Stepwise Metric Promotion for Unsupervised Video Person Re-identification

Zimo Liu[1] Dong Wang[2*] Huchuan Lu[3]
Dalian University of Technology
lzm920316@gmail.com[1] wdice[2*], lhchuan[3]@dlut.edu.cn

Abstract

The intensive annotation cost and the rich but unlabeled data contained in videos motivate us to propose an unsupervised video-based person re-identification (re-ID) method. We start from two assumptions: 1) different video tracklets typically contain different persons, given that the tracklets are taken at distinct places or with long intervals; 2) within each tracklet, the frames are mostly of the same person. Based on these assumptions, this paper propose a stepwise metric promotion approach to estimate the identities of training tracklets, which iterates between cross-camera tracklet association and feature learning. Specifically, We use each training tracklet as a query, and perform retrieval in the cross-camera training set. Our method is built on reciprocal nearest neighbor search and can eliminate the hard negative label matches, i.e., the cross-camera nearest neighbors of the false matches in the initial rank list. The tracklet that passes the reciprocal nearest neighbor check is considered to have the same ID with the query. Experimental results on the PRID 2011, ILIDS-VID, and MARS datasets show that the proposed method achieves very competitive re-ID accuracy compared with its supervised counterparts.

1. Introduction

Person re-identification (re-ID), aiming to retrieve a query identity from a gallery in a different camera view, usually relies on large volumes of labeled data. Due to the high labeling cost, this paper is devoted to the unsupervised scenario in which no identity labels are needed.

Focusing on video-based unsupervised re-ID, our work is motivated from three aspects. First, videos contain much richer information than single images, *e.g.*, the space-time cues and the pose variations, and are non-trivially available by detection and tracking. The appearance/temporal information can significantly improve the discriminative ability of the learned visual representations [48, 29, 39]. Moreover, since videos contain more prior knowledge about the scenario, the influence of background noise can be largely weakened [42].

Figure 1. A video frame on the MARS dataset [48]. We draw in this frame two tracklets (red and yellow) which are actually observed at distinct time stamps. It is intuitive to assume that the two tracklets can be treated as different identities. This observation contributes to the model initialization process.

Second, video tracklets, produced by pedestrian detection and tracking (Fig. 1), are reliable data source for unsupervised learning methods. This process is fully automatic and unsupervised. As implied in the recent survey [50], different tracklets can be treated as different identities, as long as we assume that the tracklets are captured at distinct places or with long intervals. Therefore, even if no human annotation is available, a discriminative model can be obtained via the tracklet data.

Third, feature learning using tracklets from the same view may result in low discriminative ability. In fact, *the key component in unsupervised re-ID is label estimation*. Given the tracklets that provide some supervision under a certain camera, it is important to propagate these labels across different cameras, so that cross-camera characteristics will be learned.

Framework. This paper adopts an rather intuitive unsupervised framework for video-based person re-ID. In brief, two steps are involved: 1) classifier initialization using the tracklets in the same camera [50] (see Fig. 1); 2) itera-

[0]Dr. Wang[2*] is the corresponding author.

tions between cross-camera tracklet association and feature learning. With more iterations, the learned feature becomes more discriminative, and the data association processes gets more accurate. This framework is also adopted in two contemporary works [5, 41].

Our Method. The proposed metric promotion approach iterates between model update and label estimation. For label estimation, under camera A, we use each tracklet as query to search for its k nearest neighbors (NNs) in camera B. Among these k candidates, the *best match* is selected as being associated with the query tracklet. We employ negative mining to reduce the impact of false positive matches in the k-NNs. This k-NN search process is then reversely repeated using the *best match* as query to see whether the initial query is its best match, a confirmation protocol to ensure that the initial query and the *best match* are truly associated. The the associated pairs are adopted for model updating. The main contributions of our approach are summarized as follows:

- A cross-camera framework is introduced for video-based person re-ID, in which the tracklets under one camera are used for model initialization.
- Negative Mining and label propagation are proposed for tracklet association.
- On three datasets, we report competitive performance compared with recent state of the art.

2. Related Work

Video-based re-ID. Due to the rich information contained in video, video-based re-ID [38, 9, 42, 30, 17, 54, 38] has drawn increasing attentions recently. The space-time information from image sequences has been both exploited in [22] and [38], where the former constructs to select the most discriminative image in video automatically while the latter aims to build a spatio-temporal appearance representation for walking pedestrian. Cho *et al.* [1] conduct the multi-shot person re-ID task by analyzing the camera viewpoints and estimating the pose of pedestrian. You *et al* [42] present a Top-Push Distance Learning model (TDL) [42], in which a latent space is explicitly learned to enlarge the margin between video sequences by enforcing a top-push constraint at the top rank. Some deep learning based approaches have also been presented for video-based re-ID. Niall *et al.* [30] exploit a siamese network architecture to present the pedestrian via a single feature vector which connects to all time-steps sequences. Zheng *et al.* [48] report a CNN descriptor learned via an extensive video benchmark (MARS) which shows good generalization ability on other video re-ID task upon fine-tuning.

