In Proc. European Conference on Computer Vision, Zurich, Switzerland, September 2014

Person Re-Identification by Video Ranking

Taiqing Wang[1], Shaogang Gong[2], Xiatian Zhu[2], and Shengjin Wang[1]

[1] Dept. of Electronic Engineering, Tsinghua University
[2] School of EECS, Queen Mary University of London

Abstract. Current person re-identification (re-id) methods typically rely on single-frame imagery features, and ignore space-time information from image sequences. Single-frame (single-shot) visual appearance matching is inherently limited for person re-id in public spaces due to visual ambiguity arising from non-overlapping camera views where viewpoint and lighting changes can cause significant appearance variation. In this work, we present a novel model to automatically select the most discriminative video fragments from noisy image sequences of people where more reliable space-time features can be extracted, whilst simultaneously to learn a video ranking function for person re-id. Also, we introduce a new image sequence re-id dataset (iLIDS-VID) based on the i-LIDS MCT benchmark data. Using the iLIDS-VID and PRID 2011 sequence re-id datasets, we extensively conducted comparative evaluations to demonstrate the advantages of the proposed model over contemporary gait recognition, holistic image sequence matching and state-of-the-art single-shot/multi-shot based re-id methods.

1 Introduction

In person re-identification, one matches a probe (or query) person against a set of gallery persons for generating a ranked list according to their matching similarity, typically assuming the correct match is assigned to one of the top ranks, ideally Rank-1 [50, 20, 11, 12]. As the probe and gallery persons are often captured from a pair of non-overlapping camera views at different time, cross-view visual appearance variation can be significant. Re-identification by visual matching is inherently challenging [14]. The state-of-the-art methods perform this task mostly by matching spatial appearance features (e.g. colour and intensity gradient histograms) using a pair of single-shot person images [11, 35, 20, 49]. However, single-shot appearance features are intrinsically limited due to the inherent visual ambiguity caused by clothing similarity among people in public spaces, and appearance changes from cross-view illumination variation, viewpoint difference, cluttered background and occlusions (Fig. 1). It is desirable to explore space-time information from image sequences of people for re-identification in public spaces.

Space-time information has been explored extensively for action recognition [34, 45]. Moreover, discriminative space-time video patches have also been exploited for action recognition [37]. However, action recognition approaches are

(a) Cross-view lighting variations (b) Camera viewpoint changes

(c) Clothing similarity (d) Background clutter & occlusions

Fig. 1. Person re-identification challenges in public space scenes [42].

not directly applicable to person re-identification because pedestrians in public spaces exhibit similar walking activities without distinctive and semantically categorisable actions unique to different people. On the other hand, gait recognition techniques have been developed for person recognition using image sequences by discriminating subtle distinctiveness in the style of walking [33, 38]. Different from action recognition, gait is a behavioural biometric that measures the way people walk. An advantage of gait recognition is no assumption being made on either subject cooperation (framing) or person distinctive actions (posing). These are similar to person re-id situations. However, existing gait recognition models are subject to stringent requirements on person foreground segmentation and accurate alignment over time throughout a gait image sequence (cycle). It is also assumed that complete gait cycles were captured in target image sequences [17, 31]. Most gait recognition methods do not cope well with cluttered background and/or random occlusion with unknown covariate conditions [1]. Person re-id in public spaces is inherently challenging for gait recognition (Fig. 1).

In this study, we aim to construct a discriminative video matching framework for person re-identification by selecting more reliable space-time features from videos of a person. To that end, we assume the availability of image sequences of people which may be highly noisy, i.e. with arbitrary sequence duration and starting/ending frames, unknown camera viewpoint/lighting variations during each image sequence, also with likely incomplete frames due to occlusion. We call this *unregulated* image sequences of people (Fig. 1 and Fig. 4). More specifically, we propose a novel approach to Discriminative Video fragments selection and Ranking (DVR) based on a robust space-time feature representation given unregulated image sequences of people.

The main contributions of this study are: (1) We derive a multi-fragments based space-time feature representation of image sequences of people. This representation is based on a combination of HOG3D features and optic flow energy profile over each image sequence, designed to break down automatically unregulated video clips of people into multiple fragments. (2) We propose a discrim-

inative video ranking model for cross-view re-identification by simultaneously selecting and matching more reliable space-time features from video fragments. The model is formulated using a multi-instance ranking strategy for learning from pairs of image sequences over non-overlapping camera views. This method can significantly relax the strict assumptions required by gait recognition techniques. (3) We introduce a new image sequence based person re-identification dataset called iLIDS-VID, extracted from the i-LIDS Multiple-Camera Tracking Scenario (MCTS) [42]. To our knowledge, this is the largest image sequence based re-identification dataset that is publically available.

2 Related Work

Space-time features - Space-time feature representations have been extensively explored in action/activity recognition [34, 43, 15]. One common representation is constructed based on space-time interest points [26, 10, 46, 5]. They facilitate a compact description of image sequences based on sparse interest points, but are somewhat sensitive to shadows and highlights in appearance [24] and may lose discriminative information [13]. Thus they may not be suitable to person re-id scenarios where lighting variations are unknown and uncontrolled. Alternatively, space-time volume/patch based representations [34] can be more robust. Mostly these are spatial-temporal extensions of image descriptors, e.g. HoGHoF [27], 3D-SIFT [39], HOG3D [25]. In this study, we adopt HOG3D [25] as space-time features for video fragment representation due to: (1) they can be computed efficiently; (2) they contain both spatial gradient and temporal dynamic information, therefore potentially more expressive [43, 25]; (3) they are more robust against cluttered background and occlusions [25]. The choice of a space-time feature representation is independent of our model.

