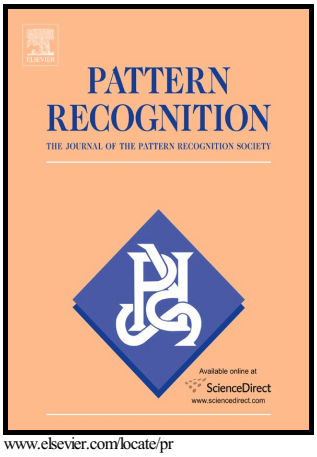
Person Re-Identification by Unsupervised Video Matching

Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, Yisheng Zhong



www.elsevier.com/locate/pr

PII: S0031-3203(16)30376-4
DOI: http://dx.doi.org/10.1016/j.patcog.2016.11.018
Reference: PR5961

To appear in: *Pattern Recognition*

Received date: 31 March 2016
Revised date: 21 November 2016
Accepted date: 21 November 2016

Cite this article as: Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie Jianming Hu, Kin-Man Lam and Yisheng Zhong, Person Re-Identification by Unsupervised Video Matching, *Pattern Recognition* http://dx.doi.org/10.1016/j.patcog.2016.11.018

Person Re-Identification by Unsupervised Video Matching

Xiaolong Ma[1,4], Xiatian Zhu[2], Shaogang Gong[2], Xudong Xie[1], Jianming Hu[1], Kin-Man Lam[3], Yisheng Zhong[1]

Abstract

Most existing person re-identification (ReID) methods rely only on the spatial appearance information from either one or multiple person images, whilst ignore the space-time cues readily available in video or image-sequence data. Moreover, they often assume the availability of exhaustively labelled cross-view pairwise data for every camera pair, making them non-scalable to ReID applications in real-world large scale camera networks. In this work, we introduce a novel video based person ReID method capable of accurately matching people across views from arbitrary *unaligned* image-sequences without any labelled pairwise data. Specifically, we introduce a new space-time person representation by encoding multiple granularities of spatio-temporal dynamics in form of time series. Moreover, a Time Shift Dynamic Time Warping (TS-DTW) model is derived for performing automatically alignment whilst achieving data selection and matching between inherently inaccurate and incomplete sequences in a unified way. We further extend the TS-DTW model for accommodating multiple feature-sequences of an image-sequence in order to fuse information from different descriptions. Crucially, this model does not require pairwise labelled training data (i.e. unsupervised) therefore readily scalable to large scale cam-

Email addresses: goup000@163.com (Xiaolong Ma), xiatian.zhu@qmul.ac.uk (Xiatian Zhu), s.gong@qmul.ac.uk (Shaogang Gong), xdxie@mail.tsinghua.edu.cn (Xudong Xie), hujm@mail.tsinghua.edu.cn (Jianming Hu), kin.man.lam@polyu.edu.hk (Kin-Man Lam), zys-dau@mail.tsinghua.edu.cn (Yisheng Zhong)

[1]Tsinghua University, China
[2]Queen Mary University of London, United Kingdom
[3]The Hong Kong Polytechnic University, Hong Kong
[4]China Academy of Electronics and Information Technology

era networks of arbitrary camera pairs without the need for exhaustive data annotation for every camera pair. We show the effectiveness and advantages of the proposed method by extensive comparisons with related state-of-the-art approaches using two benchmarking ReID datasets, PRID2011 and iLIDS-VID.

Keywords: Person re-identification, action recognition, gait recognition, video matching, temporal sequence matching, spatio-temporal pyramids, time shift.

1. Introduction

In visual surveillance, associating automatically individual people across disjoint camera views is essential. This task is known as *person re-identification* (ReID). Cross-view person ReID enables automated discovery and analysis of person-specific long-term structural activities over widely expanded areas and is fundamental to many important surveillance applications such as multi-camera people tracking and forensic search. Specifically, for performing person ReID, one matches a probe (or query) person observed in one camera view against a set of gallery people captured in another disjoint view for generating a ranked list according to their matching distance or similarity [1]. This is an inherently challenging problem [2]. Most existing approaches [3, 4, 5, 6, 7, 8, 9, 10] perform ReID by modelling spatial visual appearance (shape, texture and colour) of one or multiple person images. However, people appearance is intrinsically limited due to the inevitable visual ambiguity and unreliability caused by appearance similarity among different people and appearance variations of the same person from unknown significant cross-view changes in human pose, viewpoint, illumination, occlusion, and dynamic background clutter. This motivates the need of seeking additional visual information sources for person ReID.

On the other hand, video (or image-sequence) data are often available from visual surveillance cameras. Videos have been extensively exploited for performing action and activity recognition by extracting and modelling a variety of dynamic space-time visual features [11, 12]. However, action recognition differs fundamentally from person ReID. First, it often aims to discriminate between

(a) Appearance change across views (b) Visually similar but different people

Figure 1: The challenges of person re-identification in visual surveillance [2]. (a) The appearance of the same person may change significantly across disjoint camera views due to great cross-camera variations in illumination, viewpoint, random inter-object occlusion and complex background clutter in typically-crowded public spaces. Each blue bounding box corresponds to a specific person. (b) Different people may present largely similar visual appearance.

different action categories but tolerate the variance of the same action performed by different people. In contrast, the objective of ReID is to discriminate among different person identities regardless of actions by the person. Moreover, action recognition methods often consider a pre-defined set of action categories during both training/testing phases, whereas person ReID models are required to generalise from the training categories (identities) to previously unseen ones.

Apart from action recognition, another closely related problem is gait recognition [13]. Similar to ReID, gait recognition aims for differentiating between distinct people by characterising people's walking dynamics. Further, an advantage of gait recognition is no assumption being made on either subject cooperation or person distinctive actions. These characteristics are analogous in spirit to person ReID. Nonetheless, existing gait recognition methods are heavily subject to stringent requirements on person foreground segmentation and accurate temporal alignment throughout a gait image sequence (a walking cycle). Additionally, most gait recognition methods do not deal well with cluttered background and/or random occlusions with unknown covariate conditions [14] (Figures 1 and 2). Hence, person ReID in public spaces is inherently challenging for existing gait recognition techniques.

This work aims to develop a video based person ReID approach, without the need for exhaustively labelling people pairs across camera views. To that end, one needs to extract and model reliably person-specific space-time infor-

mation from videos. This is non-trivial, especially when the videos are captured from uncontrolled and crowded public scenes. The specific challenges include: (1) The starting/ending frames of individual videos may correspond to arbitrary walking phases. Thus, any two compared videos are mostly unaligned. This misalignment leads to inaccuracy in people matching, especially when the useful space-time information in person videos can be very subtle. (2) Person videos have varying numbers of walking cycles and a holistic matching between videos may yield suboptimal recognition. While pose estimation and walking cycle detection may help in theory, contemporary techniques [15, 16] are still rather unreliable for video data with distracting background and low imaging quality. (3) Person image-sequences captured from public places can consist of corrupted frames due to background clutter and random inter-object occlusions (see Figure 1). A blind trust and utilisation of all visual data may degrade the person matching accuracy. Following [17], we call this *unregulated* image-sequences. We wish to develop an accurate person ReID method that does not require performing explicit walking phase detection for videos neither occlusion estimation for image frames. The **main contributions** of this study are:

1. We propose an unsupervised approach to person ReID based on typical surveillance image-sequences. Our model differs significantly from most conventional static image based methods (e.g. leveraging dynamic space-time information *versus* static appearance information), and also the recent DVR video ReID model [18] (e.g. unsupervised *versus* supervised).
2. We present a new video representation particularly tailored for person ReID. Specifically, this representation is built up on existing action space-time features (e.g. histograms of oriented 3D spatio-temporal gradient [19]) and spatio-temporal pyramids [20, 21]. In contrast to most visual features for action recognition which are vectorial, our video representation is in form of sequence or time series. This is specially designed for reliable selection based person matching between cross-view unregulated video pairs with possibly ambiguous, incomplete and noisy observation.
3. We introduce an effective video matching algorithm, Time Shift Dynamic

Time Warping (TS-DTW) and its Multi-Dimensional variant MDTS-DTW, for data selective based sequence matching. Particularly, the proposed model computes the distance between two videos by iteratively (1) altering their mutual time shift relation and (2) then matching two partial segments of them. Importantly, our method is capable of simultaneously performing sequence alignment, selecting best-matched segments, and fusing diverse information for person ReID in a unified manner.

We show the effectiveness of the proposed approach on two benchmarking image-sequence ReID datasets (PRID2011 [22] and iLIDS-VID [17]) under both the closed-world and more realistic open-world scenarios [23, 9]. Extensive comparative evaluations were conducted by comparing alternative sequence-matching person recognition models including gait recognition [24] and dynamic time warping [25], and the state-of-the-art person ReID methods including SDALF [3], eSDC [6], DVR [18], RDL [26], and XQDA [27].

The remainder of this paper is organised as follows. In Section 2, we discuss broadly the related studies. In Section 3, we present an overview of our approach, followed by video representation in Section 4, video matching in Section 5, and person re-identification application in Section 6. Then, we depict the experimental settings in Section 7 and provide comparative evaluations of our proposed approach in Section 8. Finally, we conclude this study in Section 9.

2. Related Work

Gait recognition. Gait recognition [13, 28, 29, 30, 31] has been extensively exploited for people identification using video space-time features, e.g. correlation based motion feature [32], and Gait Energy Image (GEI) templates [33]. To improve gait representations, Veres et al. [34] and Matovski et al. [35] suggest feature selection and quality measure. These methods assume that image-sequences are aligned and captured in controlled environments with uncluttered background, as well as having complete gait cycles, little occlusion, and accurate

gait phase estimation. However, these constraints are often invalid in person ReID context as shown in Figures 2 and 7.

To handle often-occurring occlusion, Hofmann et al. [36] propose a specific dataset for evaluating their negative influence on gait recognition performance. Meanwhile, a number of part-based methods [37, 38, 39] are developed by assuming that matched people share common observed parts (COPs). For relaxing this assumption, Muramatsu et al. [40] reconstruct complete gait features from partially observed body parts without sharing COPs. These methods rely on accurate body part segmentation and occlusion detection, which is however overdemanding for contemporary segmentation methods [15, 16, 41] given typical ReID video data captured against uncooperative people and dynamic scenes.

Main challenges for gait recognition arise from various covariate conditions, e.g. carrying, clothing, walking surface, footwear, and viewpoint. Beyond the attempts of designing and investigating gait features invariable to specific covariates [13, 42, 43, 44, 14], more powerful learning based methods have also been presented for explicitly and accurately modelling the complex variances of gait structures. For example, Martín-Félez and Xiang [45] exploit the learning-to-rank strategy for jointly characterising a variety of covariate conditions in a unified model. Whilst a learning process may help improve the gait recognition accuracy, this strategy is heavily affected by the goodness of gait features. On person ReID videos however, gait features are likely to be extremely unreliable, as demonstrated in Figure 2.

