# Long-term Person Re-identification using True Motion from Videos

Peng Zhang Qiang Wu Jingsong Xu Jian Zhang
School of Electrical and Data Engineering, University of Technology Sydney
Peng.Zhang-2@student.uts.edu.au, {Qiang.Wu, Jingsong.Xu, Jian.Zhang}@uts.edu.au

## Abstract

*Most person re-identification approaches and benchmarks assume that pedestrians go across the surveillance network without significant appearance changes in a brief period, which explicitly restricts person re-identification to a short-term event and incurs inter-sample similarity measurement by appearance matching. However, pedestrians are likely to reappear in the surveillance network after a long-time interval (long-term) and change their wearing in many real-world scenarios. These scenarios inevitably cause appearances between subjects more ambiguous and indistinguishable. In this paper we consider these scenarios and propose a unified feature representation based on true motion cues from videos named FIne moTion encoDing (FITD). Our hypothesis is that people keep constant motion patterns under non-distraction walking condition. Therefore, the motion characteristics are more reliable than static appearance feature to describe a walking person. Particularly, we extract motion patterns hierarchically by encoding trajectory-aligned descriptors with Fisher vectors in a spatial-aligned pyramid. To verify benefits of the proposed FITD, we collect a new dataset typically for the long-term situations. Extensive experiments demonstrate the merits of our FITD especially for the long-term scenarios.*

## 1. Introduction

Locating and identifying a target pedestrian across a network of surveillance cameras at a distinct time termed as person re-identification (Re-ID) [1, 2, 5, 6, 14, 25, 27, 31], is a critical task due to its applications such as inter-camera retrieval and video monitoring. The current literature generally limit Re-ID to a short-term event by assuming that the target subject passes the surveillance network in a short period [2, 6, 14, 25, 27, 31]. However, the target of interest may reappear in the network after a long-time gap, *e.g.*, several days or even several months later, in many realistic scenarios. A typical example is that pedestrians would pass through the same or different gates in an office block or subway station on various days. We term Re-ID in these scenarios as *long-term person Re-ID*. In this paper, we thus focus on the long-term person Re-ID issue.

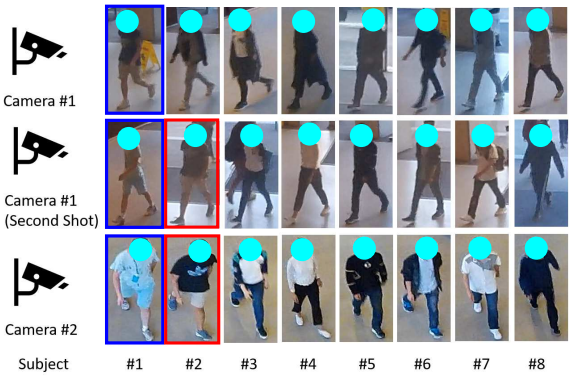


Figure 1. Illustration of long-term person Re-ID challenges. Each line of images is collected from same camera, and each column of images belongs to the same subject. Images from last two rows are captured by a long-time interval with the top row.

Currently, most of the state-of-the-art Re-ID works concentrate on the short-term Re-ID, which yields methods building on the fact that the target in the query and its corresponding gallery set are in the similar appearance, *e.g.*, clothing color and texture. This is reasonable because people rarely change their clothing within a short interval. Regarding the fact, different appearance based features have been exploited, *e.g.*, color histograms [7, 26, 31], local binary patterns (LBP) [20], ensemble of local features (ELF) [5], local maximal occurrence (LOMO) [14], spatial-temporal information such as STFV3D [15] and HOG3D [25, 27], and some high-level features learned by CNN models [8, 30]. These features achieve significant performance along with the typical metric learning models [7, 14, 27, 26] on the existing benchmark datasets. Under the long-term scenarios, however, intra-class appearances suffer larger variations, and positive pairs are easier to be imposed by other subjects due to the dressing change. Intra-class cluster based on appearance becomes more dispersive while inter-class cluster based on appearance is more ambiguity than the short-term Re-ID. Features

based on visual appearances, such as color and texture histograms are at a disadvantage when matching the same subject with distinct clothes from not matter the same camera or different cameras (see Figure 1 query subject bounded by blue line at top row is in a dark T-shirt and light short pant, but the positive matching in the gallery set at the rest rows are in blue clothing which is easily imposed by subject #2 in the gallery causing a mismatch). It implies that different feature descriptors using fewer appearance cues are essential for the Re-ID issue in long-term scenarios.

Inspired by the success of soft biometrics in gait and activity recognition across views [9, 24], we formulate a FIne moTion encoDing (FITD) model based on dynamic cues. The proposed FITD is true motion information extracted from dense trajectories, which characters dynamic motion patterns of the human body from raw footage without any scalability normalization. Especially, we adopt a patch-wise strategy [2, 13, 15, 18, 21] which divides human body into several fundamental body-action primitives. Fisher vectors [22] are then utilized to respectively summarize the trajectory-aligned descriptors, *e.g*. Histograms of Optical Flow (HOF) [10] and Motion Boundary Histogram (MBH) [23], within each body-action unit (comprised by the fundamental body-action primitives) in the predefined body-action pyramid model. By this, both local and global motion statistics are computed. And, the final unified motion representation FITD is obtained by concatenating the bag of visual descriptors from all the body-action units.