Unsupervised re-ID. Compared with supervised methods [16, 52, 34, 43, 24, 40, 44], there are fewer unsupervised methods available for re-ID. Most of them directly utilize hand-crafted descriptors [27, 8, 6, 35, 19, 49]. For example, Ma *et al.* [26] propose the BiCov by combining the Biologically Inspired Features (BIF) and Covariance descriptors. Zheng *et al.* [49] propose a Bag-of-Words (BOW) descriptor which describes each pedestrian by visual word histogram and enables global fast matching. LOMO [19] extracts the local maximal occurrence representation scheme based on HSV color histograms and Scale Invariant Local Ternary Pattern (SILTP) [20]. Tetsu *et al.* [35] describe the local region in an image via hierarchical Gaussian distribution in which both means and covariances are considered. Another saliency matching approach, Unsupervised Saliency Learning (USL) [47], matches persons by building dense correspondence between image pairs and learning human salience on patch level.

For unsupervised learning approaches, Ma *et al.* [28] propose a selective sequence matching method to match two partial segments of two sequences. In [5], k-means clustering is used for label estimation, and the ID-discriminative embedding (IDE) [48] is used for feature learning. In [41], Ye *et al.* propose a graph matching method for cross-camera label estimation, followed by metric learning to iteratively improve the accuracy of label estimation.

Label Propagation. In the semi-supervised framework, the idea of label propagation [4, 46, 45, 23] has been used to estimate the label value in many research fields, such as image annotation and patch labeling. With the assumption that data points occupying the same manifold may be very likely to share similar/same label, the label confidence associated with each sample could be spread to their nearby neighbors through an iterative process. Therefore, when given a probe data point, we can estimate the label value by computing its similarities among some reliable labeled samples.

Negative Mining. Negative mining has been applied in classification and weakly labeled data annotation [25, 33, 10]. The negative mining method in this paper is related to POP [21], which performs a post-rank optimization process by allowing a user to manually select some negative samples to refine the initial rank list. The sparse human negative feedback on-the-fly in POP steers an automatic selection of more relevant re-identification features. By merging the negative information with the initial rank list, the accuracy of matching the true positive sample at the top rank rises up. Nevertheless, POP is a re-ranking method, requiring the assistance of human, which is applicable to unsupervised learning. In this paper, given a query tracklet in camera A, we first perform k-NN retrieval in the same camera. The k nearest neighbors can thus be thought as negative samples (details to be described in Section 3.3).

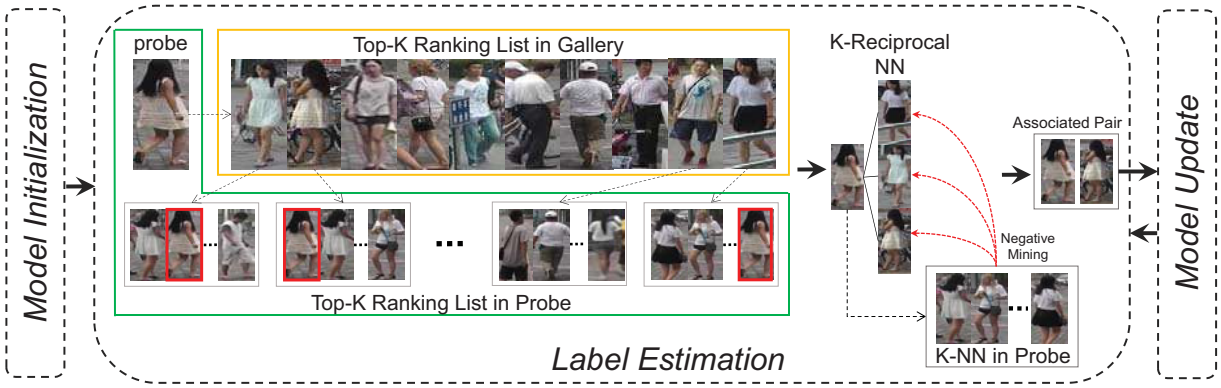


Figure 2. The pipeline of the proposed approach (best viewed in color). After model initialization, several candidates of a given probe are selected and refined via the K-reciprocal nearest neighbor searching and negative mining. Cross-camera tracklets associated with this query are located and used to updated the model. Model updating and sample association stop when no more cross-camera tracklet pairs are generated.

3. Proposed Approach

3.1. Overall Framework

The framework adopted in this paper is in essence similar to two contemporary works [5, 41]. In a nutshell, three components are involved.