Gait recognition - Space-time information of sequences has been extensively exploited by gait recognition [33, 38, 17, 31]. However, these methods often make stringent assumptions on the image sequences, e.g. uncluttered background, consistent silhouette extraction and alignment, accurate gait phase estimation and complete gait cycles, most of which are unrealistic in typical person re-id scenarios. It is challenging to extract a suitable gait representation from such re-id data. In contrast, our approach relaxes significantly these assumptions by simultaneously selecting discriminative video fragments from noisy image sequences, and matching them cross-views without temporal alignment.

Temporal sequence matching - One approach to exploiting image sequences for re-identification is sequence matching. For instance, Dynamic Time Warping (DTW) is a popular sequence matching method widely used for action recognition [29], and more recently also for person re-id [40]. However, given two sequences with unsynchronised starting/ending frames, it is difficult to align sequences for matching, especially if the image sequences are subject to significant noise caused by unknown camera viewpoint change, background clutters and significant lighting changes. Our approach is designed to address this problem

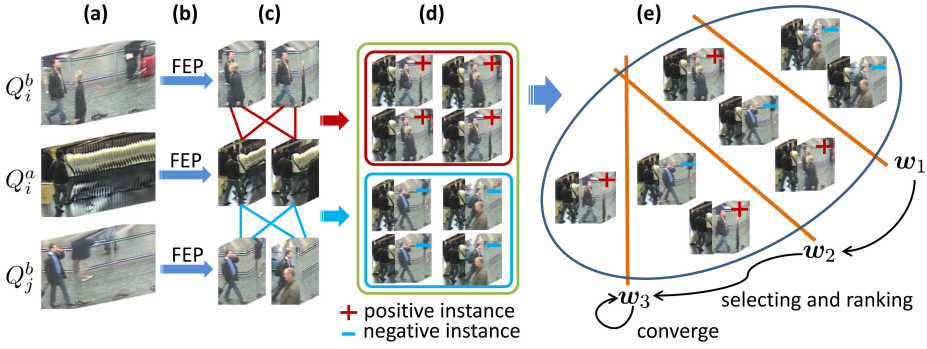


Fig. 2. Pipeline of the training phase of our DVR model. (a) Image sequences, Q_i^a denotes the image sequence of person p_i in camera a (Sec. 3.1). (b) Generating candidate fragment pools by Flow Energy Profiling (FEP) (Sec. 3.2). (c)-(d) Creating candidate fragment pairs as positive and negative instances (Sec. 3.3). (e) Simultaneously selecting and ranking the most discriminative fragment pairs (Sec. 3.3).

so to avoid any implicit assumptions on sequence alignment and camera view similarity among image frames both within and between sequences.

Multi-shot re-identification - Multiple images of a sequence have been exploited for person re-identification. For example, interest points were accumulated across images for capturing appearance variability [16]. Manifold geometric structures from image sequences of people were utilised to construct more compact spatial descriptors of people [8]. The time index of image frames and identity consistency of a sequence were used to constrain spatial feature similarity estimation [23]. There are also attempts on training an appearance model from image sets [32] or by selecting best pairs [28]. Multiple images of a person sequence were used either to enhance local image region/patch spatial feature description [12, 11, 7, 48], or to extract additional appearance information such as change statistics [2]. In contrast, the proposed model in this work aims to simultaneously select and match discriminative video space-time features for maximising cross-view ranking. Our experiments show the advantages of the proposed model over existing multi-shot models for person re-identification.

3 A Framework for Discriminative Video Ranking

We wish to construct a model capable of simultaneously selecting and matching discriminative video fragments from unregulated pairs of image sequences of people captured from two non-overlapping camera views (Fig. 2(a)). The model is based on (1) optic flow energy profiling over time in the image sequences, (2) HOG3D space-time feature extraction from video fragments and (3) a multi-instance learning strategy for simultaneous discriminative video fragments selection and cross-view matching by ranking. The learned model can then be deployed to perform person re-identification given unseen probe image sequences

against a set of gallery image sequences of people with arbitrary sequence length and unknown/unsegmented walking cycles. An overview diagram of the proposed approach is depicted in Fig. 2.

3.1 Problem Definition

Suppose we have a collection of person sequence pairs $\{(Q_i^a, Q_i^b)\}_{i=1}^N$, where Q_i^a and Q_i^b refer to the image sequences of person p_i captured by two disjoint cameras a and b, and N the number of people in a training set. Each image sequence is defined as a set of consecutive frames I obtained by an independent person tracker, e.g. [3, 18] : $Q = (I_1, ..., I_t)$, where t is not a constant as in typical surveillance videos, tracked person image sequences do not have guaranteed uniform length (arbitrary number of frames), nor number of walking cycles and starting/ending phases. For model training, we aim to learn a ranking function of image sequences $f(Q^a, Q^b)$ that satisfies the ranking constraints as:

$$f(Q_i^a, Q_i^b) > f(Q_i^a, Q_j^b), \forall i = \{1, ..., N\}, \forall j \neq i. \quad (1)$$

That is, a pair of image sequences of the same person p_i is constrained/maximised to be assigned with a top rank, i.e. the highest ranking score.

Learning a ranking function *holistically without discrimination and selection* from pairs of unsegmented and temporally unaligned person image sequences will subject the learned model to significant noise and degrade any meaningful discriminative information contained in the image sequences. This is an inherent drawback of any holistic sequence matching approach, including those with dynamic time warping applied for nonlinear mapping (see experiments in Sec. 4). Reliable human parsing/pose detection [22] or occlusion detection [47] may help, but such approaches are difficult to be scaled, especially with image sequences from crowded public scenes. The challenge is to learn a robust ranking function effective in coping with incomplete and partial image sequences by identifying and selecting most discriminative video fragments from each sequence suitable for extracting space-time features. Let us first consider generating a pool of candidates for video fragmentation.

