Video-Based Pedestrian Re-Identification by Adaptive Spatio-Temporal Appearance Model

Wei Zhang, *Member, IEEE*, Bingpeng Ma, *Member, IEEE*, Kan Liu, *Student Member, IEEE*, and Rui Huang, *Member, IEEE*

***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 spatiotemporal appearance representation for pedestrian re-identification. Particularly, given a video sequence, we exploit the periodicity exhibited by a walking person to generate a spatiotemporal 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 spatiotemporal features that only take into account local dynamic appearance information, our representation aligns the spatiotemporal appearance of a pedestrian globally. Extensive experiments on public data sets show the effectiveness of our approach compared with the state of the art.**

***Index Terms*— Pedestrian re-identification, appearance representation, spatio-temporal alignment.**

I. 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 [1]. 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 appearance, 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). *multi-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

Manuscript received May 13, 2016; revised November 15, 2016; accepted February 16, 2017. Date of publication February 19, 2017; date of current version March 17, 2017. This work was supported in part by NSFC under Grant 61573222, Grant 61572465, and Grant 61233014, in part by the Major Research Program of Shandong Province under Grant 2015ZDXX0801A02, in part The Fundamental Research Funds of Shandong University 2016JC014, and in part by SRF for ROCS, SEM. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paul Rodriguez. *(Corresponding author: Rui Huang.)*

W. Zhang and K. Liu are with the School of Control Science and Engineering, Shandong University, Jinan 250061, China.

B. Ma is with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China.

R. Huang is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: ruihuang@cuhk.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2672440

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 [2]–[5].

Unlike the previous work, in this paper we focus on investigating 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 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 adaptively divide the different body parts apart based on the pedestrian's pose. 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 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.

Besides, considering that most existing re-id datasets are designed for image-based methods (single-shot or multi-shot), we introduce a new image sequence dataset (named SDU-VID, larger than iLIDS-VID) to compensate the lack video re-id benchmark data. Using iLIDS-VID, PRID 2011 and the newly introduced dataset SDU-VID, extensive evaluations were conducted to demonstrate the superiority of the proposed spatio-temporal representation over existing methods.

In the following we will first briefly review the most relevant literature (Section II) and then explain our method in detail (Section III). We have conducted extensive experiments (Section IV) to validate our approach on two public datasets and our collected dataset, with discussions on the strength and weakness of our approach. Finally we conclude the paper with some ideas for future work (Section V).

II. 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 [1]; and 2) distance metric learning [7]. We refer the readers to [1], [6], [8]–[10] 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 [11], 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 [12], Bag-of-Words type of dictionary learning [13], 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 [14], [15], body part templates [16]), adaptive to the input data [2], [17], or learned from the training data [3], [18]. Applying a part-based body model is essentially a form of spatial alignment, which can address the pose and occlusion problem to some extent.

Another way of appearance modeling is through Deep Learning [19], which has been applied to tackle kinds of problems with success in areas such as vision [20]–[22] and audio [23]. For example, Convolutional Neural Network (CNN) [24] can well handle the image-based tasks such as image classification, image retrieval and object recognition. For image-based person re-id task, filter pairing neural network [25] is proposed to jointly handle the misalignment, photometric and geometric transforms, occlusions and background clutter. A scalable distance driven feature learning framework [26] is presented based on CNN to preserve similarity of the same person against large appearance and structure variations while discriminating different individuals.

As mentioned previously, multi-shot methods can also be used to improve appearance modeling. Early approaches often rely on the matching methods to choose the most representative features [17], [27], while more recent approaches accumulate or average the features from the multiple images into a single signature [12], [28]. When video sequences are available, i.e., the multiple images of a person are temporally related, the features taking advantage 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 [29] and 3D HOG [30] 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.

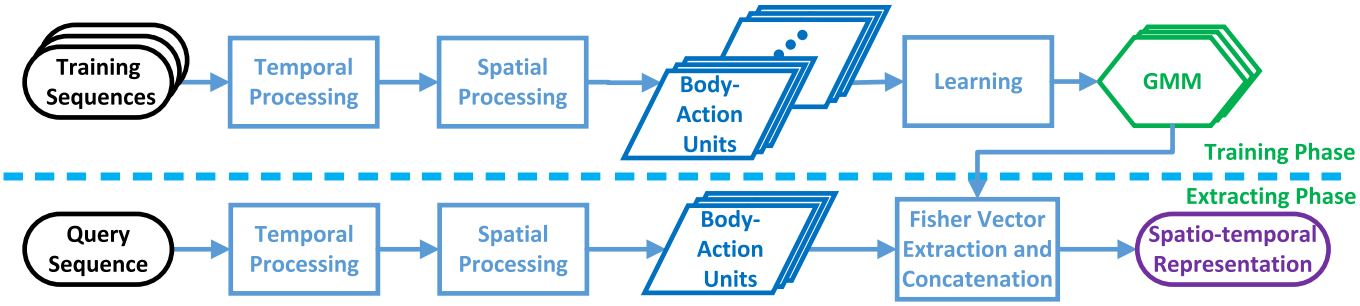


Fig. 1. Framework of the proposed spatio-temporal representation.

Gait has long been studied for video-based person identification [31]–[33]. As a biometric characterizing a person's walking style, gait is usually analyzed based on the silhouettes (model-free approaches) [34], [35] or the body part configurations (model-based approaches) [36], [37], 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, [38] 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 [2], [3] and [4]. 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 [5], 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 [39]–[43]. 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 an adaptive body part model to obtain spatio-temporally meaningful video blobs called body-action units. We then extract Fisher vectors [44] built on a concise low-level descriptor that combines color and gradients inspired by [13]. 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 [45].

