A SPATIO-TEMPORAL APPEARANCE REPRESENTATION FOR VIDEO-BASED PEDESTRIAN RE-IDENTIFICATION

Kan Liu[†] Bingpeng Ma[‡] Wei Zhang[†] Rui Huang[⋆*]
[†]School of Control Science and Engineering, Shandong University, China
[‡]School of Computer and Control Engineering, University of Chinese Academy of Sciences, China
[⋆]NEC Laboratories, China
sakuraxiafan@gmail.com, bpma@ucas.ac.cn, davidzhangsdu@gmail.com, huang-rui@nec.cn

Abstract

Pedestrian re-identification is a difficult problem due to the large variations in a person's appearance caused by different poses and viewpoints, illumination changes, and occlusions. Spatial alignment is commonly used to address these issues by treating the appearance of different body parts independently. However, a body part can also appear differently during different phases of an action. In this paper we consider the temporal alignment problem, in addition to the spatial one, and propose a new approach that takes the video of a walking person as input and builds a spatio-temporal appearance representation for pedestrian re-identification. Particularly, given a video sequence we exploit the periodicity exhibited by a walking person to generate a spatio-temporal body-action model, which consists of a series of body-action units corresponding to certain action primitives of certain body parts. Fisher vectors are learned and extracted from individual body-action units and concatenated into the final representation of the walking person. Unlike previous spatio-temporal features that only take into account local dynamic appearance information, our representation aligns the spatio-temporal appearance of a pedestrian globally. Extensive experiments on public datasets show the effectiveness of our approach compared with the state of the art.

1. Introduction

Identifying a specific person in videos is critical to many surveillance, security and multimedia applications such as on-line tracking or off-line searching a person of interest in videos. Person re-identification (re-id) has been widely used to describe such a task, i.e., *re*-identifying a person who has been previously observed in a video camera network. The entire pipeline of a re-id system may include person detection, tracking, segmentation (desirable but not necessary), feature modeling and matching. A typical re-id algorithm often focuses on feature modeling and matching, assuming that the input are cropped images containing the roughly aligned human subjects, coming from a person detector or tracker, preferably with reasonable segmentation.

Although *face* is probably the most reliable, visually accessible biometric to a person's identity, it is not always useful in video surveillance scenarios due to the low resolution and pose variations of individuals in typical surveillance footage. In such cases, body features are more useful because they can be detected and measured at lower resolution. *Gait* is a whole-body, behavioral biometric that describes the way a person walks and has long been studied for person identification. However, since gait is considered a biometric that is not affected by the appearance of a person, most state-of-the-art gait recognition methods work with silhouettes, which are difficult to extract, especially from surveillance data with cluttered background and occlusions. Therefore, in this paper we make the usual assumption that the person of interest does not change clothes between cameras, and focus on the person re-id methods that mainly use the body appearance, while also take into account the gait information to some extent.

This problem is quite challenging primarily because of the large variations in a person's appearance caused by different poses and viewpoints, illumination changes, and occlusions. A common strategy to address these issues is to exploit a body part model to take into account the non-rigid shape of the human body and treat the appearance of different body parts independently [25]. This is essentially a form of spatial alignment. However, a body part can also appear differently during different phases of an action. For instance, the arms may change appearance when swinging, sometimes may occlude the torso and change the torso's ap-

[*]Corresponding author. This work was partially supported by NSFC 61203253, 61572465, 61573222, Shandong Province RF 2015ZDXX0801A02, BS2013DX023, and SRF for ROCS, SEM.

pearance, etc. In this paper, we address the temporal alignment problem, in addition to the spatial one, of person re-id. The intuition behind our proposal is that we should not only model the appearance of different body parts independently, but also deal with the different phases of an action independently.

It is impossible to capture the varying appearance of a body part performing different action primitives using a single image (*single-shot* re-id). *Multiple-shot* approaches that use multiple images of a person to extract the appearance descriptors might work if we can obtain all the key frames corresponding to the different action primitives of an action sequence, which is not easy to achieve. Naturally we have to deal with the video-based re-id problem, because videos inherently contain more information than independent images, not only more body poses but also the underlying dynamics of a moving person, not to mention in many practical applications the input are videos to begin with. On the other hand, it is also more difficult and costlier to process videos with abundant information to obtain stable and robust appearance descriptors, and only a few studies have explored this problem [10, 4, 1, 30].

Unlike the previous work that only uses the videos to extract local spatio-temporal features, in this paper we consider a spatio-temporal representation that encodes both the spatial layout of the body parts and the temporal ordering of the action primitives, so that two pedestrians to be compared are aligned both spatially and temporally through such a representation. Our video-based pedestrian re-id algorithm assumes that the input are video sequences containing walking pedestrians. We use the term *pedestrian* to emphasize our focus on exploiting additional temporal information in *walking* for spatio-temporal appearance modeling while ignoring other complicated actions at present. This is a special case of person re-id, but also one that describes the most common and natural status of the human subjects in surveillance footage.

