


Video-Based Person Re-Identification With Unregulated Sequences

Wenjun Huang, National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China

Chao Liang, National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China

 <https://orcid.org/0000-0002-8287-8655>

Chunxia Xiao, School of Computer Science, Wuhan University, Wuhan, China

Zhen Han, School of Computer Science, Wuhan University, Wuhan, China

ABSTRACT

Video-based person re-identification (re-id) has recently attracted widespread attentions because extra space-time information and more appearance cues in videos can be used to improve the performance of image-based person re-id. Most existing approaches equally treat person video images, ignoring their individual discrepancy. However, in real scenarios, captured images are usually contaminated by various noises, especially occlusions, resulting in a series of unregulated sequences. Through investigating the impact of unregulated sequences to feature representation of video-based person re-id, the authors find a remarkable promotion by eliminating noisy sub sequences. Based on this interesting finding, an adaptive unregulated sub sequence detection and refinement method is proposed to purify original video sequence and obtain a more effective and discriminative feature representation for video-based person re-id. Experimental results on two public datasets demonstrate that the proposed method outperforms the state-of-the-art work.

KEYWORDS

Adaptive Weighting, Noise Detection, Recognition, Sequence Stability Measure, Sparse Construction

INTRODUCTION

Person re-identification, which aims at identifying a person of interest among different cameras, has become increasingly popular in the community due to its critical role in many surveillance, security and multimedia applications. Currently, major efforts towards this problem focus on the still-image-based scenario, in which each person has only one image available per camera view. Many methods have been developed to either extract discriminative features (Liao et al., 2015; Matsukawa et al., 2013; Satta et al., 2013; Shen et al., 2013) or learn effective distance metric (Hirzer et al., 2012; Köstinger et al., 2012; Liang et al., 2015; Liao et al., 2015; Yang et al., 2016) for this problem.

In spite of great research progress achieved for the still-image-based task, the real-world re-id performance is hindered by limited information extracted from a single image. Such still-image-based person re-id ignores the temporal information among person images, which leads poor feature representation of person. In practical surveillance systems, persons are always recorded by videos, which means that there are multiple consecutive frames available for an individual in each camera's view field. Thus, it is intuitive to use such sequential images to improve re-id performance, which directly motivates the investigation of video-based person re-id.

DOI: 10.4018/IJDCF.2020040104

This article, originally published under IGI Global's copyright on April 1, 2020 will proceed with publication as an Open Access article starting on January 27, 2021 in the gold Open Access journal, International Journal of Digital Crime and Forensics (converted to gold Open Access January 1, 2021), and will be distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Recently, impressive research progress has been reported in video-based person re-id. However, most approaches mentioned above generally assume that all images in each sequence are of equal importance, losing sight of their difference caused by the interference of various noises. Take iLIDS-VID dataset (Wang et al. 2014) in Figure 1 as an example, person images are always flooded with various noises, such as object occlusions or background clutters, resulting in highly noisy unregulated sequences. In our preliminary comparative experiment conducted on 199-pair unregulated person sequences of iLIDS-VID dataset, average matching accuracy on original unregulated video sequences is only 7%, ten percent lower than that obtained on filtered clean video sequences.

Occlusion is a very common case in video-based applications, taking dataset ETHZ (Schwartz et al. 2010), iLIDS-VID (Wang et al. 2014) and OTB (Wu et al., 2013) as examples, we count the number of video sequences that have heavy occlusions in these public video datasets, as shown in Table 1. From Table 1, we can see that: (1) Occlusion is universal in real world scenarios. (2) In the case of the occlusions, ‘long-term occlusion’ (e.g., the target person is occluded by another person in the whole sequence, which makes there are no clean sub sequences of target person) accounts for only a few parts. In this paper, we focus on the ‘temporary occlusion’ (e.g., another person passes by the target person temporarily or the target person is occluded by surrounding objects temporarily), which means that in most cases, there are still some clean sub sequences even if part of the sequence is occluded.

In this field, Wang et al. (Wang et al., 2014; Wang et al., 2016) first noticed the quality discrepancy problem of different person sequences, and an optical flow-based algorithm was raised to detect walking cycles to divide a video sequence into different sub fragments. Then, a ranking model was proposed to select and match video fragment pairs. However, (1) it is hard to obtain a reliable optical flow estimation without considering the occlusion (Ayvaci et al., 2010). Thus, the algorithm will eventually generate noisy sub sequences when occlusion occurs. (2) It uniformly exploits all video

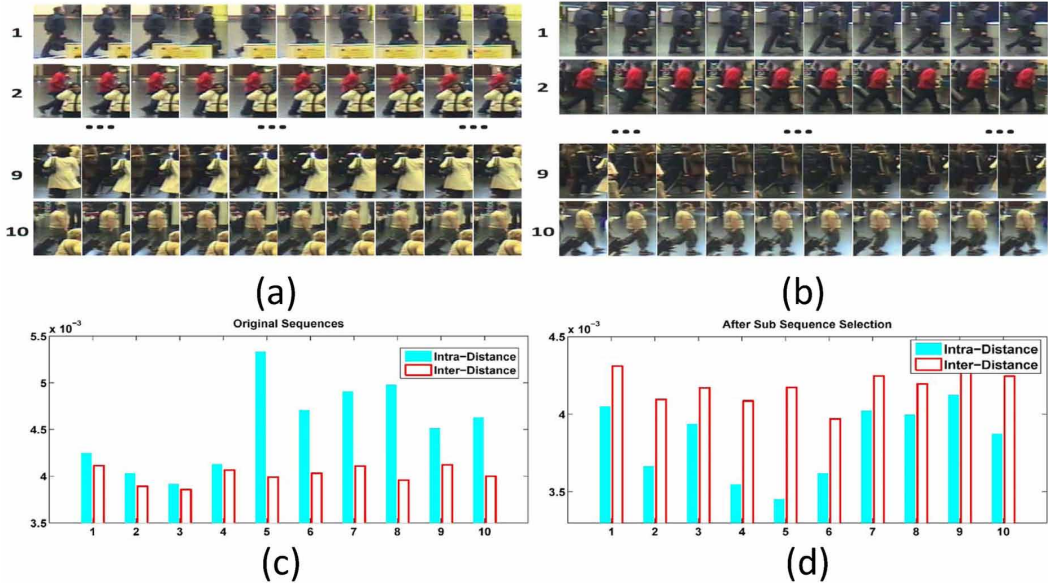
Figure 1. An example illustrating that person images are always highly noisy in practical situations