To evaluate merits of the proposed FITD descriptors, we constructed a novel long-term Re-ID dataset named Motion-ReID. As we know, this is the first dataset available for the typical long-term Re-ID task. Different from the previous benchmark Re-ID datasets, the dataset is collected using real surveillance cameras in a building block rather than self-deployed cameras and contains pairwise samples of the same subject recorded with a long timespan, *e.g*., one week or more. Thus, most of the samples are with various clothes and carrying conditions, which are typically realistic scenarios of long-term Re-ID.

To summarize, the main contributions of this paper are (1) This paper addresses Re-ID of long-term case (*i.e*., re-identifying a person after a long time intereval) which is seldom discussed in the area. (2) We develop a novel FITD model characterizing motion patterns pyramidally from both global and local body action units, which achieves significant performance for the long-term Re-ID task and aids for the classical short-term Re-ID problem. (3) We propose a novel dataset for the general long-term Re-ID problem, and this is the only available one so far. Besides, a comprehensive evaluation of the performance of popular feature representations is conducted on both short-term and long-term Re-ID tasks.

## 2. Related Work

Typically, most pipelines for person Re-ID include two major components: feature representation [5, 14, 15, 4] and metric learning [14, 26, 27]. We refer to [1, 29] for detailed reviews about classical works on this topic. Here we only concentrate on literature that is most related to our work.

In the previous works, significant efforts have been made to develop or learn better features that are at least partially robust to illumination changing, viewpoint variation, pose indeterminacy, *etc*. These features are roughly divided into two categories: image-based features and video-based features. Image-based features which are usually generated from one or multiple discontinuous frames, *e.g*., Xiong *et al*. [26] utilized fusion features of RGB, YUV, HSV color-based histograms and LBP to evaluate the effect of different kernels embedded in metric learning algorithms for Re-ID. To overcome illumination and background variations, Ma *et al*. [16] proposed a feature representation that characterized texture information using covariance descriptors.

On the other hand, video-based features [25, 27, 15] are usually extracted from consecutive walking sequences, *e.g*., both Wang *et al*. [25] and You *et al*. [27] applied HOG3D to extract spatial-temporal information from walking pedestrians. After combining them with color or texture histograms, they improved the performance with a large margin. Considering space-time alignment problem, Liu *et al*. [15] proposed a spatial-temporal appearance representation named STFV3D which encodes local descriptors by Fisher Vector with respect to a body-action unit model. The STFV3D exactly solves body misalignment caused by viewpoint change to some content. However, it is worth pointing out that both above-mentioned features in current works are closely relevant to people's appearance, which will be unreliable once one's appearance is drastically changed such as matching the same individual with different clothing or carrying conditions.

Our proposed FITD belongs to the video-based feature, which is inspired by the success of dense trajectory on activity recognition [24]. Dense trajectories depict displacement information of space-time interest points which has demonstrated momentous success on action description combining with encoding descriptors such as HOG, HOF and MBH. In person Re-ID, Gou *et al*. [4] extracted soft biometrics from motion by encoding short trajectories with Hankel matrix, which has shown advantage against appearance impaired case. However, their trajectories extracted from normalized bounding areas only indicate pixel displacement between normalized areas of interest, which are not real motions.

In contrast to aforementioned methods, our FITD leverages trajectory-based true motion patterns from raw video volumes, and trajectory-aligned descriptors are embedded before Fisher encoding to get more robust motion information. Besides, a body-action pyramid model is considered

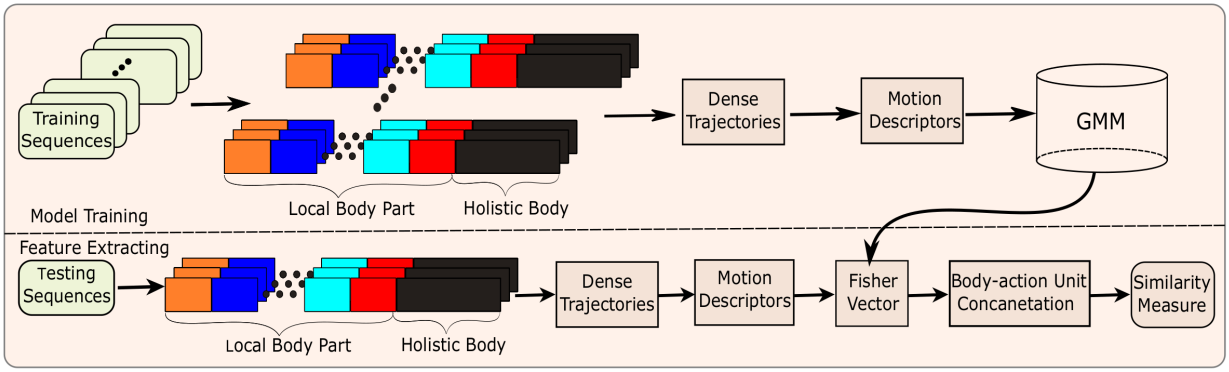


Figure 2. Framework of the proposed FITD model.

to obtain both global and local motion information to boost feature discriminability for Re-ID tasks.