- The model is first initialized, *e.g.,* using a transferred representation [5] from the source.
- Standard metric learning [14] or feature learning methods [50] are applied iteratively with label estimation. During the iterations, the learned features and estimated labels improve simultaneously.
- After model convergence, the learned features/metrics are used for testing.

Three fundamental issues are critical in this framework: (1) how to initialize a discriminative model; (2) how to accurately associate the tracklets from different camera views; and (3) how to upgrade the model with the augmented cross-camera pairs.

This paper contributes to problem 1) and problem 2), which will be elaborated in Section 3.2 and Section 3.3, respectively. As for problem 3), this paper adopts some readily available techniques for metric and feature learning, such as the XQDA [19], KISSME [14] and the deep feature learning method IDE [48]. The pipeline of the proposed approach is stated in Figure 2.

3.2. Model Initialization

Model initialization is a critical component in an unsupervised re-ID system. It is expected that a well initialized model can lead to a superior model upon convergence. In [5], the CNN model is initialized by directly deploying a fine-tuned CNN model on a source dataset. This initialization method is suitable for image-based datasets.

In this paper, we propose an alternative approach for model initialization which particularly suits the video-based re-ID task. Our key assumption is two-fold. First, different tracklets contain different person identities, as long as these video tracklets are taken at distinct places or with long time intervals. Second, within each tracklet, the frames generally depict the same person.

On the one hand, under the first assumption, we can obtain different IDs. Chance is remote that two tracklets belong to the same person, if they are captured in place and time disjoint cameras. So generally speaking, the first assumption holds. On the other hand, although tracking error sometimes appears, in which one tracklet may contain different IDs, our second assumption is usually valid considering the fast improvement of pedestrian detection and tracking.

Given the two assumptions, this paper uses the video tracklets in one camera for model initialization. More formally, we use N tracklets from one camera which represents N IDs. Let $n_i, i = 1, ..., N$ denote the number of frames for tracklet i. We thus have n_i positive samples for identity i. For example, for the iLIDS [37] and PRID [11] datasets, each ID has but one tracklet under each camera, so we select all the tracklets in one camera for initialization. For MARS [48], since each ID has more than one tracklets under each camera, we manually select one tracklet for each ID, simulating the situation when tracklets are captured with long time intervals. Tracklet selection errors are also evaluated during experiment. Using these unlabeled tracklets in a camera, a discriminative model can be initialized, which demonstrates a fair distinguishing ability even to the samples in the other views.

We validate the effectiveness of the classifier which

Figure 3. Some top-15 ranking results of camera view 2-6 on the MARS dataset via the classifier initialized by training data from camera view 1. Images framed in green correspond to the ground truth image of the probe image in the other views.

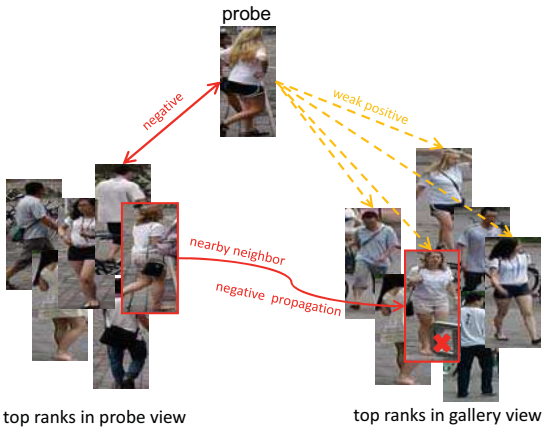


Figure 4. The overview of negative mining. Some false matches of the probe are removed by propagating negative information to their nearby neighbors in the gallery view.

learned from automatically labeled data by training a classifier using only a single view features on the MARS dataset, and employing this classifier in searching the most similar samples in the other five views. Figure 3 presents an intuitive example, where with respect to the probe the top-15 ranking results are illustrated from left to right. The images with green rectangle boxes correspond to the ground truth images in the other views. From this figure, we can see that the truly matched persons have more chance to be listed in top-ranked results.

3.3. Label Estimation with Negative Mining

The challenge for unsupervised learning in video person re-id consists in how to more accurately associate the tracklets captured under different cameras. In this paper, we propose to use negative samples for cross-camera label estimation. Specifically, given a probe tracklet (in the training set) to be associated with the gallery set (in the training set, too), the matching function generates a rank list of the gallery set by computing their similarities. The top-ranked candidates are thought to be visually similar to the probe, but the rank-1 candidate cannot be guaranteed to be the true positive match. To address this issue, we introduce a negative mining approach to trigger the refinement of the suboptimal rank list. Specifically, our method aims to utilize the nearby neighbors of the probe tracklet in the same camera, which exhibit high visual similarities but are false matches, to propagate negative information to re-order the initial rank list, as shown in Figure 4. The detailed procedure of this approach is described as follows.