This work is extended from our previous ICCV paper [46], and offers novel contributions as follows: 1) Two novel adaptive body-action units are introduced to improve the spatio-temporal alignment by encoding the temporal information into the spatial segmentation utilizing two simple functions; 2) The Deformable Parts Model (DPM) is integrated into the walking cycle extraction for the spatio-temporal segmentation; 3) Two different distance measures are investigated in the comparison of the final pedestrian representation; 4) Considering the lack of video re-id benchmark data, a new sequence named SDU-VID is introduced to evaluate the performance of state-of-the-art video-based approaches.

III. 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 III-B). 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 III-A).

A. Spatio-Temporal Body-Action Model

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 [5]. 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:

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

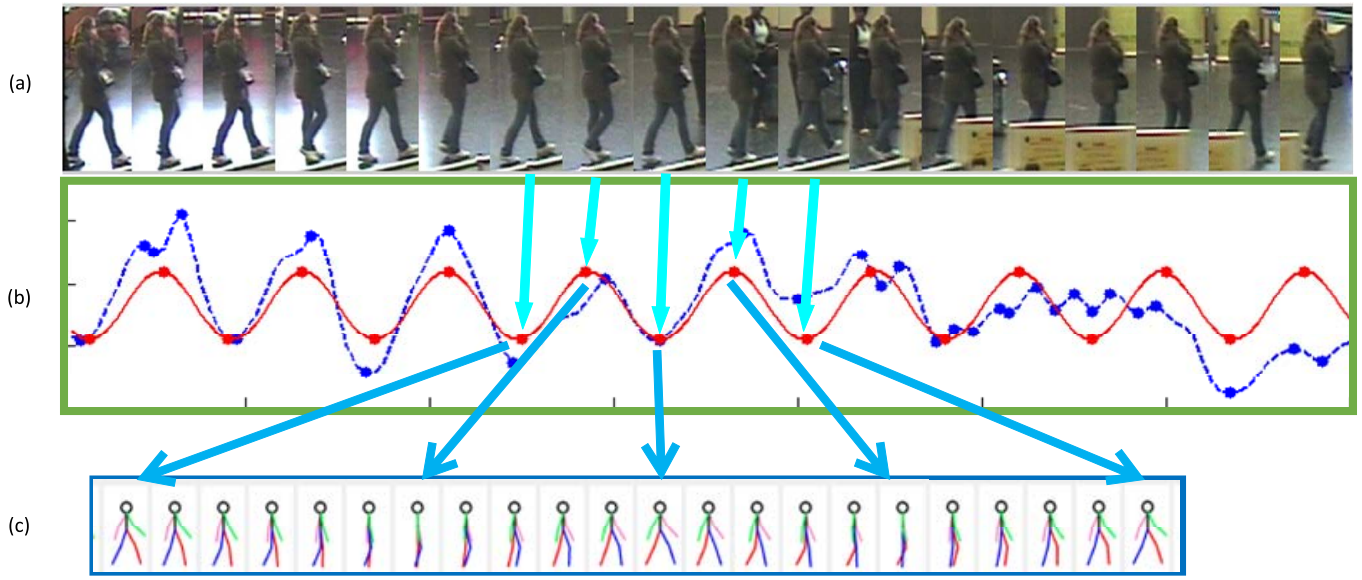


Fig. 2. Walking cycle extraction. (a) A video sequence of a pedestrian (only key frames). (b) The original FEP (blue curve) and the regulated FEP (red curve). (c) The stick figures illustrating the pedestrian poses extracted from a walking cycle. (Better viewed in color.)

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. Wang *et al.* [5] 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 [5], 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.

2) Fixed 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. 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). The template is empirically derived from the average image of the training set.

As shown in Figure 3, from the 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 persons appearance during walking. As depicted in Figure 3(b), the intensities of the bounding boxes around the whole body encode the N (e.g., $N = 4$) action primitives obtained by the temporal segmentation, while the colors of the smaller bounding boxes in Figure 3(c) encode the M (e.g., $M = 6$) body parts obtained by the spatial segmentation. Combining them, we obtain $M \times N$ spatially and temporally aligned video blobs, named body-action units, as shown in Figure 3(d):

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

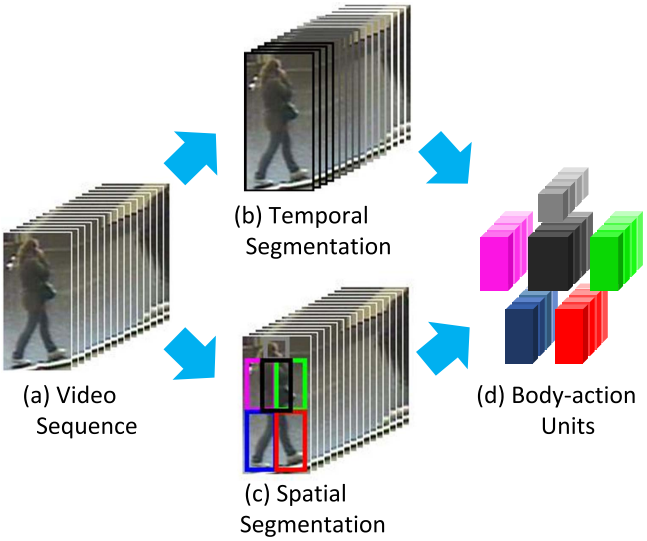


Fig. 3. (a) Input video sequence. (b) The temporal segmentation in which the intensities of the image bounding boxes encode the N (e.g., $N = 4$) action primitives. (c) The spatial segmentation in which the colors of the small bounding boxes encode the M (e.g., $M = 6$) body parts. (d) Fixed spatial-temporal body-action units in which different colors represent different body parts, and different intensities represent different action primitives. (Better viewed in color.)