3.2 Generating Candidates Pool for Video Fragmentation

Given the unregulated image sequences of people, it is too noisy to attempt holistically locating and extracting reliable discriminative space-time features from an entire image sequence. Instead, we consider breaking down each image sequence and generate a pool of video fragment candidates for a learning model to automatically select the most discriminative fragment(s) (Sec. 3.3).

It can be observed that motion energy intensity induced by the two legs of a walking person (or the activity of human muscles during walking) exhibits regular periodicity [44]. This motion energy intensity can be approximately estimated by optic flow. We call this Flow Energy Profile (FEP), see Fig. 3. This FEP signal is particularly suitable to address our video fragmentation problem

for selecting more robust and discriminative space-time features due to: (i) Its local minimum and maximum are likely to correspond to some characteristic phases in a pedestrian's walking cycle, thus helping in estimating these characteristic walking postures (e.g. one leg is about to land); (ii) Relatively robust to changes in camera viewpoint. More precisely, given a sequence $Q = (I_1, ..., I_t)$, we first compute the optic flow field (v_x, v_y) centered at each frame I. The flow energy e of I is defined as

$$e = \sum_{(x,y)\in U} \|[\, v_x(x,y), v_y(x,y)\,]\|_2, \tag{2}$$

where U is the pixel set of the lower body, e.g. the lower half of image I. The FEP $\mathcal{E}$ of Q is then defined as $\mathcal{E} = (e_1, ..., e_t)$, which is further smoothed by a Gaussian filter to suppress noise. Finally, we generate a candidate pool (set) of video fragments $S = \{s\}$ for each image sequence Q by detecting the local minima and maxima landmarks $\{t\}$ of $\mathcal{E}$ and extracting the surrounding frames $s = (I_{t-L}, ..., I_t, ..., I_{t+L})$ of each landmark as a video fragment. We fix $L = 10$ for all our experiments, determined by cross-validation on the iLIDS-VID dataset. It is worth pointing out that many of the obtained fragments of each image sequence can have similar phases of a walking cycle since the local minimum/maximum $\{t\}$ of the FEP signal are likely to correspond to certain characteristic walking postures (Fig. 3). This increases the possibility of finding aligned video fragment pairs (i.e. centred at similar walking postures) given a pair (S^a, S^b) of video fragment sets, facilitating discriminative video fragments selection and matching during model learning (Sec. 3.3).

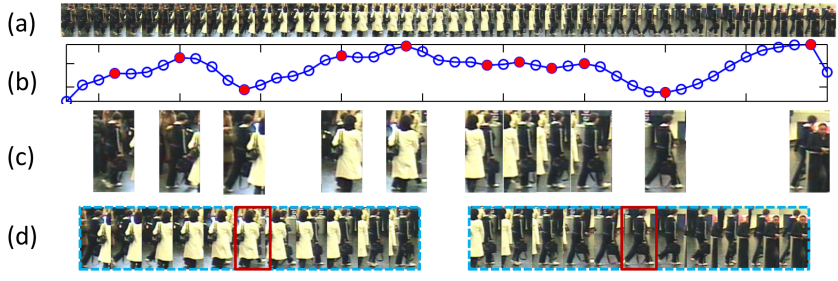


Fig. 3. (a) A person sequence of 50 frames is shown, with the motion intensity of each frame shown in (b). The red dots in (b) denote automatically detected local minima and maxima temporal landmarks in the motion intensity profile, of which the corresponding frames are shown in (c). (d) Two example video fragments (shown every 2 frames) with the landmark frames highlighted by red bounding boxes.

Video fragment space-time feature representation - We exploit HOG3D for space-time feature representation of a video fragment, due to its advantages demonstrated for applications in action and activity recognition [25]. In order

to capture spatially more localised space-time information of a person in motion, e.g. body parts such as head, torso, arms and legs, we decompose a video fragment spatially into 2×5 even cells. To encode separately the information of sub-intervals before and after the characteristic walking posture (Fig. 3 (d)) potentially situated in the middle of a video fragment, the fragment is further divided temporally into two smaller sections. Two adjacent cells have 50% overlap for increased robustness to possible spatio-temporal fragment misalignment. A space-time gradient histogram is computed in each cell and then concatenated to form the HOG3D descriptor $\boldsymbol{x}$ of the fragment s. We denote by $X = \{\boldsymbol{x}\}$ the HOG3D feature set of a fragment set $S = \{s\}$.

3.3 Selecting and Ranking the Most Discriminative Fragment Pairs

Given the candidate fragment sets $\{(X_i^a, X_i^b)\}_{i=1}^N$ represented by HOG3D features, the next problem for re-identification is how to select and match the most discriminative fragment pairs from cross-view fragment *sets*. Inspired by multi-instance classification [9] where each training sample is a bag (or set) of instances, we formulate a similar strategy for *multi-instance ranking* of video fragment candidates for automatic cross-view fragment selection and matching. More specifically, we aim for person re-identification by automatically selecting the most discriminative cross-view fragment pairs such that the selected fragments optimise cross-view re-id ranking score. Formally, we denote two cross-view fragments of person p_i captured in camera a and b as $\boldsymbol{x}_i^a \in X_i^a$ and $\boldsymbol{x}_i^b \in X_i^b$. The objective is to learn a linear ranking function on the absolute difference of cross-view fragment pairs:

$$h(\boldsymbol{x}_i^a, \boldsymbol{x}_i^b) = \boldsymbol{w}^\top |\boldsymbol{x}_i^a - \boldsymbol{x}_i^b| \tag{3}$$

that prefers the most discriminative cross-view fragment pair of the same person p_i over those of two different persons, i.e.