Temporal sequence matching. Temporal sequence matching is another alternative strategy. The Dynamic Time Warping (DTW) model [25, 46, 47] and its variants including derivative DTW [48, 49], weighted DTW [50], are common sequence matching algorithms widely used in data mining and pattern recognition. Given two temporal sequences, it searches for the optimal non-linear warp path between the sequences that minimises the matching distance. However, the conventional DTW models assume that the two sequences have the same number of temporal cycles (phases) and are aligned at the starting and end-

Figure 2: Example GEI features of PRID2011 [22] (top) and iLIDS-VID [17] (bottom) videos.

ing points/elements. These conditions are difficult to be met in person videos from typical surveillance scenes. Hence, directly using DTW variants to holistically match these unregulated videos may be suboptimal. To further compound the problem, there are often unknown occlusions and background clutters that can lead to corrupted video frames with missing and/or noisy observation thus potentially inaccurate distance measurement.

In case of cyclic sequences, e.g. closed curves, the starting element is often unknown and may be located by a greedy search or some heuristic method [51]. However, there can exist more than one starting elements for periodic sequences like people walking videos. Whilst continuous dynamic programming or spotting [52] identifies both starting/ending elements, it requires a good predefined threshold, which however is not available in our person ReID problem.

Single/multi-shot and video based person ReID. Most existing ReID methods [4, 5, 6, 7, 8, 53, 54, 55, 56, 57] only consider one-shot image per person per view. This is inherently weak when multi-shot are available, due to the intrinsically ambiguous and noisy people appearance and large cross-view appearance variations (Figure 1). There are efforts on multi-shot ReID. For example, Hamdoun et al. [58] propose to employ the interest points cumulated across a number of images; Cong et al. [59] utilise the data manifold geometric structures of multiple images for constructing more compact spatial appearance

description. Other attempts include training a robust appearance model using image sets [60] and enhancing local image region/patch spatial feature representation [61, 3, 62, 63]. In contrast to all these methods focusing on exploiting spatial appearance information, this work explores space-time information from available videos for person ReID.

Previous efforts of exploiting space-time dynamics for person ReID are built on either gait recognition or action recognition. Specifically, gait features are exploited for enriching appearance ReID representations in [64, 65, 66, 67]. But these methods naturally share similar limitations of gait recognition models, e.g. severely suffering from feature noises inherent in ReID data. Recently, Wang et al. [17, 18] partly solve this problem by formulating a discriminative video ranking (DVR) model using the space-time HOG3D feature [19]. However, this fragment-based DVR model is limited as only a few local fragments from each person image-sequence is exploited whilst the remaining data is totally discarded. Critically, the DVR model is supervised, i.e. its model construction requires a large number of cross-view matched people for each camera pair. This renders DVR non-scalable for large-scale networks with many camera pairs. Other video based ReID methods [68, 69] are also supervised and thus subject to the similar scalability limitation as DVR.

Space-time visual features. Our person video representation is inspired by existing successful action features and the DVR model [18], e.g. histograms of oriented 3D spatio-temporal gradient (HOG3D) [19]. In contrast to most feature vector based action representations [70, 71, 72, 73, 74, 75, 21, 76], we represent person videos with temporal sequences based representations. This design is capable of (1) not only encoding the dynamic temporal structures of motion, (2) but also selectively matching unregulated person videos (see Section 5). While some action recognition models also regard videos as sequences of observation [77, 78, 79, 80, 81], their focus is coarse temporal structure modelling alone.

To extract different granularities of localised temporal ordering dynamics, we adopt the notion of temporal pyramids (see Figure 4(b)). Instead of using

temporal sub-sampling to construct a temporal pyramid [82, 83], we segment videos with different sequence-element lengths for preserving all possible dynamic information at all levels as in [21, 84]. However, our representation is different significantly from the latter two because: (1) They use vector based representations *whilst* ours are sequential or temporal series; (2) They assume well segmented videos as input (e.g. one action per video) *whilst* our person videos can contain a varying numbers of walking action periods without any temporal segmentation; (3) We *additionally* consider spatial pyramid [20] at each temporal granularity and importantly data selection in video matching.

3. Approach Overview

Unlike most action recognition methods that represent each video with a feature vector [11] or the image-sequence based person re-identification (ReID) approach that describes each video with a set of independent vectors [17, 18], we consider person videos as sequences of localised space-time dynamics for performing ReID. This allows to: (1) Explicitly represent and model localised temporal motion dynamics; (2) Flexibly achieve temporal alignment between different videos; (3) Facilitate data driven selective matching without any supervision (see Section 5). All these capabilities are desired and helpful for reliable person ReID by accurately characterising and exploiting space-time dynamic information of person's walking behaviour recorded in unregulated videos with random inter-object occlusions, arbitrary video duration and uncertain starting/ending phases, and uncontrolled background clutter.

However, it is non-trivial to automatically detect and exploit identity-sensitive space-time information from noisy video data, particularly in an unsupervised manner. Critically, one needs to address the problems of (1) how to extract rich dynamics information of people's walking motion, and (2) how to suppress the negative influence of unknown noisy observation, e.g. various types of occlusion and clutter in the background. This is beyond solving the more common temporal misalignment problem in video matching. To this end, we formulate a

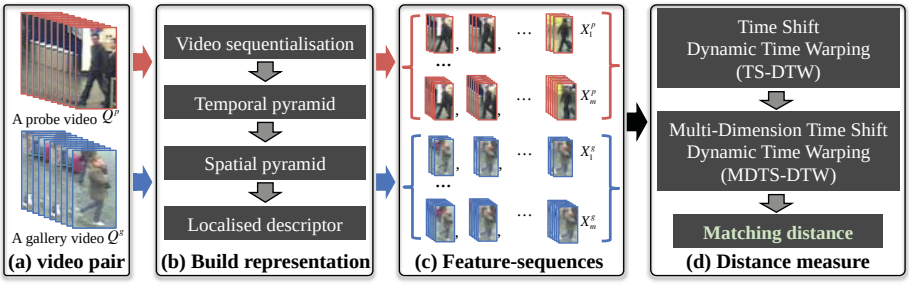


Figure 3: Overview of the proposed unsupervised video matching approach for person ReID. (a) An input pair of person videos; (b) Construct video representation by video sequentialisation (Section 4.1), temporal pyramid (Section 4.2), spatial pyramid (Section 4.3), and localised space-time descriptor computation (Section 4.4); (c) Obtained feature-sequences; (d) Video matching by the proposed TS-DTW (Section 5.2) and MDTS-DTW (Section 5.3) models.

novel unsupervised person re-identification method capable of extracting multi-scale spatio-temporal structure information (Section 4), automatically aligning sequence pairs and adaptively selecting/employing informative visual data (Section 5) from noisy person videos captured in non-overlapping camera views. This allows to relax the stringent assumptions of existing gait recognition methods and overcome the limitations of previous temporal sequence matching models, and result in more accurate person recognition, particularly with incomplete and noisy person videos captured in public spaces. Compared with the state-of-the-art DVR re-id model, our method is able to extract and employ much richer space-time cues from videos. Moreover, the proposed method is unsupervised, as opposite to DVR which needs a large number of cross-view matching pairs for every camera pair. Therefore, our proposed method is more scalable to the real-world applications involving large surveillance camera networks. Additionally, we further consider information fusion from multiple feature-sequences each capturing some different aspects of person video data. An overview diagram of the proposed approach is presented in Figure 3.

4. Structured Video Representation

4.1. Video Sequentialisation

Suppose we have a collection of video (or image-sequence) pairs $\{(Q_i^p, Q_i^g)\}_{i=1}^n$, where Q_i^p and Q_i^g denote the videos of person i captured by two disjoint cameras p and g, and n the number of people. Each video is defined as a set of consecutive frames $\mathbf{I}$ (e.g. obtained by an independent person tracking process [85] with simple post-processing or not): $Q = \{\mathbf{I}_1, \mathbf{I}_2, ...\}$, where the video length $|Q|$ is varying as in typical surveillance settings, independently extracted person videos do not guarantee to have a uniform duration (arbitrary frame number), nor the number of walking cycles and starting/ending phases.

Given varying-long videos with unknown and random noise, it is ineffective to perform matching between two image-sequences holistically. A possible strategy [18] is: (1) Segmenting each video into multiple independent fragments; (2) Selecting the optimal/best fragment pairs for matching. This method, however, may lose potentially useful information encoded in the discarded fragments. In this work, we instead consider a richer representation for exploiting as much space-time information from inherently noisy videos as possible.

Specifically, we divide uniformly each individual video Q into multiple temporally localised *slices* with a small number l of image frames. Different slice lengths l correspond to different temporal granularities. Each slice encodes localised space-time information about the walking characteristics of the corresponding person. As a result, a video can be converted into a *space-time slice-sequence* $S = \{s_1, s_2, \dots\}$ (Figure 4). This localised slice-based sequence representation has three advantages over the bag-of-fragments model [17]: (1) It keeps the original sequential data form, whilst DVR only considers each fragment of a sequence as an isolated instance without temporal ordering among fragments. This allows us to enjoy the merits of existing sequence matching algorithms, e.g. non-linear dynamic time warping for handling the misalignment problem. (2) Alignment between sequences (e.g. starting/ending with the same walking phases) is made more robust due to the existence of a large

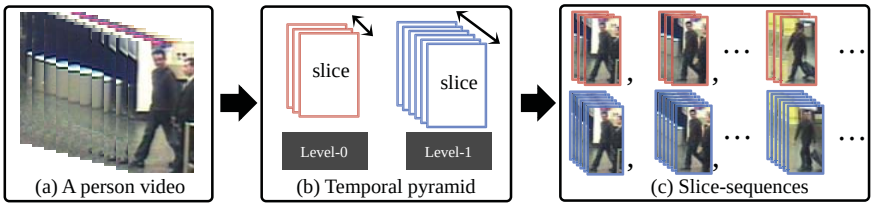


Figure 4: Illustration of temporal pyramid and video sequentialisation. Note the colour-coded correspondence between (b) the temporal pyramid level and (c) the slice-sequence.

number of short localised slices corresponding to various walking phases. In contrast, the bag-of-fragments strategy may suffer from fragilely aligned fragment pairs at times when only a small number of fragments are available from a video and the starting/ending phases of fragments are not sufficiently diverse to match. (3) It provides more flexible opportunities for selecting and exploring informative localised space-time information irregularly distributed across the original image-sequences, e.g. not only in the form of isolated fragments. This is difficult for the bag-of-fragments representation in DVR due to its hard video fragmentation and coarse fragment selection mechanism.

