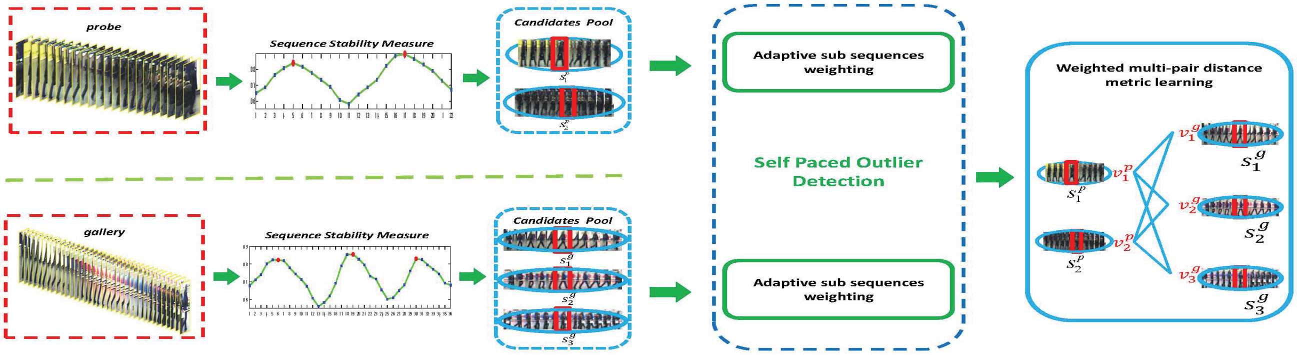

probe
Sequence Stability Measure
Candidates Pool
s_1^p
s_2^p
gallery
Sequence Stability Measure
Candidates Pool
s_1^g
s_2^g
s_3^g
Adaptive sub sequences weighting
Self Paced Outlier Detection
Adaptive sub sequences weighting
Weighted multi-pair distance metric learning
v_1^g
s_1^g
s_1^p
v_1^p
v_2^g
s_2^g
s_2^p
v_2^p
v_3^g
s_3^g

Figure 2: The overall scheme of the proposed SPW method. In the video sequence segmentation part, a video sequence is divided into a series of sub sequences by detecting *stationary points* (red dots in the figure, each of which corresponds to a local stable state of a person) of SSM signals. Then, the SPOD method is adopted to evaluate the sub sequences noise degree V, the lager the noise level $v_i^{(k)}$ is, the more likely this sub sequence posses noise, thus resulting a larger distance. Finally, a weighted multi-pair distance metric learning approach is adopted to measure the distance of two person image sequences.

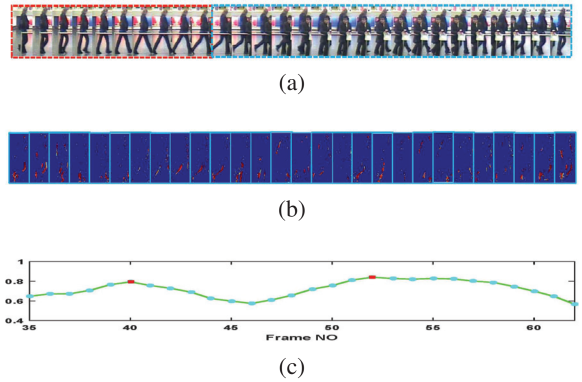
Proposed Approaches

The overall scheme of the proposed SPW method is shown in Fig. 2. It is divided into three stages. In the video sequence segmentation stage, SSM is defined to detect the state change of video sequences and automatically break down unregulated video sequences into multiple fragments. In the outlier detection stage, we propose a SPOD method to evaluate the noise degree of each sub sequence. Then, in the distance measure stage, a weighted multi-pair distance metric learning approach is proposed to jointly measure the distance between two person image sequences.