$$\max_{\boldsymbol{x}_i^a \in X_i^a, \boldsymbol{x}_i^b \in X_i^b} h(\boldsymbol{x}_i^a, \boldsymbol{x}_i^b) > h(\boldsymbol{x}_i^a, \boldsymbol{x}_j^b), \forall\, j \neq i. \tag{4}$$

For notation simplicity, we define $\boldsymbol{y}^+ = |\boldsymbol{x}_i^a - \boldsymbol{x}_i^b|$ as the *positive* instance (the absolute difference between two cross-view fragments of the same person), and $\boldsymbol{y}^- = |\boldsymbol{x}_i^a - \boldsymbol{x}_j^b|$ as the *negative* instance (the absolute difference between two cross-view fragments of two different persons). By enumerating all possible cross-view combinations between fragment sets, for every person p_i, we form a *positive* bag $B_i^+ = \{\boldsymbol{y}^+\}$ with all $\boldsymbol{y}^+$, and a *negative* bag $B_i^- = \{\boldsymbol{y}^-\}$ with all $\boldsymbol{y}^-$. After redefining the ranking function $h(\boldsymbol{x}^a, \boldsymbol{x}^b) = g(|\boldsymbol{x}^a - \boldsymbol{x}^b|) = g(\boldsymbol{y})$, Eqn. (4) can be rewritten as

$$\max_{\boldsymbol{y}^+ \in B_i^+} g(\boldsymbol{y}^+) > g(\boldsymbol{y}^-), \forall\, \boldsymbol{y}^- \in B_i^-. \tag{5}$$

With this constraint Eqn. (5), we aim to automatically discover and locate the most informative (most discriminative) cross-view video fragment pair within the positive bag B_i^+ for each person p_i during optimisation. To that end, we

introduce a binary selection variable $\boldsymbol{v}_i$ with each entry being 0 or 1 and of unity ℓ_0 norm for each person p_i, and obtain

$$g(\boldsymbol{Y}_i\boldsymbol{v}_i) > g(\boldsymbol{y}^-), \forall\, \boldsymbol{y}^- \in B_i^-, \tag{6}$$

where each column of $\boldsymbol{Y}_i$ corresponds to one $\boldsymbol{y}^+ \in B_i^+$.

Optimising $\boldsymbol{w}$, $\boldsymbol{v}$ - For the optimisation of the ranking function Eqn. (3) under the constraint Eqn. (6), we relax $\boldsymbol{v}$ to be continuous with non-negative entry and unity ℓ_1 norm as in [4]. In particular, to optimise $\boldsymbol{w}$ (Eqn. (3)) subject to Eqn. (6), we formulate our problem into the standard max-margin framework,

$$\begin{aligned} \boldsymbol{w}^* = \arg\min_{\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{v}} \frac{1}{2}||\boldsymbol{w}||^2 + C\boldsymbol{e}^\top\boldsymbol{\xi} \\ \text{s.t. } \boldsymbol{v}_i^\top\boldsymbol{Y}_i^\top\boldsymbol{w} - \boldsymbol{y}_k^\top\boldsymbol{w} \geq 1 - \xi_{i,k},\ \forall\, \boldsymbol{y}_k \in B_i^-,\ \xi_{i,k} \geq 0, \\ \boldsymbol{e}^\top\boldsymbol{v}_i = 1, \boldsymbol{v}_i \geq 0,\ \ i \in \{1,\ldots,N\},\ \ k \in \{1,\ldots,|B_i^-|\}, \end{aligned} \tag{7}$$

where $\boldsymbol{e}$ refers to the vector of all ones and N the number of persons in the training set; $\boldsymbol{\xi}$ is the slack vector, with entry $\xi_{i,k}$; $\boldsymbol{v}$ denotes the selection vector for all training persons, a concatenation of personwise selector vectors $\boldsymbol{v}_i$. We solve Eqn. (7) by optimising $\boldsymbol{w}$ and $\boldsymbol{v}$ iteratively between two steps. In the *ranking* step, we fix $\boldsymbol{v}$ to optimise $\boldsymbol{w}$: this Quadratic Programming problem can be solved by the interior point algorithm. In the *selection* step, we fix $\boldsymbol{w}$ to estimate $\boldsymbol{v}$: the simplex algorithm is used to solve this Linear Programming problem. During this iterative optimisation, a pair of two well aligned cross-view fragments of the same person with less partial/missing frames and noises is more likely to be selected since they can share more similarity in the space-time feature space and thus induce a higher ranking score (Eqn. (3)). Such more discriminative video fragment pairs are favoured by model optimisation.

Initially each $\boldsymbol{v}_i$ is set to $\frac{1}{|B_i^+|}\boldsymbol{e}$. The iteration terminates when $\boldsymbol{v}$ is converged e.g. $\|\boldsymbol{v}^{(q)} - \boldsymbol{v}^{(q-1)}\|_2 \leq 10^{-8}$, with q the current iteration index. For efficiency consideration, only 10% instances out of each B_i^- are employed during training.