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) Adaptive Body-Action Units: Ideally different frames may have different body part segmentation, i.e., P is dependent on time. This makes the fixed model inappropriate for the dynamic object, especially for the walking persons in videos. For example, when a pedestrian is observed from sideview, the poses of the swinging arms and legs always dependent on time. From this point, it is better to adopt temporal adaptive template for spatial alignment.

However, different pedestrians have different gait cycles. In order to extract proper body-action units for all cases, we intend to encode the temporal information obtained above (Section III-A.1) into the spatial segmentation. As shown in Figure 4(a), the proposed spatial segmentation of the fixed body-action units is proper for the body parts when the feet of the pedestrian both touch the ground. However, when the feet overlap each other as shown in Figure 4(b), the same template does not hold as in the former case. Hence, a new template is proposed to handle the spatial segmentation for this situation. As shown in Figure 4(c), the blue bounding box and the red bounding box overlap each other. Now, we have two different types of spatial segmentation for these two key action primitives.

In order to cover the whole walking cycle, we propose to generate temporal adaptive templates between the two key

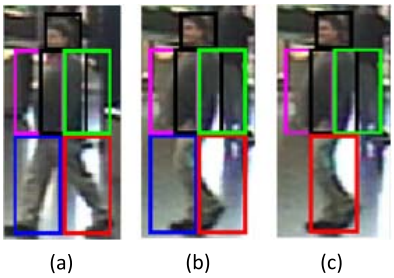


Fig. 4. Different spatial segmentation for the two key action primitives: (a) Spatial segmentation of the fixed body part model when feet both touch the ground; (b) Improper spatial segmentation when feet overlap each other; and (c) Adaptive body part model when feet overlapped (the blue bounding box and the red bounding box overlap each other).

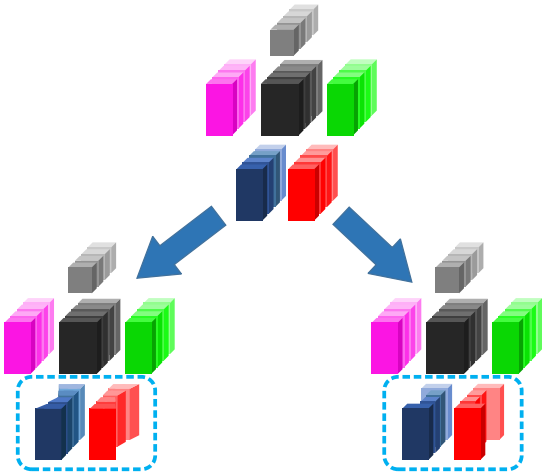


Fig. 5. Adaptive spatial-temporal body-action units based on the interpolation function (bottom left) and the step function (bottom right).

action primitives by two functions: interpolation function and step function. The interpolation function uniforms the variation of the templates between neighboring frames, while the step function sets the same templates based on the nearest key action primitive. Thus we obtain two precise and adaptive body action models according to the temporal segmentation results, as shown in Figure 5. The adaptive body-action units are more suitable for pedestrian representation and less sensitive to the background. Note that the poses of the upper body do not change obviously in a walking cycle, so we mainly adjust the lower body-action units for the adaptive body part model.

Besides, we also attempt to combine the walking cycle extraction with the classical spatial segmentation algorithm such as Deformable Parts Model (DPM), and compare it with our proposed body-action units. The performance of spatial segmentation using the above four methods during a walking cycle can be visualized in Figure 6. The comparison on re-id between DPM method and our proposed body-action units will be given in Section IV-C.

B. 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

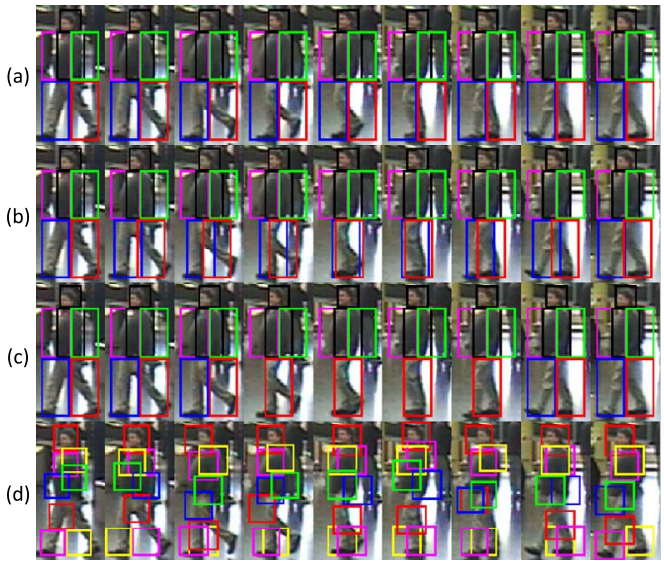


Fig. 6. Different spatial segmentation models. (a) The fixed body part model; (b) Adaptive body part model with the interpolation function; (c) Adaptive body part model with the step function; (d) Segmentation model with DPM.

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 [44] 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.

C. 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 [30]. For the re-id problem, however, a higher level of alignment accuracy is desirable. [5] 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.