More specifically, given a video sequence of a walking person (roughly cropped out in each frame), we first extract the individual walking cycles. For each walking cycle, we divide the chunk of video data both spatially and temporally. In the temporal dimension, we split the sequence into a couple of segments corresponding to different phases of a walking cycle; and in the spatial domain, we divide the different body parts apart. We then obtain multiple video blobs based on the spatial and temporal segmentation, and each video blob is a small chunk of data corresponding to a certain action primitive of a certain body part, which is named a *body-action unit*. Based on the spatio-temporally meaningful body-action units we then train visual vocabularies and extract Fisher vectors, a generalized Bag-of-Words (BoW) type of feature. Finally we concatenate the Fisher vectors extracted from all the body-action units to form a fixed-length feature vector to represent the appearance of a walking person.

The benefits of such a representation are: 1) It describes a person's appearance during a walking cycle, hence covers almost the entire variety of poses and shapes; 2) It aligns the appearance of different people both spatially and temporally; 3) The formation of each body-action unit can be very flexible and different for each person, while Fisher vectors can work with any volume topologies, so the final representation is a consistent feature vector. In the following we will first briefly review the most relevant literature (Section 2) and then explain our method in detail (Section 3). We have conducted extensive experiments (Section 4) to validate our approach on two public datasets, with discussions on the strength and weakness of our approach. Finally we conclude the paper with some ideas for future work (Section 5).

2. Related work

Person re-id has been an active research topic in the past few years. It faces great challenges caused by different poses and viewpoints, illumination changes, and occlusions. In general, most recent work focuses on two aspects of the solution [6]: 1) appearance modeling [25]; and 2) distance metric learning [35]. We refer the readers to [8, 31, 25, 6, 11] for comprehensive reviews on this topic. In this section, we give a brief review of the studies most related to our work.

For appearance modeling, the most often used low-level features are color, texture, gradient, and naturally, the combination of these features [17], extracted either from the whole body area (*global* features) or from the points/regions of interest (*local* features). On top of the low-level features, many methods build more discriminative appearance descriptors using learning algorithms, e.g., boosting [2], Bag-of-Words type of dictionary learning [18], etc.

To alleviate the misalignment caused by pose variations, appearance modeling typically exploits part-based body models to take into account the non-rigid shape of the human body and treat the appearance of different body parts independently. Such body part models can be manually designed (e.g., horizontal stripes [23, 37], body part templates [33]), adaptive to the input data [10, 9], or learned from the training data [4, 32]. Applying a part-based body model is essentially a form of spatial alignment, which can address the pose and occlusion problem to some extent.

As mentioned previously, multiple-shot methods can also be used to improve appearance modeling. Early approaches often rely on the matching methods to choose the most representative features [20, 9], while more recent approaches accumulate or average the features from the multiple images into a single signature [2, 3]. When video sequences are available, i.e., the multiple images of a person are temporally related, the features taking advantage

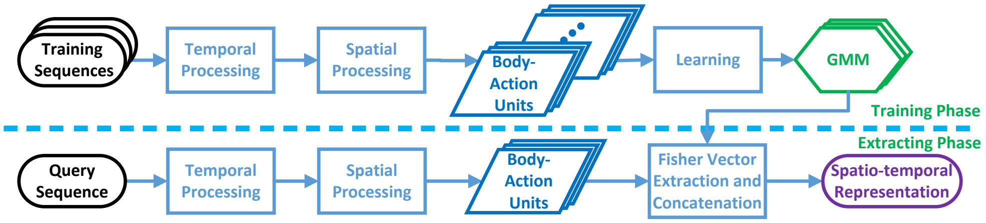

Figure 1. Framework of the proposed spatio-temporal representation.

of such temporal correlation are called spatio-temporal features. Many approaches have exploited the third dimension of the video data to build spatio-temporal representations. For instance, 3D SIFT [26] and 3D HOG [14] are both 3D extensions to the widely used 2D features. However they are usually used for action recognition because they are mostly based on gradients with little color information.

Gait has long been studied for video-based person identification [7, 21, 27]. As a biometric characterizing a person's walking style, gait is usually analyzed based on the silhouettes (model-free approaches) [29, 12] or the body part configurations (model-based approaches) [28, 34], without making use of the person's appearance. These approaches often require accurate silhouette extraction or body part segmentation, which are still open problems. Therefore, [5] proposed to incorporate gait features with colors only if the silhouette extraction is successful by some measurement.

Our approach is partly inspired by the spatio-temporal appearance models such as [10, 4, 1]. Although these approaches treat the video data as 3D volumes, they do not align the sequences from different people temporally using the available action information, such as the intrinsic periodicity property exhibited by a walking person. We want to further exploit the global temporal information contained in the actions for the re-id problem, in the form of temporal alignment through a series of action primitives, analogous to spatial alignment through a body part model.