4.2. Temporal Pyramid

Since variations in walking styles may exist over various local temporal extends, it is suboptimal to utilise video slices of a uniform length. Also, fine-to-coarse localised temporal information is possible to complement each other in expressing temporal structure dynamics, as demonstrated in existing action recognition studies [21, 84]. In light of these considerations, we enrich our representation of person videos by imposing a temporal pyramid structure, motivated by pyramid match kernel [86] and its spatial extension [20].

Specifically, we use a set of video slice length for video sequentialisation as:

$$L = \{2^0 l,\ \ldots,\ 2^{(h_t-1)} l\} \tag{1}$$

which corresponds to a temporal pyramid with h_t levels/layers. Given a video Q_i, we generate a separate slice-sequence at each temporal pyramid level. Thus, a total of h_t slice-sequences $\{S_i^l\}_{l=0}^{h_t-1}$ can be produced for each video Q_i after

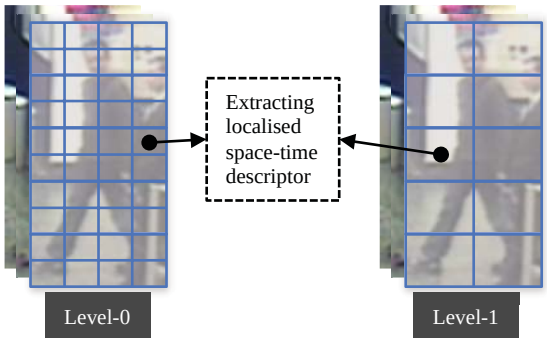


Figure 5: Spatial pyramid structures on a temporally-localised video slice.

applying this temporal pyramid (Figure 4(c)). During sequentialising a video, at any temporal pyramid level, we discard the last few image frames of person videos if they are not sufficient to form a slice. For example, suppose there is 56 frames in a person video and the slice length is 10, we drop/ignore the last 6 frames as they are not enough for a complete slice of 10 frames.

4.3. Spatial Pyramid

After obtaining slice-sequences $S = \{s_1, \ldots, s_i, \ldots\}$ of person video, we need to consider how to represent their localised video slices s_i. This is the same as deriving video representation for action recognition [11] in that each slice can be considered as a tiny action video. We want to capture localised spatio-temporal dynamic structures of people's walking. Apparently, the style or characteristics of walking motion is closely related to the action of different body parts, e.g. head, torso, arms, legs. Hence, we spatially decompose every slice into a grid of 2×5 uniform cells which approximately correspond to the layout of all body parts (Figure 5(right)). This division allows to encode roughly detailed spatial cues of individual parts into video slices.

Additionally, accurate ReID may need more fine-grained and subtle spatially structured cues of people's walking behaviour. This is because finer spatial decomposition provides more detailed information and potentially complements coarse divisions. To that end, we adopt the spatial pyramid match kernel [20], due to its superior expressive capability shown in action recognition [72]. In particular, we further split each cell into 2×2 smaller ones, resulting in a grid of

40 cells on each slice (Figure 5(left)). By repeating this process, we can obtain a h_s-level spatial pyramid. Together with temporal pyramid, we call our video representation as "Spatio-Temporal Pyramidal Sequence" (**STPS**). Next, we describe the dynamic feature descriptor for numerically representing localised space-time cells below.

4.4. Localised Space-Time Descriptor

We consider the HOG3D feature [19] for representing video slices due to its strong expressiveness for recognising different activities [87] and importantly for distinguishing between distinct people [17, 18]. Particularly, given a specific spatial division on any video slice s, we first extract the space-time gradient histogram from each cell where 3D gradient orientations are quantised using regular polyhedrons [19], then concatenate them to form a HOG3D feature vector $\mathbf{x}$ for the slice s. Note that there is 50% overlap between any two adjacent cells for increasing robustness against tracking/annotation errors. As such, we obtain a HOG3D feature-sequence $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots\}$ for a slice-sequence $S = \{s_1, s_2, \dots\}$. Finally, we apply histogram equalisation for reducing the effect of uneven illuminations. While other space-time descriptors, such as motion boundary histograms (MBH) [88], are considerable, it is beyond our scope to exhaustively discuss and evaluate a variety of different space-time descriptors.

5. Unsupervised Video Matching

In this section, we describe the details of the proposed sequence/video matching model for person ReID. We aim to formulate an unsupervised model. As a result, the expensive cross-camera pairwise labelling process for every camera pair can be eliminated for realising good deployment scalability in reality. To that end, we select the well-known Dynamic Time Warping (DTW) algorithm [25, 89] as the basis of our model due to: (1) Its great success and popularity in sequence based data analysis; (2) Its simple but elegant modelling.

Specifically, we derive a new sequence matching algorithm based on the DTW model, called *Time Shift Dynamic Time Warping* (TS-DTW), and further

generalise TS-DTW to the multi-dimensional setting, i.e. with multiple feature-sequences per person video. This formulation is motivated by works in time delay based studies [90, 91], multi-dimension fusion [92], and neural networks (or deep learning) [93]. This proposed model is characterised with alignment free, data selection, and information fusion. Before detailing our method, let us first briefly describe the conventional DTW model.

5.1. Conventional DTW

In general, the DTW model [25, 46, 47, 89] aims at measuring the distance or similarity between two temporal-sequences by searching for the optimal non-linear warp path. Formally, given two feature-sequences $X^p = \{s_1^p, \dots, s_i^p, \dots\}$ and $X^g = \{s_1^g, \dots, s_j^g, \dots\}$, we define a warp path as:

$$W = \{\mathbf{w}_1, \dots, \mathbf{w}_d\} \tag{2}$$

where the k-th entry $\mathbf{w}_k = (w_k^p, w_k^g)$ indicates that the w_k^p-th element from X^p and w_k^g-th element from X^g are matched. The warp path length holds as: $\max(|X^p|, |X^g|) \leq d < |X^p| + |X^g|$. The symbol $|\cdot|$ denotes the set size. We then define the sequence matching distance $\text{dist}_{\text{dtw}}(X^p, X^g)$ between X^p and X^g as:

$$\text{dist}_{\text{dtw}}(X^p, X^g) = \frac{1}{d}\sum_{k=1}^{d} \text{dist}_{\text{el}}(\mathbf{x}^p_{w_k^p}, \mathbf{x}^g_{w_k^g}) \tag{3}$$

with $\text{dist}_{\text{el}}(\cdot, \cdot)$ as the distance metric between two elements (or slices), e.g. L_1 or L_2 norm, and $d = |W|$ the warp path length. The objective of DTW is to find the optimal warp path W^* such that

$$W^* = \text{argmin}_{W \in \Omega} \text{dist}_{\text{dtw}}(X^p, X^g) \tag{4}$$

where Ω is the set of all possible warp paths. This optimisation can be realised using dynamic programming [94] subject to three constraints: (1) bounding constraint: $\mathbf{w}_1 = (1, 1)$ and $\mathbf{w}_d = (|X^p|, |X^g|)$; (2) monotonicity constraint: $w_1^p \leq w_2^p \leq ... \leq w_d^p$ and $w_1^g \leq w_2^g \leq ... \leq w_d^g$; and (3) step-size constraint: $\mathbf{w}_{k+1} - \mathbf{w}_k \in (1, 0), (0, 1), (1, 1)$ for $k \in [1 : d - 1]$.

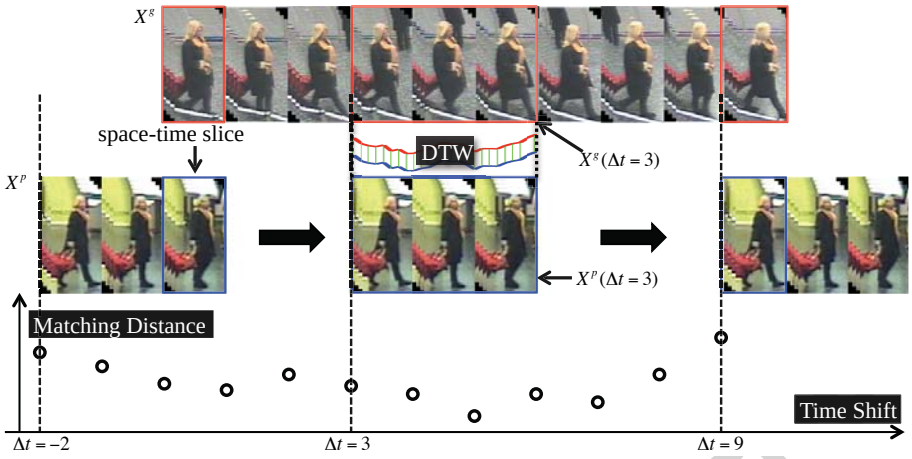


Figure 6: Overview of our proposed time shift driven sequence alignment and matching.

As indicated in the above bounding constraint, DTW assumes that the starting and ending data elements of the two sequences are aligned. However, this is mostly invalid in videos available for person ReID as aforementioned. Moreover, DTW utilises all sequence element data for distance computation, regardless the quality of individual elements. This is likely to make the obtained distance sensitive to data noise often present in typical ReID videos.

5.2. Time Shift Driven Alignment and Selective Matching

To overcome the above limitations of DTW, we develop a new model, Time Shift Dynamic Time Warping **(TS-DTW)**, by introducing additionally the notions of time shift and max-pooling into sequence matching. Instead of matching two sequences (X^p, X^g) holistically at one time as DTW, we perform iterative and partial matching. An illustration of this time shift driven sequence alignment and matching is depicted in Figure 6. Specifically, given two feature-sequences X^p (probe) and X^g (gallery), we temporally shift one sequence (say X^p) against the other (X^g) from the beginning position (where only the rightmost slice of X^p is utilised in matching with the leftmost slice of X^g, e.g. $\Delta t = -2$ as in Figure 6), to the ending position (where the rightmost slice of X^g is matched with the leftmost slice of X^p, e.g. $\Delta t = 9$ as in Figure 6, and black dotted vertical lines indicate several (not all) shift positions attempted

during the entire shifting process). At any shift Δt, the alignment between partial segments $X^p(\Delta t)$ and $X^g(\Delta t)$ (highlighted by the corresponding blue and red bounding box in Figure 6) is performed by the conventional DTW algorithm [89]. As such, a set of local matching distances $D = \{\text{dist}_{\text{dtw}}(X^p, X^g, \Delta t)\}_{\Delta t \in T}$ (indicated as the black hollow circles in Figure 6) can be generated over all time shifts T. Finally, we obtain the person video matching distance by taking together all local ones as

$$\text{dist}_{\text{ts}}(X^p, X^g) = \min_{\Delta t \in T}\{\text{dist}_{\text{dtw}}(X^p, X^g, \Delta t)\} \tag{5}$$

i.e. selecting the best-matched result. This *time shift ensemble* model is inspired by the max-pooling layer in neural networks which aim at summarising the responses of neighbouring groups of neurons [93]. We cope with a similar situation if sequence-element is thought of as neuron and sequence-segment as group of neurons. Critically, the max-pooling operation has data selection capability for guiding the supervised learning of neurons in neural network learning. Whereas our objective is to achieve data selective sequence matching or recognition in an unsupervised way, enjoying similar spirit but with a different learning strategy.