D. Incorporating Supervised Learning

To obtain better performance we combine our appearance representation with supervised distance metric learning methods such as KISSME [45]. The KISSME method is based on the assumption that different pairwise features are Gaussian distributed, which acquired desirable performance for person re-identification on several databases. To make the paper self-sufficient, we give a brief introduction of KISSME in this section.

Given the pedestrian representation $\mathbf{x}$ with n_x feature vectors $\mathbf{x}_i : i = 1, \ldots, n_x$ from the query set and representation $\mathbf{y}$ with n_y feature vectors $\mathbf{y}_j : j = 1, \ldots, n_y$ from the gallery

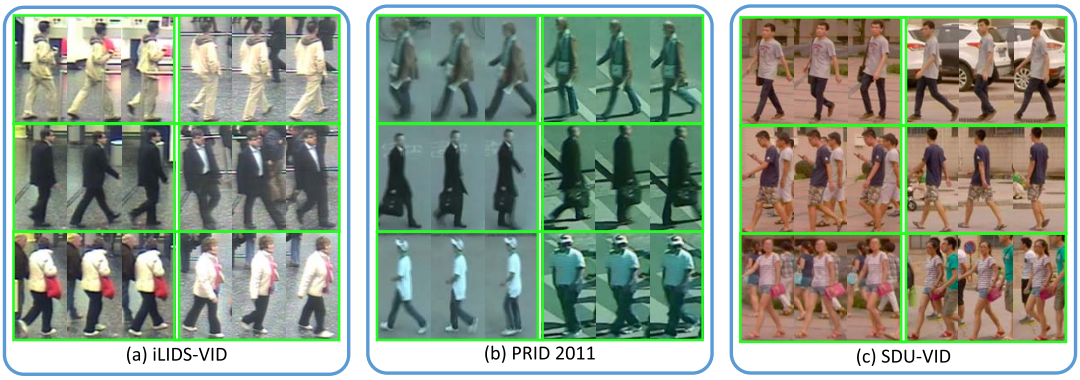


Fig. 7. Example pairs of the same people in different camera views from three person re-id datasets: a) iLIDS-VID dataset [5]; b) PRID 2011 dataset [48]; c) The newly introduced SDU-VID dataset.

set, the difference between the two feature vectors $\mathbf{x}_i$ and $\mathbf{y}_j$ is defined as $\mathbf{d}_{ij} = \mathbf{x}_i - \mathbf{y}_j$. Furthermore, we denote the covariance matrix of intrapersonal difference $\mathbf{d}_{ij}$ as $\mathbf{\Sigma}_I$, and the covariance matrix of extrapersonal difference $\mathbf{d}_{ij}$ as $\mathbf{\Sigma}_E$. So the logarithm of the ratio between the two feature vectors can be written as

$$\delta\left(\mathbf{d}_{ij}\right) = \mathbf{d}_{ij}^T \left(\mathbf{\Sigma}_I^{-1} - \mathbf{\Sigma}_E^{-1}\right) \mathbf{d}_{ij}. \tag{9}$$

In the experiment, the covariance matrices are estimated as follows:

$$\hat{\mathbf{\Sigma}}_I = \frac{1}{L_I} \sum_{l_{ij}=0} \mathbf{d}_{ij} \mathbf{d}_{ij}^T = \frac{1}{L_I} \sum_{l_{ij}=0} \left(\mathbf{x}_i - \mathbf{y}_j\right) \left(\mathbf{x}_i - \mathbf{y}_j\right)^T \tag{10}$$

$$\hat{\mathbf{\Sigma}}_E = \frac{1}{L_E} \sum_{l_{ij}=0} \mathbf{d}_{ij} \mathbf{d}_{ij}^T = \frac{1}{L_E} \sum_{l_{ij}=0} \left(\mathbf{x}_i - \mathbf{y}_j\right) \left(\mathbf{x}_i - \mathbf{y}_j\right)^T \tag{11}$$

where l_{ij} is defined as the indicated variable of $\mathbf{x}_i$ and $\mathbf{y}_j$: $l_{ij} = 1$ if $\mathbf{x}_i$ and $\mathbf{y}_j$ belong to the same person, otherwise $l_{ij} = 0$. We denote the number of same and different feature pairs as L_I and L_E respectively.

It is obviously that $\mathbf{\Sigma}_I$ and $\mathbf{\Sigma}_E$ are semi-positive definite matrices. Therefore, their eigenvalues are non-negative. Let $\mathbf{\Psi} = \hat{\mathbf{\Sigma}}_I^{-1} - \hat{\mathbf{\Sigma}}_E^{-1}$. Therefore, the derived distance function between $\mathbf{x}_i$ and $\mathbf{y}_j$ is

$$d(\mathbf{x}_i, \mathbf{y}_j) = \delta\left(\mathbf{d}_{ij}\right) = \mathbf{d}_{ij}^T \mathbf{\Psi} \mathbf{d}_{ij}, \tag{12}$$

where $\mathbf{\Psi}$ is the KISSME metric matrix.