One of the very few studies that have addressed this problem is [30], which breaks down an image sequence based on the motion energy intensity, and generates a pool of video fragment candidates for a learning model to automatically select the most discriminative fragments. Although it is not explicitly guaranteed, the learned ranking model is more likely to choose the temporally aligned video fragments. This approach belongs to the distance metric learning based approaches that focus on learning appropriate distance metrics to maximize the matching accuracy, regardless of the choice of appearance modeling [23, 37, 15, 22, 36, 16]. However, these approaches rely on a set of training data from a fixed set of cameras for supervised learning, which might be an impractical requirement in many real-world applications.

In summary, our method belongs to the appearance modeling category of the person re-id approaches. We first propose a method to temporally divide the image sequence into small segments corresponding to the action primitives of walking cycles, and combine the temporal segmentation with a simple manually designed fixed body part model to obtain spatio-temporally meaningful video blobs called body-action units. We then extract Fisher vectors [24] built on a concise low-level descriptor that combines color and gradients inspired by [18]. While our focus is on a better representation that encodes both the spatial layout of the body parts and the temporal ordering of the action primitives of a walking person, we will also show in the experiments that our approach can be further improved by distance metric learning methods, in particular a Mahalanobis metric [15].

3. Proposed method

In this section, we introduce a new spatio-temporal representation of a pedestrian's appearance in a video. Given a video sequence $Q = (I_1, I_2, ..., I_t)$ obtained from a person tracking algorithm, our goal is to extract a feature vector that encodes the spatially and temporally aligned appearance of the person in a walking cycle, or a set of such feature vectors, depending on how many walking cycles can be found in the video. The entire framework, as depicted in Figure 1, includes a training phase to learn the probabilistic visual vocabulary, e.g., Gaussian Mixture Models (GMMs), and a feature extraction phase to generate the actual feature vectors, e.g., Fisher vectors (Section 3.2). Both the dictionary learning and feature extraction phases are performed with respect to the body-action units corresponding to the action primitives of the body parts (Section 3.1).

3.1. Spatio-temporal body-action model

3.1.1 Walking cycle extraction

In this module, we are trying to extract individual walking cycles from the given video $Q = (I_1, I_2, ..., I_t)$. We first extract the Flow Energy Profile (FEP) as proposed in [30]. The FEP is a one dimensional signal $E = (e_1, e_2, ..., e_t)$, which approximates the motion energy intensity profile of the consecutive frames in Q using the optic flow field. For each frame I:

$$e = \sum_{(x,y)\in U} \|[v_x(x,y), v_y(x,y)]\|_2, \qquad (1)$$

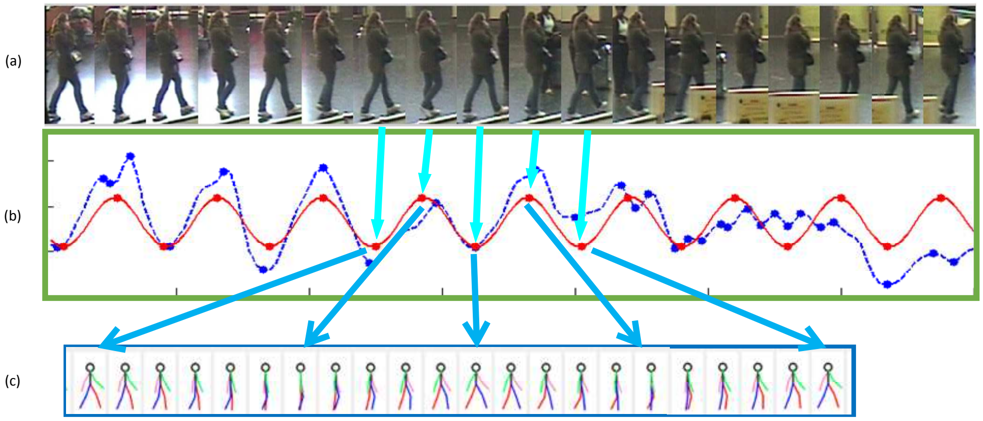

Figure 2. Walking cycle extraction. (a) A video sequence of a pedestrian (only key frames). (b) The original FEP (blue curve) and the regulated FEP (red curve). (c) The stick figures illustrating the pedestrian poses extracted from a walking cycle. (Better viewed in color.)

where U is the lower half of the image containing the lower body of a pedestrian (because the movement of the lower body is the most prominent and consistent), and v_x, v_y are the optic flows on the horizontal and vertical direction. It is worth pointing out that we find that for many video sequences the horizontal optic flow v_x alone is more effective.

Ideally, the local maxima of E correspond to the postures when the person's two legs overlap while at the local minima the two legs are the farthest away. However, the signal is often perturbed by noisy background and occlusions, so some local maxima/minima may not appear as expected. In addition, we sometimes observe small dips around the local maxima, which are not as stable as the local minima (Figure 2(b), blue dotted curve). It is difficult to extract walking cycles from the unregulated FEP. In [30] the authors simply extracted fixed-length fragments around the local maxima/minima of E and relied on the learning method to choose the most discriminative fragments.