Video Sequence Segmentation

Given an unregulated video sequence with heavy occlusions, as shown in Fig. 1, it is too noisy to directly applying unregulated sequences. To effectively utilize clean sub sequences in video sequences, it is necessary to divide the entire unregulated sequences into fragments, in which more emphasis should be put on clean sub sequences in re-id process. Wang *et al.* (Wang et al. 2016a; 2014) try to find aligned sub sequence pairs by detecting motion information, however, it is hard to obtain a reliable optical flow estimation without considering the occlusions. Different from Wang's work (Wang et al. 2016a; 2014), we divide the video sequence into a series of fragments by using occlusion information between consecutive frames instead of optical flow information, which can better reflect the state change between consecutive frames in complex scenarios. Besides, by detecting the occlusion information, image frames within the same sub sequence always have unified state (i.e., noisy or clean, as mentioned earlier), which is crucial for our approach.

To better reflect the state change between video frames, we use the occlusion information between consecutive frames (Ayvaci, Raptis, and Soatto 2010) to measure the state change within each video sequence. In fact, the occlusion information reflects the inter-frame occlusion, which indicates the occlusion condition of current frame with respect to its previous frame, as shown in Fig. 3(b). Let o_i denotes the occlusion information between frame I_i and I_{i-1}, the larger the o_i is, the state change between frame I_i and I_{i-1} is more vigorous. From a certain point of view, o_i expresses some kind of stability between two frames. On this observation, the stability measure ϕ_i is defined, which indicates the state difference between each frame and its previous frame.


(a)

(b)

(c)

Figure 3: (a) An example video fragment in iLIDS-VID dataset. (b) The state change information between consecutive frames of (a). (c) The Sequence Stability Measure Signal of (a).

Given a video sequence $M = (I_1, I_2, \cdots, I_N)$, we use I_i to denote the $i-th$ frame for simplicity, the occlusion information o_i between consecutive frames is computed, then, the stability of each frame is defined as follows

$$\phi_i = \exp(\frac{-||o_i||_2}{c}), i = 2, 3, ..., N, \qquad (1)$$

where c is a constant factor. Then, the *sequence stability measure* (SSM) ϵ of M is defined as $\epsilon = (\phi_2, \cdots, \phi_N)$, which is further smoothed by a Gaussian filter to suppress noise, as shown in Fig. 3(c).

It can be observed that the local maximum $\{t\}$ of the SSM signal ϵ corresponds to a characteristic state of the person, which means that the surrounding frames of a local maximum have consistent state (i.e., the surrounding frames are all occluded or all clean), thus helping in estimating the state discrepancy between consecutive video fragments. And we name those maxima landmarks of ϵ *stationary point* (red dots in Fig. 3(b)). Finally, the video sequence M is split into a set of sub sequences $S = \{s_i\}, i = 1, \cdots, z$ by detecting stationary point $\{t\}$ of ϵ and extracting the surrounding frames $s = (I_{t-L}, \cdots, I_t, \cdots, I_{t+L})$ of each stationary point as a sub sequence, as shown in Fig. 3(a). The temporal range L is adaptively determined by the range between the local maximum and local minimum.

Self Paced Outlier Detection

Given video sub sequences $S = \{s_i\}, i = 1, \cdots, z$, the next step is to distinguish the sub sequences with heavy noise from those which are relatively clean. Motivated by the curriculum learning idea of self paced learning (SPL)(Kumar, Packer, and Koller 2010), whose core idea is to generally start with learning easier aspects of a task, and then gradually take more complex examples into consideration, we attempt to adopt a similar thresholding and weighting scheme to adaptively learn samples ranking scores and their noise degree. Particularly, different from classic SPL methods that mainly proposed for classification problems, this paper raises a new Self Paced Outlier Detection (SPOD) algorithm whose loss function is defined on pairwise samples, having the ability to formulate ranking relations among samples.

Problem In the video-based person re-identification problem, there are a probe sequence p and n gallery sequences. By dividing each video sequence into a series of sub sequences, all sub sequences of probe sequence p and n gallery sequences can be collectively denoted as $S = (S^{(1)}, \cdots, S^{(i)}, \cdots, S^{(K)})$, $K = n + 1$, where columns of $S^{(i)}$ correspond to the sub sequences in the $i - th$ video sequence. Let m_i denotes the number of sub sequences in the $i - th$ video sequence, then the total sub sequences m in S can be obtained as $m = \sum_{i=1}^{K} m_i$. Accordingly, we denote the weight vector as $V = [v^{(1)}, \cdots, v^{(i)}, \cdots, v^{(K)}]$, where $v^{(i)} = (v_1^i, \cdots, v_{m_i}^i)^T$.