Discussion The data selection capability in our proposed matching algorithm above is significant to accurately matching sequences, especially for unregulated ReID videos from uncontrolled camera viewing conditions. We summarise the key points for data selection below. First, we automatically select the starting/ending walking poses, in contrast to DTW which enforces the first and last elements of compared sequences to be aligned so potentially introduces weak or noisy alignments into distance computation. Moreover, we attempt many different partial segments of X^p and X^g, and select the best-aligned parts for distance estimation, different from DTW that uses all observed data regardless of how good the constituent elements are. Thus, noisy elements can be possibly suppressed in distance computation. These two abilities are achieved by successively varying Δt, since the element data of $X^p(\Delta t)$ and $X^g(\Delta t)$ changes over time shifts. Apparently, the two benefits are complementary to each other and their combination allows us to more accurately match incomplete and noisy

surveillance videos for person ReID in an unsupervised manner, as demonstrated by our experimental evaluations in Section 8.

5.3. Generalisation to the Multi-Dimensional Setting

The TS-DTW model presented in Section 5.2 assumes one feature-sequence per person video. This is the single-dimensional setting, a special case of the multi-dimensional setting, e.g. $\geq$2 feature-sequences per video [92]. The term "dimension" here can be understood as a specific way of extracting feature-sequence from videos. Our setting is multi-dimensional (Figure 4). Specifically, defining a dimension in our context is related to one of the two aspects: (i) temporal pyramid (h_t levels); and (ii) spatial pyramid (h_s levels); Thus, we have a total of $h_t \times h_s$ dimensions (feature extraction ways). Note that, two feature-sequences at different dimensions for the same video may have different lengths, e.g. those extracted at different temporal pyramid levels (Section 4.2).

Generally, there are two strategies to combine information from multiple dimensions of sequences: (1) *dependent*, and (2) *independent*. We will generalise our TS-DTW model to the multi-dimensional setting using both strategies as detailed below.

(I) Dependent fusion. The dependent fusion strategy assumes that: (1) feature-sequences of a given video at different dimensions have the same length; (2) different dimensions are strongly correlated one another, i.e. their warping paths should be identical. Due to condition (1), we can not perform fusion of multiple dimensions across different temporal pyramid levels with this strategy. Consequently, we can only combine the h_s dimensions from different spatial divisions within each individual temporal pyramid level, those extracted from the same slice-sequence.

Formally, when matching two slice-sequences of the same temporal pyramid level: $S^p = \{s_1^p, \ldots, s_i^p, \ldots\}$ from video Q^p, and $S^g = \{s_1^g, \ldots, s_j^g, \ldots\}$ from video Q^g, we perform a joint sequence alignment by using the feature data of

all dimensions to compute the distance between two elements s_i^p and s_j^g as

$$\mathrm{dist}_{\mathrm{el}}^{\mathrm{D}}(s_i^p, s_j^g) = \sum_{k=1}^{\kappa} \alpha_k \times \mathrm{dist}_{\mathrm{el}}(\mathbf{x}_{(i,k)}^p, \mathbf{x}_{(j,k)}^g) \tag{6}$$

where $\mathbf{x}_{(i,k)}^p$ and $\mathbf{x}_{(j,k)}^g$ are the feature data in the k-th dimension for s_i^p and s_j^g, respectively, κ is the total number of dimensions to be fused, and α_k defines the weight of the k-th dimension. To incorporate the fine-to-coarse spatial information encoded in walking motion, we relate the value of α_k to the structure of spatial pyramid by setting

$$\alpha_k = 2^{\varepsilon_k} \tag{7}$$

where $\varepsilon_k \in [0, 1, \ldots, h_s - 1]$ denotes the spatial pyramid level of the k-th dimension (see Figure 5). This design is similar in spirit to pyramid kernel matching [86]. All fused dimensions are at the same level of the temporal pyramid whose structure is thus not considered here.

By replacing the single-dimensional distance $\mathrm{dist}_{\mathrm{el}}(\cdot,\cdot)$ of DTW with Eqn. (6), our TS-DTW model can be readily generalised to the multi-dimensional scenario and performs dimension fusion dependently. We call this dependently generalised model "**MDTS-DTW$_{\mathbf{D}}$**".

(II) Independent fusion. In contrast to the dependent fusion policy, the independent counterpart assumes independent alignment behaviours among individual dimensions by performing information combination in the distance level. Importantly, this strategy is more flexible than the former as it allows each dimension having their respective sequence structure, e.g. the sequence length. Therefore, sequences across different temporal pyramid levels can be combined in this fusion way. Similarly, we further take into account temporal fine-to-coarse structures and combine all dimensions to generate the final matching sequence distance between two videos Q^p and Q^g via

$$\mathrm{dist}^{\mathrm{I}}(Q^p, Q^g) = \sum_{k=1}^{\kappa} \beta_k \times \alpha_k \times \mathrm{dist}_k(Q^p, Q^g) \tag{8}$$

where $\beta_k = 2^{\tau_k}, \tau_k \in \{0, 1, \ldots, h_t - 1\}$ is the temporal pyramid level of the k-th dimension (see Figure 4), and $\mathrm{dist}_k(Q^p, Q^g)$ the corresponding matching

distance using our TS-DTW model, i.e. Eqn. (5). The parameters κ and α_k are same as in Eqns. (6) and (7). We call this model "**MDTS-DTW$_{\mathbf{I}}$**"

Usually, the two fusion strategies yield different matching results over the same dimensions. This is because each dimension may capture different aspects of video data and produce non-identical alignment solutions, and thus result in different distance values. We will evaluate and discuss their performances for person ReID in Section 8.

5.4. Model Complexity

We analyse the video matching complexity of our TS-DTW model. Formally, given two feature-sequences X^p and X^g, we need to compute the matching distance between $X^p(\Delta t)$ and $X^g(\Delta t)$ with the time shift $\Delta t \in T = \{-|X^p|+1, \ldots, |X^p|+|X^g|-1\}$. $|X^p(\Delta t)|$ (or $|X^g(\Delta t)|$) lies in the range of $[1, \min(|X^p|, |X^g|)]$ (see Figure 6). Therefore, the total matching complexity ψ_{tsdtw} of our TS-DTW model is

$$\psi_{\text{tsdtw}} = \sum_{\Delta t \in T} \psi_{\text{dtw}}(|X^p(\Delta t)|) \tag{9}$$

where $\psi_{\text{dtw}}(|X^p(\Delta t)|)$ refers to the matching complexity of DTW, which is $O(|X^p(\Delta t)|^2)$ by the standard DTW model [89] and $O(|X^p(\Delta t)|)$ by fast variants [95]. As person feature-sequences are typically short (e.g. <25 on PRID2011 and <40 on iLIDS-VID), the entire matching process is still efficient. Moreover, we can parallelise easily the matching process over individual time shifts for further reducing the running time, as they are independent against each other.

6. Person Re-Identification

Given a probe person video $Q^p \in P$ and a gallery set $G = \{Q_i^g\}$ captured from two non-overlapping cameras, person ReID aims to find the true identity match of Q^p in G. To achieve this, we first compute the space-time feature based distance $\text{dist}^{\text{st}}(Q^p, Q_i^g)$ between Q^p and every gallery video Q_i^g with our TS-DTW (Eqn. (5)) or MDTS-DTW$_{\text{D}}$ (Eqn. (6)) or MDTS-DTW$_{\text{I}}$ (Eqn. (8))

model. In this way, we can obtain all cross-camera pairwise video matching distances $\{\text{dist}^{\text{st}}(Q^p, Q_i^g)\}_{i=1}^{|G|}$. Finally, we generate a ranked list of all the gallery people in ascendant order of their matching distances, where the rank-1 gallery video is considered to be the most likely true match of Q^p.

Combination with the spatial appearance methods. The ReID matching distances computed by the proposed model can be readily fused with those by other spatial appearance models. In particular, we incorporate our results $\text{dist}^{\text{st}}(Q^p, Q_i^g)$ into other appearance based distance measures $\{\text{dist}_k^{\text{sp}}\}$ as

$$\text{dist}^{\text{fused}}(Q^p, Q_i^g) = \text{dist}^{\text{st}}(Q^p, Q_i^g) + \sum_k c_k \times \text{dist}_k^{\text{sp}}(Q^p, Q_i^g) \tag{10}$$

where c_i is a weighting assigned to the k-th method. Instead of cross-validation, we simply set $c_k = 1$ for generality consideration since in practice it is not always valid to assume the availability of pairwise labelled data which is required by cross-validation. As matching distances by distinct methods may lie in different ranges, we normalise all per-probe pairwise distances $\text{dist}^{\text{st}}(Q^p, Q_i^g)/\text{dist}_k^{\text{sp}}(Q^p, Q_i^g)$ to $[0, 1]$ per method separately before performing fusion. Specifically, given any matching distance $\text{dist}^* \in \{\text{dist}^{\text{st}}, \text{dist}_1^{\text{sp}}, \ldots, \text{dist}_k^{\text{sp}}, \ldots\}$, we rescale all distances $\{\text{dist}^*(Q^p, Q_i^g)\}_{i=1}^{|G|}$ with respect to a probe Q^p as

$$\widehat{\text{dist}}^*(Q^p, Q_i^g) = \frac{\text{dist}^*(Q^p, Q_i^g)}{\max(\{\text{dist}^*(Q^p, Q_i^g)\}_{i=1}^{|G|})} \tag{11}$$

where $\max(\cdot)$ returns the maximal value of a set. Then, the final fused distance can be expressed as

$$\widehat{\text{dist}}^{\text{fused}}(Q^p, Q_i^g) = \widehat{\text{dist}}^{\text{st}}(Q^p, Q_i^g) + \sum_k \widehat{\text{dist}}_k^{\text{sp}}(Q^p, Q_i^g) \tag{12}$$

We will evaluate the complementary effect between space-time and appearance features based person ReID methods in Section 8.