Efficient learning - As the number of ranking constraint (Eqn. (6)) grows quadratically with the number of fragments per sequence, an efficient version of the ranking step is necessary to make model learning scalable. Motivated by [6], we relax the ranking step into a non-constrained primal problem that can be more efficiently solved by the linear conjugate gradient method as

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w},\boldsymbol{\xi}} \frac{1}{2}||\boldsymbol{w}||^2 + C \sum_{\boldsymbol{y}_k \in \{B_i^-\}} \ell\left(0, 1 - (\boldsymbol{v}_i^\top\boldsymbol{Y}_i^\top - \boldsymbol{y}_k^\top)\boldsymbol{w}\right)^2, \tag{8}$$

where ℓ refers to the hinge loss function. Our method is thus called Primal Max-Margin Multi-Instance Ranking (PM3IR).

3.4 Re-Identification by Discriminative Video Fragment Ranking

The learned ranking model can be deployed to perform person re-id by matching a probe person image sequence Q^p observed in one camera against a gallery set

$\{Q^g\}$ in another camera. Formally, the ranking score of a gallery person sequence Q^g with respect to the probe Q^p is computed as

$$f(Q^p, Q^g) = \max_{\boldsymbol{x}_i \in X^p, \boldsymbol{x}_j \in X^g} \boldsymbol{w}^\top |\boldsymbol{x}_i - \boldsymbol{x}_j|, \tag{9}$$

where X^p and X^g are the HOG3D feature sets of the video fragments in sequence Q^p and Q^g, respectively. The gallery persons are then sorted in descending order of their assigned matching scores to generate a ranked list.

Combination with existing spatial feature based models - Our approach can complement existing spatial feature based re-id approaches. In particular, we incorporate Eqn. (9) into the ranking scores γ_i obtained by other models as

$$\hat{f}(Q^p, Q^g) = \sum_i \alpha_i \gamma_i(Q^p, Q^g) + f(Q^p, Q^g). \tag{10}$$

where α_i refers to a weighting assigned to the i-th method, which is estimated by cross-validation.

4 Experiments

We conducted extensive experiments on two image sequence datasets designed for person re-identification, the PRID 2011 dataset [19] and our newly introduced dataset named iLIDS-VID [1].

iLIDS-VID dataset - A new iLIDS-VID person sequence dataset has been created based on two non-overlapping camera views from the i-LIDS Multiple-Camera Tracking Scenario (MCTS) [42], which was captured at an airport arrival hall under a multi-camera CCTV network. It consists of 600 image sequences for 300 randomly sampled people, with one pair of image sequences from two camera views for each person. Each image sequence has variable length consisting of 23 to 192 image frames, with an average number of 73. This dataset is very challenging due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and occlusions (Fig. 1 and Fig. 4 (a)).

PRID 2011 dataset - The PRID 2011 re-identification dataset [19] includes 400 image sequences for 200 people from two camera views that are adjacent to each other. Each image sequence has variable length consisting of 5 to 675 image frames [2], with an average number of 100. Compared with the iLIDS-VID dataset, it was captured in uncrowded outdoor scenes with relatively simple and clean background and rare occlusions (Fig. 4 (b)).

Evaluation settings - From both datasets, the total pool of sequence pairs is randomly split into two subsets of equal size, one for training and one for testing. Following the evaluation protocol on the PRID 2011 dataset [19], in the

[1] The iLIDS-VID dataset is available at `http://www.eecs.qmul.ac.uk/~xz303/downloads_qmul_iLIDS-VID_ReID_dataset.html`

[2] Sequences with more than 21 frames from 178 persons are used in our experiments.

(a) iLIDS-VID (b) PRID 2011

Fig. 4. Example pairs of image sequences of the same people appearing in different camera views from (a) the iLIDS-VID dataset, (b) the PRID 2011 dataset. Only every 3rd frame is shown and the total number of frames for each sequence is not identical.

testing phase, the sequences from the first camera are used as the probe set while the ones from the other camera as the gallery set. The results are shown in Cumulated Matching Characteristics (CMC) curves. To obtain stable statistical results, we repeat the experiments for 10 trials and report the average results.

Comparison with gait recognition and temporal sequence matching - We compared the proposed DVR model with contemporary gait recognition and temporal sequence matching methods for person (re-)identification:
(1) Gait recognition (GEI+RSVM) [31]: A state-of-the-art gait recognition model using Gait Energy Image (GEI) [17] (computed from pre-segmented silhouettes in their datasets) as sequence representation and RankSVM [6] for recognition. A challenge for applying gait recognition to unregulated person sequences in re-id scenarios is to generate good gait silhouettes as input. To that end, we first deployed the DPAdaptiveMedianBGS algorithm in the BGSLibrary [41] to extract silhouettes from video sequences in the iLIDS-VID dataset [3]. This approach produces better foreground masking than other alternatives.
(2) Colour&LBP+DTW, HoGHoF+DTW: We applied Dynamic Time Warping [36] to compute the similarity of two sequences, using either Colour&LBP [20] or HoGHoF [27] as per-frame feature descriptor. This is similar to the approach of Simonnet et al. [40], except that they only used colour features. In comparison, Colour&LBP is a stronger representation as it encodes both colour and texture. Alternatively, HoGHoF encodes both texture and motion information.

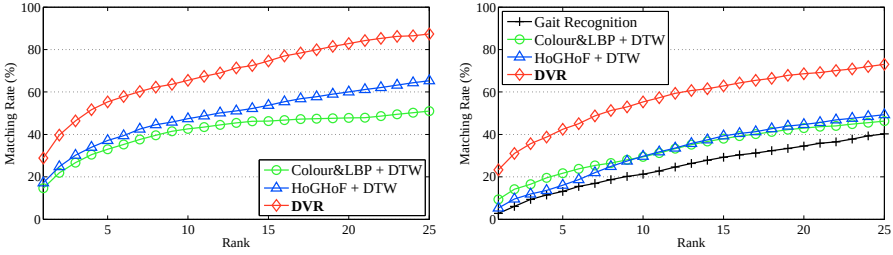
Fig. 5 and Table 1 show comparative results between DVR, GEI+RSVM (gait), Colour&LBP+DTW and HoGHoF+DTW. It is evident that the proposed DVR outperforms significantly all others on both datasets (gait was not applied to PRID 2011 for reasons stated above).