In the comparison of the final representation, we investigate the classification performance using different distance measures. As depicted in Equation (13), the first classifier measures the minimum distance among all $n_x \cdot n_y$ feature pairs as the distance between them, namely the nearest neighbors distance. For the second classifier, we utilize the distance mentioned in [47]. This distance is defined as the average of the minimum distance for each feature vector, as shown in Equation (14).

$$d(\mathbf{x}, \mathbf{y}) = \min_{i,j} d(\mathbf{x}_i, \mathbf{y}_j). \tag{13}$$

$$d(\mathbf{x}, \mathbf{y}) = \frac{\sum_i \min_j d(\mathbf{x}_i, \mathbf{y}_j)}{2n_x} + \frac{\sum_j \min_i d(\mathbf{x}_i, \mathbf{y}_j)}{2n_y}. \tag{14}$$

TABLE I
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
SDU-VID	300	2	130	64×128

To interpret the second distance measure, let's first define the distance between a query feature vector $\mathbf{x}_i$ and representation $\mathbf{y}$ as $d(\mathbf{x}_i, \mathbf{y}) = \min_j d(\mathbf{x}_i, \mathbf{y}_j)$. The distance between the representations $\mathbf{x}$ and $\mathbf{y}$ is then computed by simply averaging these minimum distance in both direction. In the following experiments, it is found that this distance measure leads to superior performance combined with the distance metric learning.

IV. EXPERIMENTS

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

A. Datasets and Settings

Experiments were conducted on three person re-id datasets: iLIDS-VID dataset [5], PRID 2011 dataset [48] and the newly introduced dataset named SDU-VID, as shown in Figure 7. The details of the three datasets are depicted in Table I.

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 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.

2) 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,

TABLE II
PERFORMANCE OF DIFFERENT LOW-LEVEL DESCRIPTORS

Dataset	iLIDS-VID				PRID 2011				SDU-VID			
Rank R	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$
STFV3D(12)	27.0	55.7	71.6	84.7	42.1	71.9	84.4	91.6	58.0	81.3	86.7	89.3
STFV3D(24)	37.0	64.3	77.0	86.9	21.6	46.4	58.3	73.8	62.0	81.3	88.0	92.7
STFV3D+KISSME(12)	34.9	63.0	76.0	86.3	62.4	84.9	87.1	91.4	70.7	88.0	**95.3**	**96.0**
STFV3D+KISSME(24)	**44.3**	**71.7**	**83.7**	**91.7**	**64.1**	**87.3**	**89.9**	**92.0**	**73.3**	**92.7**	**95.3**	**96.0**

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.

3) SDU-VID Dataset: The SDU-VID dataset contains 600 image sequences for 300 randomly sampled people. Each image sequence has variable length consisting of 16 to 346 image frames, with an average number of 130 which is larger than those of iLIDS-VID and PRID 2011. It has two non-overlapping camera views captured in outdoor scenes. As stated above, our proposed method mainly focuses on the appearance representation utilizing the spatial and temporal alignment. The more frames the image sequence contains, the better results of walking cycle extraction we may obtain. However, the current datasets like iLIDS-VID and PRID 2011 contain relatively short sequences, which provide limited information in temporal. Thus we introduced the new dataset to measure our method¡¯s performance with more image frames in each video sequence. We also believe this is reasonable in some real-world applications where relatively longer sequences are captured for each pedestrian. Although this newly introduced dataset has more information than the two public datasets in temporal, it is still very challenging due to the cluttered background, occlusions and viewpoint variations across camera views. The dataset will be publicly available on the project webpage: http://www.vsislab.com/projects/MLAI/PedestrianRepresentation/.

4) 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 all datasets, the performance is measured by the average Cumulative Matching Characteristics (CMC) curves after 10 trails with Rank *R* rate which indicates the expectation of the true match being found within the first *R* ranks.

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.

B. 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 III-B) do not work well with such data. We performed a series experiments to investigate the effectiveness of the low-level descriptors. In Table II, 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 and SDU-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 [45], and repeated the experiments (the last two rows in Table II). As we expected, supervised learning can take care of feature selection quite well, and the 24-dimensional richer low-level descriptors perform better on all datasets. In the following experiments, we use the 12-dimensional descriptor on the PRID 2011 dataset when no supervised learning is employed.

C. Evaluation of the Body-Action Models

To validate the effects of various body-action models, we investigate the performance of different spatial and temporal segmentation methods.

The aforementioned four different spatial templates (in Section III-A.2 and Section III-A.3) are firstly evaluated combining with the proposed temporal segmentation. We denote the fixed body-action units as **STFV3D**, the adaptive body-action units based on interpolation as $\mathbf{S}_{int}$**TFV3D**, the adaptive body-action units based on step function as $\mathbf{S}_{stp}$**TFV3D**, and the units with DPM spatial segmentation as $\mathbf{S}_{dpm}$**TFV3D**. The experimental results are explicated in Table III. We can observe that the body-action units S_{int}TFV3D and S_{stp}TFV3D reach better performance

TABLE III
PERFORMANCE OF DIFFERENT SPATIAL TEMPLATES

Dataset	iLIDS-VID				PRID 2011				SDU-VID			
Rank R	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$
STFV3D	37.0	64.3	77.0	86.9	42.1	**71.9**	84.4	**91.6**	62.0	81.3	88.0	92.7
S_{int}TFV3D	**41.0**	67.0	**79.3**	88.7	43.0	70.6	**84.8**	90.9	**68.7**	85.3	90.0	93.3
S_{stp}TFV3D	39.0	**67.3**	**79.3**	**89.0**	**44.4**	71.0	**84.8**	90.6	65.3	84.0	**90.7**	93.3
S_{dpm}TFV3D	35.3	65.7	75.3	85.7	38.6	68.0	82.6	90.9	66.0	**86.7**	**90.7**	**94.7**