First, given a probe tracklet x_p (in the training set) and the gallery set (in the training set), the initial ranking list can be obtained by computing their Mahalanobis distance. It is intuitive to associate the top match in the ranking list with the probe, but this method may be prone false matches. In this paper, we view the top-K samples as possible candidates. This idea has also been revealed in [53] that the K-reciprocal nearest neighbors are more related to the probe. Therefore, we adopt the K-reciprocal nearest neighbors as candidates.

Second, to further refine the ranking list and improve the ratio of correctly associated samples, we introduce negative mining into the model. The similar but label different samples (M Nearest Neighbor) in the same view of the probe tracklet x_p are defined as $x_p^1, x_p^2, ..., x_p^M$, where $Y_L = \{y_p, y_p^1, ..., y_p^M\}$ are the corresponding class labels. These samples are taken as the negative pairs and be used to re-order the ranking list. Let $X = \{x_p, x_p^1, ..., x_p^M\}$, $X_g = \{x_g^1, x_g^2, ..., x_g^K\}$ be the top ranks of x_p in the gallery view where the corresponding label $Y_U = \{y_g^1, ..., y_g^K\}$ are unobserved, the problem is to estimate Y_U from X and Y_L.

An affinity graph is created among all the labeled and unlabeled data points, where the edge between any nodes i, j is weighted so that the closer the nodes are in Mahalanobis distance, $d_{i,j}$, the larger the weight $w_{i,j}$. The weights are computed by

$$w_{i,j} = exp(-\frac{d_{i,j}}{\sigma^2}), \tag{1}$$

where σ is the controlled parameter. The probabilistic transition matrix P is defined as

$$P_{i,j} = \frac{w_{i,j}}{\sum_{k=1}^{K+M+1} w_{k,j}} \tag{2}$$

where $P_{i,j}$ denotes the probability of jumping from node j to i. Also define a label matrix Y, whose i-th row represents the label probabilities of node i, and the top $(K+1)$ rows are Y_L and the remaining rows the Y_U. All nodes propagate labels for one step is computed by $Y \leftarrow PY$. The Y_L should never change for keeping the labeled data not "fade

away", and we split the P after the $(K+1)$-th row and $(K+1)$-th column into 4 sub-matrices

$$P = \begin{bmatrix} P_{LL} & P_{LU} \\ P_{UL} & P_{UU} \end{bmatrix} \tag{3}$$

We attach the label of the probe X_p to the unlabeled instance with the highest similarity in Y_U, where

$$Y_U = (I - P_{UU})^{-1} P_{UL} Y_L. \tag{4}$$

A reverse directional process is also operated, and the tracklet pairs who share each other as the top rank are associated together. The associated pairs are then gathered for model updating.

In this paper, we extend the XQDA approach to unsupervised learning (U-XQDA), where we consider to learn a unified projection W and distance metric M for all the training samples with a close-form solution. In XQDA, the corresponding Generalized Rayleigh Quotient for projection direction W is written as

$$J(W) = \frac{W^\top \Sigma_E W}{W^\top \Sigma_I W} \tag{5}$$

, where Σ_I and Σ_E are the intra-personal variations and the extra-personal variations. Compared to XQDA, the proposed approach utilize both the label auto-marked set S_1 and the label-estimated tracklet pairs S_2 to update the model. Therefore, the objective function can be written as

$$J(W) = \frac{W^\top (\Sigma_{E,S_1} + \Sigma_{E,S_2}) W}{W^\top (\Sigma_{I,S_1} + \Sigma_{I,S_2}) W}. \tag{6}$$

, where the $\Sigma_{E,S_1}, \Sigma_{E,S_2}$ and $\Sigma_{I,S_1}, \Sigma_{I,S_2}$ are the extra-personal variations and the intra-personal variations. This maximization problem can be also solved by the generalized eigenvalue decomposition approach. The updated model is then used to estimate the label and associate cross-view tracklet pairs, and this iterative process of the model optimization and label estimation step stops when no more cross-view tracklets pairs are generated. In Figure 5, we show the pipeline of the method for better understanding.

3.4. Test Stage

Currently, almost all the video-based person re-identification approaches measure person similarity utilizing the point-to-point distance via a max-pooling strategy (a sequence of descriptors is remodeled into a vector), as shown in the left part of Figure 6. However, in some video-based recognition or multi-view object recognition tasks [36], measuring the distance between two image sets seems to be a much proper way by computing the average distance among all images of the tracklets, as shown

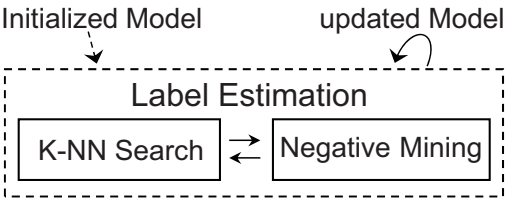


Figure 5. The pipeline of the proposed method. The initialized (in unsupervised manner) or the further updated(via the estimated label) model are used to measure the similarity in the label estimation part.