In particular, gait recognition [31] achieves the worst re-identification accuracy on the iLIDS-VID dataset. This is largely due to very noisy GEI features avaliable from person sequences. This is evident from the examples shown in

[3] We can only evaluate on the iLIDS-VID dataset because the original image sequences are not included in the PRID 2011 dataset.

Table 1. Comparison with gait recognition and temporal sequence matching methods.

Dataset	PRID 2011				iLIDS-VID			
Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
Gait Recognition [31]	-	-	-	-	2.8	13.1	21.3	34.5
Colour&LBP[20]+DTW[36]	14.6	33.0	42.6	47.8	9.3	21.7	29.5	43.0
HoGHoF[27]+DTW[36]	17.2	37.2	47.4	60.0	5.3	16.1	29.7	44.7
DVR (ours)	**28.9**	**55.3**	**65.5**	**82.8**	**23.3**	**42.4**	**55.3**	**68.4**

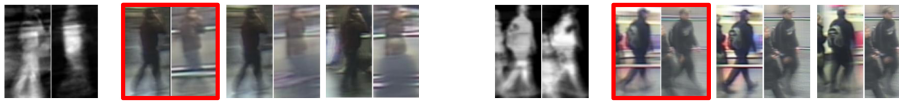


(a) PRID 2011 (b) iLIDS-VID

Fig. 5. Comparing CMC curves of the DVR model, gait recognition and temporal sequence matching based methods.

Fig. 6 : The extracted gait foreground masks tend to be affected by other moving objects in the scene, whilst our DVR model trains itself by simultaneously selecting and ranking only those fragments of image sequences which suffer the least from occlusion and noise. Moreover, DTW based matching for re-id using either Colour&LBP+DTW or HoGHoF+DTW features also suffer notably from the inherent uncertain nature of re-id sequences and perform significantly poorer than the proposed DVR approach. This is largely due to: (1) Person sequences have different durations with arbitrary starting/ending frames, also potentially different walking cycles. Therefore, attempts to match holistically entire sequences inevitably suffer from mismatching with erroneous similarity measurement. (2) There is no clear (explicit) mechanism to avoid incompleteness/missing data, typical in busy scenes. (3) Direct sequence matching is less discriminative than learning an inter-camera discriminative mapping function explicitly, which is built into the DVR model by exploring multi-instance ranking.

Comparison with single-shot and multi-frame/multi-shot spatial feature representations - To evaluate the effectiveness of discriminative video fragmentation and ranking using space-time features for person re-identification, we compared the proposed DVR model against a wide range of contemporary re-id models using spatial features, either in single-shot or as multiple frames (multi-shot). In order to process the iLIDS-VID dataset for the experiments, we mainly considered those methods with both their code available publically and being contemporary. They include (1) SDALF [11] (both single-shot and



(a) (b)

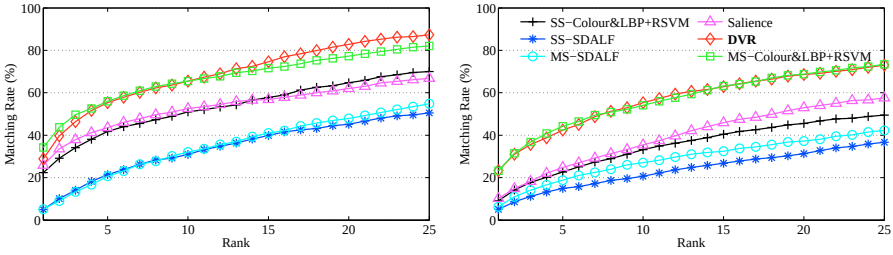
Fig. 6. (a) and (b) show two examples of the GEI gait features and our video fragment pairs. In both (a) and (b), the leftmost thumbnail shows GEI gait features, while the remaining thumbnails present some examples of fragment pairs, with the automatically selected pairs marked by red bounding boxes. A fragment is visualized as the weighted average of all its frames with emphasis on its central frame.

Table 2. Comparing spatial feature methods (SS: Single-Shot; MS: multi-shot)

Dataset	PRID 2011				iLIDS-VID			
Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
SS-Colour&LBP[20]+RSVM	22.4	41.8	51.0	64.7	9.1	22.6	33.2	45.5
SS-SDALF [11]	4.9	21.5	30.9	45.2	5.1	14.9	20.7	31.3
MS-SDALF [11]	5.2	20.7	32.0	47.9	6.3	18.8	27.1	37.3
Salience [49]	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9
DVR (ours)	28.9	55.3	**65.5**	**82.8**	**23.3**	42.4	**55.3**	68.4
MS-Colour&LBP+RSVM	**34.3**	**56.0**	**65.5**	77.3	23.2	**44.2**	54.1	**68.8**

multi-shot versions), (2) Salience [49], (3) a combination of colour and texture (Colour&LBP) [20] with RankSVM [6] as the distance metric. (4) Moreover, we also extended a Colour&LBP single-shot model to multi-shot by averaging the Colour&LBP features of each frame over a person sequence to focus on stable appearance cues and suppress noises, in a similar approach to [21]. We call this method MS-Colour&LBP+RSVM. Table 2 and Fig. 7 show the results. It is evident that the proposed DVR model outperforms significantly all the spatial feature based methods except our extended multi-shot MS-Colour&LBP+RSVM model, which offers slight advantage on PRID 2011 and very close performance on iLIDS-VID. This can be explained by that the DVR model with HOG3D space-time feature representation *only* utilises spatio-temporal gradient information *without* benefiting from any colour information. As colour information can often play an important role in person re-id [30], it is rather significant that using only space-time texture information (HOG3D), the proposed DVR model outperforms significantly most spatial feature based models, e.g. 10.7% and 128.4% Rank 1 improvement over Salience on PRID 2011 and iLIDS-VID, respectively. For a further analysis on the DVR model when colour information is incorporated, more details are discussed next.