Given $Y = [y_i] \in \mathbb{R}^{m\times 1}$ and the final ranking score $f = [f_i] \in \mathbb{R}^{m\times 1}$, where the sub sequences in the same sequence have same ranking score f. The pairwise ranking loss $l_{i,j}$ between sub sequence s_i and s_j can be formulated under manifold assumption (Zhou et al. 2003) as

$$l_{i,j} = \omega_{i,j}(f_i - f_j)^2 + \alpha_i(f_i - y_i)^2 + \alpha_j(f_j - y_j)^2, \quad (2)$$

where $\omega_{i,j} = \exp(-dist(s_i, s_j))$ and pairwise data affinity can be represented as $W = [\omega_{i,j}] \in \mathbb{R}^{m\times m}$. The first term is a smoothness regularization term that requires nearby samples having similar ranking scores (Liang et al. 2015), and last two terms are fitting constraint terms that require ranking scores should not change too much from their comprehensive ranking (Chen, Liu, and Gong 2014). $\alpha = [\alpha_i] \in R^{m\times 1}$ is the weight parameters to balance the smoothness and fitting terms. On this basis, the proposed problem can be modeled as a constraint optimization problem

$$\begin{aligned} (f^*, V^*) &= \arg\min_{f,V} E(f, V) \\ &s.t. -1 \le f \le 1 \\ &0 \le V \le 1. \end{aligned} \quad (3)$$

Similar to the self paced learning with diversity (SPLD) (Jiang et al. 2014), optimization objective $E(f, V)$ takes the form of

$$E(f, V) = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}(1 - v_i)(1 - v_j)[l_{i,j} - \beta] - \gamma||V||_{2,1}, \quad (4)$$

where $V = [v_i^{(k)}]$ reflects the noise level of sub sequences, whose values are determined by comparative results of ranking loss $L = [l_{i,j}]$ and self pace thresholds β. Specifically, when ranking loss $l_{i,j}$ is smaller than the threshold β, their corresponding noise level v_i and v_j will approach to 0 so that the prefix weight $(1 - v_i)(1 - v_j)$ is close to 1. Conversely, larger $l_{i,j}$ will result in larger v_i and v_j, indicating that the sub sequences s_i and s_j are very likely to be outliers, which means s_i and s_j have larger noise level. As for the diversity term (the negative $l_{2,1} - norm : -||V||_{2,1}$), which has a counter-effect to group-wise sparsity (Jiang et al. 2014), i.e. non-zero entities of V tend to be scattered across different sequences, is defined as

$$-||V||_{2,1} = -\sum_{k=1}^{K} ||v^{(k)}||_2 \,. \quad (5)$$

Solution To solve the proposed problem above, an alternative optimization method can be adopted.

Step1 (Fix V) First, fix noise level matrix V, the optimization objective of problem (Eq. (4)) changes into

$$E_V(f) = \sum_{i=1}^{m}\sum_{j=1}^{m}(1 - v_i)(1 - v_j)l_{i,j} \,. \quad (6)$$

Compared to Eq. (4), Eq. (6) abandons self paced β term and negative $l_{2,1} - norm$ term that are solely dependent on V, while only keeps the loss function term $l_{i,j}$ that are dependent on f . Thus, the final optimization can be solved by regular convex optimization methods.