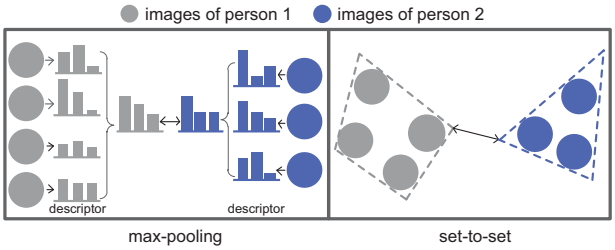


Figure 6. Max-pooling distance and set-to-set distance.

in the right part of Figure 6. We employ the set-to-set distance measurement rather the max-pooling strategy to compute the similarity between tracklet pairs in both the training (i.e., unlabeled data association) and testing stage. The performance comparisons of the max-pooling strategy and set-to-set distance measurement will be reported in the experimental section.

4. Experiments

4.1. Datasets

In this experiment, we compare our method with other related algorithms by using three public video datasets for person re-identification, including the PRID 2011 [11], ILIDS-VID [37], and most recent MARS [48] datasets.

PRID 2011: The image sequences of the PRID 2011 dataset [11] are captured from two static surveillance cameras for different views, which have different illumination, background, and camera characteristics. One camera view shows 385 persons and another one captures 749 persons, and only the first 200 persons appear in both views. Each video consists of 5 to 675 frames, with an average number of 100. To guarantee the effective length of the video, we select 178 persons that appear more than 27 frames. Then, 89 persons are randomly selected for training and the remaining ones are adopted for testing. This partition procedure is repeated 10 times and then the average results are reported.

ILIDS-VID: The ILIDS-VID dataset [37] consists of 300 distinct individuals observed in two non-overlapping camera views. In each view, each person has 23 to 192 images (the average number per person is 73). Since this dataset is captured at an airport arrival hall under a multi-camera CCTV network, it is challenging because of clothing similarities among persons as well as lighting and viewpoint variations across views. In our experiments, the ILIDS-VID dataset is randomly partitioned into two subsets with the same size, i.e., 150 persons for training and 150 persons for testing. The cumulative matching characteristic (CMC) curve is used to measure the performance of different algorithms. The partition procedure is also repeated 10 times for reporting a reliable result.

MARS: The MARS dataset [48] is the largest and most recent video dataset for person re-identification, which is captured from six near-synchronized cameras on Tsinghua campus. A total of 20,478 tracklets are automatically generated via the DPM [7] detector and GMMCP [3] tracker. Among them, there exist 3,278 distractor tracklets exist because of false detection and association. This dataset contains 1,261 different persons who appear in at least 2 cameras and average 13.2 tracklets for each person. The entire dataset is partitioned into 625 persons for training and the others for testing. In our experiments, we adopt the same partition to evaluate our algorithm and other competing ones. Each person from each view corresponds to multiple tracklets, from which we randomly select one tracklet to represent the person. Since a query responds to multiple ground truths in the MARS system, it does not fully reflect the true ranking performance to merely use the CMC rule. Thus, we apply both CMC score and mean Average Precision (mAP) as the evaluation criterion in our experiments.

4.2. Feature Representation and Parameter Setting

In this work, the proposed method is implemented with the recently proposed Local Maximal Occurrence (LOMO) feature [19]. The LOMO feature includes HSV color and SILTP histograms, resulting in 26,960 dimensions in total for each person image with the normalized 128x48 size. The PCA method is further exploited to reduce the feature dimension into 600 in our implementation, and the nearby neighbor number is set to 10 in all the reported datasets.

4.3. Experimental Results

Results on the PRID 2011 and ILIDS-VID datasets. We compare with some existing state-of-art methods and two baseline method (Euclidean distance of LOMO and GOG feature) on these two datasets. In addition, deep-learning-based methods [30, 48] and an unsupervised saliency matching algorithm [47] are also compared.

The CMC curve and the rank-n (n = 1, 5, 10, 20) matching rates are shown in Figure 7(a), Figure 7(b) and Table 1. These results show that: (1) the proposed approach outperforms the existing unsupervised methods, with rank-1 matching rate achieving 80.9% and 41.7% on the PRID2011 and ILIDS dataset, respectively. (2) compared to the supervised deep learning methods, our approach still achieves satisfactory performance.