Instead we try to obtain more accurate walking cycles assuming the dominant periodicity contained in the FEP of a walking sequence is caused by the walking cycles. Therefore, we transform the original FEP signal E into the frequency domain using the discrete Fourier transform, filter out all the frequencies except the dominant one, and obtain the regulated FEP signal E' using the inverse discrete Fourier transform on the remaining frequency (Figure 2(b), red curve). As one can see the local maxima/minima of E' are better indicators of the walking cycles.

We then split the whole video sequence into segments according to these local maxima/minima. Due to the symmetry of the walking action, a full cycle contains two consecutive sinusoid curves, one step from each leg. However it is extremely difficult to distinguish between the two, hence we treat each sinusoid curve, i.e., a single step, as a walking cycle (with a little abuse of terminology). Unlike the fixed-length fragments in [30], each person may have a different pace. To temporally align different walking cycles, we further divide a cycle into smaller segments $S = (s_1, s_2, ..., s_N)$, where s_i is a set of consecutive indices of Q, corresponding to an action primitive. Walking is a relatively simple action, so we have $N = 4$ segments for each walking cycle in this work.

3.1.2 Body-action units

As to spatial alignment, we need to find the proper parts of the human body, $P = (p_1, p_2, ..., p_M)$, where p_i is an area in a frame I, corresponding to a body part. Ideally different frames may have different body part segmentation, i.e., P is dependent on time. In practice, however, we find that a fixed body part model works fine at a very low computational cost. In particular, to take advantage of the common spatial configuration of walking pedestrians (e.g., mostly standing upright, often appearing symmetric) without using sophisticated part matching algorithms, we describe the entire human body area with $M = 6$ smaller rectangles roughly corresponding to the six human body parts (i.e., head, torso, left and right arms, left and right legs), as shown in Figure 3. The template is empirically derived from the average image of the training set.

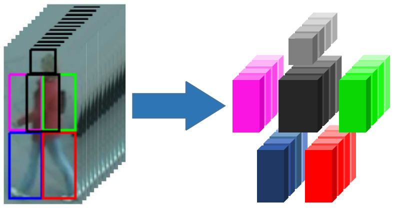

Figure 3. Spatial-temporal body-action units. (Color encodes the body parts, and intensity encodes the action primitives. Better viewed in color.)

From above spatial and temporal segmentation of the input video sequence, we obtain both the spatial bounding boxes corresponding to the body parts and the temporal segments corresponding to the action primitives of a person's appearance during walking. Combining them, we obtain $M \times N$ spatially and temporally aligned video blobs, named body-action units, as shown in Figure 3:

$$W_{mn} = \{(x, y, t) | (x, y) \in P_m, t \in S_n\},\\ m = 1, ..., M, n = 1, ..., N, \tag{2}$$

where P_m denotes the area of the m^{th} body part and S_n denotes the n^{th} temporal segment within the walking cycle.

It is worth noting that a body-action unit W_{mn} neither has to be a regular volume such as a cuboid, nor be the same size for different people. Feature extraction and model training are performed with respect to each body-action unit separately. For clarity, we limit the following discussion in a single unit. The complete feature or model is a concatenation of the features or models from all the units.

3.2. Fisher vector learning and extraction

In order to characterize the appearance of each body-action unit, we extract Fisher vectors built upon low-level feature descriptors. The low-level feature we used is a very concise local descriptor that combines color, texture, and gradient information:

$$f(x, y, t) = [\tilde{x}, \tilde{y}, \tilde{t}, I(x, y, t), \frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial I}{\partial t}, \frac{\partial^2 I}{\partial x^2}, \frac{\partial^2 I}{\partial y^2}, \frac{\partial^2 I}{\partial t^2}], \tag{3}$$

where $\tilde{x}$, $\tilde{y}$ and $\tilde{t}$ are the relative coordinates of the pixel within the unit. $I(x, y, t)$ is the pixel intensity, and the rest are the first and second derivatives. In practice, there are usually three color channels for each pixel, e.g., we use HSV in our implementation, so in total there are D = 3 (relative coordinates) + 7 (color/gradient features) × 3 (color channels) = 24 dimensions for a descriptor on each pixel.

The Fisher vector [24] is an image representation which is usually used in visual classification and has seen success in person re-id. Given the training images for a body-action unit W, we learn a GMM using the extracted D-dimensional local descriptors. The learned model is denoted by $\Theta = \{(\mu_k, \sigma_k, \pi_k) : k = 1, \ldots, K\}$, where μ_k, σ_k and π_k are the mean, covariance and prior probability of the k-th Gaussian component, respectively. Thus we have:

$$\mathcal{N}(f; \mu_k, \sigma_k) = \frac{1}{(2\pi)^{D/2} |\sigma_k|^{1/2}} \exp\{-\frac{1}{2}(f - \mu_k)' \sigma_k^{-1} (f - \mu_k)\}, \tag{4}$$

where $\mathcal{N}(f; \mu_k, \sigma_k)$ denotes the k-th Gaussian component and f is the low-level local descriptor mentioned above. In our implementation, K is empirically set to 32 for each body-action unit and σ_k is diagonal.