Step2 (Fix f) Given solved f , the optimization objective of problem (Eq. (4)) changes into

$$\begin{aligned}E_f(V) &= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}(1-v_i)(1-v_j)[l_{i,j}-\beta]-\gamma||V||_{2,1}\\ &=\sum_{i,j,k}\frac{1}{{m_k}^2}(1-v_i^{(k)})(1-v_j^{(k)})[l_{i,j}^{(k)}-\beta]-\sum_{k=1}^{K}\frac{\gamma}{m}||v^{(k)}||_2\\ &=\sum_{k=1}^{K}\left[\frac{1}{{m_k}^2}\sum_{i,j}(1-v_i^{(k)})(1-v_j^{(k)})[l_{i,j}^{(k)}-\beta]\right]-\sum_{k=1}^{K}\frac{\gamma\cdot||v^{(k)}||_2}{m_k}\\ &=\sum_{k=1}^{K}E_f(v^{(k)})\ .\end{aligned} \tag{7}$$

Then, the above problem can be solved with CCCP algorithm (Yuille and Rangarajan 2002) and the noise level matrix V can be obtained. And alternation and iterative strategy is adopted until the algorithm converges.

Weighted Multi-pair Distance Measure

As mentioned in (Wang et al. 2014), learning a ranking function holistically without discrimination and selection from pairs of unsegmented and temporally unaligned person image sequences will subject the learned model to significant noise and degrade any meaningful discriminative information contained in the image sequences. In this paper, a weighted multi-pairs distance metric learning approach is adopted to measure the distance of two person image sequences. Specifically, the query sequence $S^{(p)}$ and gallery sequence $S^{(g)}$, which contain a series of sub sequences, are denoted as: $S^{(p)} = [s_1^p, ..., s_{n_1}^p]$, $S^{(g)} = [s_1^g, ..., s_{n_2}^g]$, n_1 and n_2 are the number of sub sequences of $S^{(p)}$ and $S^{(g)}$ accordingly. Then, the distance between $S^{(p)}$ and $S^{(g)}$ is measured as follows:

$$\begin{aligned}dist(S^{(p)}, S^{(g)}) &= \frac{\sum\limits_{i}\min\limits_{j} z_{i,j}||s_i^p - s_j^g||_2}{2n_1}\\ &+ \frac{\sum\limits_{j}\min\limits_{i} z_{i,j}||s_i^p - s_j^g||_2}{2n_2}\ .\end{aligned} \tag{8}$$

where

$$z_{i,j} = \frac{1}{1+\exp(-v_i^p \cdot v_j^g)}. \tag{9}$$

and v_i^p and v_j^g are the noise level value of sub sequences s_i^p and s_j^g.

The basic idea of this distance measure method is that more emphasis should be put on clean sub sequences when measuring the distance between video sequences. In fact, from Eq.(9), we can see that when $v_i^p \cdot v_j^g$ is very large, meaning that s_i^p and s_j^g are highly noisy, then this distance $||s_i^p - s_j^g||_2$ is further expand. Thus, s_i^p and s_j^g have almost no effect when calculating the distance between $S^{(p)}$ and $S^{(g)}$. In other words, when a sub sequence is highly noisy, it would be eliminated to some extent when measuring distance between video sequences.

Experiments

Datasets and Settings

Our experiments are conducted on two publicly available video datasets for video-based person re-id: the PRID 2011 dataset (Hirzer et al. 2011) and the iLIDS-VID dataset (Wang et al. 2014).

PRID 2011 dataset. The PRID 2011 dataset consists of video pairs recorded from two different cameras. 385 persons are recorded in camera view A, and 749 persons in camera view B. Among all persons, 200 persons are recorded in both camera views. Each video sequence contains 5 to 675 image frames, with an average number of 84. The dataset has two adjacent camera views captured in uncrowded outdoor scenes with rare occlusions and clean background.

iLIDS-VID dataset. The iLIDS-VID dataset includes 600 video sequences for 300 people based on two non-overlapping camera views. The length of image sequences varies from 23 to 192, with an average number of 73. Compared with the PRID 2011 dataset, this dataset is more challenging due to environment variations especially complicated background, occlusions, clothing similarities and viewpoint variations across camera views, as shown in Fig. 1.