Table 1. Occlusion condition investigation

Dataset	Occlusion	Long-Term Occlusion	Temporary Occlusion
ETHZ	45.21%	15.15%	84.85%
OTB	58.05%	10.03%	89.97%
iLIDS-VID	66.33%	10.05%	89.95%

Figure 2. A preliminary comparative experiment illustrates the negative impact of unregulated sequences. Video sequences containing: (a) Heavy occlusion sub sequences; and (b) Little occlusions sub sequences in iLIDS-VID dataset. Same rows in both (a) and (b) corresponds to the same person. (c) The intra distance (blue bar) and inter distance (red bar) of 10 noisy video sub sequences in (a). Their intra distances are greater than inter distances on all 10 video sub-sequences with heavy occlusion. (d) After eliminating the sub sequences with heavy occlusions, the inter distance is greater than intra distance.



sub sequences to identify persons, while ignoring their quality discrepancy caused by object motions and occlusions, which eventually incurs significant performance degradation.

To investigate the impact of unregulated sequences to feature representation of video-based person re-id, we further conduct a preliminary comparative experiment. Ten video sequence pairs of ten different persons from two camera views are randomly selected from 199 person sequence pairs that have heavy occlusions in the iLIDS-VID dataset, as shown in Figure 2(a). In this condition, we observe that intra distance (the distance of a pair of video sequences of same person captured by two cameras) is greater than inter distance (the minimum distance between a video sequence in probe and video sequences of other persons in gallery) as shown in Figure 2(c). This validates negative impacts of directly applying unregulated sequences to video-based person re-id. Furthermore, for comparison, we manually eliminate sub sequences with heavy occlusions from these ten video sequences and preserve clean sub sequences, as shown in Figure 2(b), resulting in a nearly perfect result, where inter distance is larger than intra distance, as shown in Figure 2(d). This interesting finding reveals that eliminating occluded sub sequences is an effective approach to improve the performance of video-based person re-id.

Motivated by the above findings, we propose a novel video-based person re-id method by purifying the unregulated person image sequence. In our work, two key issues need to be addressed. (1) The local consistency of video sequences should be well utilized to decompose a video sequence into a series of sub sequences. (2) The impact of noisy sub sequences on feature representation for each video sequence should be well handled.

For the first issue, in order to measure the state of each frame within video sequences, we define a Sequence Stability Measure (SSM), reflecting the state change of a video sequence. Then, the local maximum of SSM signals, named stationary point, corresponds to a local stable state of a person. Thus, frames around a stationary point may have unified state, which are regarded as a sub sequence. For the second issue, motivated by the effectiveness of sparse representation in handling occlusions

(Lu et al., 2012; Wright et al., 2009), it is reasonable to assume that if a sub sequence is heavily occluded, it is hard to represent the images within this sub sequence by other sub sequences in the video sequence, which causes larger reconstruction error. On this basis, we propose an unregulated sub sequences detection method based on sparse representation to detect occluded sub sequences in video sequences. At last, we eliminate sub sequences with heavy occlusions and use remaining sub sequences to obtain a more discriminative feature representation of person.

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

- We find an interesting phenomenon where remarkable performance can be achieved by eliminating noisy sub sequences from their clean counterparts in the feature representation of person object;
- We propose an adaptive sub sequences selection method based on sparse representation to obtain a more effective and discriminative feature representation of persons;
- Experimental results on two public datasets demonstrate that the proposed method outperforms the state-of-the-art work.

RELATED WORK

Space-Time Features

Space-time feature representations have been extensively explored in action/activity recognition (Poppe et al., 2010; Wang et al., 2009). In action recognition, an image sequence is viewed as a 3-dim space-time volume. One common representation is constructed based on space-time interest points (Laptev et al., 2003; Willems et al., 2008; Bregonzio et al., 2009). These approaches facilitate a compact description of image sequences using the sparse interest points, which are sensitive to variations such as viewpoint, speed, scale and may lose discriminative information (Ke et al. 2010, Gilbert et al. 2009). A recent trend is to incorporate temporal cues in the model, space-time volume/patch based representations can be richer and more robust. Popular descriptors include HOG3D (Klaser et al., 2008), 3D-SIFT (Scovanner et al., 2007), HoGHoF (Laptev et al., 2008), etc. which can be viewed as extensions of their corresponding 2-dim versions. In 2015, Liu et al. (Liu et al., 2015) proposed a spatio-temporal body-action model, in which fisher vectors were learned from individual body-action units and concatenated into the final representation of the walking person.

Multi-Shot Person Re-Identification

Multiple images from a sequence of the same person have been exploited for person re-identification. In (Hirzer et al., 2012), multiple shots are used to train a discriminative boosting model based on a set of covariance features. In (Hamdoun et al., 2008), the SURF local feature is used to detect and describe interest points within short video sequences that are in turn indexed in the KD-tree to speed up matching. In (Gheissari et al., 2006), a spatial-temporal graph is generated to identify spatial-temporal stable regions for foreground segmentation. The local descriptions are then calculated using a clustering method over time to improve matching performance. Truong et al. (Truong et al., 2009) employ the manifold geometric structures from video sequences to construct more compact spatial descriptors with color-based features. Karanam et al. (Karanam et al. 2015) make use of multi-shots for a person and propose that the probe feature be presented as a linear combination of the same person in the gallery. Karaman et al. (Karanam et al., 2012) propose using the conditional random field (CRF) to incorporate constraints in the spatial and temporal domains.

However, all these methods mentioned above are based on a key assumption that all images of person video sequence are equally treated, losing sight of negative influences of noises. However, in real scenarios, captured person images are usually contaminated by various noises, especially occlusions, resulting in a series of unregulated sequences. In this study, we investigate the impact

of unregulated sequences on feature representation of video-based person re-id and show that a remarkable promotion can be obtained by eliminating noisy sub sequences.