Once we have learned the probabilistic visual vocabulary, defined as GMMs, we can compute the posterior probability γ_{ik} of a local descriptor f_i being generated by the kth Gaussian component:

$$\gamma_{ik} = p(k | f_i; \mu_k, \sigma_k) = \frac{\pi_k \mathcal{N}(f_i; \mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(f_i; \mu_j, \sigma_j)}, \tag{5}$$

and the Fisher vector is the concatenation of the deviation vectors w_k, u_k and v_k, i.e., $\Phi(W) = [w_1, u_1, v_1, \ldots, w_K, u_K, v_K]^\top$, where

$$w_k = \frac{1}{|W|\sqrt{\pi_k}} \sum_{i \in idx(W)} (\gamma_{ik} - \pi_k) \tag{6}$$

$$u_k = \frac{1}{|W|\sqrt{\pi_k}} \sum_{i \in idx(W)} \gamma_{ik} \frac{f_i - \mu_k}{\sigma_k} \tag{7}$$

$$v_k = \frac{1}{|W|\sqrt{2\pi_k}} \sum_{i \in idx(W)} \gamma_{ik} \left[\left(\frac{f_i - \mu_k}{\sigma_k} \right)^2 - 1 \right] \tag{8}$$

Note that w_k is a scalar while u_k and v_k both have the same dimensionality as the low-level feature descriptor, therefore the Fisher vectors are $(2D + 1)K$ dimensional. The final representation of the pedestrian's appearance is the concatenation of the Fisher vectors of all the body-action units, hence is $(2D + 1)KMN$ dimensional.

3.3. Differences to other spatio-temporal features

Many spatio-temporal features simply add the extra temporal dimension to the original two dimensional image space, without considering the alignment problem. Such features are simply local 3D features. From a global point of view, to align two volumes of video data, i.e., to encode the spatial and temporal layout of the local features, a simple strategy of dividing the volume with a regular grid is somewhat effective, as used in features like 3D HOG [14]. For the re-id problem, however, a higher level of alignment accuracy is desirable. [30] advocates the alignment of the key postures, and builds a fixed-size block around the key frame for extracting 3D HOG. Our representation takes a step further in this direction, and aligns the appearance of different pedestrians both spatially and temporally. The formation of each body-action unit can be flexible and different for each person. It is even possible to use different body part models for different action primitives, or vice versa, as long as the number of parts and primitives are fixed, resulting in a very flexible joint body-action model, yet the final representation is a consistent feature vector across different people for easy comparisons.

4. Experiments

In this section, we validate our method and compare it to other state-of-the-art approaches on two public datasets.

4.1. Datasets and Settings

Experiments were conducted on two person re-id datasets: the iLIDS-VID dataset [30] and the PRID 2011 dataset [13], as shown in Figure 4 and Table 1.

iLIDS-VID dataset. The iLIDS-VID dataset includes 600 image sequences for 300 randomly sampled people, which is created based on two non-overlapping camera

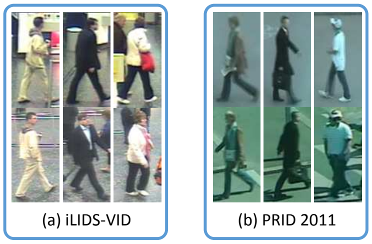

Figure 4. Example pairs of the same people in different camera views from two datasets.

Table 1. Dataset Information

Dataset	# of people	# of cameras	Average length	Image size
iLIDS-VID	300	2	73	64×128
PRID 2011	200	2	100	64×128

views. Each image sequence has variable length consisting of 23 to 192 image frames, with an average number of 73. Due to cluttered background, occlusions, clothing similarities and viewpoint variations across camera views, this dataset is very challenging.

PRID 2011 dataset. The PRID 2011 dataset consists of 400 image sequences for 200 people, and each image sequence has variable length consisting of 5 to 675 image frames, with an average number of 100. In our experiments, the sequence pairs with less than 20 frames are ignored due to the requirement on the sequence length for extracting walking cycles. The dataset has two adjacent camera views captured in uncrowded outdoor scenes with rare occlusions and clean background. However the color inconsistency between the two camera views is obvious, and the shadows are severer in one of the views.

Settings. To evaluate our method, we equally split the whole pool of sequence pairs into two subsets for each dataset, one for training and the other for testing. The query set consists of the sequences from the first camera while the gallery set from the other one. For both datasets, the performance is measured by the average Cumulative Matching Characteristics (CMC) curves after 10 trails.