Settings. In our experiments, all datasets are randomly split into two subsets, half for training and half for testing. In the testing stage, the video sequences from the first camera are used as the probe set, and the ones from the other camera as the gallery set. According to the formulations in Eq.(2) and Eq.(4), there are three key parameters in our SPOD model: α, β and γ. α is the weight coefficient to emphasize the fitting terms about ranking score f and sub sequences score Y. β is the self pace threshold to discriminate the ranking quality of each $< i, j >$sample pair. γ is the weight coefficient to emphasize the negative $l_{2,1} - norm$ of noise level V. These three parameters are determined by cross validation on iLIDS-VID dataset and fixed for all test sequences during the experiments. We use the cumulative matching characteristic (CMC) curve (Wang et al. 2007) to measure the performance on both datasets, the average CMC curves of 10 trails is adopted to obtain a more reliable results.

The Effectiveness of Our Method

In this section, to evaluate the effectiveness of our method, we integrate our method with some basic feature representations which are widely used in video-based person re-identification (Wang et al. 2014) to make a comparison with the original basic descriptors. We conduct 8 groups of experiments to make a more comprehensive comparison.

HOG3D. The 3D HOG features from volumes of video data are extracted similar to (Wang et al. 2014). More specifically, for each local maximum/minimum of the FEP signal E, 10 frames immediately before and after the central frame are taken as a fragment, divided into $2 \times 5(spatial) \times 2$ (temporal) cells with 50% overlap. A spatial-temporal gradient histogram is computed in each cell and then concatenated to form the HOG3D descriptor (Klaser, Marszałek,

Table 1: Comparison with different feature representations with metric learning XQDA (Liao et al. 2015) and TDL (You et al. 2016) on PRID 2011 and iLIDS-VID datasets. For each group of experiment, we compare the original approach with our approach, $\surd$ denotes that our method SPW is adopted, performance improvement is shown in red font. Results are shown as matching rates (%) at Rank = 1, 5, 20. Best result of each rank is in boldface font.

Method				PRID 2011			iLIDS-VID		
Experiment	Feature	Metric	SPW	Rank-1	Rank-5	Rank-20	Rank-1	Rank-5	Rank-20
	HOG3D	XQDA	–	44.9	69.9	88.4	33.8	58.1	83.4
exp 1	HOG3D	XQDA	√	51.5 (↑ **6.6**)	75.2 (↑ **5.3**)	93.6(↑ **5.2**)	40.7(↑ **6.9**)	62.8(↑ **4.7**)	88.9(↑ **5.5**)
	Color	XQDA	–	53.7	76.0	93.0	38.5	64.0	82.9
exp 2	Color	XQDA	√	59.0(↑ **5.3**)	80.8(↑ **4.8**)	98.5(↑ **5.5**)	45.5(↑ **7.0**)	70.3(↑ **6.3**)	88.9(↑ **6.0**)
	HOG3D + Color	XQDA	–	60.4	82.4	94.9	41.3	66.7	87.0
exp 3	HOG3D + Color	XQDA	√	67.5(↑ **7.1**)	85.3(↑ **2.9**)	97.1(↑ **2.2**)	55.4(↑ **14.1**)	78.1(↑ **11.4**)	91.9(↑ **4.9**)
	IDE	XQDA	–	78.8	94.4	97.9	52.2	73.0	87.9
exp 4	IDE	XQDA	√	82.3(↑ **3.5**)	96.1(↑ **1.7**)	98.8(↑ **0.9**)	58.7(↑ **6.6**)	79.5(↑ **6.5**)	90.3(↑ **2.6**)
	HOG3D	TDL	–	47.3	73.5	92.5	39.3	66.9	88.7
exp 5	HOG3D	TDL	√	50.7(↑ **3.4**)	74.0(↑ **0.5**)	92.8(↑ **0.3**)	45.3(↑ **6.0**)	73.3(↑ **6.4**)	91.3(↑ **2.6**)
	Color	TDL	–	54.8	76.7	92.4	40.2	70.3	87.2
exp 6	Color	TDL	√	61.6(↑ **5.8**)	81.2(↑ **4.5**)	95.9(↑ **3.5**)	54.0(↑ **13.8**)	77.3(↑ **7.0**)	92.7(↑ **5.6**)
	HOG3D + Color	TDL	–	62.8	85.7	97.9	44.4	72.1	91.3
exp 7	HOG3D + Color	TDL	√	69.1(↑ **6.3**)	90.0(↑ **4.3**)	99.0(↑ **1.2**)	59.3(↑ **14.9**)	83.3(↑ **11.2**)	96.0(↑ **4.7**)
	IDE	TDL	–	79.1	94.7	98.7	61.8	84.7	92.5
exp 8	IDE	TDL	√	**83.5(↑ 4.4)**	**96.3(↑ 1.6)**	**100.0(↑ 1.3)**	**69.3(↑ 7.5)**	**89.6(↑ 4.9)**	**98.2(↑ 5.7)**

and Schmid 2008). Finally, each person video is described by the average pooling of all video fragments.