## 3. Fine Motion Encoding

In this section, we present a novel spatio-temporal motion representation for person Re-ID specific in long-term scenarios. As depicted in Figure 2, the proposed framework includes two phases at which *model-training* learns a feature codebook consisting of discriminative motion primitives and *feature-extraction* encodes motion vocabularies to generate unified feature vectors. Particularly, each stage is performed on the basis of trajectory-aligned motion statistics with respect to body-action units corresponding to various levels of motion primitives in the predefined body-action pyramid (see section 3.1).

### 3.1. Body-action Pyramid Model

The human body is a non-rigid object which generates complex movement traces with respect to its flexible joints while walking. This causes movement of the human body varies from part to part as in Figure 3, which would suffer great losses of local information if we only consider motions of human body from a global view. Due to this, our body-action pyramid model (BPM) takes motion patterns of body-action units from various levels into account.

Inspired by successes of some body-part based models [15, 18, 13], we define a BPM which divides human body from coarse to fine sub-regions with respect to some prior knowledge of geometry structure and kinematical characteristics of the human body. In specific, we depict the entire human body in three levels, each of which includes a unique number of patches corresponding to various combinations of the neighbored action primitives, *i.e.* head (20%), upper torso (20%), lower torso (15%), upper leg (15%) and lower leg (30%). The action primitive template is empirically derived on the basis of the spatial structure of walking pedestrians from multiple benchmarks and fine-tuned in terms of motion characteristics. As shown in Figure 3, the top level depicts the entire body, which is divided into two horizontal strips locating upper body (55%) and leg (45%) respectively due to the motion characteristics. Further, the upper body is subdivided into three sections corresponding to the first three action primitives whilst the leg section is subdivided into two parts corresponding to the last two action primitives, so as to characterize the motion patterns in a finer way. The total eight parts from three levels compromise our BPM.

From above segmentation of the input video sequence, the patches corresponding to pyramids of body parts are subsequently divided into eight body-action units, as shown in Figure 3,

$$\begin{aligned} P_m &= \{(x_t, y_t) | (x_t, y_t) \in P_{m,t}\}, \\ m &= 1, 2, \cdots, 8; t = 1, 2, \cdots, N \end{aligned} \tag{1}$$

where $P_{m,t}$ denotes the $m$-th patch of the $t$-th frame, $(x_t, y_t)$ is the absolute position in the input video sequence.

In practice, the obtained patches are with irregular size with respect to tracking and annotation results. And they are just used for restricting regions and identifying whether an untracked feature point in the region was appended to the tracking process. Usage about the BPM will be introduced in the next section.

### 3.2. Motion Trajectories for Re-ID

To capture motion patterns of a walking pedestrian, we extract dense trajectories with respect to each body-action unit separately. For ease of discussion, we take one single unit as an example. Inspired by the framework of dense trajectories [24], we tracked the sampled feature points $(x_t, y_t) \in P_{m,t}$ to the next frame $t+1$ in a dense optical flow field $\omega = (\mu_t, \nu_t)$.

$$\begin{aligned} (x_{t+1}, y_{t+1}) &= (x_t, y_t) + (M * \omega)|_{(\bar{x}_t, \bar{y}_t)} \\ s.t. (x_{t+1}, y_{t+1}) &\in padding(P_{m,t+1}) \end{aligned} \tag{2}$$

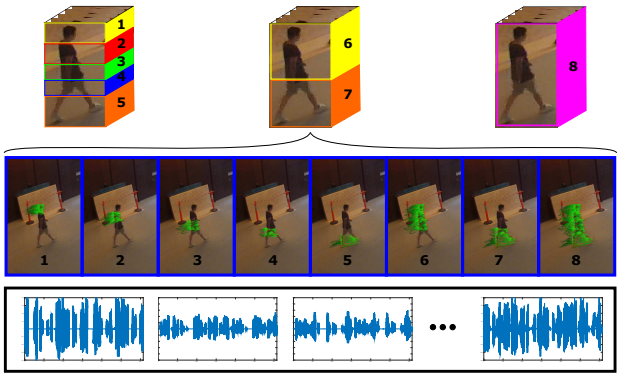


Figure 3. Top: our proposed body-action pyramid model consists of eight body-action units, which is labelled from 1 to 8. Middle: dense trajectories in each body-action unit. Bottom: Fisher vectors correspond to the eight units. **Best viewed in color.**

where $padding(P_{m,t+1})$ denotes extending the patch $P_{m,t+1}$ by padding pixels from the neighbored areas, here we set to extend with extra horizonal strips of 10% of total body height, $M$ is the filter kernel, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of $(x_t, y_t)$.

It is worth noting that our feature point detection and tracking process are restricted to different areas. By this, feature points in the body-action units are not only completely tracked but also avoided to drifting too much. Since the tracking process is only related to optical flow, static trajectories related to homogeneous image areas (background) are pruned. Moreover, we consider bounding area of interest in our framework, which largely suppresses effects of other moving objects and generates pure trajectories of target pedestrian. As shown in Figure 3, dense trajectories are mainly spanned on pre-defined body-action units of the moving pedestrian. These motion trajectories model abundant of soft-biometric characteristics which provide distinctiveness of distinct human motion patterns such as walking speed and stride.

### 3.3. Trajectory-aligned Motion Statistics

Local descriptors have been proved effectiveness to character motion information in dense trajectories for many activity recognition approaches. To utilize motion information in our trajectories, we consider both HOF and MBH, which are popular to represent action characteristics. When formulating the above descriptors, we follow the setting in [24] to describe descriptors around the trajectories in a space-time volume.