Proposed Approaches

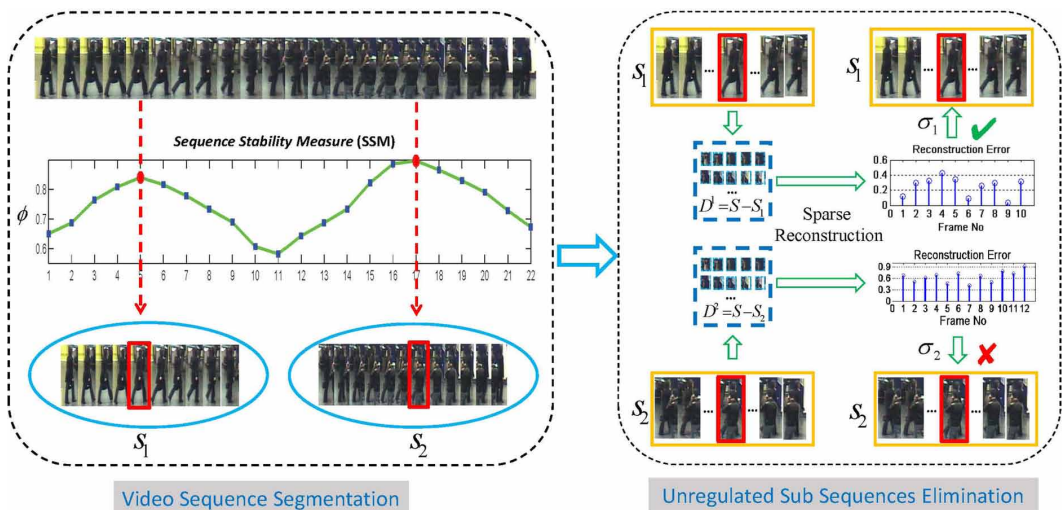
The Figure 3 illustrates the framework of the proposed sequence segmentation and elimination framework. The framework consists of two parts, the video sequence segmentation and the unregulated sub sequences elimination. In the video sequence segmentation part, a video sequence is divided into a series of sub sequences by detecting stationary points (red dots in the Figure 3, each of which corresponds to a local stable state of a person) of SSM signals. In the unregulated sub sequences elimination part, we use the sparse representation based method to eliminate the occluded sub sequences, the larger the reconstruction error is, the more likely this sub sequence will be occluded, e.g., we eliminate the sub sequence s_2 , which has larger construction error. Finally, sub sequences without occlusions are used to construct the appearance model of a person.

Video Sequence Segmentation

Given an unregulated video sequence with heavy noises, as shown in Figure 1, it is too noisy to extract discriminative features from entire video sequence. In (Wang et al., 2014; Wang et al., 2016), Wang et al. 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, we attempt to divide a video sequence into a series of fragments by using occlusion information between consecutive frames, which can better reflect the state change between consecutive frames in complex scenarios compared with optical flow.

Occlusion detection plays an important role in priming visual recognition of detached objects, and recently it is often used together with optical flow. In (Ayvaci et al., 2010), Ayvaci et al. formulated occlusion detection and optical flow estimation between two images as a joint optimization problem. Under assumptions of Lambertian reflection and static illumination, the task can be posed as a variational optimization problem, and its solution approximated using convex minimization. By assembling a function that penalizes the (unknown) optical flow residual in the (unknown) co-visible regions, as well as the area of the occluded region, the resulting optimization problem can be solved jointly with respect to the unknown optical flow field and the indicator function of the occluded

Figure 3. The framework of the proposed sequence segmentation and elimination method



region. We use the occlusion information o_i in (Ayvaci et al., 2010) between consecutive frames to measure the state change within each video sequence.

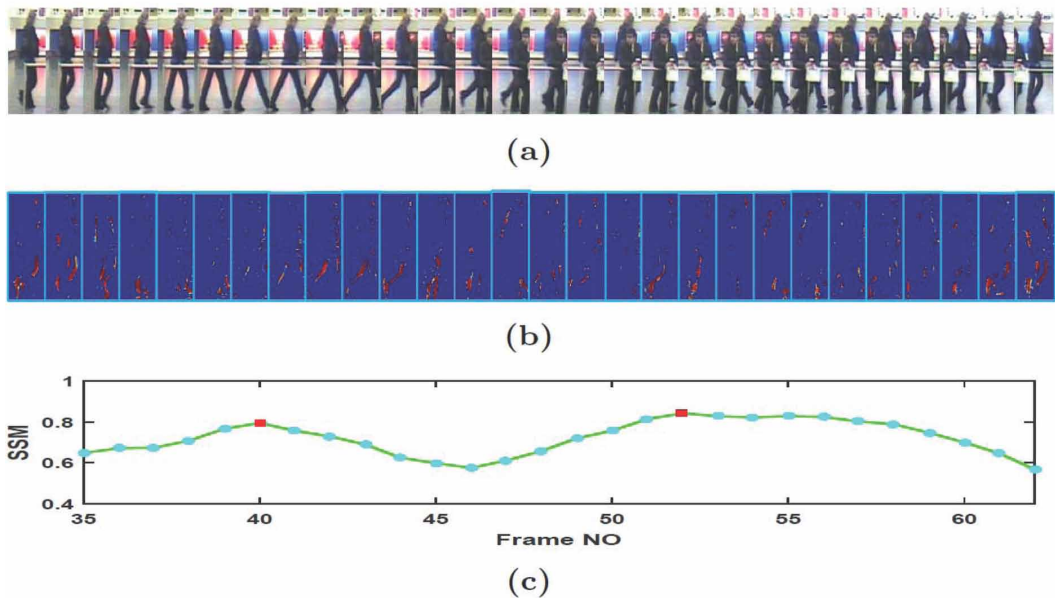
In fact, the occlusion information o_i reflects the inter-frame occlusion, which indicates the occlusion condition of current frame with respect to its previous frame, as shown in Figure 4. Thus, from a certain point of view, o_i expresses some kind of state change between two frames. For example, the Figure 4(b) shows the state change between consecutive frames of the sub sequence in Figure 4(a). On this basis, a measure of stability ϕ_i is defined, which indicates the state difference between each frame and its previous frame. Given a video sequence $M = (I_1, I_2, \dots, I_N)$, we use I_i to denote the i^{th} frame for simplicity, let o_i denote the occlusion information between frame I_i and I_{i-1} as mentioned in (Ayvaci et al., 2010), then, the stability of each frame is defined as follows:

$$\phi_i = \exp\left(-\frac{\|o_i\|^2}{c}\right), i = 2, 3, \dots, N \quad (1)$$

where c is a constant factor. Then, the Sequence Stability Measure (SSM) ε of M is defined as $\varepsilon = (\phi_2, \dots, \phi_N)$, which is further smoothed by a Gaussian filter.