7. Experimental Settings

7.1. Datasets

Two benchmark image sequence based person ReID datasets (PRID2011 [22] and iLIDS-VID [17]) were utilised for evaluating the performance of the pro-

(a) PRID2011 (b) iLIDS-VID

Figure 7: Example person videos from the (a) PRID2011 [22] and (b) iLIDS-VID [17] datasets. In each dataset, every blue bounding box contains two videos from the same person captured by two non-overlapping camera views.

posed approach. Both datasets are challenging due to the large cross-view covariates in view point, illumination condition, and background noises. The dataset details are given below.

1. *PRID*2011. The PRID2011 dataset [22] includes 400 image sequences captured from 200 different people under two disjoint outdoor camera views. Each image sequence contains 5 to 675 image frames[5] (Figure 7a).
2. *iLIDS-VID.* The iLIDS-VID dataset [17] contains a total of 600 image sequences from 300 randomly sampled people, each with one pair of image sequences from two indoor camera views. Every image sequence has a variable length, e.g. consisting of 22 to 192 image frames (Figure 7b). Compared with PRID2011, this dataset has more complex occlusion and background clutter.

7.2. Baseline Methods

We compared our method with related state-of-the-art methods as follows:

1. *GEI-RSVM* [24]: A state-of-the-art gait recognition model using Gait Energy Image (GEI) feature [33] and the ranking SVM [96] model.

[5]For a fair comparison with existing methods, we followed the setting in [17], i.e. sequences of more than 21 frames from 178 people were selected and utilised in our evaluations.

2. *DTW* [89]: The widely used sequence matching algorithm - Dynamic Time Warping. DTW measures the distance between two sequences based on the optimal non-linear warping of elements across sequences.
3. *DDTW* [49]: In contrast to DTW directly comparing feature values of elements that can be sensitive to diverse variations, DDTW considers the global shape of sequences by matching the first derivative of the original sequences. Besides, DDTW allows to avoid singularities, i.e. a single element of one sequence may map with a large partition of another sequence, which may lead to pathological measures [48].
4. *WDTW* [50]: The weighted form of DTW model that also takes into account the shape similarity between two sequences. Specifically, WDTW introduces a multiplicative weight penalty on the warping distance between elements during distance estimation. This may suppress the negative influence of some outlier elements that are far away in element index but happen to be well matched. This model usually prefers close warping. We utilised a logistic weight function of the warping index-difference $\text{abs}(w_k^p - w_k^q)$ as: $f(w_k^p, w_k^q) = \frac{1}{1+\exp(-(\text{abs}(w_k^p - w_k^q)-\mu)/2)}$, where μ is the half average-length of two sequences Q^p and Q^g; w_k^p and w_k^q are the corresponding aligned element index of the k-th warp path entry (Eqn. (2)).
5. *SDALF* [3]: A classic hand-crafted visual appearance ReID feature. Both single and multiple shot cases are considered.
6. *eSDC* [6]: A state-of-the-art unsupervised spatial appearance based ReID method, which is able to learn localised appearance saliency statistics for measuring local patch importance.
7. *Iterative Sparse Ranking* (ISR) [97]: A contemporary weighted dictionary learning based algorithm that iteratively extends sparse discriminative classifiers in a transductive learning manner.
8. *Regularised Dictionary Learning* (RDL) [26]: The most recent dictionary learning based unsupervised ReID model. It iteratively learns the dictionary with the regularisation term updated in each iteration so that the cross-view noisy correspondence can be improved gradually.

9. *SS-ColLBP* [5]: A ranking SVM model [96] based ReID method with one of the most effective features Colour&LBP [5].
10. *MS-ColLBP* [17]: A multi-shot extension of SS-ColLBP. Specifically, the averaged Colour&LBP feature [5] over all image frames of a video is used to represent the spatial appearance of the person.
11. *L_1/L_2-norm*: The basic common distance metrics that can be very competitive with other complex metrics in many cases [98]. For matching two sequences, we remove the tail part of the longer one to make the two sequences have an equal duration.
12. *Kernelised Cross-View Discriminant Component Analysis* (KCVDCA) [99]: A competitive asymmetric distance learning method capable of inducing camera-specific projections for transforming unmatched visual features from different camera views to a shared subspace wherein discriminative features can be then learned and extracted.
13. *Cross-View Quadratic Discriminant Analysis (XQDA)* [27]: A state-of-the-art static appearance feature based supervised person ReID approach. Specifically, the XQDA algorithm learns simultaneously a discriminant low dimensional subspace and a QDA metric on the derived subspace.
14. *DVR* [18]: The state-of-the-art image-sequence based person ReID model which achieves the most competitive performance. In particular, this supervised model is characterised by discriminative fragment selection and exploitation for learning an effective space-time ranking function.

7.3. Person ReID Scenarios

We evaluated two person ReID scenarios, closed-world and open-world:

1. *Closed-World ReID*: In this setting, all probe people are assumed to exist in the gallery. In evaluations, we followed the data partition setting as [17, 18]. Specifically, for either PRID2011 or iLIDS-VID, we split the entire dataset into two partitions: one half for training, and the other half for testing. Note that our model does not utilise the training partition since it is unsupervised.

2. *Open-World ReID*: In addition, we evaluated a more realistic scenario called open-world ReID [23]. Specifically, its key difference from the closed-world case is that a probe person $i \in P$ is not assumed to appear *necessarily* in the gallery G under the open-world setting. This situation is more plausible to real-world ReID applications since we generally have no prior knowledge about whether one person (in gallery) re-appears in certain (probe) camera views in most applications, e.g. due to the complex topology structure of camera networks. That is, P and G may be just partially overlapped in different camera views. Similar data partitions as the closed-world case were utilised, with the only difference that the gallery set of the testing partition is reduced by one third ($\frac{1}{3}$) of randomly selected people (they are considered as imposters, only appearing in the probe set), i.e. 60 gallery people on PRID2011 and 100 on iLIDS-VID.

7.4. Evaluation Metrics

For closed-world ReID, the conventional Cumulated Matching Characteristics (CMC) curves were utilised for a quantitative performance comparison between different methods [1]. For open-world ReID, two separate steps are involved in performance evaluation under the open-world setting [23]: (1) Detection - decide if a probe person $Q^p \in P$ exists in the gallery or not; For convenience, we define $\bar{P} = P \setminus G$, the probe people that are not included in the gallery G. (2) Identification - compute the truly matched rates over only *accepted* target people. Specifically, we utilised detection and identification rate (DIR) and false accept rate (FAR) defined as:

$$\mathrm{DIR}(\tau, k) = \frac{|\{Q^p | \hat{Q}^g \in G, \mathrm{rank}(Q^p) \leq k,\ \mathrm{dist}(\hat{Q}^g, Q^p) \leq \tau\}|}{|G|} \tag{13}$$

$$\mathrm{FAR}(\tau) = \frac{|\{Q^p | Q^p \in \bar{P}, \min_{Q^g \in G} \mathrm{dist}(Q^g, Q^p) \leq \tau\}|}{|\bar{P}|} \tag{14}$$

where $\mathrm{dist}(\cdot, \cdot)$ refers to the cross-view distance score induced by some person ReID model, $\hat{Q}^g$ the gallery person having the same identity (i.e. true match) as the probe person Q^p, and τ the decision threshold. $\mathrm{rank}(\hat{Q}^g) = k$ means that

the true match $\hat{Q}^g$ is ranked at k in the ranking list. Thus, given a rank k, a Receiver Operating Characteristic (ROC) curve can be obtained by varying τ.

7.5. Implementation Details

Since video slices are localised over time, the value of l (the shortest slice length) should be small and related to the walking cycle length. We fixed $l = 5$ in that the process of a walking step takes around $2l = 10$ frames. Whilst the size h_t of the temporal pyramid largely depends on video length, e.g. an over-large h_t may lead to discarding many frames during sequentialisation (thus causing potentially much information loss), or very few slices produced for videos (with little temporal ordering dynamics). Thus, h_t is set to 2 accordingly. We utilised a 2-level spatial pyramid, i.e. $h_s = 2$. This is because, our empirical experiments suggest that the addition of one more spatial pyramid level slightly degrades the model performance possibly due to the local patch misalignment problem in over fine-grained spatial decomposition. The distance metric between sequence elements $\text{dist}_{\text{el}}(\cdot, \cdot)$ is set as L_1.

For obtaining stable statistics, we evaluated both person ReID scenarios with 10 folds of experiments with different random training/testing partitions on each dataset, and reported the averaged results.

8. Experimental Results

8.1. Evaluation on Our Proposed Approach

We evaluated the detailed aspects of the proposed video representation and sequence matching models for person ReID in the common *closed-world* scenario, i.e. the ReID accuracies of our TS-DTW and MDTS-DTW models using different parts of the proposed STPS features. The results are reported in Table 1. It is evident that both temporal and spatial pyramids are effective for person ReID and their fusion with the proposed method can improve significantly the matching accuracy. This is consistent with the finding in scene and action recognition [20, 21].

Table 1: The *closed-world* person ReID performance of the proposed TS-DTW (single-dimensional) and MDTS-DTW (multi-dimensional) model with different parts of our STPS video representation. (TPL: Temporal Pyramid Level; SPL: Spatial Pyramid Level)

Dataset	PRID2011 [22]				iLIDS-VID [17]			
Rank R (%)	1	5	10	20	1	5	10	20
TS-DTW(TPL^0,SPL^0)	36.7	59.1	73.5	84.7	23.3	51.5	65.2	79.6
TS-DTW(TPL^0,SPL^1)	32.5	63.8	75.4	84.9	12.3	37.0	53.2	68.5
MDTS-DTW$_\text{D}$(TPL^0)	37.1	60.2	73.7	85.7	25.1	51.9	66.5	79.9
MDTS-DTW$_\text{I}$(TPL^0)	39.2	60.8	75.3	86.6	25.9	52.7	67.1	79.1
TS-DTW(TPL^1,SPL^0)	34.2	58.9	74.4	86.1	23.8	49.5	62.7	78.4
TS-DTW(TPL^1,SPL^1)	32.4	61.7	77.0	87.2	16.5	40.7	53.4	68.7
MDTS-DTW$_\text{D}$(TPL^1)	36.2	60.3	74.8	86.3	23.8	50.0	62.5	78.6
MDTS-DTW$_\text{I}$(TPL^1)	37.2	61.7	75.2	87.0	24.3	50.1	62.4	78.5
MDTS-DTW$_\text{I}$(full)	**41.7**	**67.1**	**79.4**	**90.1**	**31.5**	**62.1**	**72.8**	**82.4**

Specifically, given either of the two temporal pyramid levels, when comparing with the coarse spatial pyramid level (SPL-1), the fine-grained spatial division (SPL-0) produces similar result on PRID2011, but significantly better accuracy on the more challenging iLIDS-VID. In contrast, with the same SPL, two temporal pyramid levels (TPL-0 and TPL-1) produce similar results. The plausible reason is that larger spatial regions are more likely to be contaminated by random noise in a crowded public space. When combining the matching results from different dimensions/feature-sequences of the same temporal pyramid level by either MDTS-DTW$_\text{D}$ or MDTS-DTW$_\text{I}$, the ReID accuracy can be improved similarly on both datasets. This suggests largely the independence property among distinct sequence dimensions, i.e. modelling their dependence does not bring any benefit in enhancing ReID. Moreover, after the results from different temporal granularities are fused by MDTS-DTW$_\text{I}$, ReID accuracies are further increased (note, MDTS-DTW$_\text{D}$ is not able to fuse image sequences of different lengths, see Section 5.3). These evidences show good complementary effect of different spatio-temporal pyramid levels and effectiveness of our model in fusing information from multiple localised motion patterns with different space-time extends. In the remaining evaluations, we utilised our MDTS-DTW$_\text{I}$ model and the full STPS video representation for comparison with the baseline methods.