TABLE IV
PERFORMANCE OF DIFFERENT TEMPORAL SEGMENTATIONS

Dataset	iLIDS-VID				PRID 2011				SDU-VID			
Rank R	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$
STFV3D	37.0	64.3	77.0	86.9	42.1	71.9	84.4	91.6	62.0	81.3	88.0	92.7
ST_{gt}FV3D	41.7	70.7	79.7	90.3	46.2	75.7	84.8	91.7	62.0	84.7	89.3	96.7
STFV3D+KISSME	44.3	71.7	83.7	91.7	64.1	87.3	**89.9**	**92.0**	73.3	92.7	95.3	96.0
ST_{gt}FV3D+KISSME	**47.0**	**77.0**	**88.0**	**94.0**	**69.8**	**87.6**	89.1	90.9	**84.0**	**97.3**	**98.0**	**98.0**

TABLE V
PERFORMANCE OF DIFFERENT DISTANCE MEASURES

Dataset	iLIDS-VID				PRID 2011				SDU-VID			
Rank R	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$
STFV3D(C_{nn})	37.0	64.3	77.0	86.9	42.1	71.9	84.4	**91.6**	62.0	81.3	88.0	92.7
STFV3D(C_{avg})	35.7	65.3	77.0	87.0	41.5	71.7	84.1	90.6	65.3	85.3	92.0	93.3
STFV3D+KISSME(C_{nn})	44.0	73.7	82.7	90.7	64.4	85.8	88.0	88.7	73.3	92.7	**95.3**	**96.0**
STFV3D+KISSME(C_{avg})	**49.7**	**78.3**	**84.7**	**91.7**	**66.2**	**87.3**	**88.4**	89.4	**85.3**	**94.7**	**95.3**	**96.0**

than STFV3D, especially for the iLIDS-VID and SDU-VID datasets. This is mainly because the adaptive body-action units change its spatial template according to the extracted walking cycle and thus can produce more precise body parts. The improvement of S_{int}TFV3D and S_{stp}TFV3D on the PRID 2011 dataset is not as impressive as the other two datasets in which the inconsistent video sequences bring trouble to the walking cycle extraction. The body-action units with spatial segmentation based on DPM, on the other hand, only work well on the new dataset. This is partly because the variation of the viewpoints on the first two datasets make the DPM difficult to describe pedestrian appearance. DPM may also lack the coherence between neighboring frames for each blob. Hence, with more proper spatial segmentation to obtain more precise body parts of pedestrians, it may reach superior performance. We will carry on our work based on STFV3D units for its simplicity in the following experiments.

Next, we investigate the influence of the temporal segmentation. The fixed body-action units utilizing the proposed walking cycles extraction **STFV3D** is compared to the one using the groundtruth labeled walking cycles **ST_{gt}FV3D**. We then repeat the experiments combining these methods with the KISSME distance metric learning algorithm. As shown in Table IV, from the first two rows we can find that ST_{gt}FV3D with labeled walking cycles has superior results than STFV3D with extracted walking cycles. This means that if the walking cycles of pedestrians are available, our method could reach a higher result. The same conclusion can be obtained from the last two rows of the table combining with the KISSME method. However, due to the innumerable video sequences, it is difficult to obtain the groundtruth of pedestrians' walking cycles. Thus our proposed temporal segmentation is more feasible in practice.

D. Evaluation of the Distance Measures

As stated in Section III-D, we intend to test two classifiers with different distance measures in pedestrian re-id, i.e., the nearest neighbor classifier (denoted as C_{nn}) and the averaging neighbor classifier (denoted as C_{avg}). It is observed that, without combining distance metric learning, the latter classifier C_{avg} produced competitive results to that of the former one C_{nn}. As shown in Table V, after combining the supervised distance metric learning, the performance of the classifiers C_{nn} and C_{avg} can be improved significantly, and the superior performance is obtained by using the classifier C_{avg}. However, to make a fair comparison with other methods, we employ the nearest neighbor classifier C_{nn} for re-id in the following experiments.

E. 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 [5]. 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.

FV2D, which is a multi-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 [13].

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

TABLE VI
COMPARISON OF DIFFERENT FEATURE DESCRIPTORS

Dataset	iLIDS-VID				PRID 2011				SDU-VID			
Rank R	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$
HOG3D	8.3	28.7	38.3	60.7	20.7	44.5	57.1	76.8	10.7	22.7	34.0	44.0
FV2D	18.2	35.6	49.2	63.8	33.6	64.0	76.3	86.0	46.0	78.0	86.7	**94.0**
FV3D	25.3	54.0	68.3	**87.3**	38.7	71.0	80.6	90.3	50.7	78.0	86.7	92.7
TFV3D	27.3	58.7	68.7	**87.3**	39.6	69.5	82.5	89.8	52.3	78.0	86.7	91.3
SFV3D	34.0	61.3	70.7	84.7	41.8	71.4	83.5	91.2	57.3	**82.7**	86.7	92.0
STFV3D	**37.0**	**64.3**	**77.0**	86.9	**42.1**	**71.9**	**84.4**	**91.6**	**62.0**	81.3	**88.0**	92.7

TABLE VII
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 [23]	2.8	13.1	21.3	34.5	-	-	-	-
HOG3D+DVR [37]	23.3	42.4	55.3	68.4	28.9	55.3	65.5	82.8
Color+LFDA [28]	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

TFV3D, which is the extension of FV3D with temporal alignment of the walking sequence only.

SFV3D, which is the extension of FV3D with spatial alignment of the walking sequence only.

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 VI. From the results we can observe that in general STFV3D performs the best, and more specifically:

1) 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).

2) Temporal Alignment vs. Spatial Alignment: Both SFV3D and TFV3D outperform FV3D, which demonstrates the effectiveness of the proposed spatial and temporal alignment for a walking sequence. It is interesting to see that SFV3D achieves better performance than TFV3D, which implies that the spatial alignment is more effective than the temporal alignment. This is probably because temporal alignment is more challenging, as it is dealing with the different phases of an action in multiple shots. By combining the spatial and temporal alignment together, STFV3D outperforms both SFV3D and TFV3D, which utilize the spatial or temporal alignment only, respectively.