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 noisy or all clean). And we name those maxima landmarks of ε stationary points (red dots in Figure 4(c)). Finally, the video sequence M is split into a set of sub sequences $S = \{s_i\}, i = 1, \dots, m$ by detecting stationary point $\{t\}$ of ε and extracting the surrounding frames $s_t = (I_{t-L}, \dots, I_t, \dots, I_{t+L})$ of each stationary point as a sub sequence. The temporal range L is adaptively determined by the range between the local maximum and local minimum.

Figure 4. The framework of the proposed sequence segmentation and elimination method



Adaptive Sub Sequences Selection

As mentioned before, by detecting the stationary point of ε , a video sequence is decomposed into a series of sub sequences $S = \{s_i\}, i = 1, \dots, m$. Then, the next step is to distinguish the sub sequences with heavy noises from those which are relatively clean. Motivated by the widely successful applications of sparse representation in many tasks and its effectiveness in handling occlusions (Lu et al. 2012, Wright et al. 2009), in this paper, the sparse representation is exploited to measure the noise degree of video sub sequences.

As mentioned in Table 1, ‘temporary occlusion’ occupies the majority in video sequences, which means that in most cases, there are still some clean sub sequences in video sequences. So, it is reasonable for us to assume that the clean sub sequences occupy the majority of the whole video sequence. Under the circumstances, we can use the reconstruction error of each frame within the sub sequence to evaluate the occlusion degree of each sub sequence. In order to get a more reliable evaluation of occlusion degree, we exclude the frames of the same sub sequence when we evaluate each frame. This is because if a frame is occluded, its surrounding frames in the same sub sequence would be probably occluded, so these frames can’t be used to evaluate the occlusion condition of current frame.

Given a sub sequence $s_i = (I_1, I_2, \dots, I_N)$, we sample K local image patches inside each image I of s_i with a spatial layout. After sampling local patches, they are used to form corresponding dictionaries for encoding local patches. In order to get a more reliable evaluation of occlusion degree, we exclude the frames of the same sub sequence when we evaluate each frame. Specifically, for the k^{th} patch of I , the corresponding dictionary $D^k = \{s_j^k\}$ can be obtained, where $j = 1, \dots, m, j \neq i$ and s_j^k denotes the k^{th} patch of images in s_j .

For any patch P of image I , the sparsity-based representation model is denoted as:

$$\min_{\beta} \|P - D\beta\|_2^2 + \lambda \|\beta\|_1 \quad (2)$$

where β represents the sparse coefficient vector of P and can be computed by the optimization function of the sparsity-based representation model and λ is regularization factor. Then the reconstruction error of P is defined as: $\|P - D\beta\|_2^2$. Thus, for a sub sequence s_i , the reconstruction error e_i of s_i can be defined as:

$$e_i = \frac{1}{K} \sum_{k=1}^K \|Z^k - D^k X^k\|_2^2 \quad (3)$$

Here, $Z^k \in R^{d \times t}$, each column of Z^k denotes the k^{th} patch of images in sub sequence s_i , t is the number of images in s_i . X^k is the sparse coefficient matrix and D^k is the dictionary of k^{th} patch in s_i , which is obtained as described before.

Based on the reconstruction error, the noise degree of sub sequence s_i , which indicates the usability of s_i , is defined as:

$$\sigma_i = \frac{\exp(e_i^2)}{\sum_{i=1}^m \exp(e_i^2)} \quad (4)$$

- **Elimination:** From the above equation, it can be observed that the noise degree σ_i becomes larger when the reconstruction error e_i of s_i is getting larger. Thus, the larger σ_i is, the more likely s_i possesses noises, such as occlusions or background clusters, which indicates that the usability of this sub sequence is relatively low, as shown in Figure 4. Then, based on σ_i Q defined before, we eliminate the sub sequences from S , whose σ values are greater than the threshold θ ;
- **Adaptive weighting:** After eliminating the sub sequences that are heavily noisy, a sub sequences pool $Q = \{s_1, \dots, s_N\}$ is obtained, in which sub sequences are relatively clean. To obtain more abundant and robust feature for representing a person, all sub sequences in are used to construct the appearance model for person. Let f_i denote the feature representation of sub sequence s_i , then the feature representation of Q is defined as:

$$F_Q = \{f_1, \dots, f_N\} \quad (5)$$

Finally, the feature representation of video M is defined as:

$$F_M = \sum_{i=1}^N \omega_i f_i \quad (6)$$

where ω_i denotes the weight of s_i , which is defined as:

$$\omega_i = \frac{1 + \exp(-\sigma_i)}{\omega^*} \quad (7)$$

where ω^* is a normalization factor.

DISCUSSION

In our framework, both two key components: the video sequence segmentation and the unregulated sub sequences elimination are based on an important assumption that in most cases, there is still some clean sub sequences in the whole sequences even if the sequence is noisy. It will degrade the performance of our method on both two components if the target person is occluded in the whole sequence. In the video sequence segmentation stage, it will make the segmentation of video sequence useless when the whole sequence is noisy. Because the most important purpose of the video sequence segmentation is to split a video sequence into a series of sub sequences so that noisy images are separated from clean images. In the unregulated sub sequences elimination, when the whole sequence is noisy, the sparse-based method couldn't effectively evaluate the noisy degree of each sub sequence.

Thus, in order to investigate the rationality of this hypothesis, we investigate several public video datasets, such as ETHZ (Schwartz et al. 2010), iLIDS-VID (Wang et al. 2014) and OTB (Wu et al. 2013), as shown in Table 1. As we observed in Table 1, in real world scenarios, 'long-term occlusion' accounts for only a few parts, which makes our method very effective in general.

Besides, the experimental results further demonstrate the effectiveness of our method. Specifically, in Figure 7, we conduct an experiment on 199 occluded video sequence pairs in iLIDS-VID dataset, which demonstrates the effectiveness of our method.