Table 2: Comparing gait recognition and sequence matching methods (closed-world scenario).

Dataset	PRID2011 [22]				iLIDS-VID [17]			
Rank R (%)	1	5	10	20	1	5	10	20
GEI-RSVM [24]	20.9	45.5	58.3	70.9	2.8	13.1	21.3	34.5
DTW [89]	19.9	41.2	53.6	65.8	15.9	32.1	41.5	55.5
DDTW [49]	5.4	18.2	27.5	38.5	2.9	10.1	18.1	31.5
WDTW [50]	4.2	13.7	20.9	29.4	5.1	11.5	16.0	23.9
MDTS-DTW$_{\mathbf{I}}$	**41.7**	**67.1**	**79.4**	**90.1**	**31.5**	**62.1**	**72.8**	**82.4**

Computational cost: Apart from person re-id accuracy, we also evaluated the computational cost of our MDTS-DTW$_{\mathrm{I}}$ model on matching cross-view person videos for ReID. Time was measured on a work station (Intel i7-4770K CPU at 3.50 GHz and memory of 16 GB) with Matlab implementation in Windows OS. Time analysis was conducted under the same experimental setting as above. On average, matching each probe video against the gallery set takes 5.26 seconds on PRID (89 gallery people) and 9.50 seconds on iLIDS-VID (150 gallery people). That is, the average matching time for two person sequences is around 0.06 second. Note that, the whole process above can be conducted in parallel over a cluster of machines to further speed up model deployment.

8.2. Evaluation on Closed-World Person ReID

In this conventional setting, we performed comparative evaluations with gait recognition, temporal sequence matching, and person ReID approaches.

8.2.1. Comparing Gait Recognition and Temporal Sequence Matching Methods

In Table 2, we compared our MDTS-DTW$_{\mathrm{I}}$ model with a number of state-of-the-art gait recognition and dynamic programming based sequence matching methods. It is evident that the proposed model outperforms both alternative strategies by a large margin on each dataset. Specifically, the gait recognition method produces much better ReID accuracy on PRID2011 than on iLIDS-VID. This is because, the image sequences from the latter contain more background noise such as clutter and occlusion which can contaminate the gait feature heavily (see Figure 2). By automatically aligning starting/ending walking phases

and selecting best-matched sequence parts, our TS-DTW model allows to better overcome this challenge. On the other hand, conventional temporal sequence matching algorithms, e.g. DTW and its variants, can only provide much weaker results than the proposed MDTS-DTW. This is largely owing to: (1) ReID image sequences have different lengths with arbitrary starting/ending phases, and incomplete/noisy frames. Hence, attempts to match and utilise entire sequences inevitably suffer from mismatching with erroneous similarity measurement; (2) there is no explicit mechanism to avoid incomplete/missing data, typical in crowded surveillance scenes.

8.2.2. Comparing Person ReID Methods

We compared our MDTS-DTW$_{\mathrm{I}}$ method with contemporary unsupervised and supervised ReID methods, and further evaluated the complementary effect between appearance and space-time feature based approaches.

Table 3: Comparing unsupervised person ReID methods (closed-world scenario).

Dataset	PRID2011 [22]				iLIDS-VID [17]			
Rank R (%)	1	5	10	20	1	5	10	20
L_1-norm	26.4	47.5	57.8	73.7	19.3	39.2	51.9	66.5
L_2-norm	23.3	46.7	57.5	73.6	15.6	37.7	49.0	63.1
SS-SDALF [3]	4.9	21.5	30.9	45.2	5.1	14.9	20.7	31.3
MS-SDALF [3]	5.2	20.7	32.0	47.9	6.3	18.8	27.1	37.3
ISR [97]	17.3	38.2	53.4	64.5	7.9	22.8	30.3	41.8
eSDC [6]	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9
RDL [26]	29.1	53.6	66.2	76.1	11.5	26.2	34.3	46.3
MDTS-DTW$_{\mathrm{I}}$	**41.7**	**67.1**	**79.4**	**90.1**	**31.5**	**62.1**	**72.8**	**82.4**

Comparing unsupervised methods. Table 3 shows the comparison among unsupervised ReID approaches. The proposed MDTS-DTW$_{\mathrm{I}}$ outperforms significantly all competitors on PRID2011 and iLIDS-VID. Specifically, space-time feature based methods (e.g. ours and L_1/L_2-norm) produce better ReID accuracies than the remaining spatial appearance based methods, particularly on the more challenging iLIDS-VID dataset. This suggests the inherent challenge caused by the ambiguous and unreliable nature of people's appearance in person ReID applications, and simultaneously the exceptional effectiveness of

space-time cues for people matching when expressed and exploited effectively. In addition, the weak performance by SDALF is largely because of the intrinsic difficulty in designing general identity-discriminative hand-crafted appearance feature given unknown cross-camera covariates. Through iteratively learning and extending discriminative classifiers in ISR or modelling localised saliency statistics in eSDC or exploiting iteratively cross-view soft-correspondence in RDL, person ReID performance is greatly improved. However, due to relying on static appearance information alone, they are inherently sensitive to cross-camera viewing conditions, e.g. with a severe perform degradation from PRID2011 to iLIDS-VID. In contrast, our method mitigates this challenge by properly designing and effectively exploiting dynamic space-time features, another information source which presents better stability than the widely-used appearance features.

Table 4: Comparing supervised person ReID methods (closed-world scenario).

Dataset	PRID2011 [22]				iLIDS-VID [17]			
Rank R (%)	1	5	10	20	1	5	10	20
SS-ColLBP [5]	22.4	41.8	51.0	64.7	9.1	22.6	33.2	45.5
MS-ColLBP [5]	34.3	56.0	65.5	77.3	23.2	44.2	54.1	68.8
DVR [18]	40.0	71.7	84.5	92.2	**39.5**	61.1	71.7	81.0
KCVDCA [99]	43.8	69.7	76.4	87.6	16.7	43.3	54.0	70.7
XQDA [27]	**46.3**	**78.2**	**89.1**	**96.3**	16.7	39.1	52.3	66.8
MDTS-DTW$_{\mathbf{I}}$	41.7	67.1	79.4	90.1	31.5	**62.1**	**72.8**	**82.4**

Comparing supervised methods. We present the comparison between our unsupervised MDTS-DTW$_{\mathrm{I}}$ and previous supervised methods in Table 4. It is found that space-time feature based methods (i.e. DVR & ours) are less sensitive to crowded background than other appearance feature based models particularly XQDA and KCVDCA, when comparing the ReID performance on PRID and iLIDS-VID (more busy and crowded, see Figure 7). This is partially attributed to the selective matching strategy in the former models for extracting more reliable space-time representations. Moreover, it is observed that our method surpasses appearance based SS-/MS-ColLBP on two datasets and XQDA/KCVDCA on iLIDS-VID, and produces competitive results as video

based DVR. Note that the DVR model exploits both space-time and colour information in the price of exhaustive pairwise labelling whilst our MDTS-DTW$_{\mathrm{I}}$ method only utilises dynamic space-time cues without the need for cross-view pairwise labelling. These comparisons demonstrate the advantage and capability of our STPS video representation and selective matching model in extracting and exploiting identity-discriminative space-time information from noisy person videos for relaxing the label availability assumption and making better use of unregulated video data.

Table 5: Evaluating the complementary effect between space-time and appearance feature based person ReID methods (closed-world scenario).

Dataset	PRID2011 [22]				iLIDS-VID [17]			
Rank R (%)	1	5	10	20	1	5	10	20
DVR [18]	40.0	71.7	84.5	92.2	39.5	61.1	71.7	81.0
MDTS-DTW$_{\mathrm{I}}$	41.7	67.1	79.4	90.1	31.5	62.1	72.8	82.4
eSDC[6]	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9
eSDC+DVR[18]	44.3	68.4	78.2	91.1	29.5	54.0	66.4	78.4
eSDC+**MDTS-DTW$_{\mathrm{I}}$**	48.0	69.9	82.0	91.8	33.5	64.1	74.2	83.5
ISR [97]	17.3	38.2	53.4	64.5	7.9	22.8	30.3	41.8
ISR+DVR	43.8	63.3	72.5	81.3	30.0	46.0	55.1	63.6
ISR+**MDTS-DTW$_{\mathrm{I}}$**	46.2	66.7	72.6	83.3	33.1	51.5	58.7	69.7
RDL [26]	29.1	53.6	66.2	76.1	11.5	26.2	34.3	46.3
RDL+DVR	58.9	79.7	87.5	93.6	31.7	56.9	67.7	80.5
RDL+**MDTS-DTW$_{\mathrm{I}}$**	59.2	82.7	88.4	94.9	35.3	63.4	73.9	83.3
MS-ColLBP [5]	34.3	56.0	65.5	77.3	23.2	44.2	54.1	68.8
MS-ColLBP+DVR	44.8	66.9	77.1	89.9	39.5	61.0	72.7	82.8
MS-ColLBP+**MDTS-DTW$_{\mathrm{I}}$**	47.8	67.5	79.9	91.0	44.1	69.9	79.1	88.8
KCVDCA [99]	43.8	69.7	76.4	87.6	16.7	43.3	54.0	70.7
KCVDCA+DVR	65.7	88.1	93.4	97.3	**54.9**	76.8	83.7	91.3
KCVDCA+**MDTS-DTW$_{\mathrm{I}}$**	71.0	89.0	93.8	97.5	50.6	**77.0**	**85.6**	**92.6**
XQDA [27]	46.3	78.2	89.1	96.3	16.7	39.1	52.3	66.8
XQDA+DVR	**77.4**	**93.9**	**97.0**	**99.4**	51.1	75.7	83.9	90.5
XQDA+**MDTS-DTW$_{\mathrm{I}}$**	69.6	89.4	94.3	97.9	49.5	75.7	84.5	91.9

Evaluating complementary effect. We further evaluated how well spatial appearance and space-time feature based ReID methods complement each other. To this end, we integrated contemporary unsupervised (eSDC, ISR and RDL) and supervised (MS-ColLBP, KCVDCA and XQDA) appearance based approaches with DVR and our MDTS-DTW$_{\mathrm{I}}$ model (Eqn. (10)), respectively.