3) Video-Based Approaches vs. Independent Multi-Shot: 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.

4) FV3D vs. HOG3D: Both 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.

5) iLIDS-VID vs. PRID 2011 vs. SDU-VID: The above observations hold for all datasets. We would like to point out again that, for the PRID 2011 dataset, we only used the 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 that can learn the relationship between two cameras. Nonetheless our proposed appearance modeling still shows its merit in an unsupervised manner. This is particularly important when dealing with the videos from multiple cameras unseen before. Another notable difference among the results is that FV2D performs better on the PRID 2011 and SDU-VID datasets than on the iLIDS-VID dataset, considering its relative performance to the 3D approaches. This is probably because the iLIDS-VID dataset has more cluttered background and considerable occlusions, which increase the re-id difficulty for 2D approaches.

F. 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 on the two public datasets. To achieve the best performance we combine STFV3D with supervised distance metric learning methods such as KISSME [45] and Local Fisher Discriminant Analysis (LFDA [49]). 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 VII, the first three rows show the performance of the state-of-the-art approaches, namely, Gait Energy Image (GEI)+Rank SVM (RSVM) [50], HOG3D+Discriminative Video Ranking (DVR) [5], Color+LFDA [49]. The second and third group of methods are variants of our proposal. From these results we can see that distance metric learning can

TABLE VIII

COMPARISON OF THE PROPOSED METHODS WITH AND WITHOUT DISTANCE METRIC LEARNING ON SDU-VID

Dataset	SDU-VID			
Rank R	$R=1$	$R=5$	$R=10$	$R=20$
FV3D	50.7	78.0	86.7	92.7
STFV3D	62.0	81.3	88.0	92.7
FV3D+KISSME	62.7	88.7	94.0	**97.3**
STFV3D+KISSME	**73.3**	**92.7**	**95.3**	96.0

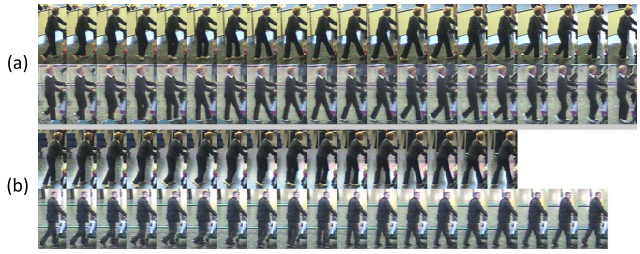


Fig. 8. Failure example 1.

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. Besides, we also test the performance of the proposed STFV3D and FV3D, with and without distance metric learning KISSME on the new SDU-VID dataset. Similar conclusions can be obtained as shown in Table VIII.

G. Limitations and Failure Examples

Finally we discuss the limitations of our approach, and show some failure examples (Figure 8 and Figure 9). Each figure contains the matching results of the same person 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 temporally based on walking cycle extraction. In both cases, FV3D finds the correct match while STFV3D does not. In Figure 8, the color inconsistency is causing trouble for both representations, and the matching is probably more affected by pose and shape. Figure 8(b) shows that the cluttered background causes inaccurate walking cycle extraction in STFV3D, and hence incorrect matching between the query and gallery. In Figure 9, 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 9(b)).

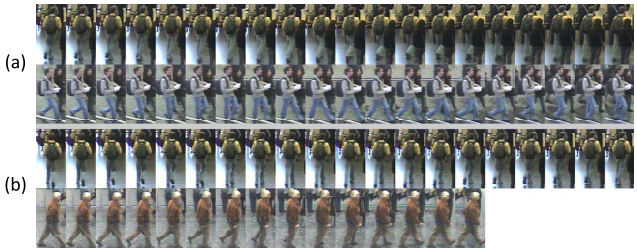


Fig. 9. Failure example 2.

V. 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, and the performances of the proposed adaptive body-action units appear to be more impressive, while the final representation is still 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 the pedestrian's poses can be evaluated from the input videos and the appearance representation can be built based on the extracted poses, which is more effective for the video-based re-id problem.

REFERENCES

[1] R. Satta. (2013). "Appearance descriptors for person re-identification: A comprehensive review." [Online]. Available: https://arxiv.org/abs/1307.5748

[2] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2006, pp. 1528–1535.

[3] A. Bedagkar-Gala and S. K. Shah, "Multiple person re-identification using part based spatio-temporal color appearance model," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 1721–1728.

[4] S. Bąk, G. Charpiat, E. Corvee, F. Brémond, and M. Thonnat, "Learning to match appearances by correlations in a covariance metric space," in *Computer Vision—ECCV*. New York, NY, USA: Springer, 2012, pp. 806–820.

[5] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Computer Vision—ECCV*. New York, NY, USA: Springer, 2014, pp. 688–703.

[6] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image Vis. Comput.*, vol. 32, no. 4, pp. 270–286, 2014.

[7] L. Yang, "Distance metric learning: A comprehensive survey," Michigan State Univ., Lansing, MI, USA, Tech. Rep., vol. 2, no. 2, 2006.

[8] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Appearance-based person reidentification in camera networks: Problem overview and current approaches," *J. Ambient Intell. Humanized Comput.*, vol. 2, no. 2, pp. 127–151, 2011.

[9] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3–19, 2013.

[10] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person Re-Identification*. New York, NY, USA: Springer, 2014.