For each walking cycle extracted from the video sequences, we divided it into 24 body-action units (6 spatial body parts and 4 temporal action primitives). In each unit, we first extract the low-level local descriptors. The Fisher vector model learning and feature extraction are then performed. We observed that the performance was not very sensitive to the number of GMM components, which was set to 32 in all of our experiments. The 24 descriptors are then concatenated into the complete representation, which is $(2 \times 24 + 1) \times 32 \times 24 = 37632$ dimensional.

Because different sequences may contain different numbers of walking cycles, for each sequence we may extract a different number of spatio-temporal descriptors. We use all of them as query or gallery descriptors and apply the nearest neighbor classifier to determine the distance between two sets of descriptors extracted from two sequences.

4.2. Evaluation of the low-level descriptor

As we pointed out above, the image sequences in the PRID 2011 dataset have significant color inconsistency under the two cameras, we have found that the color and second-order derivatives in the low-level descriptor (Section 3.2) do not work well with such data. We performed a series experiments to investigate the effectiveness of the low-level descriptors. In Table 2, the first two rows show the different performances of our representation (denoted STFV3D) based on two variants of the low-level descriptor (i.e., the original 24-dimensional one and the 12-dimensional one with color and second derivatives omitted). For iLIDS-VID the original descriptor works better, while for PRID 2011 the 12-dimensional one works better. This shows that even though the unsupervised Fisher vector learning can produce a good representation, the extracted features are not necessarily optimal for classification. Empirical feature selection in this case is helpful. We then combined STFV3D with a supervised distance metric learning method, the KISSME algorithm [15], and repeated the experiments (the last two rows in Table 2). As we expected, supervised learning can take care of feature selection quite well, and the 24-dimensional richer low-level descriptors perform better on both datasets. In the following experiments, we use the 12-dimensional descriptor on the PRID 2011 dataset when no supervised learning is employed.

4.3. Comparison to other representations

In this section, we compare our STFV3D, without distance metric learning, to three other description methods:

HOG3D, which extracts 3D HOG features from volumes of video data collected similar to [30]. More specifically, for each local maximum/minimum of the FEP signal E, 10 frames from 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.

FV3D, which is similar to HOG3D but we replace the HOG features with Fisher vectors.

FV2D, which is a multiple-shot approach treating the video sequences as multiple independent images using Fisher vectors as the features. This is one of the state-of-the-art approaches for image-based person re-id [18].

Note that for these methods we extract descriptors at every local maxima/minima of the FEP, which generates considerably more descriptors for matching than our own walking cycle based approach. The experimental results are shown in Table 3. From the results we can observe that in general STFV3D performs the best, and more specifically:

Table 2. Performance of different low-level descriptors

Dataset	iLIDS-VID				PRID 2011			
Rank R	$R = 1$	$R = 5$	$R = 10$	$R = 20$	$R = 1$	$R = 5$	$R = 10$	$R = 20$
STFV3D(12)	27.0	55.7	71.6	84.7	42.1	71.9	84.4	91.6
STFV3D(24)	37.0	64.3	77.0	86.9	21.6	46.4	58.3	73.8
STFV3D+KISSME(12)	34.9	63.0	76.0	86.3	62.4	84.9	87.1	91.4
STFV3D+KISSME(24)	**44.3**	**71.7**	**83.7**	**91.7**	**64.1**	**87.3**	**89.9**	**92.0**

Table 3. Comparison of different feature descriptors

Dataset	iLIDS-VID				PRID 2011			
Rank R	$R = 1$	$R = 5$	$R = 10$	$R = 20$	$R = 1$	$R = 5$	$R = 10$	$R = 20$
HOG3D	8.3	28.7	38.3	60.7	20.7	44.5	57.1	76.8
FV2D	18.2	35.6	49.2	63.8	33.6	64.0	76.3	86.0
FV3D	25.3	54.0	68.3	**87.3**	38.7	71.0	80.6	90.3
STFV3D	**37.0**	**64.3**	**77.0**	86.9	**42.1**	**71.9**	**84.4**	**91.6**

Body-action units vs. regular grid: STFV3D outperforms FV3D, which means the spatio-temporal segmentation of the video data improves the re-id performance over simple regular grid based 3D schemes (as used by most previous spatio-temporal representations, especially on the temporal dimension).

Video-based approaches vs. independent multi-shot: Both STFV3D and FV3D outperform FV2D, which means the additional effort made to model the temporal correlation paid off. It is worth noting that we find it impractical to use all the images due to the computational complexity, therefore in our experiment FV2D only used the images corresponding to the local maxima/minima of the FEP signal.