The results are presented in Table 5. It is observed that by fusing space-time feature based ReID results of either DVR or ours, the matching accuracies of existing appearance based methods can be significantly boosted. This confirms the similar finding by [18] that, the combination of appearance and space-time motion information sources can be very effective for person ReID as they are largely independent in nature. Overall, XQDA+DVR achieves the best performance on PRID2011 whilst KCVDCA+Ours and KCVDCA+DVR perform similarly best on iLIDS-VID. This is as expected because the combination with DVR doubly benefits much from effective modelling on labelled data which contain strong discriminative information but very expensive to acquire for every camera pair in reality. Once removing the label availability assumption, the best results are obtained by eSDC+Ours on iLIDS-VID and RDL+Ours on PRID2011. Under the unsupervised setting, we observed a similar complementary effect as XQDA/KCVDCA+DVR/Ours. This validates the efficacy of our ReID method in deriving dynamic identity information from unregulated videos, independent of and completing well the commonly used spatial appearance.

8.3. Evaluation on Open-World Person ReID

In this section, we evaluated the open-world ReID problem, a more practical scenario compared to the above closed-world setting. Different single ReID methods and their combinations were assessed and reported in Table 6. The performance evaluation metric is Detection and Identification Rate (DIR, Eqn. (13)) with $k = 1$ (e.g. Rank-1) at given False Accept Rates (FAR, Eqn. (14)). For the performance of single models, largely similar situations are found as in the closed-world case. Particularly, for iLIDS-VID, the supervised space-time ReID method DVR obtains the best results followed by our approach and KCVDCA but ours is unsupervised. On PRID2011, our method has the best DIR scores given low ($\leq 10\%$) FAR rates (corresponding to small τ in Eqn. (14)). That means, our method can recognise more accurately the true match at rank-1 when the false accept rate is required to be small. This situation is mostly ignored in the current ReID literature but very important in real-world

Table 6: Comparing the *open-world* ReID performance. Metric: Detection and Identification Rate (DIR, Eqn. (13) with $k = 1$) over four False Accept Rates (FAR, Eqn. (14)).

Dataset	PRID2011 [22]				iLIDS-VID [17]			
FAR (%)	1	10	50	100	1	10	50	100
L_1-norm	4.3	8.7	18.5	28.3	1.0	5.2	15.6	22.9
MS-SDALF [3]	0.5	1.0	4.5	6.3	0.2	0.5	3.3	8.4
ISR [97]	0.0	18.0	18.2	18.8	0.0	8.9	8.9	10.6
eSDC [6]	5.2	9.7	20.8	28.3	1.4	4.2	8.3	12.4
RDL [26]	9.3	13.3	27.5	33.0	2.1	4.9	10.4	13.9
MS-ColLBP [5]	4.3	6.7	24.3	39.8	1.1	4.8	15.6	25.9
DVR [18]	4.0	12.3	34.7	46.8	4.2	**14.1**	**31.8**	**43.7**
KCVDCA [99]	14.5	20.2	**43.0**	49.5	**7.1**	12.1	20.8	24.8
XQDA [27]	11.5	19.8	40.3	**51.7**	1.3	4.3	11.5	21.2
MDTS-DTW$_\mathbf{I}$	**17.5**	**25.5**	38.2	46.5	3.4	8.7	26.4	37.0
eSDC+DVR	13.3	25.2	43.3	48.5	7.2	14.4	27.7	34.6
eSDC+**MDTS-DTW$_\mathbf{I}$**	16.8	28.2	44.7	51.3	6.3	12.0	31.5	39.6
ISR+DVR	15.0	27.8	42.7	47.7	10.7	20.3	29.3	32.9
ISR+**MDTS-DTW$_\mathbf{I}$**	25.2	36.0	46.8	49.7	11.3	17.6	32.6	35.7
RDL+DVR	26.7	39.3	58.8	62.7	8.5	15.4	30.1	37.3
RDL+**MDTS-DTW$_\mathbf{I}$**	21.8	38.5	59.7	63.7	9.2	18.7	33.7	41.7
MS-ColLBP+DVR	25.5	29.2	45.8	50.0	16.3	22.6	38.5	43.3
MS-ColLBP+**MDTS-DTW$_\mathbf{I}$**	27.7	33.2	49.7	51.2	11.6	21.3	43.8	50.0
KCVDCA+DVR	31.7	55.2	72.5	75.3	17.0	29.7	50.9	56.4
KCVDCA+**MDTS-DTW$_\mathbf{I}$**	42.7	52.8	72.5	73.5	16.8	30.2	51.5	56.4
XQDA+DVR	**46.8**	**58.3**	**78.3**	**79.7**	**17.3**	29.1	49.9	**57.8**
XQDA+**MDTS-DTW$_\mathbf{I}$**	42.7	55.2	70.5	72.8	12.7	**32.6**	**51.8**	57.3

applications, particularly when a large number of probe people are given and high FARs are not acceptable.

When fusing appearance and space-time feature based ReID methods, the recognition scores across all FARs are greatly improved, similar to the early observations. In particular, the best ReID accuracies are obtained by the combination of XQDA/KCVDCA and DVR/Ours, assuming truth match labels are accessible. In the unsupervised setting, RDL+Ours is the best on both PRID201 and iLIDS-VID. Clearly, most findings in the closed-world scenario can be reflected in the open-world setting, whilst some new different observations emerge especially under strict false accept rate conditions. In general, all comparisons above extensively validate the advantages and effectiveness of the proposed video representation and selective matching models for person ReID.

9. Conclusion and Future Work

Conclusion. In this work, we presented a video matching based person ReID framework. This is achieved by (1) developing an effective spatio-temporal pyramids based video representation, called Spatio-Temporal Pyramid Sequence (STPS), for encoding more effective and complete space-time information available in person video data; and (2) formulating a novel Time Shift Dynamic Time Warping (TS-DTW) model and its Multi-Dimensional extension named MDTS-DTW for selective matching between pairs of inherently incomplete and noisy image sequences from two disjoint camera views. Our method also shows significant complementary effect on previous spatial appearance based ReID approaches for obtaining favourable ReID accuracies. Importantly, our model is unsupervised and does not require exhaustive cross-view pairwise data annotation for every camera pair in model building. Under both the closed-world and open-world ReID scenarios, extensive comparative evaluations have demonstrated clearly the advantages of the proposed approach over a wide range of contemporary state-of-the-art gait recognition, temporal sequence matching, supervised and unsupervised ReID methods.

Future work. Our future work for the unsolved person ReID problem includes: (1) How to introduce other complementary schemes beyond time shift based data selection for further suppressing noisy observations caused by background distractions; (2) How to exploit effectively extra types of information (e.g. semantic text from human or correlated sources) as computing constraints for improving the matching performance.

Acknowledgements

This work was partially supported by National Basic Research Program of China (973 Project) 2012CB725405, the national science and technology support program(2014BAG03B01), National Natural Science Foundation China 61273238, Beijing Municipal Science and Technology Project (D15110900280000), Tsinghua University Project (20131089307) and the Foundation of Beijing Key

Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control. Xiatian Zhu and Xiaolong Ma equally contributed to this work.

References

[1] S. Gong, M. Cristani, S. Yan, C. C. Loy, Person re-identification, Springer, 2014.

[2] S. Gong, M. Cristani, C. C. Loy, T. M. Hospedales, The re-identification challenge, in: Person Re-Identification, Springer, 2014, pp. 1–20.

[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2360–2367.

[4] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by support vector ranking, in: British Machine Vision Conference, 2010.

[5] M. Hirzer, P. M. Roth, M. Köstinger, H. Bischof, Relaxed pairwise learned metric for person re-identification, in: European Conference on Computer Vision, 2012, pp. 780–793.

[6] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3586–3593.

[7] A. Bhuiyan, A. Perina, V. Murino, Person re-identification by discriminatively selecting parts and features, in: Workshop of European Conference on Computer Vision, 2014, pp. 147–161.

[8] C. Liu, S. Gong, C. C. Loy, On-the-fly feature importance mining for person re-identification, Pattern Recognition 47 (2014) 1602–1615.

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

[10] L. Zhang, T. Xiang, S. Gong, Learning a discriminative null space for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[11] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al., Evaluation of local spatio-temporal features for action recognition, in: British Machine Vision Conference, 2009.

[12] R. Poppe, A survey on vision-based human action recognition, Image and Vision Computing 28 (2010) 976–990.

[13] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, K. W. Bowyer, The humanid gait challenge problem: Data sets, performance, and analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2) (2005) 162–177.

[14] K. Bashir, T. Xiang, S. Gong, Gait recognition without subject cooperation, Pattern Recognition Letters 31 (2010) 2052–2060.

[15] Y. Yang, D. Ramanan, Articulated human detection with flexible mixtures of parts, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (12) (2013) 2878–2890.

[16] W. Ouyang, X. Chu, X. Wang, Multi-source deep learning for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2337–2344.

[17] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: European Conference on Computer Vision, 2014, pp. 688–703.

[18] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by discriminative selection in video ranking, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (12) (2016) 2501–2514.

[19] A. Klaser, M. Marszalek, A spatio-temporal descriptor based on 3d-gradients, in: British Machine Vision Conference, 2008.

[20] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 2169–2178.

[21] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2847–2854.

[22] M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Scandinavian Conference on Image Analysis, 2011.

[23] S. Liao, Z. Mo, Y. Hu, S. Z. Li, Open-set person re-identification, arXiv preprint (2014) 1–16.

[24] R. Martín-Félez, T. Xiang, Gait recognition by ranking, in: European Conference on Computer Vision, 2012, pp. 328–341.

[25] L. R. Rabiner, B.-H. Juang, Fundamentals of speech recognition, Vol. 14, PTR Prentice Hall Englewood Cliffs, 1993.

[26] E. Kodirov, T. Xiang, S. Gong, Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification, in: British Machine Vision Conference, 2015.

[27] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.

[28] D. Xu, Y. Huang, Z. Zeng, X. Xu, Human gait recognition using patch distribution feature and locality-constrained group sparse representation, IEEE Transactions on Image Processing 21 (1) (2012) 316–326.

[29] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, G. Rigoll, The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits, Journal of Visual Communication and Image Representation 25 (1) (2014) 195–206.