[11] B. Ma, Y. Su, and F. Jurie, "BiCov: A novel image representation for person re-identification and face verification," in *Proc. Brit. Mach. Vis. Conf.*, 2012, p. 11.

[12] S. Bąk, E. Corvee, F. Bremond, and M. Thonnat, "Boosted human re-identification using Riemannian manifolds," *Image Vis. Comput.*, vol. 30, nos. 6–7, pp. 443–452, 2012.

[13] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by Fisher vectors for person re-identification," in *Computer Vision—ECCV 2012. Workshops and Demonstrations*. New York, NY, USA: Springer, 2012, pp. 413–422.

[14] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. BMVC*, vol. 2. 2010, p. 6.

[15] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 649–656.

[16] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 937–940.

[17] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2360–2367.

[18] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu, "Human re-identification by matching compositional template with cluster sampling," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3152–3159.

[19] W. Ouyang, X. Zeng, and X. Wang, "Learning mutual visibility relationship for pedestrian detection with a deep model," *Int. J. Comput. Vis.*, vol. 120, no. 1, pp. 14–27, 2016.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[21] W. Zhang, C. Qu, L. Ma, J. Guan, and R. Huang, "Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network," *Pattern Recognit.*, vol. 59, pp. 176–187, Nov. 2016.

[22] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, "Multimodal learning for facial expression recognition," *Pattern Recognit.*, vol. 48, no. 10, pp. 3191–3202, 2015.

[23] A.-R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[24] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA, USA: MIT Press, vol. 3361. 1995.

[25] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[26] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, Oct. 2015.

[27] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio, "Full-body person recognition system," *Pattern Recognit.*, vol. 36, no. 9, pp. 1997–2006, 2003.

[28] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 898–903, 2012.

[29] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. 15th Int. Conf. Multimedia*, 2007, pp. 357–360.

[30] A. Klaser, M. Marszaek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. 19th Brit. Mach. Vis. Conf. (BMVC)*, 2008, pp. 1–275.

[31] N. V. Boulgouris, D. Hatzinakos, and K. N. Plataniotis, "Gait recognition: A challenging signal processing technology for biometric identification," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 78–90, Nov. 2005.

[32] M. S. Nixon and J. N. Carter, "Automatic recognition by gait," *Proc. IEEE*, vol. 94, no. 11, pp. 2013–2024, Nov. 2006.

[33] J. Wang, M. She, S. Nahavandi, and A. Kouzani, "A review of vision-based gait recognition methods for human identification," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, 2010, pp. 320–327.

[34] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1505–1518, Dec. 2003.

[35] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[36] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 149–158, Feb. 2004.

[37] C. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, 2004.

[38] A. Bedagkar-Gala and S. K. Shah, "Gait-assisted person re-identification in wide area surveillance," in *Computer Vision—ACCV 2014 Workshops*. New York, NY, USA: Springer, 2014, pp. 633–649.

[39] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 591–606, Mar. 2015.

[40] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2528–2535.

[41] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 356–370, Feb. 2017.

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

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

[44] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[45] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2288–2295.

[46] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3810–3818.

[47] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 945–953.

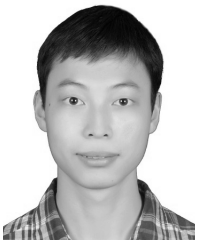
[48] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*. New York, NY, USA: Springer, 2011, pp. 91–102.

[49] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3318–3325.

[50] R. Martín-Félez and T. Xiang, "Gait recognition by ranking," in *Computer Vision—ECCV*. New York, NY, USA: Springer, 2012, pp. 328–341.

Wei Zhang (S'06–M'11) received the Ph.D. degree in electronic engineering from The Chinese University of Hong Kong in 2010. He is currently an Associate Professor with the School of Control Science and Engineering, Shandong University, China. He has authored about 40 papers in international journals and refereed conferences. His research interests include computer vision, image processing, pattern recognition, and robotics. He served as a Program Committee Member and a Reviewer for various international conferences and journals in image processing, computer vision, and robotics.

Bingpeng Ma received the B.S. degree in mechanics and the M.S. degree in mathematics from the Huazhong University of Science and Technology in 1998 and 2003, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 2009. He was a Post-Doctoral Researcher with the University of Caen, France, from 2011 to 2012. He joined the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, in 2013, and he is currently an Associate Professor. His research interests cover computer vision, pattern recognition, and machine learning. He especially focuses on face recognition, person re-identification, and the related research topics.



Kan Liu (S'13) received the B.S. degree in control science and engineering from the Huazhong University of Science and Technology in 2011 and the M.S. degree in control science and engineering from Shandong University in 2016. He is currently pursuing the Ph.D. degree with the Tsinghua-Berkeley Shenzhen Institute, Tsinghua University. His research interests include computer vision, image processing, pattern recognition, and computational photography.

Rui Huang received the B.Sc. degree from Peking University in 1999, the M.Eng. degree from the Chinese Academy of Sciences in 2002, and the Ph.D. degree from Rutgers University in 2008. He was a Post-Doctoral Researcher with Rutgers University, before going back to China in 2010, and joined the Huazhong University of Science and Technology as a Faculty Member. From 2012 to 2016, he was a Research Scientist with NEC Laboratories China. He is currently an Associate Professor with The Chinese University of Hong Kong, Shenzhen. He has been involved in various research topics, including subspace analysis, deformable models, probabilistic graphical models, and their applications in computer vision, pattern recognition, and medical image analysis. He has authored over 50 papers in related areas and has been the Principal Investigator of various research grants. His current research interests include machine learning methods for intelligent video surveillance and computer vision for robotics.