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. Following (Wang et al. 2014), in our experiments, sequences with more than 21 frames from 178 persons are used. So, the probe and gallery both have 89 identities. The dataset has two adjacent camera views captured in uncrowded outdoor scenes with rare occlusions and clean background.

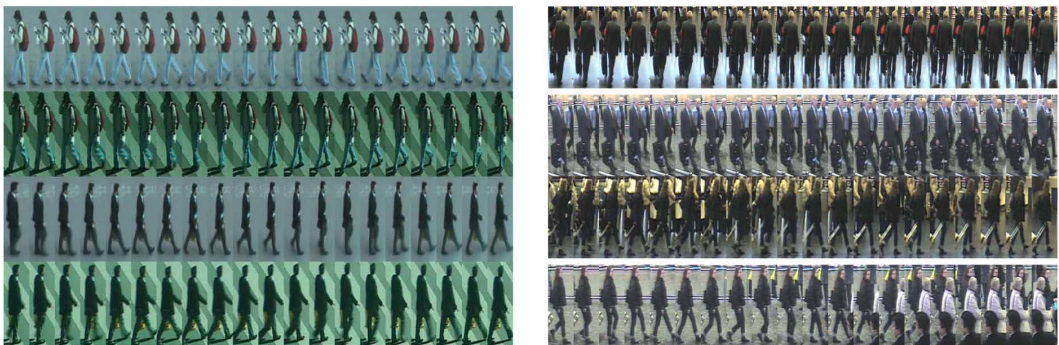
iLIDS-VID Dataset

The iLIDS-VID dataset includes 600 video sequences for 300 randomly sampled 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 Figure 5.

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. The number K of local image patches inside each image is set to 6 and the threshold θ is set to the mean value of reconstruction errors of all sub

Figure 5. The example sequences of the two publicly available video datasets



sequences. We use the cumulative matching characteristic (CMC) curve (Wang et al. 2007) to measure the performance on both datasets, and we use the average CMC curves of 10 trails to obtain a more reliable result.

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 combine each of them with two different distance metric learning method XQDA (Liao et al., 2015) and TDL (You et al., 2016) to make a more comprehensive comparison.

HOG3D

The 3D HOG features from volumes of video data is extracted in a way 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 2×5 (*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 et al., 2008). In the end, 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 16x32 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.

Figure 6. The CMC curves of average matching rates with different feature representations combined with metric learning XQDA and TDL on the PRID 2011 and iLIDS-VID datasets

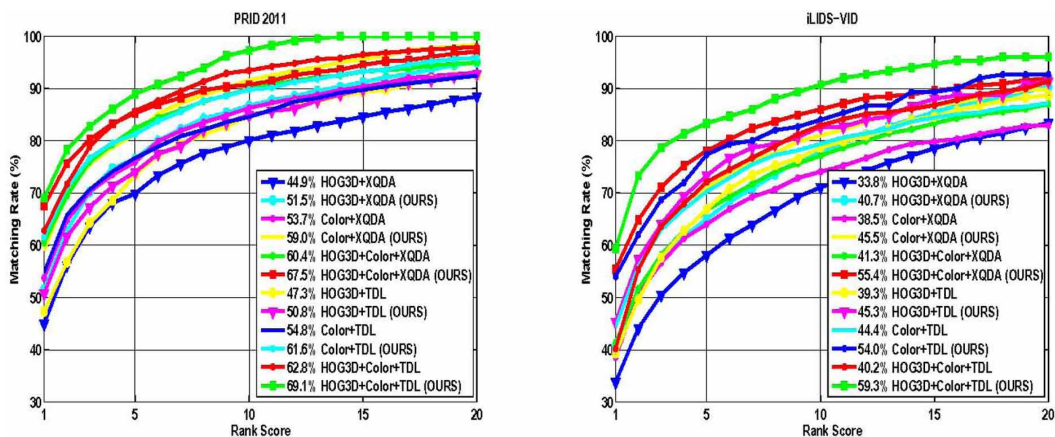


Table 2. Comparison with different feature representations with metric learning XQDA and TDL on PRID 2011 dataset. For each combination, we compare the original approach with our approach. Results are shown as matching rates (%) at Rank = 1, 5, 10, 20. The improvement result of each rank is in boldface font.

Methods	PRID 2011			
	Rank-1	Rank-5	Rank-10	Rank-20
HOG3D+XQDA	44.9	69.9	80.1	88.4
HOG3D+XQDA(OURS)	51.5 (↑ 6.6)	75.2 (↑ 5.3)	85.8 (↑ 5.7)	93.6 (↑ 5.2)
Color+XQDA	53.7	76.0	86.3	93.0
Color+XQDA(OURS)	59.0 (↑ 5.3)	80.8 (↑ 4.8)	90.9 (↑ 4.6)	98.5 (↑ 5.5)
HOG3D+Color+XQDA	60.4	82.4	89.9	94.9
HOG3D+Color+XQDA(OURS)	67.5 (↑ 7.1)	85.3 (↑ 2.9)	90.7 (↑ 0.8)	97.1 (↑ 2.2)
HOG3D+TDL	47.3	73.5	84.7	92.5
HOG3D+TDL(OURS)	50.7 (↑ 3.4)	74.0 (↑ 0.5)	84.9 (↑ 0.2)	92.8 (↑ 0.3)
Color+TDL	54.8	76.7	84.6	92.4
Color+TDL(OURS)	61.6 (↑ 5.8)	81.2 (↑ 4.5)	89.8 (↑ 5.2)	95.9 (↑ 3.5)
HOG3D+Color+TDL	62.8	85.7	93.4	97.9
HOG3D+Color+TDL(OURS)	69.1 (↑ 6.3)	90.0 (↑ 4.3)	97.3 (↑ 3.9)	100.0 (↑ 2.1)

HOG3D + Color Histograms

The combination of HOG3D and color appearance feature, all the images of each sequence are used equally. Specifically, for each video sequence, we exploited the combination of HOG3D (space-time feature) and the average pooling of color histograms (appearance feature). Then, each video sequence was represented by a 5892-dimensional feature vector.