FV3D vs. HOG3D: FV3D and FV2D outperform HOG3D, which is not a surprise because the Fisher vectors based on our local descriptors are more sophisticated and suitable for the re-id problem, even though a lot more HOG3D descriptors are used as the gallery and query.

iLIDS-VID vs. PRID 2011: The above observations hold for both datasets. We would like to point out again that on the PRID 2011 dataset we used only 12-dimensional low-level features without the HSV values because of the significant color inconsistency. We will later show how this empirical feature selection problem can be addressed by supervised distance metric learning methods that can learn the relationship between the two cameras. Nonetheless our sophisticated appearance modeling still shows its merit, especially under the unsupervised setting. This is particularly important when we are dealing with the videos from multiple cameras unseen before. Another notable difference between the results on the two datasets is that FV2D performs better on the PRID 2011 dataset than on the iLIDS-VID dataset, considering its relative performance to the 3D approaches. We believe this is because the iLIDS-VID dataset has more cluttered background and considerable occlusions, which probably causes more trouble for the 2D approaches.

4.4. Comparison to the state of the art

In this section we compare our method with the state-of-the-art video-based person re-id approaches. To achieve the best performance we combine STFV3D with supervised distance metric learning methods such as KISSME [15] and Local Fisher Discriminant Analysis (LFDA [22]). In both methods, PCA is first performed to reduce the dimension of our original representation. We have empirically chosen the reduced dimension as 150 in our implementation.

In Table 4, the first three rows show the performance of the state-of-the-art approaches, namely, Gait Energy Image (GEI)+Rank SVM (RSVM) [19], HOG3D+Discriminative Video Ranking (DVR) [30], Color+LFDA [22]. The second and third group of methods are variants of our proposal. From these results we can see that distance metric learning can further improve the performance of our appearance modeling approach. The performance boost is largely because that distance metric learning can bridge the gap of color and viewpoint variations across camera views, which are difficult for unsupervised appearance modeling methods to handle. This effect is more obvious on PRID 2011 because of the significant color inconsistency in this dataset. Interestingly, the improvement due to distance metric learning decreases when the rank number increases. We believe that it is partly because our appearance modeling method already performs pretty well at the higher rank numbers, and the distance metric learning algorithms can pull reasonably similar pairs closer but does not have much effect on really distant pairs. Our appearance modeling approach combined with the KISSME algorithm achieved the overall best performance, and the gait features alone do not perform well on these datasets.

4.5. Limitations and failure examples

Finally we discuss the limitations of our approach, and show some failure examples (Figure 5 and Figure 6). Each figure contains the matching results of the same person

Table 4. Comparison of our proposed methods and the state of the art

Dataset	iLIDS-VID				PRID 2011			
Rank R	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$
GEI+RSVM [19]	2.8	13.1	21.3	34.5	-	-	-	-
HOG3D+DVR [30]	23.3	42.4	55.3	68.4	28.9	55.3	65.5	82.8
Color+LFDA [22]	28.0	55.3	70.6	88.0	43.0	73.1	82.9	90.3
FV3D	25.3	54.0	68.3	87.3	38.7	71.0	80.6	90.3
FV3D+LFDA	32.0	59.3	78.6	88.6	47.2	76.2	84.1	90.6
FV3D+KISSME	36.6	69.3	82.6	91.3	62.3	83.8	86.0	**92.4**
STFV3D	37.0	64.3	77.0	86.9	42.1	71.9	84.4	91.6
STFV3D+LFDA	38.3	70.1	83.4	90.2	48.1	81.2	85.7	90.1
STFV3D+KISSME	**44.3**	**71.7**	**83.7**	**91.7**	**64.1**	**87.3**	**89.9**	92.0

by two approaches, (a) FV3D and (b) STFV3D. For each approach, we show a pair of video segments that is the best matched pair of query (top) and gallery (bottom) using the nearest neighbor classifier. Note that FV3D uses fixed-length segments while STFV3D uses flexible segments based on walking cycle extraction. In both cases, FV3D finds the correct match while STFV3D does not. In Figure 5, the color inconsistency is causing trouble for both representations, and the matching is probably more affected by pose and shape. Figure 5(b) shows that the cluttered background causes inaccurate walking cycle extraction in STFV3D, and hence incorrect matching between the query and gallery. In Figure 6, the viewpoint of the query sequence is significantly different from our *sideview* assumption, which also causes inaccuracy of both spatial and temporal alignment in STFV3D (Figure 6(b)).

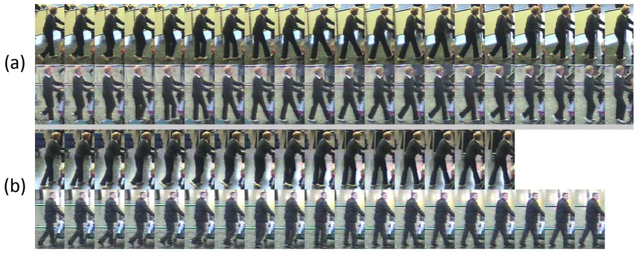

Figure 5. Failure example 1.

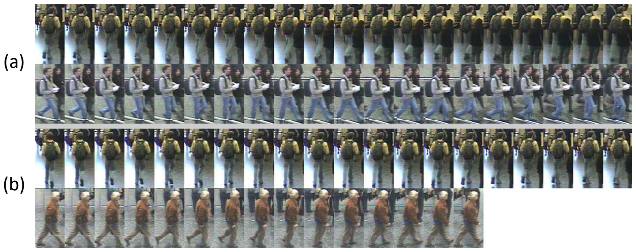

Figure 6. Failure example 2.