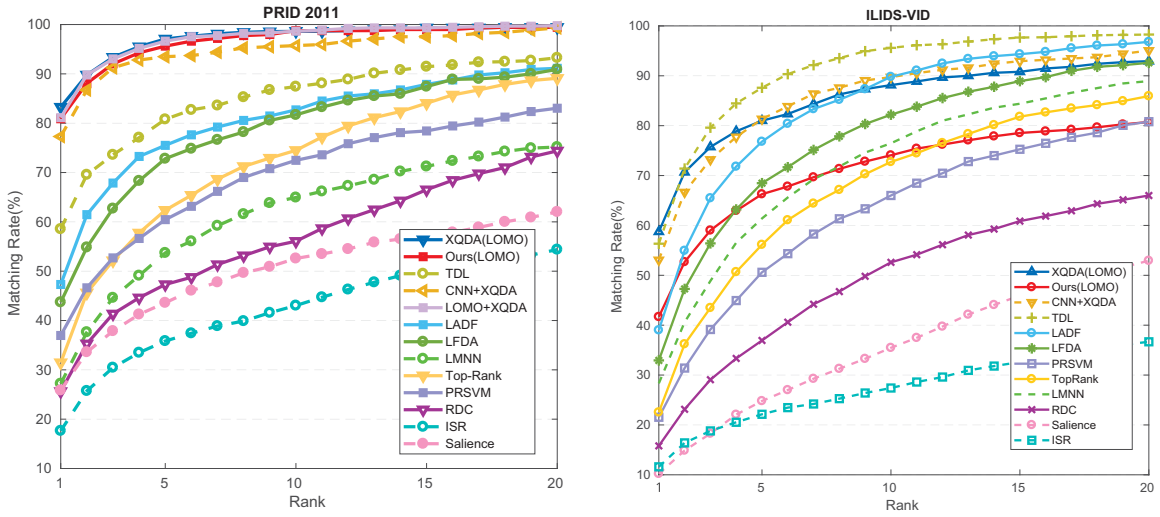
Results on the MARS dataset. This dataset is the largest and most practical video dataset for person re-identification. Since it is a most recent dataset, the results of few methods are available. In this experiment, we compare the proposed algorithm with a baseline method based on the Euclidean distance with the LOMO feature and the supervised method (XQDA [19]). The detailed results are reported in Table 2, from which we can see that the proposed unsupervised algorithm performs significantly better than the baseline method and slightly worse than the compared supervised approaches, with rank-1 matching rate and mAP value 3.33% and 1.78% lower than the XQDA approach, respectively.

4.4. Analysis of the Proposed Method

Algorithm Convergence: In this work, the discriminative power of the proposed method is improved gradually with more positive cross-view pairs generated. The iterative process is terminated when no more cross-view pair is generated through the negative label propagation step. To investigate the convergence effect of the proposed method, in Figure 8 (a), the changes of the objective function during the iterations on the PRID 2011 dataset is visualized, which shows that the proposed method can be converged after a few iterative steps. Figure 8 (b) shows the growth of the generated cross-view pair number during iterations (in one trail), including the correctly associated number (red) and the total associated number (blue). In Figure 8 (c), the improvement of rank-n (n = 1, 5, 10) matching rates (marked in red, blue and yellow respectively) with the iterations are illustrated, where the rank-1 matching rate improves by around 30% via our iterative process.

Set-to-Set *vs* Maxpooling: To fully exploit the abundant information within video, we exploit the set-to-set distance rather than the traditional max-pooling scheme in the test phase, which computes the tracklets similarities by the average distance among all images from the tracklets. Figure 8 (d) shows the CMC curves of both set-to-set measurement and max-pooling strategies on the PRID 2011 dataset via the proposed approach and the XQDA method. From this figure, we can see that the set-to-set distance achieves much better performance than the max-pooling scheme for video-based person re-identification.

Analysis of parameter K : As described in former, extending the nearest neighbor number from 1 to K may lead



(a) iLIDS-VID (b) PRID-2011

Figure 7. Comparison with the state-of-the-art methods on (a) PRID-2011 and (b) ILIDS-VID datasets. The CMC curves are shown.

Table 1. Matching results by the proposed algorithm and other competing methods on the PRID 2011 and ILIDS-VID datasets. The CMC scores (%) of rank-1, 5, 10, 20 are reported.