As we mentioned earlier, we split the whole video sequence into a series of sub sequences based on SSM signals and use a sparse representation-based selection method to eliminate the sub sequences with heavy occlusions. So, we eliminate sub sequences with heavy occlusions and use the remaining sub sequences to make a feature representation of person. That is to say, each remain sub sequence is described with the basic feature representations mentioned above respectively. Finally, each person video is represented with a feature vector by using weighted summation of remaining sub sequences as described in Equation (6).

With all parameters being the same in each representation, Figure 6 shows the CMC curves on both datasets of the compared features, and Table 2 and Table 3 show the detailed Rank-1, Rank-5, Rank-10, and Rank-20 matching rates of all the compared features. We can observe that our method achieves better matching rates in each rank compared to the competing basic descriptor. 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 demonstrate the effectiveness of our method. Compared with the competing feature representations, the main advantage of our method is that our method eliminates sub sequences with heavy occlusions in video sequences, which makes

a more robust feature representation of person and generates a better rank list. Actually, we find that our method improves the performance a lot compared to the original feature representation no matter which basic descriptor we use.

iLIDS-VID vs. PRID 2011

From Table 2 and Table 3, we can see our method achieves large performance improvement on both datasets. And an interesting but indeed fact 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 19.1% (59.3%-40.2%) (When we use HOG3D+Color as the basic feature representation and use TDL as the metric method). Compared with the PRID 2011 dataset, this dataset is more challenging due to environment variations especially complicated background and heavy occlusions, as shown in Figure 1. As we mentioned before, it helps a lot to improve the re-id performance by a sub sequences selection approach, which makes our method more effective on such more challenging circumstance like the iLIDS-VID dataset and achieve a great performance improvement.

Integration With Deep Feature

To make a more comprehensive evaluation, we integrate our method with deep feature, which we use pre-trained ResNet-50 as an implicit feature extractor. To prove the independence of our method and specific metrics, we integrate our method with metric XQDA and TDL which are widely used in video-based person re-identification. Table 4 and Table 5 show the detailed Rank-1, Rank-5, Rank-10, and Rank-20 matching rates with feature IDE. We can observe that our method achieves better

Table 3. Comparison with different feature representations with metric learning XQDA and TDL on iLIDS-VID dataset. For each combination, we compare the original approach with our approach. Results are shown as matching rates (%) at Rank = 1, 5, 10, 20. The improvement result of each rank is in boldface font.

Methods	iLIDS-VID			
	Rank-1	Rank-5	Rank-10	Rank-20
HOG3D+XQDA	33.8	58.1	71.1	83.4
HOG3D+XQDA(OURS)	40.7 (↑ 6.9)	62.8 (↑ 4.7)	76.7 (↑ 5.6)	88.9 (↑ 5.5)
Color+XQDA	38.5	64.0	74.1	82.9
Color+XQDA(OURS)	45.5 (↑ 7.0)	70.3 (↑ 6.3)	81.7 (↑ 7.6)	88.9 (↑ 6.0)
HOG3D+Color+XQDA	41.3	66.7	77.1	87.0
HOG3D+Color+XQDA(OURS)	55.4 (↑ 14.1)	78.1 (↑ 11.4)	86.0 (↑ 8.9)	91.9 (↑ 4.9)
HOG3D+TDL	39.3	66.9	78.7	88.7
HOG3D+TDL(OURS)	45.3 (↑ 6.0)	73.3 (↑ 6.4)	82.7 (↑ 4.0)	91.3 (↑ 2.6)
Color+TDL	44.4	70.3	79.4	87.2
Color+TDL(OURS)	54.0 (↑ 9.6)	77.3 (↑ 7.0)	84.0 (↑ 4.6)	92.7 (↑ 5.6)
HOG3D+Color+TDL	40.2	72.1	82.9	91.3
HOG3D+Color+TDL(OURS)	59.3 (↑ 19.1)	83.3 (↑ 11.2)	90.7 (↑ 7.8)	96.0 (↑ 4.7)

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 is improved by 4.9%.

From the Table 4 and Table 5, we can see that our method still outperforms the competing approaches when use deep model as feature extractor. This further demonstrates the effectiveness of our method. Furthermore, use pre-trained deep model as feature extractor, our method has lower algorithm complexity than other more complex CNN-based approaches.

Further Evaluation

To further evaluate the effectiveness of our method for unregulated sequences, we select all 199 video pairs of 199 different persons captured by two cameras in iLIDS-VID dataset which have heavy occlusions to make a more comprehensive experiment. We compare the performance of our method that eliminates the occluded sub sequences with the approach where all sub sequences are equally used. The same feature representation for each sub sequence is used for comparison as mentioned above (HOG3D+Color (pooling)), and the Euclidean distance is used to calculate the distance between probe and gallery sequences. The CMC curves on the 199 sequences of iLIDS-VID dataset are shown in Figure 7. It can be observed that with our method, all rank matching rates on 199 occluded video sequences are greatly improved, which implies that it is necessary and effective for improving the re-id performance to eliminate sub sequences with heavy occlusions in video sequences. And also, it reveals a fact that traditional models in which all the images of each sequence are treated equally may result in a significant loss of performance when occlusion occurs in video sequences.

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 DVSR (Wang et al., 2016), DVDL (Karanam et al., 2015), Spatial-Temporal Fisher Vector Representation (STFV3D+KISSME) (Liu et al., 2015) and the Deep Feature Guided

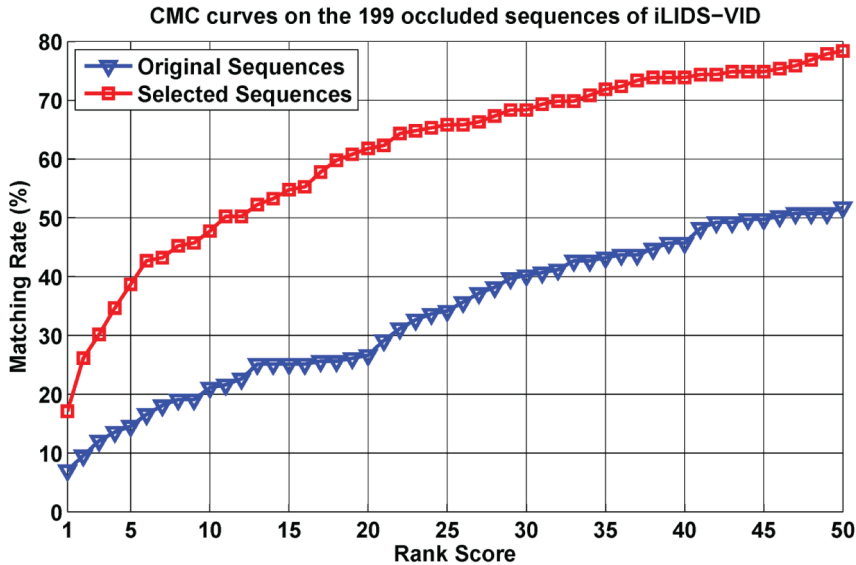
Table 4. Comparison with feature extracted from pre-trained Resnet-50 (He et al. 2016) with metric learning XQDA and TDL on PRID 2011 dataset. For each combination, we compare the original approach with our approach.