Complementary to existing spatial feature representations - We further evaluated the effects from both adding additional colour information into the DVR model and combining the DVR model with existing colour and texture feature representations. The results are shown in Table 3. It is evident that significant performance gain was achieved by incorporating the DVR ranking score (Eqn. (10)). More specifically, the Rank-1 re-id performance of using multi-



(a) PRID 2011 (b) iLIDS-VID

Fig. 7. CMC curve comparison between the proposed DVR model (without colour information) and existing spatial feature based models (SS: Single-Shot, MS: multi-shot).

Table 3. Performance from combining DVR with spatial features (MS: Multi-Shot)

Dataset	PRID 2011				iLIDS-VID			
Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
MS-Colour+RSVM	29.7	49.4	59.3	71.1	16.4	37.3	48.5	62.6
MS-Colour+DVR	**41.8**	63.8	76.7	88.3	32.7	56.5	67.0	77.4
MS-Colour&LBP+RSVM	34.3	56.0	65.5	77.3	23.2	44.2	54.1	68.8
MS-Colour&LBP+DVR	37.6	63.9	75.3	**89.4**	**34.5**	**56.7**	**67.5**	**77.5**
MS-SDALF [11]	5.2	20.7	32.0	47.9	6.3	18.8	27.1	37.3
MS-SDALF+DVR	31.6	58.0	70.3	85.3	26.7	49.3	61.0	71.6
Salience [49]	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9
Salience+DVR	41.7	**64.5**	**77.5**	88.8	30.9	54.4	65.1	77.1

shot colour feature (MS-Colour) was boosted by 40.7% and 99.4% on PRID 2011 and iLIDS-VID respectively; Rank 1 score of MS-SDALF feature was boosted by 507.7% and 323.8% on PRID 2011 and iLIDS-VID respectively; and Rank 1 of Salience feature was boosted by 61.6% and 202.9% on PRID 2011 and iLIDS respectively.

Evaluation of space-time fragment selection - To evaluate the space-time video fragment selection mechanism in the proposed DVR model, we implemented two baseline methods without this mechanism: (1) SS-HOG3D+RSVM: Each person sequence is represented by the HOG3D descriptor of a single fragment randomly selected from the sequence; (2) MS-HOG3D+RSVM: Each person sequence is represented by the averaged HOG3D descriptors of four fragments uniformly selected from the sequence. In both these baseline methods, RankSVM [6] is used to rank the person sequence representations. The results are presented in Table 4. On the PRID 2011 dataset, the DVR model outperforms SS-HOG3D+RSVM and MS-HOG3D+RSVM in Rank 1 by 160.4% and 49.0% respectively. The performance advantage is even greater on the more challenging iLIDS-VID dataset, i.e. in Rank-1 by 206.6% and 92.6% respectively. It demonstrates clearly that in the presence of significant noise and given unregulated person image sequences, it is indispensable to automatically select discrim-

Table 4. The effect of space-time video fragment selection (SS: Single-Shot, MS: Multi-Shot)

Dataset	PRID 2011				iLIDS-VID			
Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
SS-HOG3D+RSVM	11.1	30.0	41.1	57.1	7.6	18.7	29.1	46.5
MS-HOG3D+RSVM	19.4	44.9	59.3	72.2	12.1	29.3	41.5	56.3
DVR (ours)	**28.9**	**55.3**	**65.5**	**82.8**	**23.3**	**42.4**	**55.3**	**68.4**

inatively space-time features from raw image sequences in order to construct a more robust model for person re-id. It is also noted that MS-HOG3D+RSVM outperforms SS-HOG3D+RSVM by suppressing noises benefited from temporal averaging. Although such a straightforward temporal averaging approach can have some benefits over single-shot methods, it loses important discriminative space-time information when applying uniformly temporal smoothing.

5 Conclusion

We have presented a novel DVR framework for person re-identification by video ranking using discriminative space-time feature selection. Our extensive evaluations show that this model outperforms a wide range of contemporary techniques from gait recognition, temporal sequence matching, to state-of-the-art single-shot/multi-shot/multi-frame spatial feature representation based re-id models. In contrast to existing approaches, the proposed method is capable of capturing more accurately space-time information that are discriminative to person re-identification through learning a cross-view multi-instance ranking function. This is made possible by the ability of our model to automatically discover and exploit the most reliable video fragments extracted from inherently incomplete and inaccurate person image sequences captured against cluttered background, and without any guarantee on person walking cycles and starting/ending frame alignment. Moreover, the proposed DVR model complements (improves) significantly existing spatial appearance features when combined for person re-identification. Extensive comparative evaluations were conducted to validate the advantages of the proposed model with the introduction of a new image sequence re-id dataset iLIDS-VID, which to our knowledge is currently the largest image sequence re-id dataset in the public domain.