In practice, HOF is applied to estimate local motion information with 9 bins covering all the orientations. It well describes the latent motion cues of people's walking styles, due to its properties that HOF is invariant to directions of motion and scalability. While HOF describes absolute motion, MBH encodes the relative motion between feature points. In particular, MBH descriptor treats horizontal (MBHx) and vertical (MBHy) components of optical flow separately, which yields motion information in both directions. This implicitly reflects motion boundaries in both directions differently, as we know, motions are drastic in horizontal direction while subtle in the vertical direction. In this paper, we equally quantize orientations into 8 bins for each component, and totally two 96 dimensional descriptors are obtained. MBH expresses human walking effectively which yields excellent Re-ID performances, especially the horizonal component as shown in section 5.

Other than the above two pure motion-based encoding descriptors, HOG [3] is also considered due to its powerful gradient representation. However, HOG is commonly considered as a type of appearance-based feature. Different from the previous methods which apply HOG directly to an image, we implement it to a space-time video volume around dense trajectories embedding to our model. This is implicitly related to human's motion patterns.

### 3.4. Fisher Vector Encoding of Motions

The Fisher Vector [22] was first proposed to describe an image for large-scale visual classification and has gained remarkable success in many applications, *e.g.*, activity recognition, image retrieval and even person Re-ID. Given a body-action unit $P_m$ in the proposed BPM, we describe the unit of a sample with $N$ aforementioned descriptors, denoting as $X = \{x_n | x_n \in R^D, n = 1, \ldots, N\}$. To make the descriptors compact, we model the descriptors with $K$ probabilistic visual vocabularies (PVVs) which make the body-action unit complying a distinct distribution $P(X|\Theta)$, where $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$ is the parameters for the $K$ PVVs. In this paper, the $K$ PVVs are learned by a Gaussian Mixture Model (GMM) $\Psi$ with parameters $\theta_k = \{\mu_k, \sigma_k^2, \rho_k\}$, where $\mu_k$, $\sigma_k$, $\rho_k$ are respectively the mean vector, standard deviation and mixture weight,

$$\Psi = \sum_{k=1}^{K} \rho_k \psi_k, s.t. \rho_k \geq 0, \sum_{k=1}^{K} \rho_k = 1 \tag{3}$$

$$\psi_k(x; \mu_k, \sigma_k) = \frac{1}{(2\pi)^{D/2}|\sigma_k|} exp\{-\frac{1}{2}\|\sigma_k^{-1}(x - \mu_k)\|_2^2\} \tag{4}$$

where $\psi_k$ is the $k$-th Gaussian component, $\sigma_k$ is a diagonal matrix, and $x$ is the feature descriptor mentioned above such as HOF and MBH. Once the GMMs are obtained, the Fisher vector of the sample in the body-action unit $P_m$ is a concatenation of the deviations $\alpha_k^X$, $\mu_k^X$ and $\varepsilon_k^X$, i.e., $f(X) = [\alpha_1^X; \mu_1^X; \varepsilon_1^X; \ldots; \alpha_K^X; \mu_K^X; \varepsilon_K^X]$, where

$$\alpha_k^X = \frac{1}{N\sqrt{\rho_k}} \sum_{n=1}^{N} (\gamma_{nk} - \rho_k) \tag{5}$$

$$\mu_k^X = \frac{1}{N\sqrt{\rho_k}} \sum_{n=1}^{N} \gamma_{nk} \rho_k^{-1} (x_n - \mu_k) \quad (6)$$

$$\varepsilon_k^X = \frac{1}{N\sqrt{2\rho_k}} \sum_{n=1}^{N} \gamma_{nk} \{\sigma_k^{-2} (x_n - \mu_k)^2 - e\} \quad (7)$$

here $\gamma_{nk}$ is the posterior probability which determines whether descriptor $x_n$ is generated by the $k$-th component or not, $e$ is a $D$ dimensional vector whose elements are all 1. By concatenating Fisher vectors in of all the body-action units, we obtained our final high-level feature which depicts human's motion characteristics in a fine-grained way.

### 3.5. Feature Fusion

Feature fusion plays a significant role when multiple features are available. One strategy is concatenating our descriptors in the feature level, which uses the mixed feature to train the GMM model and generate our Fisher vectors.

Another fusion strategy is aggregating the similarity metrics in the score-level, which refers to sum up the weighted similarity scores from various descriptors,

$$s_j = \sum_k \omega_k s_j^{(k)}, \forall_k : \omega_k \geq 0, \sum_k \omega_k = 1 \quad (8)$$

where $s_j^{(k)}$ is the similarity score between query sample and the $j$-th gallery sample with the k-th descriptor, $\omega_k$ weights the contribution of the $k$-th descriptor. For the sake of ease, we leverage Euclidean distance as our similarity score. Before fusion, the distances between samples with a certain descriptor are first normalized to $[0, 1]$ by a min-max manner as in [4].

## 4. Dataset

As person Re-ID is firstly proposed to track people among multiple non-overlap cameras, this implicitly restricts Re-ID to the short-term scenarios and yields many corresponding benchmarks, *e.g.*, VIPeR [5], CUHK01-03 [11, 12], PRID2011 [6], iLIDS-VID [25], MARS [28]. However, these benchmarks are insufficient to cover our case re-identifying a subject with long-term intervals, which explicitly increases the difficulties, *e.g.*, more drastic illumination variation and clothing changes. This invokes us to construct a new dataset specific to the long-term Re-ID problem. In this section, we first briefly review the existing benchmarks, especially video-based PRID2011. Then, our new long-term Re-ID dataset is introduced as well as the evaluation protocols are defined.