	PRID 2011				ILIDS-VID			
Methods	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
Ours(LOMO)	80.9	95.6	98.8	99.4	41.7	66.3	74.1	80.7
Ours(GOG)	80.8	96.0	98.3	99.3	33.2	55.7	64.4	72.5
XQDA(LOMO) [19]	83.3	97.1	98.7	99.4	58.7	81.0	88.1	92.9
XQDA(GOG)	86.9	97.2	98.8	99.3	52.7	78.3	85.7	92.2
Euclidean(LOMO)	9.0	29.6	48.31	73.2	7.1	24.5	33.1	48.6
Euclidean(GOG)	42.1	67.4	77.2	87.2	23.3	45.7	55.2	66.3
Salience [47]	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9
CSPL [2]	83.0	97.8	99.4	99.9	48.7	77.9	87.3	93.7
DTDL [12]	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9
RCN [30]	70.0	90.0	95.0	97.0	58.0	84.0	91.0	96.0
CNN + XQDA [48]	77.3	93.5	-	99.3	53.0	81.4	-	95.1
CNN + KISSME [48]	69.9	90.6	-	98.2	48.8	75.6	-	92.6
DGM(MLAPG,LOMO)[41]	73.1	92.5	96.7	99.0	37.1	61.3	72.2	82.0
DGM(XQDA,LOMO)[41]	82.4	95.4	98.3	99.8	31.3	55.3	70.7	83.4
DGM(IDE)[41]	56.4	81.3	88.0	96.4	36.2	62.8	73.6	82.7
LFDA [31]	43.7	72.8	81.7	90.9	32.9	68.5	82.2	92.6
TDL [42]	56.7	80.0	87.6	93.6	56.3	87.6	95.6	98.3
LADF [18]	47.3	75.5	82.7	91.1	39.0	76.80	89.0	96.8
RDC [51]	25.6	47.3	56.1	74.4	15.8	36.9	52.6	66.0
TopRank [15]	31.7	62.2	75.3	89.4	22.5	56.1	72.7	86.0
PRSVM [32]	37.0	60.5	72.5	83.0	21.5	50.6	66.0	80.8
STA [22]	64.0	87.0	90.0	92.0	44.0	72.0	84.0	92.0
SRID [13]	35.1	59.4	69.8	79.7	24.9	44.5	55.6	66.2

to better performance. Since more reciprocal pairs are introduced via the increasing K, more positive pairs may be searched. We conducted experiments on the PRID2011 dataset of a random trail with different K value (setting as 1, 5, 10, 20, 50). In Table 3 , we state the rank-n (n = 1, 5, 10) matching rate and the corresponding score of Recall.

Table 2. Comparison of the proposed approach and some supervised methods using LOMO on the MARS dataset .

Methods	Rank			mAP
	rank-1	rank-5	rank-20	
Ours	23.59	35.81	44.90	10.54
XQDA	26.92	38.13	50.66	12.32
Eucl	12.78	21.16	31.57	4.29

Table 3. Comparison of different K on the PRID 2011 dataset.

K	rank-1	rank-5	rank-10	Recall
K=1	70.8	92.1	97.8	0.5168
K=5	80.9	93.3	97.8	0.7303
K=10	80.9	93.3	97.8	0.8314
K=20	84.27	96.6	98.9	0.8764
K=50	78.65	95.5	96.6	0.8202

Table 4. The Precision, Recall and F-score value on the three reported datasets.

Dataset	Recall	Precision	F-score
PRID2011	0.8494	0.9767	0.9086
ILIDS-VID	0.5120	0.5413	0.6644
MARS	0.5972	0.7082	0.6480

As can be seen, the Recall achieves 0.8314 when K is extended to 10 which outweigh 0.3146 than only selecting one possible candidate.

Performance of Label Estimation: We adopt the Precision-Recall value and F-score to evaluate the performance of the proposed label estimation and report the results in Table 4, where the F-score reported in this paper is computed by

$$F\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (7)$$

The results on the PRID2011 and ILIDS-VID datasets are the average result via 10 trails. The proposed approach achieves F-score at around 0.91, 0.66 and 0.65 for the PRID2011, ILIDS-VID and MARS dataset, respectively.

Running Cost: We conduct the proposed approach with Matlab implementation on a desktop PC with E5-2650 v3 @2.30GHz CPU, and the reported running time is averaged via 10 trails on the PRID 2011 dataset. The computation time of our iterative training process is 113.48 seconds, which demonstrates that our iterative process is actually effective. For the testing phase, it costs 0.00025 seconds to compute the similarity for each two tracklets, which indicates good applicability of the proposed approach in real system.

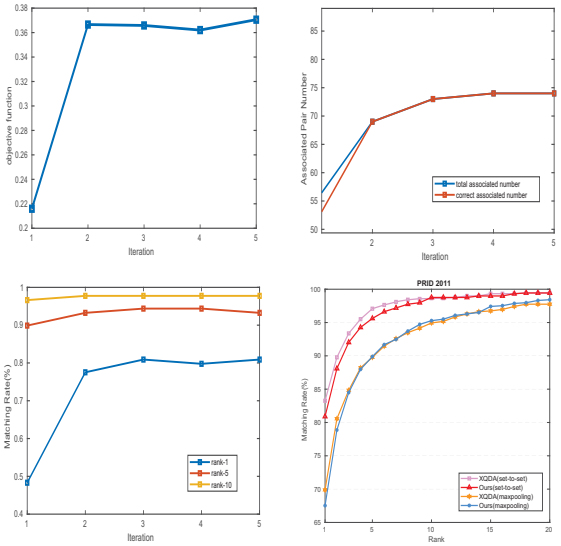


Figure 8. (a) The changes of objective functions between iterations. (b) The growth of the total/correct associated cross-view tracklet pairs. (c) The rank-n (n = 1, 5, 10) matching rates improvement with iterations. (d) Comparison of the set-to-set measurement and the max-pooling stratedgy.