[30] P. Chattopadhyay, A. Roy, S. Sural, J. Mukhopadhyay, Pose depth volume extraction from rgb-d streams for frontal gait recognition, Journal of Visual Communication and Image Representation 25 (1) (2014) 53–63.

[31] S. D. Choudhury, T. Tjahjadi, Robust view-invariant multiscale gait recognition, Pattern Recognition 48 (3) (2015) 798–811.

[32] T. Kobayashi, N. Otsu, Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation, in: IEEE International Conference on Pattern Recognition, Vol. 3, 2004, pp. 741–744.

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

[34] G. V. Veres, L. Gordon, J. N. Carter, M. S. Nixon, What image information is important in silhouette-based gait recognition?, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2004, pp. II–776.

[35] D. S. Matovski, M. S. Nixon, S. Mahmoodi, T. Mansfield, On including quality in applied automatic gait recognition, in: IEEE International Conference on Pattern Recognition, 2012, pp. 3272–3275.

[36] M. Hofmann, S. Sural, G. Rigoll, Gait recognition in the presence of occlusion: A new dataset and baseline algorithms, in: International Conference on Computer Graphics, Visualization and Computer Vision, 2011, pp. 99–104.

[37] N. V. Boulgouris, Z. X. Chi, Human gait recognition based on matching of body components, Pattern Recognition 40 (6) (2007) 1763–1770.

[38] M. A. Hossain, Y. Makihara, J. Wang, Y. Yagi, Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control, Pattern Recognition 43 (6) (2010) 2281–2291.

[39] S. H. Shaikh, K. Saeed, N. Chaki, Gait recognition using partial silhouette-based approach, in: IEEE International Conference on Signal Processing and Integrated Networks, 2014, pp. 101–106.

[40] D. Muramatsu, Y. Makihara, Y. Yagi, Gait regeneration for recognition, in: IAPR International Conference on Biometrics, 2015, pp. 1–8.

[41] J. Xiao, H. Cheng, H. Sawhney, C. Rao, M. Isnardi, Bilateral filtering-based optical flow estimation with occlusion detection, in: European Conference on Computer Vision, 2006, pp. 211–224.

[42] S. Yu, D. Tan, T. Tan, Modelling the effect of view angle variation on appearance-based gait recognition, Asian Conference on Computer Vision (2006) 807–816.

[43] X. Yang, Y. Zhou, T. Zhang, G. Shu, J. Yang, Gait recognition based on dynamic region analysis, Signal Processing 88 (9) (2008) 2350–2356.

[44] S. Singh, K. Biswas, Biometric gait recognition with carrying and clothing variants, in: Pattern Recognition and Machine Intelligence, 2009, pp. 446–451.

[45] R. Martín-Félez, T. Xiang, Uncooperative gait recognition by learning to rank, Pattern Recognition 47 (12) (2014) 3793–3806.

[46] P. Senin, Dynamic time warping algorithm review, Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA (2008) 1–23.

[47] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, E. Keogh, Searching and mining trillions of time series subsequences under dynamic time warping, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 262–270.

[48] E. J. Keogh, M. J. Pazzani, Derivative dynamic time warping, in: SIAM International Conference on Data Mining, Vol. 1, 2001, pp. 5–7.

[49] F. Gullo, G. Ponti, A. Tagarelli, S. Greco, A time series representation model for accurate and fast similarity detection, Pattern Recognition 42 (11) (2009) 2998–3014.

[50] Y.-S. Jeong, M. K. Jeong, O. A. Omitaomu, Weighted dynamic time warping for time series classification, Pattern Recognition 44 (9) (2011) 2231–2240.

[51] J.-H. Horng, J. T. Li, An automatic and efficient dynamic programming algorithm for polygonal approximation of digital curves, Pattern Recognition Letters 23 (1–3) (2002) 171–182.

[52] R. Oka, Spotting method for classification of real world data, The Computer Journal 41 (8) (1998) 559–565.

[53] Z. Wu, Y. Li, R. J. Radke, Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (5) (2015) 1095–1108.

[54] Y.-C. Chen, W.-S. Zheng, J. Lai, Mirror representation for modeling view-specific transform in person re-identification, in: International Joint Conference of Artificial Intelligence, 2015, pp. 3402–3408.

[55] H. Wang, S. Gong, X. Zhu, T. Xiang, Human-in-the-loop person re-identification, in: European Conference on Computer Vision, 2016, pp. 405–422.

[56] H. Wang, X. Zhu, T. Xiang, S. Gong, Towards unsupervised open-set person re-identification, in: IEEE International Conference on Image Processing, 2016.

[57] E. Kodirov, T. Xiang, Z. Fu, S. Gong, Person re-identification by unsupervised l1 graph learning, in: European Conference on Computer Vision, 2016.

[58] O. Hamdoun, F. Moutarde, B. Stanciulescu, B. Steux, Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences, in: ACM International Conference on Distributed Smart Cameras, 2008, pp. 1–6.

[59] D. N. T. Cong, C. Achard, L. Khoudour, L. Douadi, Video sequences association for people re-identification across multiple non-overlapping cameras, in: International Conference on Image Analysis and Processing, 2009, pp. 179–189.

[60] C. Nakajima, M. Pontil, B. Heisele, T. Poggio, Full-body person recognition system, Pattern Recognition 36 (2003) 1997–2006.

[61] N. Gheissari, T. B. Sebastian, R. Hartley, Person reidentification using spatiotemporal appearance, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1528–1535.

[62] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification, in: British Machine Vision Conference, 2011.

[63] Y. Xu, L. Lin, W.-S. Zheng, X. Liu, Human re-identification by matching compositional template with cluster sampling, in: IEEE International Conference on Computer Vision, 2013.

[64] A. Roy, S. Sural, J. Mukherjee, A hierarchical method combining gait and phase of motion with spatiotemporal model for person re-identification, Pattern Recognition Letters 33 (14) (2012) 1891–1901.

[65] R. Kawai, Y. Makihara, C. Hua, H. Iwama, Y. Yagi, Person re-identification using view-dependent score-level fusion of gait and color features, in: IEEE International Conference on Pattern Recognition, 2012, pp. 2694–2697.

[66] A. Bedagkar-Gala, S. K. Shah, Gait-assisted person re-identification in wide area surveillance, in: Workshop of Asian Conference on Computer Vision, 2014, pp. 633–649.

[67] Z. Liu, Z. Zhang, Q. Wu, Y. Wang, Enhancing person re-identification by integrating gait biometric, Neurocomputing 168 (30) (2015) 1144–1156.

[68] J. You, A. Wu, X. Li, W.-S. Zheng, Top-push video-based person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[69] N. McLaughlin, J. Martinez del Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[70] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: IEEE International Conference on Pattern Recognition, Vol. 3, 2004, pp. 32–36.

[71] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.

[72] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[73] T.-K. Kim, R. Cipolla, Canonical correlation analysis of video volume tensors for action categorization and detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (8) (2009) 1415–1428.

[74] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: ACM International Conference on Multimedia, 2007, pp. 357–360.

[75] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: European Conference on Computer Vision, 2008, pp. 650–663.

[76] Y. Zhu, S. Lucey, Convolutional sparse coding for trajectory reconstruction, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (3) (2015) 529–540.

[77] S. Nowozin, G. Bakir, K. Tsuda, Discriminative subsequence mining for action classification, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[78] K. Schindler, L. Van Gool, Action snippets: How many frames does human action recognition require?, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[79] J. C. Niebles, C.-W. Chen, L. Fei-Fei, Modeling temporal structure of decomposable motion segments for activity classification, in: European Conference on Computer Vision, 2010, pp. 392–405.

[80] A. Gaidon, Z. Harchaoui, C. Schmid, Actom sequence models for efficient action detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3201–3208.

[81] A. Gaidon, Z. Harchaoui, C. Schmid, A time series kernel for action recognition, in: British Machine Vision Conference, 2011.

[82] L. Zelnik-Manor, M. Irani, Event-based analysis of video, in: IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[83] M. Irani, P. a. Anandan, J. Bergen, R. Kumar, S. Hsu, Efficient representations of video sequences and their applications, Signal Processing: Image Communication 8 (4) (1996) 327–351.

[84] J. Choi, W. J. Jeon, S.-C. Lee, Spatio-temporal pyramid matching for sports videos, in: ACM International Conference on Multimedia Information Retrieval, 2008, pp. 291–297.

[85] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, M. Shah, Visual tracking: An experimental survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (7) (2014) 1442–1468.

[86] K. Grauman, T. Darrell, The pyramid match kernel: Efficient learning with sets of features, The Journal of Machine Learning Research 8 (2007) 725–760.

[87] F. Shi, R. Laganiere, E. Petriu, H. Zhen, Lpm for fast action recognition with large number of classes, in: Workshop of IEEE International Conference on Computer Vision, 2013.

[88] H. Wang, D. Oneata, J. Verbeek, C. Schmid, A robust and efficient video representation for action recognition, International Journal of Computer Vision (2015) 1–20.

[89] D. J. Berndt, J. Clifford, Using dynamic time warping to find patterns in time series., in: Workshop of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Vol. 10, 1994, pp. 359–370.

[90] A. M. Fraser, H. L. Swinney, Independent coordinates for strange attractors from mutual information, Physical Review A 33 (2) (1986) 1134.

[91] C. C. Loy, T. Xiang, S. Gong, Time-delayed correlation analysis for multi-camera activity understanding, International Journal of Computer Vision 90 (2010) 106–129.

[92] M. Shokoohi-Yekta, J. Wang, E. Keogh, On the non-trivial generalization of dynamic time warping to the multi-dimensional case, in: SIAM International Conference on Data Mining, 2015, pp. 39–48.

[93] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[94] M. Müller, Dynamic time warping, Information retrieval for music and motion (2007) 69–84.

[95] S. Salvador, P. Chan, Fastdtw: Toward accurate dynamic time warping in linear time and space, in: Workshop of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.

[96] O. Chapelle, S. S. Keerthi, Efficient algorithms for ranking with svms, Information Retrieval 13 (2010) 201–215.

[97] G. Lisanti, I. Masi, A. Bagdanov, A. Del Bimbo, Person re-identification by iterative re-weighted sparse ranking, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (8) (2015) 1629–1642.

[98] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, E. Keogh, Querying and mining of time series data: experimental comparison of representations and distance measures, Proceedings of the Very Large Data Bases Endowment 1 (2) (2008) 1542–1552.

[99] Y.-C. Chen, W.-S. Zheng, P. C. Yuen, J. Lai, An asymmetric distance model for cross-view feature mapping in person re-identification, in: IEEE Transactions on Circuits and Systems for Video Technology, Vol. PP, 2015, pp. 1–1.