Method	Rank-1	Rank-5	Rank-10	Rank-20
Resnet-50 +XQDA	78.8	94.4	96.0	97.9
Resnet-50 +XQDA (OURS)	81.2	95.6	96.6	98.1
Resnet-50 +TDL	79.1	94.7	97.3	98.7
Resnet-50 +TDL (OURS)	82.3	96.1	97.9	99.3

Table 5. Comparison with feature extracted from pre-trained Resnet-50 (He et al., 2016) with metric learning XQDA and TDL on iLIDS-VID dataset. For each combination, we compare the original approach with our approach. "IDE" represents the feature extracted from pre-trained Resnet-50 (He et al. 2016).

Method	Rank-1	Rank-5	Rank-10	Rank-20
Resnet-50 +XQDA	52.2	73.0	80.7	87.9
Resnet-50 +XQDA (OURS)	57.1	77.4	83.1	90.4
Resnet-50 +TDL	61.8	84.7	90.6	92.5
Resnet-50 +TDL (OURS)	65.1	88.6	93.5	96.6

Figure 7. The CMC curves on the 199 sequence pairs with heavy occlusions from the iLIDS-VID dataset. Euclidean distance is used to calculate the distance between probe and gallery.



Pooling (DFGP) (Li 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 spatial-temporal appearance model which exploits temporal information on the action level. We also compare our method with SRID (Karanam et al., 2015), which formulates the re-identification problem as a block sparse recovery problem. To achieve the better performance, we combine our method with the distance metric learning method TDL (You et al., 2016).

The performance of each method on both datasets is shown in Table 6. The results demonstrate that with our method, the matching rate performance is improved a lot on both datasets, especially on the iLIDS-VID dataset. For instance, our method improves the Rank-1 matching rate by 15% (59.3%-44.3%) compared to the second best method STFV3D+KISSME on the iLIDS-VID dataset (To make a fair comparison, this result is computed according to the comparison with approaches with handcraft feature). 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 need to be noticed that our method achieves much better performance on both datasets compared with (Wang et al., 2016). As we mentioned earlier, in (Wang et al. 2016), 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 sub sequences selection method, which handles the occlusions in video sequences very well and achieves a very considerable performance improvement.

CONCLUSION

In this paper, we propose an effective method for video-based person re-identification. Through defining a sequence stability measure (SSM), we split each video sequence into a series of sub sequences, in which images in the same sub sequence have unified state. Besides, we propose

Table 6. 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.

Method	PRID 2011				iLIDS-VID			
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
DVR	28.9	55.3	65.5	82.8	23.3	42.4	55.3	68.4
DVDL	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9
SIFV3D	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
DFGP	51.6	83.1	91	95.5	34.5	63.3	74.5	84.4
OURS	69.1	90.9	97.3	100.0	59.3	83.3	90.7	96.0

a sparse-based sub sequences elimination method to obtain a more effective and discriminative feature representation of persons. Extensive experiments illustrate that huge benefits are obtained by eliminating the occluded sub sequences in video sequences and demonstrate the effectiveness of our proposed approach.

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. Since each person has multiple consecutive frames, 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.

ACKNOWLEDGEMENT

The research was supported by the National Nature Science Foundation of China under Grant U1903214 and 61876135, Nature Science Foundation of Hubei Province under Grant 2019CFB472, and Hubei Province Technological Innovation Major Project under Grant 2018AAA062 and 2018CFA024. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

REFERENCES

- Ayvaci, A., Raptis, M., & Soatto, S. (2010). Occlusion detection and motion estimation with convex optimization. *The British Journal of Medical Psychology*, 51(4), 335–342.
- Bregonzio, M., Gong, S., & Xiang, T. (2009). Recognising action as clouds of space-time interest points. *Computer Vision and Pattern Recognition*, 1, 1948–1955.
- Gheissari, N., Sebastian, T. B., & Hartley, R. (2006). Person Reidentification Using Spatiotemporal Appearance. In *Proceedings of the 2006 IEEE Computer Vision and Pattern Recognition*. IEEE.
- Gilbert, A., Illingworth, J., & Bowden, R. (2009). Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Proceedings of the International Conference on Computer Vision* (Vol. 30, pp. 925–931). IEEE. doi:10.1109/ICCV.2009.5459335
- Hamdoun, O., Moutarde, F., Stanculescu, B., & Steux, B. (2008). Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras* (pp.1-6). IEEE. doi:10.1109/ICDSC.2008.4635689
- Hirzer, M., Beleznaï, C., Roth, P. M., & Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *Proceedings of the Scandinavian Conference on Image Analysis* (pp. 91-102). Springer. doi:10.1007/978-3-642-21227-7_9
- Hirzer, M., Roth, P. M., Stinger, M., & Bischof, H. (2012). Relaxed pairwise learned metric for person re-identification. In *Proceedings of the European Conference on Computer Vision* (pp. 780-793). Academic Press. doi:10.1007/978-3-642-33783-3_56
- Karaman, S., & Bagdanov, A. D. (2012). Identity inference: generalizing person re-identification scenarios. In *Proceedings of the International Conference on Computer Vision* (Vol. 42, pp. 443-452). Springer.
- Karanam, S., Li, Y., & Radke, R. J. (2015). Person Re-Identification with Discriminatively Trained Viewpoint Invariant Dictionaries. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4516-4524). IEEE. doi:10.1109/ICCV.2015.513
- Karanam, S., Li, Y., & Radke, R. J. (2015). Sparse re-id: Block sparsity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 33-40). IEEE. doi:10.1109/CVPRW.2015.7301392
- Ke, Y., Sukthankar, R., & Hebert, M. (2010). Volumetric features for video event detection. *International Journal of Computer Vision*, 88(3), 339–362. doi:10.1007/s11263-009-0308-z
- Klaser, A. (2008). A spatiotemporal descriptor based on 3D-gradients. In *Proceedings of the British Machine Vision Conference*. Academic Press.
- Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2288-2295). IEEE Computer Society. doi:10.1109/CVPR.2012.6247939
- Laptev, I., & Lindeberg, T. (2003). On space-time interest points. *International Journal of Computer Vision*, 64(2-3), 432–439.
- Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008* (pp.1-8). IEEE. doi:10.1109/CVPR.2008.4587756
- Li, Y., Zhuo, L., Li, J., Zhang, J., Liang, X., & Tian, Q. (2017). Video-Based Person Re-identification by Deep Feature Guided Pooling. In *Proceedings of the Computer Vision and Pattern Recognition Workshops* (pp. 1454–1461). IEEE.
- Liang, C., Huang, B., Hu, R., Zhang, C., Jing, X., & Xiao, J. (2015, October). A unsupervised person re-identification method using model based representation and ranking. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 771-774). ACM.

- Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person re-identification by Local Maximal Occurrence representation and metric learning. *Computer Vision and Pattern Recognition*, 8, 2197–2206.
- Liao, S., & Li, S. Z. (2015). Efficient PSD Constrained Asymmetric Metric Learning for Person Re-Identification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3685–3693). IEEE. doi:10.1109/ICCV.2015.420
- Liu, K., Ma, B., Zhang, W., & Huang, R. (2015). A Spatio-Temporal Appearance Representation for Video-Based Pedestrian Re-Identification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3810–3818). IEEE. doi:10.1109/ICCV.2015.434
- Lu, H., Jia, X., & Yang, M. H. (2012). Visual tracking via adaptive structural local sparse appearance model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 157, pp. 1822–1829). IEEE Computer Society.
- Matsukawa, T., Okabe, T., Suzuki, E., & Sato, Y. (2016). *Hierarchical Gaussian Descriptor for Person Re-identification*. In *Proceedings of the Computer Vision and Pattern Recognition* (pp. 1363–1372). IEEE.
- Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6), 976–990. doi:10.1016/j.imavis.2009.11.014
- Satta, R. (2013). *Appearance descriptors for person re-identification: a comprehensive review*. Eprint Arxiv.
- Schwartz, W. R., & Davis, L. S. (2010). Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics & Image Processing* (pp. 322–329). IEEE.
- Scovanner, P., Ali, S., & Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the Society Conference on DBLP* (Vol. 2, pp. 1528–1535). IEEE.
- Shen, Y., Lin, W., Yan, J., Xu, M., Wu, J., & Wang, J. (2016). Person Re-Identification with Correspondence Structure Learning. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3200–3208). IEEE.
- Truong Cong, D. N., Achard, C., Khoudour, L., & Douadi, L. (2009). Video Sequences Association for People Re-identification across Multiple Non-overlapping Cameras. *Image Analysis and Processing*, 27, 179–189.
- Wang, H., Ullah, M. M., Kläser, A., Laptev, I., & Schmid, C. (2009). Evaluation of Local Spatio-temporal Features for Action Recognition. In *Proceedings of the British Machine Vision Conference, BMVC*. Academic Press. doi:10.5244/C.23.124
- Wang, T., Gong, S., Zhu, X., & Wang, S. (2014). Person Re-identification by Video Ranking. In *Proceedings of the European Conference on Computer Vision* (pp. 688–703). Cham: Springer.
- Wang, T., Gong, S., Zhu, X., & Wang, S. (2016). Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12), 2501–2514. doi:10.1109/TPAMI.2016.2522418 PMID:26829777
- Wang, X., Doretto, G., Sebastian, T., Rittscher, J., & Tu, P. (2007). Shape and Appearance Context Modeling. In *Proceedings of the International Conference on Computer Vision* (pp.1–8). IEEE.
- Willems, G., Tuytelaars, T., & Gool, L. (2008). An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In *Proceedings of the European Conference on Computer Vision* (pp. 650–663). Springer. doi:10.1007/978-3-540-88688-4_48
- Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2), 210–227. doi:10.1109/TPAMI.2008.79 PMID:19110489
- Wu, Y., Lim, J., & Yang, M. H. (2013). Online Object Tracking: A Benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 9, pp. 2411–2418). IEEE Computer Society. doi:10.1109/CVPR.2013.312

Yang, Y., Liao, S., Lei, Z., & Li, S. Z. (2016). Large scale similarity learning using similar pairs for person verification. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 3655-3661). AAAI Press.

You, J., Wu, A., Li, X., & Zheng, W. S. (2016). Top-push video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1345-1353).

Wenjun Huang received a B.S. degree from Wuhan University, Wuhan, China, in 2015 and he is currently a master's student of Wuhan University.

Chao Liang received the Ph.D degree from National Lab of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2012. He is currently working as an associate professor at National Engineering Research Center for Multimedia Software (NERCMS), Computer School of Wuhan University, Wuhan, China. His research interests focus on multimedia content analysis and retrieval, computer vision and pattern recognition, where he has published over 30 papers, including premier conferences such as CVPR, ACM MM and honorable journals like TMM and TCSVT, and won the best paper award of PCM 2014.

Chunxia Xiao received his BSc and MSc degrees from the Mathematics Department of Hunan Normal University in 1999 and 2002, respectively, and his PhD degree from the State Key Lab of CAD & CG of Zhejiang University in 2006. Currently, he is a professor in the School of Computer, Wuhan University, China. From October 2006 to April 2007, he worked as a postdoc at the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, and during February 2012 to February 2013, he visited University of California-Davis for one year. His main interests include computer graphics, computer vision and machine learning. He is a member of IEEE.

Zhen Han received the B.S. degree in computer science and technology and Ph.D. degree in computer application technology from Wuhan University, Wuhan, China, in 2002 and in 2009 respectively. Now he is an associate professor in school of computer, Wuhan University. His research interests include image/video compressing and processing, computer vision and artificial intelligence.