5. Conclusion

In this paper, we develop a novel cross-view stepwise metric promotion algorithm for video-based person re-identification in an unsupervised manner. Motivated by our empirical observation that the classifier trained in a given view also has much discriminant power for the other views, the proposed method firstly initializes a view-specific classifier for each individual view and then introduces the cross-view information by allowing the negative samples to propagate negative information. After that, a metric learning process is exploited to improve the basic classifiers from the automatically labeled training samples and the former associated cross-view pairs based on an iteration process. The final classifiers are obtained after iteration convergence and then combined with a set-to-set distance scheme to match persons across different views. Other feature learning methods or deep models can also be used. Numerous experimental results on the PRID 2011, ILIDS-VID, and MARS datasets demonstrate that the proposed method not only outperforms the other unsupervised methods but also achieves competitive performance relative to many supervised algorithms.

Acknowledgement This paper is supported by the Natural Science Foundation of China #61502070 , #61472060 and #61528101.

References

[1] Y. Cho and K. Yoon. Improving person re-identification via pose-aware multi-shot matching. In *CVPR*, 2016.

[2] J. Dai, Y. Zhang, and H. Lu. Cross-view semantic projection learning for person re-identification. In *PR*, 2017.

[3] A. Dehghan, S. M. Assari, and M. Shah. GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, 2015.

[4] W. F and Z. C. Label propagation through linear neighborhoods. In *TKDE*, 2008.

[5] H. Fan, L. Zheng, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *arXiv*, 2017.

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

[7] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[9] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ICDSC*, 2008.

[10] J. F. Henriques, J. Carreira, R. Caseiro, and J. Batista. Beyond hard negative mining: Efficient detector learning via block-circulant decomposition. In *ICCV*, 2013.

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

[12] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015.

[13] S. Karanam, Y. Li, and R. J. Radke. Sparse re-id: Block sparsity for person re-identification. In *CVPR*, 2015.

[14] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[15] N. Li, R. Jin, and Z. Zhou. Top rank optimization in linear time. In *NIPS*, 2014.

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

[17] Y. Li, Z. Wu, S. Karanam, and R. J. Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *BMVC*, 2015.

[18] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013.

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[20] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *CVPR*, 2010.

[21] C. Liu, C. C. Loy, S. Gong, and G. Wang. POP: person re-identification post-rank optimisation. In *ICCV*, 2013.

[22] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*, 2015.

[23] W. Liu and T. Zhang. Bidirectional label propagation over graphs. *IJSI*, 2013.

[24] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*, 2014.

[25] B. M, H. S, and J. F. Hard negative mining for metric learning based zero-shot classification. In *ECCV Workshops*, 2016.

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

[27] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *IVC*, 2014.

[28] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *PR*, 2017.

[29] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016.

[30] M. Niall, M. del Rincon Jesus, and M. Paul. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016.

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

[32] B. J. Prosser, W. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010.

[33] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labeled data. In *ECCV*, 2012.

[34] C. Sun, D. Wang, and H. Lu. Person re-identification via distance metric learning with latent variables. 2017.

[35] M. Tetsu, O. Takahiro, S. Einoshin, and S. Yoichi. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016.

[36] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, 2008.

[37] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[38] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *TPAMI*, 2016.

[39] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, 2016.

[40] M. Ye, C. Liang, Y. yu, and etal. Person re-identification via ranking aggregation of similarity pulling and dissimilarity pushing. In *TMM*, 2016.

[41] M. Ye, J. Ma, J. Li, L. Zheng, and P. Yuen. Label graph matching for unsupervised video re-identification. In *ICCV*, 2017.

[42] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *CVPR*, 2016.

[43] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016.

[44] Y. Zhang, B. Li, and H. L. A. I. X. Ruan. Sample-specific svm learning for person re-identification. In *CVPR*, 2016.

[45] Z. Zhang, X. Jing, and T. Wang. Label propagation based semi-supervised learning for software defect prediction. *ASE*, 2017.

[46] Z. Zhang, M. Zhao, and T. W. S. Chow. Label propagation and soft-similarity measure for graph based constrained semi-supervised learning. In *IJCNN*, 2014.

[47] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

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

[49] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[50] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv*, 2016.

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

[52] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[53] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[54] X. Zhu, X. Jing, F. Wu, and H. Feng. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *IJCAI*, 2016.