### 4.1. Benchmark Datasets

By far, the Re-ID problem has experienced several milestones and derived multiple directions such as single-shot Re-ID [7, 17, 21], multi-shot Re-ID [2, 6, 30] and video-based Re-ID [27, 15, 25]. According to the specific directions, various kinds of datasets are constructed, which greatly promotes the development of Re-ID research. We refer readers to [29] for a comprehensive review.

Among the existing benchmarks, we take PRID2011 as an example to evaluate our proposed FITD for classical short-term Re-ID. This is because PRID2011 is the only one releasing raw video and annotation information, which is essential for our proposed FITD to extract true motion trajectories. PRID2011 dataset is captured under two disjoint cameras, where 385 and 749 identities are recorded for each camera respectively. Among them, only the first 200 subjects appear in both cameras. Since the dataset is collected under outdoor environment, multiple factors are included, *e.g.*, viewpoint variance, lighting condition and background difference. In this paper, we follow the protocol in [4, 25, 27], which only 178 of first 200 subjects with more than 25 frames are used, due to the requirement for extracting dense trajectories.

### 4.2. Motion-ReID Dataset

Since our proposed FITD is specific to solve Re-ID problem in long-term scenarios, we collect and annotate a new dataset named Motion-ReID, some samples are shown in Figure 1. It includes video sequences extracted from two disjoint static surveillance cameras deployed in an office building, which covers the field of two distinct entrance gates respectively. We have collected total 240 video clips from 30 persons, which half of them are captured by camera #1 and the rests are captured by camera #2. In particular, each subject is recorded twice under the same camera with a long-time interval which is at least one week. For clarity, we list the recording timeline of one subject as in Figure 4. Opposite walking directions are separately recorded for one recording such as entering and exiting a door (We use *front* and *back* to represent the distinct directions in the following sections). Each video sequence includes approximately 20 to 204 frames with an average 102 frames which cover at least one walking cycle. Bounding boxes are manually labeled in each frame with varying size, which makes the dataset easy to evaluate Re-ID algorithms.

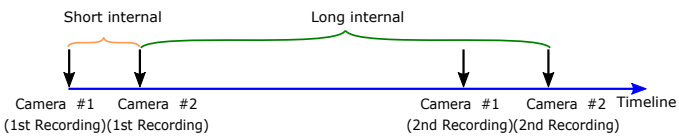


Figure 4. Video recording timeline of a subject.

Considering the specific task, we develop eight

challenging validation sets in terms of camera $C = \{camera\#1, camera\#2\}$, walking direction $D = \{back, front\}$ and recording time $R = \{1st, 2nd\}$. For all the sets, gallery and probe sets are with significantly different recording time. In particular, the gallery and probe sets in the first four validation sets are recorded by same camera in same walking direction, *i.e.* $S_{1-4} = \{(C_1, D_1, R_1; C_1, D_1, R_2), (C_1, D_2, R_1; C_1, D_2, R_2), (C_2, D_1, R_1; C_2, D_1, R_2), (C_2, D_2, R_1; C_2, D_2, R_2)\}$, whist the gallery and probe set in the rest sets are recorded by different cameras in same walking direction, *i.e.*, $S_{5-8} = \{(C_1, D_1, R_1; C_2, D_1, R_2), (C_1, D_2, R_1; C_2, D_2, R_2), (C_1, D_1, R_2; C_2, D_1, R_1), (C_1, D_2, R_2; C_2, D_2, R_1)\}$. The validation setting covers all the long-term Re-ID situations which is critical to evaluate algorithms specific for long-term Re-ID. Here we do not consider walking directions since entering and exiting a door are exactly opposite and has slight influence to Re-ID algorithms.

Compared to current benchmarks, the dataset is more challengeable, because (1) The dataset is collected by real surveillance cameras rather than self-deployed ones, which causes image quality and camera viewpoint more challengeable; (2) Our dataset is specially collected for long-term Re-ID task, which brings out new challenges, *e.g.*, wearing (clothing style and color) and carrying condition changes.

# 5. Experiments

In this section, we evaluate our FITD on both benchmark PRID2011 dataset and the proposed long-term Motion-ReID dataset.

## 5.1. Experiments Setting

To highlight the importance of true motion cues for Re-ID, we conduct all experiments without using any supervised metric learning method as in [4]. For PRID2011, we randomly split the dataset into two equal subsets and compute the ranking scores for 10 trials. Average cumulative matching characteristics (%) are reported for comparison [25]. We also conduct the evaluation for wearing similar clothes named PRID BK as in [4], which picks 35 samples with dark clothing forming testing set and 89 from the rest samples forming the training set. For Motion-ReID, we evaluate our method on all the 8 subsets. Due to the small size of the dataset, we utilize 5-fold cross-validation method and repeat our experiments for 10 times. Average rank-1 accuracies (%) are reported for all the validation sets.