Color. The appearance feature on image level is widely used to express person video. Specifically, at the image level, each frame of the person video is divided into 49 patches with size 16 x 32 with 50% overlap both in the horizontal and vertical directions. For each patch, histograms of color channels in HSV and LAB color spaces are computed. We concatenate all the patch descriptors to form the appearance feature on image level, and the average pooling of image level descriptor over all image frames of the video sequence is taken. Finally, we describe the appearance feature on image level of each video with a 3332-dimensional feature vector.

HOG3D+Color. The combination of HOG3D and color appearance feature, each image in video sequences is used equally.

IDE. The IDE extractor is pre-trained on ResNet-50 (He et al. 2016). It generates a 2,048-dim vector for each image, which is effective in large-scale re-ID datasets.

As we mentioned earlier, we split the whole video sequence into a series of sub sequences based on SSM signals and use a SPOD method to evaluate the noise level of sub sequences. Then, we eliminate sub sequences with heavy occlusions and use the remaining sub sequences to jointly measure the distance between person image sequences. That is to say, each sub sequence is described with the basic feature representations mentioned above respectively. Finally, the distance of two person image sequences is measured with a weighted multi-pair distance metric learning approach, as described in Eq.(8).

With all parameters being the same in each group, Table 1 shows the detailed Rank-1, Rank-5 and Rank-20 matching rates of all the compared methods. We can observe that our method achieves better matching rates in each rank compared to the competing basic approaches. Using XQDA method as an example, in the iLIDS-VID dataset, the Rank-1 matching rate of HOG3D is improved by 6.9%, for Color, the Rank-1 matching rate is improved by 7%, and for HOG3D+Color, the Rank-1 matching rate is improved by 14.1%. The experimental results demonstrates the effectiveness of our method. Compared with the competing approaches, the main advantage of our method is that our method eliminate sub sequences with heavy occlusions in video sequences and use a weighted multi-pair distance metric method, which makes a more robust distance metric between person images and generates a better rank list. Actually, we find that our method improves the performance a lot compared to the original approach no matter which basic descriptor we use.

iLIDS-VID vs. PRID 2011: From Table 1, we can see our method achieves large performance improvement on both datasets. And an interesting can be observed that our method outperformed others much better on the iLIDS-VID dataset, the Rank-1 matching rate on the iLIDS-VID dataset is improved by **14.9%(59.3%-44.4%)**. Compared with the PRID 2011 dataset, this dataset is more challenging due to environment variations especially complicated background and heavy occlusions, as shown in Fig. 1. As we mentioned earlier, it helps a lot to improve the re-id performance by a SPOD approach, which makes our method more effective on such more challenging circumstance like the iLIDS-VID dataset and achieve a great performance improvement.

Table 2: Comparison with the state-of-the-art methods on PRID 2011 and iLIDS-VID datasets. Results are shown as matching rates (%) at Rank = 1, 5, 10, 20. Best results are in boldface font. The literatures in the first block are using traditional methods, while the second block contains deep neural network based methods.