References

1. Bashir, K., Xiang, T., Gong, S.: Gait recognition without subject cooperation. PRL 31, 2052–2060 (2010)
2. Bedagkar-Gala, A., Shah, S.K.: Part-based spatio-temporal model for multi-person re-identification. PRL 33, 1908–1915 (2012)
3. Ben Shitrit, H., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: ICCV. pp. 137–144 (2011)

4. Bergeron, C., Zaretzki, J., Breneman, C., Bennett, K.P.: Multiple instance ranking. In: ICML. pp. 48–55 (2008)
5. Bregonzio, M., Gong, S., Xiang, T.: Recognising action as clouds of space-time interest points. In: CVPR. pp. 1948–1955 (2009)
6. Chapelle, O., Keerthi, S.S.: Efficient algorithms for ranking with svms. Information Retrieval 13, 201–215 (2010)
7. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: BMVC (2011)
8. Cong, D.N.T., Achard, C., Khoudour, L., Douadi, L.: Video sequences association for people re-identification across multiple non-overlapping cameras. In: ICIAP. pp. 179–189 (2009)
9. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence pp. 31–71 (1997)
10. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. pp. 65–72 (2005)
11. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR. pp. 2360–2367 (2010)
12. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: CVPR. pp. 1528–1535 (2006)
13. Gilbert, A., Illingworth, J., Bowden, R.: Fast realistic multi-action recognition using mined dense spatio-temporal features. In: ICCV. pp. 925–931 (2009)
14. Gong, S., Cristani, M., Loy, C., Hospedales, T.: The re-identification challenge. In: Person Re-Identification, pp. 1–20. Springer (2014)
15. Gong, S., Xiang, T.: Visual analysis of behaviour: from pixels to semantics. Springer (2011)
16. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: ICDSC. pp. 1–6 (2008)
17. Han, J., Bhanu, B.: Individual recognition using gait energy image. TPAMI 28, 316–322 (2006)
18. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. In: ICCV. pp. 263–270 (2011)
19. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: SCIA (2011)
20. Hirzer, M., Roth, P.M., Köstinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: ECCV, pp. 780–793 (2012)
21. John, V., Englebienne, G., Krose, B.: Solving person re-identification in non-overlapping camera using efficient gibbs sampling. In: BMVC (2013)
22. Kanaujia, A., Sminchisescu, C., Metaxas, D.: Semi-supervised hierarchical models for 3d human pose reconstruction. In: CVPR. pp. 1–8 (2007)
23. Karaman, S., Bagdanov, A.D.: Identity inference: generalizing person re-identification scenarios. In: ECCV Workshops. pp. 443–452 (2012)
24. Ke, Y., Sukthankar, R., Hebert, M.: Volumetric features for video event detection. IJCV 88, 339–362 (2010)
25. Klaser, A., Marszalek, M.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC (2008)
26. Laptev, I.: On space-time interest points. IJCV 64, 107–123 (2005)

27. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. pp. 1–8 (2008)
28. Li, W., Wang, X.: Locally aligned feature transforms across views. In: CVPR. pp. 3594–3601 (2013)
29. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: ICCV. pp. 444–451 (2009)
30. Liu, C., Gong, S., Loy, C.C.: On-the-fly feature importance mining for person re-identification. PR 47, 1602–1615 (2014)
31. Martín-Félez, R., Xiang, T.: Gait recognition by ranking. In: ECCV, pp. 328–341 (2012)
32. Nakajima, C., Pontil, M., Heisele, B., Poggio, T.: Full-body person recognition system. PR 36, 1997–2006 (2003)
33. Nixon, M.S., Tan, T., Chellappa, R.: Human identification based on gait, vol. 4. Springer (2010)
34. Poppe, R.: A survey on vision-based human action recognition. IVC 28, 976–990 (2010)
35. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVC (2010)
36. Rabiner, L.R., Juang, B.H.: Fundamentals of speech recognition, vol. 14. PTR Prentice Hall Englewood Cliffs (1993)
37. Sapienza, M., Cuzzolin, F., Torr, P.: Learning discriminative space-time actions from weakly labelled videos. In: BMVC (2012)
38. Sarkar, S., Phillips, P.J., Liu, Z., Vega, I.R., Grother, P., Bowyer, K.W.: The humanid gait challenge problem: Data sets, performance, and analysis. TPAMI 27, 162–177 (2005)
39. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: ACM MM, pp. 357–360 (2007)
40. Simonnet, D., Lewandowski, M., Velastin, S.A., Orwell, J., Turkbeyler, E.: Re-identification of pedestrians in crowds using dynamic time warping. In: ECCV Workshops. pp. 423–432 (2012)
41. Sobral, A.: BGSLibrary: An opencv c++ background subtraction library. In: WVC. Rio de Janeiro, Brazil (2013)
42. UK Home Office: i-LIDS Multiple Camera Tracking Scenario Definition (2008)
43. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C., et al.: Evaluation of local spatio-temporal features for action recognition. In: BMVC (2009)
44. Waters, R., Morris, J.: Electrical activity of muscles of the trunk during walking. Journal of Anatomy 111, 191 (1972)
45. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. CVIU 115, 224–241 (2011)
46. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: ECCV, pp. 650–663 (2008)
47. Xiao, J., Cheng, H., Sawhney, H., Rao, C., Isnardi, M.: Bilateral filtering-based optical flow estimation with occlusion detection. In: ECCV, pp. 211–224 (2006)
48. Xu, Y., Lin, L., Zheng, W.S., Liu, X.: Human re-identification by matching compositional template with cluster sampling. In: ICCV (2013)
49. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR. pp. 3586–3593 (2013)
50. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. TPAMI 35, 653–668 (2013)