Considering the periodicity of walking and tracking drift problem, we set trajectory length $L = 15$ and $L = 12$ respectively for PRID2011 and Motion-ReID which roughly equal to half of a walking cycle. We find that the number of GMM components has little impact on the performance, thus we simply set to 32 in all our experiments. In practice, only $\mu_k^X$ and $\varepsilon_k^X$ are reserved to construct Fisher vectors and thus the length of Fisher vectors for DT, HOG, HOF, MBHx and MBHy descriptors in one body-action unit are respectively 2048, 6144, 6912, 6144 and 6144 dimensions. The final FITD is the concatenation of all the body-action units in a fixed order. During testing, we simply set even weights in score-level fusion and leverage nearest neighbourhood classifier based on Euclidean distance to calculate the matching scores for all the methods.

## 5.2. Effectiveness of FITD for Short-term Re-ID

In this section, we evaluate the proposed FITD on benchmark PRID2011 and the cropped PRID BK datasets. To achieve stable performance, we use our body-action pyramid model and concatenate Fisher vectors of all the units. A comprehensive comparison of different types of features for the short-term Re-ID case is conducted, and the results are reported in Table 1.

In Table 1, the features are roughly divided into three categories. Rows 1-3: a single appearance-based component, *e.g.*, *color or texture*; rows 4-8: ensemble appearance-based feature; rows 9-12: spatial-temporal feature. Noting that we use different trajectory-aligned descriptors for PRID 2011 and PRID BK, *i.e.*, FITD with HOG encoding descriptor for PRID 2011 and FITD with HOGMBHx fused in score level for the PRID BK dataset, this is determined by different properties of the two datasets (see supplementary).

*Single Appearance-based Components:* Among the three single appearance-based feature component, using color (Row 1) achieves better performance than that using histogram of color (Row 2) and LBP (Row 3) on both the PRID 2011 and PRID BK datasets.

*Ensemble Appearance-based Features:* Rows 4-8 are state-of-the-art appearance-based features used in short-term person Re-ID. Among them, LOMO achieves the best performance. The conclusion can be interpreted in two-fold. First, LOMO leverages Retinex images, which weakens illumination and color gaps across cameras. This makes LOMO can extract refined color features than extracting from raw images; and second, LOMO also extracts texture features using Scale Invariant Local Ternary Pattern (SILTP) which are more robust to noises than LBP. As expected, when taking the smaller testing set into account, the performances of appearance-based features which utilize color cues drop notably, *e.g.* ColorHist, Color & LBP, ColorHist & LBP and ELF.

*Spatial-temporal Features:* All of the four spatial-temporal features are developed to extracting information from videos. Notably, performances of all the spatial-temporal features do not decline when applying to the impaired PRID BK dataset. Both STFV3D and DynFV utilize Fisher vectors to describe features of a human, however, STFV3D represents more appearance-based feature, *e.g.* pixel value and gradient, while DynFV focuses more

| Dataset | PRID2011 | | | | PRID BK | | | |
|---|---|---|---|---|---|---|---|---|
| Rank | R-1 | R-5 | R-10 | R-20 | R-1 | R-5 | R-10 | R-20 |
| Color [7] | **9.33** | **29.78** | **39.21** | **60.00** | **12.86** | **31.43** | **41.43** | **70.00** |
| ColorHist [19] | 2.36 | 10.45 | 19.21 | 35.62 | 2.86 | 22.86 | 32.86 | 65.71 |
| LBP [19] | 3.03 | 14.49 | 21.91 | 35.62 | 7.14 | 25.71 | 32.86 | 68.57 |
| Color & LBP [7] | 10.22 | 27.19 | 38.65 | 60.79 | 7.14 | 25.71 | 41.43 | 65.71 |
| ColorHist & LBP [26] | 13.26 | 29.55 | 40.79 | 55.62 | 11.43 | 30.00 | 48.57 | 67.14 |
| ELF [5] | 2.36 | 11.01 | 21.80 | 33.93 | 1.43 | 14.29 | 32.86 | 62.86 |
| LOMO [14] | **22.81** | **61.46** | **77.19** | **88.31** | **40.00** | **67.14** | **82.86** | **94.29** |
| LDFV [17] | 14.27 | 34.16 | 46.97 | 60.45 | 14.29 | 32.86 | 47.14 | 65.71 |
| HOG3D [25] | 22.92 | 46.52 | 59.78 | 73.15 | 22.86 | 50.00 | 64.29 | 85.71 |
| STFV3D [15] | 42.10 | 71.90 | 84.40 | 91.60 | 40.00 | 68.57 | 82.86 | 91.43 |
| DynFV [4] | 17.63 | 47.54 | 65.00 | 83.85 | 40.57 | 79.57 | 90.57 | 99.86 |
| FITD **(Ours)** | **58.65** | **81.91** | **89.33** | **95.17** | **54.29** | **82.86** | **97.14** | **100** |

Table 1. A comparison of proposed FITD with other popular features on PRID2011 dataset and PRID BK dataset.

on motion patterns. Thus, DynFV is less discriminative than STFV3D on PRID2011 whilst more powerful on PRID BK. Compared to STFV3D, our FITD with HOG extracts texture in space-time volume around dense trajectory, thus leading to higher performances on PRID2011. Compared to DynFV, our FITD with HOGMBHx extracts texture and motion from true video volume rather than normalized image sequences, and we use trajectory-aligned descriptors instead of raw trajectories. This explains why our FITD outperforms the DynFV by a large margin.