	PRID 2011				iLIDS-VID			
Methods	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
DVR	37.6	63.9	75.3	89.4	34.5	56.7	67.5	77.5
DVDL	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9
STFV3D+KISSME	64.1	87.3	89.9	92.0	44.3	71.7	83.7	91.7
SRID	35.1	59.4	69.8	79.7	24.9	44.5	55.6	66.2
DVSR	48.3	74.9	87.3	94.4	41.3	63.5	72.7	83.1
TDL	56.7	80.0	87.6	93.6	56.3	87.6	95.6	98.2
SI^2DL	76.7	95.6	96.7	98.9	48.7	81.1	89.2	97.3
TDNN	70.0	90.0	95.0	97.0	58.0	84.0	91.0	96.0
Zhou et al.	79.4	94.4	-	99.3	55.2	86.5	-	97.0
ASTPN	77.0	95.0	99.0	99.0	62.0	86.0	94.0	98.0
SPW (OURS)	**83.5**	**96.3**	**98.5**	**100.0**	**69.3**	**89.6**	**95.7**	**98.2**

Comparison with the State-of-the-art

In this section, we report the comparison of our method with several state-of-the-art video-based person re-id approaches, including Discriminative Video Fragments Selection and Ranking (DVR) (Wang et al. 2014) and its journal version DVSR (Wang et al. 2016a), DVDL (Karanam, Li, and Radke 2015a) and the Spatial-Temporal Fisher Vector Representation (STFV3D+KISSME) (Liu et al. 2015), SRID (Karanam, Li, and Radke 2015b), TDL (You et al. 2016), SI^2DL (Zhu et al. 2016). We also compare our method with deep neural network based methods TDNN (McLaughlin et al. 2016), Zhou's work (Zhou et al. 2017) and ASTPN (Xu et al. 2017). DVR is a method based on ranking model, which tries to select aligned video fragment pairs from candidates pool. DVDL is a dictionary learning method based on multi-shot re-id datasets. STFV3D is a spatio-temporal appearance model which exploits temporal information on the action level. SRID (Karanam, Li, and Radke 2015b) formulates the re-identification problem as a block sparse recovery problem. TDNN (McLaughlin et al. 2016) extracts features and matches sequences with a recurrent neural network architecture. Zhou et al. (Zhou et al. 2017) integrates a temporal attention model within an end-to-end deep neural network architecture. ASTPN (Xu et al. 2017) proposed a novel deep architecture with jointly attentive spatial-temporal pooling.

The performance of each method on both datasets is shown in Table 2. The results demonstrate that with our method, the matching rate performance is improved a lot on both datasets, especially on the iLIDS-VID dataset. The reason why our model outperforms the others much better on the iLIDS-VID is mainly because that most of video sequences in the iLIDS-VID dataset are unregulated (66.33%). And this causes a significant loss of performance in the traditional models in which all sub sequences are used equally when occlusion occurs in video sequences. Another fact that needs to be noticed that our method achieves much better performance on both datasets compared with (Wang et al. 2016a). As we mentioned earlier in (Wang et al. 2016a), Wang *et al.* tried to select and match aligned video fragment pairs from candidates pool without eliminating the highly noisy video fragments, resulting in a significant performance degradation. On the contrary, in this paper, we effectively address this problem with an adaptive SPW method, which handles the noise in video sequences very well and achieves a very considerable performance improvement.

Conclusion

In this paper, a simple yet effective method for video-based person re-id is proposed. Through investigating the impact of unregulated sequences on video-based person re-id, we find that significant performance improvement can be made via an self paced weighting (SPW) method. Based on this delightful finding, we propose an adaptive self paced outlier detection method to evaluate the noise level of each sub sequence and a weighted multi-pair distance metric learning approach is adopted to measure the distance of two person image sequences. Extensive experiments illustrate that huge benefits can be obtained via SPW approach and demonstrate the effectiveness of our proposed method.

In the future, there are several ways to extend this work. First, a more effective way can be adopted to construct the feature representation of persons, it is still a feasible topic to find discriminative information within video frames. Moreover, an unsupervised metric learning method can be combined with our method, which can deal with the real-life datasets more effectively.

Acknowledgments

The research was supported by the National Nature Science Foundation of China under Grant U1611461, 61303114 and 61671336, by the National High Technology Research and Development Program of China under Grant 2015AA016306, by the Technology Research Program of Ministry of Public Security under Grant 2016JSYJA12, by the Hubei Province Technological Innovation Major Project under Grant 2016AAA015, 2017AAA123, by the Natural Science Foundation of Hubei Province under Grant

2016CFB573, and by the Nature Science Foundation of Jiangsu Province under Grant BK20160386.