5. Discussions

In this paper we proposed a novel video-based pedestrian re-id framework. Unlike most previous spatio-temporal modeling approaches that only explore the temporal correlation locally, we are trying to exploit temporal information on the action level, that is, dividing a video sequence into small segments corresponding to the action primitives. Combined with body part segmentation, we obtain a series of video blobs, named body-action units, corresponding to different action primitives of different body parts. Fisher vectors are learned and extracted in each unit and concatenated into the final representation. Such a representation describes a person's appearance during an action, e.g., in this paper a walking cycle, hence covers a large variety of poses and shapes. It effectively aligns the dynamic appearance of different people both spatially and temporally. The formation of each video blob can be flexible and different for each person, as opposed to a fixed-size grid, but the final representation is a consistent feature vector for easy comparison of two persons' appearance.

There are some interesting directions for further improvement of our framework. From the spatial alignment point of view, the publicly available data for video-based re-id we are dealing with contain mostly sideview pedestrians, while in practice the pedestrians in a video may walk in any direction. Even for simple actions like walking, the change of viewpoints can still cause serious problems in spatial alignment. We are investigating better body part models to address the pose/viewpoint problem. From the temporal alignment viewpoint, although we have chosen to tackle the pedestrian re-id problem at present because walking is a relatively simple periodic action, the generalization ability of our framework is limited by action analysis, which itself is still an open problem. Nonetheless, there is great potential in our model. We are experimenting a more efficient body-action model where different body parts can have different action primitives, while different action primitives involve different body parts. For instance, during walking, the head may have fewer meaningful action primitives than the arms and legs due to its relatively simple motion, while on the other hand, the stance phase may involve fewer body parts than the swing phase because of the self-occlusion of body parts.

References

[1] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *ECCV*. 2012. 2, 3

[2] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Boosted human re-identification using Riemannian manifolds. *Image and Vision Computing*, 30(6), 2012. 2

[3] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 33(7), 2012. 2

[4] A. Bedagkar-Gala and S. K. Shah. Multiple person re-identification using part based spatio-temporal color appearance model. In *ICCV Workshops*, 2011. 2, 3

[5] A. Bedagkar-Gala and S. K. Shah. Gait-assisted person re-identification in wide area surveillance. In *Computer Vision-ACCV 2014 Workshops*, 2014. 3

[6] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4), 2014. 2

[7] N. V. Boulgouris, D. Hatzinakos, and K. N. Plataniotis. Gait recognition: a challenging signal processing technology for biometric identification. *Signal Processing Magazine, IEEE*, 22(6):78–90, 2005. 3

[8] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, 2(2), 2011. 2

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2

[10] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006. 2, 3

[11] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*. Springer, 2014. 2

[12] J. Han and B. Bhanu. Individual recognition using gait energy image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(2):316–322, 2006. 3

[13] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*. 2011. 5

[14] A. Klaser, M. Marszaek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 3, 5

[15] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 3, 6, 7

[16] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 3

[17] B. Ma, Y. Su, and F. Jurie. BiCov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012. 2

[18] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by Fisher vectors for person re-identification. In *ECCV Workshops*, 2012. 2, 3, 6

[19] R. Martin-Felez and T. Xiang. Gait recognition by ranking. In *ECCV*. 2012. 7, 8

[20] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern recognition*, 36(9), 2003. 2

[21] M. S. Nixon and J. N. Carter. Automatic recognition by gait. *Proceedings of the IEEE*, 94(11):2013–2024, 2006. 3

[22] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 3, 7, 8

[23] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, 2010. 2, 3

[24] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal on Computer Vision*, 105(3), 2013. 3, 5

[25] R. Satta. Appearance descriptors for person re-identification: a comprehensive review. *arXiv preprint arXiv:1307.5748*, 2013. 1, 2

[26] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In *ACM MM*, 2007. 3

[27] J. Wang, M. She, S. Nahavandi, and A. Kouzani. A review of vision-based gait recognition methods for human identification. In *DICTA*, 2010. 3

[28] L. Wang, H. Ning, T. Tan, and W. Hu. Fusion of static and dynamic body biometrics for gait recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(2):149–158, 2004. 3

[29] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(12):1505–1518, 2003. 3

[30] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*. 2014. 2, 3, 4, 5, 6, 7, 8

[31] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1), 2013. 2

[32] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, 2013. 2

[33] Y. Xu, B. Ma, R. Huang, and L. Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *ACM MM*, 2014. 2

[34] C. Yam, M. S. Nixon, and J. N. Carter. Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, 37(5):1057–1072, 2004. 3

[35] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy Technical Report*, 2006. 2

[36] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 3

[37] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 2, 3