### 5.3. Effectiveness of FITD for Long-term Re-ID

In this section, we evaluate our FITD on the proposed Motion-ReID dataset. Table 2-4 report our results on all the eight subsets. To better prove benefits of our FITD, we evaluate it from three aspects: trajectory-aligned methods, fusion strategies and feature representations.

**Trajectory-aligned Methods:** Table 2 compares different trajectory-aligned methods on all the eight subsets. Out of these methods, FITD with motion descriptors achieves better performances than HOG in the first four subsets while FITD with HOG performs best in the last four subsets. This is not surprising because 1) Motion patterns are more discriminative in the first four subsets since video sequences from both gallery and probe in the first four subsets are captured from the same camera and clothing variation is the leading influential factor. 2) For the last four subsets, huge view difference between cameras affects motion seriously and consequently causes motion-based features less discriminative.

**Fusion Strategies:** Table 3 shows results of different types of fusion methods, Typically, Row 1-5 are fusion at the feature level, and Row 6-10 are fusion at the score level. As see the table, fusions in the score level outperform fusions in the feature level in most cases. Compared to the performance of FITD using single descriptor, the fusion methods are more stable and improve the overall performance to some extent. Considering differences between first four subsets and last four subsets, we extract features by fusion representations HOGHOFMBH and HOGHOF in score level respectively for the two scenarios.

| $S_i$ | # 1 | # 2 | # 3 | # 4 | #5 | #6 | #7 | #8 |
|---|---|---|---|---|---|---|---|---|
| DT | 55.5 | 60.0 | 40.3 | 42.0 | 20.7 | 19.3 | 19.0 | 22.3 |
| HOG | 56.7 | 55.0 | 57.7 | 48.3 | 27.3 | **27.3** | **24.3** | 18.3 |
| HOF | 52.0 | 58.7 | 54.7 | 49.0 | **28.0** | 20.7 | 19.0 | 22.7 |
| MBHx | 62.7 | **65.0** | **59.7** | **55.7** | 18.0 | 18.0 | 22.7 | 20.7 |
| MBHy | **65.0** | 60.7 | 58.3 | 50.7 | 16.7 | 14.7 | 17.7 | **24.0** |

Table 2. A comparison of proposed FITD with different trajectory-aligned descriptors.

**Feature Representations:** Table 4 compares our FITD model with some state-of-the-art feature representations as in the last section.

*Single Appearance-based Components:* Different from results on PRID2011, LBP achieves the best performance when gallery and probe samples are from the same camera, while Color is more discriminative when gallery and probe samples are from different cameras with huge viewpoint difference. However, performances of the appearance-based feature using no matter color or texture degrade significantly with camera changing and view enlarging. This is because camera changing and viewpoint variation impact texture and color differently.

*Ensemble Appearance-based Feature:* Similar to single appearance-based components, performances of the commonly used ensemble appearance-based features also decline drastically. Out of these features, LDFV achieves the best performance in the first four subsets as in [4] where the gallery and query samples are obtained from the same camera. However, the performance of LDFV drops more sharply than other appearance features, which is the least discriminative among these features.

| | Subset $S_i$ | #1 | # 2 | #3 | # 4 | #5 | # 6 | #7 | # 8 |
|---|---|---|---|---|---|---|---|---|---|
| Feature Fusion | HOGHOF | 55.00 | **65.00** | 55.67 | 53.00 | **22.67** | 20.33 | 23.33 | 19.67 |
| | MBH | **64.67** | 64.67 | **60.67** | **56.67** | 9.67 | 22.33 | 20.33 | **26.00** |
| | HOFMBH | 56.33 | 63.33 | 59.00 | 53.67 | 22.00 | 19.33 | 21.33 | 25.67 |
| | HOGMBHx | 59.00 | 60.67 | 58.00 | 54.67 | 21.33 | **24.33** | 24.33 | 23.33 |
| | HOGHOFMBH | 59.00 | **65.00** | 59.33 | 55.00 | 18.67 | 18.67 | **27.33** | 24.67 |
| Score Fusion | HOGHOF | 60.33 | 66.00 | 55.33 | 53.00 | **30.33** | **26.33** | 21.67 | 24.00 |
| | MBH | 64.33 | 64.67 | **60.67** | 54.00 | 18.00 | 16.00 | 20.33 | 22.67 |
| | HOFMBH | 61.33 | 66.00 | 60.33 | 55.67 | 21.67 | 15.00 | 18.33 | **24.67** |
| | HOGMBHx | 62.00 | **66.67** | 59.00 | **56.67** | 22.00 | 22.33 | **23.33** | 21.67 |
| | HOGHOFMBH | **65.67** | **66.67** | **60.67** | 55.33 | 24.00 | 18.00 | 21.33 | 24.00 |

Table 3. A comparison of proposed FITD with different fusion methods.