References

Auguste, R.; Martinet, J.; and Tirilly, P. 2015. Space-time histograms and their application to person re-identification in tv shows. In *ICMR*, 91–97.

Ayvaci, A.; Raptis, M.; and Soatto, S. 2010. Occlusion detection and motion estimation with convex optimization. In *NIPS*, 100–108.

Bedagkar-Gala, A., and Shah, S. K. 2014. A survey of approaches and trends in person re-identification. *IVC* 32(4):270–286.

Chen, C. L.; Liu, C.; and Gong, S. 2014. Person re-identification by manifold ranking. In *ICIP*, 3567–3571.

Gao, C.; Wang, J.; Liu, L.; Yu, J. G.; and Sang, N. 2016. Temporally aligned pooling representation for video-based person re-identification. In *ICIP*, 4284–4288.

Han, J., and Bhanu, B. 2006. Individual recognition using gait energy image. *TPAMI* 28(2):316–322.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

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

Jiang, L.; Meng, D.; Yu, S. I.; Lan, Z.; Shan, S.; and Hauptmann, A. 2014. Self-paced learning with diversity. In *NIPS*, 2078–2086.

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

Karanam, S.; Li, Y.; and Radke, R. J. 2015b. Sparse re-id: Block sparsity for person re-identification. In *CVPR Workshop*, 33–40.

Klaser, A.; Marszałek, M.; and Schmid, C. 2008. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 275–1.

Kumar, M. P.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. In *NIPS*, 1189–1197.

Liang, C.; Huang, B.; Hu, R.; Zhang, C.; Jing, X.; and Xiao, J. 2015. A unsupervised person re-identification method using model based representation and ranking. In *ACM MM*, 771–774.

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

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

Liu, Z.; Chen, J.; and Wang, Y. 2016. A fast adaptive spatio-temporal 3d feature for video-based person re-identification. In *ICIP*, 4294–4298.

Loy, C. C.; Xiang, T.; and Gong, S. 2009. Multi-camera activity correlation analysis. In *CVPR*, 1988–1995.

Lu, J.; Meng, D.; Zhao, Q.; Shan, S.; and Hauptmann, A. G. 2015. Self-paced curriculum learning. In *AAAI*, 2694–2700.

Matsukawa, T.; Okabe, T.; Suzuki, E.; and Sato, Y. 2016. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 1363–1372.

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

Wang, X.; Doretto, G.; Sebastian, T.; Rittscher, J.; and Tu, P. 2007. Shape and appearance context modeling. In *ICCV*, 1–8.

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

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2016a. Person re-identification by discriminative selection in video ranking. *TPAMI* 38(12):2501–2514.

Wang, Z.; Hu, R.; Jiang, J.; Jiang, J.; Liang, C.; and Wang, J. 2016b. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *IJCAI*, 2669–2675.

Wang, Z.; Hu, R.; Liang, C.; Yu, Y.; Jiang, J.; Ye, M.; Chen, J.; and Leng, Q. 2016c. Zero-shot person re-identification via cross-view consistency. *TMM* 18(2):260–272.

Wang, Z.; Hu, R.; Chen, C.; Yu, Y.; Jiang, J.; Liang, C.; and Satoh, S. 2017. Person reidentification via discrepancy matrix and matrix metric. *IEEE Transactions on Cybernetics* PP(99):1–15.

Xu, S.; Cheng, Y.; Gu, K.; Yang, Y.; Chang, S.; and Zhou, P. 2017. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*.

Ye, M.; Liang, C.; Yu, Y.; and et al. 2016. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. In *TMM*. IEEE.

Ye, M.; Ma, A. J.; Zheng, L.; Li, J.; and Yuen, P. C. 2017. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*.

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

Yuille, A. L., and Rangarajan, A. 2002. The concave-convex procedure (cccp). In *NIPS*, 1033–1040.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 868–884. Springer.

Zhou, D.; Bousquet, O.; Lal, T. N.; and Weston, J. 2003. Learning with local and global consistency. In *NIPS*, 321–328.

Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; and Tan, T. 2017. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*.

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