| Subset $S_i$ | #1 | # 2 | #3 | # 4 | #5 | # 6 | #7 | # 8 |
|---|---|---|---|---|---|---|---|---|
| Color [7] | 33.00 | 35.67 | 33.33 | 29.33 | **26.00** | **25.00** | **26.33** | **23.33** |
| ColorHist [19] | 33.33 | 37.00 | 28.67 | **33.33** | 23.33 | 17.33 | 17.33 | 14.33 |
| LBP [19] | **51.67** | **39.33** | **34.33** | 27.67 | 19.00 | 19.67 | 18.00 | 20.33 |
| Color & LBP [7] | 38.67 | 38.67 | **35.67** | 30.67 | 18.67 | **24.67** | **26.67** | **23.00** |
| ColorHist & LBP [26] | 39.67 | 40.67 | 29.67 | 34.67 | **24.67** | 24.00 | 19.00 | 18.00 |
| ELF [5] | 31.67 | 36.00 | 33.00 | 32.67 | 22.67 | 22.00 | 20.33 | 16.67 |
| LOMO [14] | 27.67 | 35.00 | 23.22 | 27.33 | 20.00 | 18.00 | 14.33 | 17.33 |
| LDFV [17] | **49.33** | **41.67** | 34.00 | **35.33** | 18.67 | 19.33 | 15.33 | 16.33 |
| HOG3D [25] | 39.67 | 30.33 | 37.33 | 34.67 | 13.67 | 14.33 | 17.67 | 20.67 |
| STFV3D [15] | 39.00 | 44.33 | 31.67 | 39.33 | 15.67 | 26.00 | 17.00 | 20.33 |
| DynFV [4] | 48.33 | 45.00 | 46.67 | 37.67 | 22.00 | 18.00 | 21.00 | 20.00 |
| FITD **(Ours)** | **65.67** | **66.67** | **60.67** | **55.33** | **30.33** | **26.33** | **21.67** | **24.00** |

Table 4. A comparison of proposed FITD with other popular features on Motion-ReID dataset.

*Spatial-temporal Features:* Among the four spatial-temporal features, motion-based features, *e.g*. DynFV and our FITD, outperform appearance-based features by a large margin in the first four subsets. The results prove the effectiveness of motion-based feature for long-term Re-ID. Since our FITD model extracts dense trajectory from raw/true video sequence rather than the normalized bounding area and encodes the trajectories with trajectory-aligned descriptors, our FITD model achieves better performance than DynFV. However, motion patterns are more easily affected by camera view changing which causes performance of motion-based features declining sharply. This is also one of our future research points, which aims to solve view difference problem when using motion-based features. Noting that appearance-based features also perform regularly in the subsets, it is because several targets wear the same clothes and some partially change their clothes in the dataset.

## 6. Conclusion

Up to now, most state-of-the-art Re-ID methods tried to solve the Re-ID problem in the short-term scenarios. These methods assumed the target subjects keep constant wearing conditions across cameras and relied heavily on appearance-based features extracted from one or several frames. However, these appearance-based features are not reliable to some appearance impaired scenarios, *e.g*., similar wearing between subjects and wearing changing of the same subject. In this paper, we focused on the impaired scenarios especially the long-term Re-ID and introduced the first available long-term Re-ID dataset. In specific, we proposed to solve the Re-ID task using motion patterns from true/raw video sequences named FITD. The proposed FITD model characters motion patterns by the trajectory-aligned descriptors in a three-level body-action pyramid and benefits from the Fisher vector encoding. Comprehensive experiments show that our FITD with appropriate trajectory-aligned descriptors benefits for the person Re-ID, especially the extremely wearing similar scenarios and long-term scenarios. This exactly fills the research blank in the field of long-term Re-ID. However, motion-based features suffer some new challenges, *e.g*., large walking view and camera viewpoint differences, partial occlusion and background movement. These problems will be our core research points in the future.

## References

[1] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision*

*Computing*, 32(4):270–286, 2014. 1, 2

[2] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, pages 1335–1344, 2016. 1, 2, 5

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005. 4

[4] M. Gou, X. Zhang, A. Rates-Borras, S. Asghari-Esfeden, M. Sznaier, and O. Camps. Person re-identification in appearance impaired scenarios. In *BMVC*, 2016. 2, 5, 6, 7, 8

[5] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. Springer, 2008. 1, 2, 5, 7, 8

[6] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person Re-Identification by Descriptive and Discriminative Classification. In *SCIA*, 2011. 1, 5

[7] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, pages 780–793. Springer, 2012. 1, 5, 7, 8

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1

[9] W. Kusakunniran. Recognizing gaits on spatio-temporal feature domain. *IEEE Transactions on Information Forensics and Security*, 9(9):1416–1423, 2014. 2

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1–8. IEEE, 2008. 2

[11] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601, 2013. 5

[12] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 5

[13] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, pages 2194–2200, 2017. 2, 3

[14] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 1, 2, 7, 8

[15] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*, pages 3810–3818, 2015. 1, 2, 3, 5, 7, 8

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

[17] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV Workshops*, pages 413–422. Springer, 2012. 5, 7, 8

[18] L. Ma, H. Liu, L. Hu, C. Wang, and Q. Sun. Orientation driven bag of appearances for person re-identification. *arXiv preprint arXiv:1605.02464*, 2016. 2, 3

[19] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, pages 2666–2672. IEEE, 2012. 7, 8

[20] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 1

[21] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010. 2, 5

[22] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013. 2, 4

[23] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013. 2

[24] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, 2013. 2, 3, 4

[25] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2501–2514, 2016. 1, 2, 5, 6, 7, 8

[26] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, pages 1–16. Springer, 2014. 1, 2, 7, 8

[27] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *CVPR*, pages 1345–1353, 2016. 1, 2, 5

[28] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*. Springer, 2016. 5

[29] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 2, 5

[30] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian. Person re-identification in the wild. In *CVPR*, 2017. 1, 5

[31] W.-S. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):591–606, 2016. 1