MARS: A Video Benchmark for Large-Scale Person Re-Identification

Liang Zheng[1,3], Zhi Bie[1], Yifan Sun[1], Jingdong Wang[2],
Chi Su[4], Shengjin Wang[1](✉), and Qi Tian[3]

[1] Tsinghua University, Beijing, China
liangzheng06@gmail.com, wgsgj@tsinghua.edu.cn
[2] Microsoft Research, Beijing, China
[3] UTSA, San Antonio, USA
[4] Peking University, Beijing, China

Abstract. This paper considers person re-identification (re-id) in videos. We introduce a new video re-id dataset, named **M**otion **A**nalysis and **R**e-identification **S**et (MARS), a video extension of the Market-1501 dataset. To our knowledge, MARS is the largest video re-id dataset to date. Containing 1,261 IDs and around 20,000 tracklets, it provides rich visual information compared to image-based datasets. Meanwhile, MARS reaches a step closer to practice. The tracklets are automatically generated by the Deformable Part Model (DPM) as pedestrian detector and the GMMCP tracker. A number of false detection/tracking results are also included as distractors which would exist predominantly in practical video databases. Extensive evaluation of the state-of-the-art methods including the space-time descriptors and CNN is presented. We show that CNN in classification mode can be trained from scratch using the consecutive bounding boxes of each identity. The learned CNN embedding outperforms other competing methods considerably and has good generalization ability on other video re-id datasets upon fine-tuning.

Keywords: Video person re-identification · Motion features · CNN

1 Introduction

Person re-identification, as a promising way towards automatic VIDEO surveillance, has been mostly studied in pre-defined IMAGE bounding boxes (bbox). Impressive progress has been observed with image-based re-id. However, rich information contained in video sequences (or tracklets) remains under-explored. In the generation of video database, pedestrian detectors [11] and offline trackers [7] are readily available. So it is natural to extract tracklets instead of single (or multiple) bboxes. This paper, among a few contemporary works [25,29,36,38,41], makes initial attempts on video-based re-identification.

The dataset and codes are available at http://www.liangzheng.com.cn.

B. Leibe et al. (Eds.): ECCV 2016, Part VI, LNCS 9910, pp. 868–884, 2016.
DOI: 10.1007/978-3-319-46466-4_52

With respect to the "probe-to-gallery" pattern, there are four re-id strategies: image-to-image, image-to-video, video-to-image, and video-to-video. Among them, the first mode is mostly studied in literature, and previous methods in image-based re-id [5,24,35] are developed in adaptation to the poor amount of training data. The second mode can be viewed as a special case of "multi-shot", and the third one involves multiple queries. Intuitively, the video-to-video pattern, which is our focus in this paper, is more favorable because both probe and gallery units contain much richer visual information than single images. Empirical evidences confirm that the video-to-video strategy is superior to the others (Fig. 3).

Currently, a few video re-id datasets exist [4,15,28,36]. They are limited in scale: typically several hundred identities are contained, and the number of image sequences doubles (Table 1). Without large-scale data, the scalability of algorithms is less-studied and methods that fully utilize data richness are less likely to be exploited. In fact, the evaluation in [43] indicates that re-id performance drops considerably in large-scale databases.

Moreover, image sequences in these video re-id datasets are generated by hand-drawn bboxes. This process is extremely expensive, requiring intensive human labor. And yet, in terms of bounding box quality, hand-drawn bboxes are biased towards ideal situation, where pedestrians are well-aligned. But in reality, pedestrian detectors will lead to part occlusion or misalignment which may have a non-ignorable effect on re-id accuracy [43]. Another side-effect of hand-drawn box sequences is that each identity has one box sequence under a camera. This happens because there are no natural break points inside each sequence. But in automatically generated data, a number of tracklets are available for each identity due to miss detection or tracking. As a result, in practice one identity will have multiple probes and multiple sequences as ground truths. It remains unsolved how to make use of these visual cues.

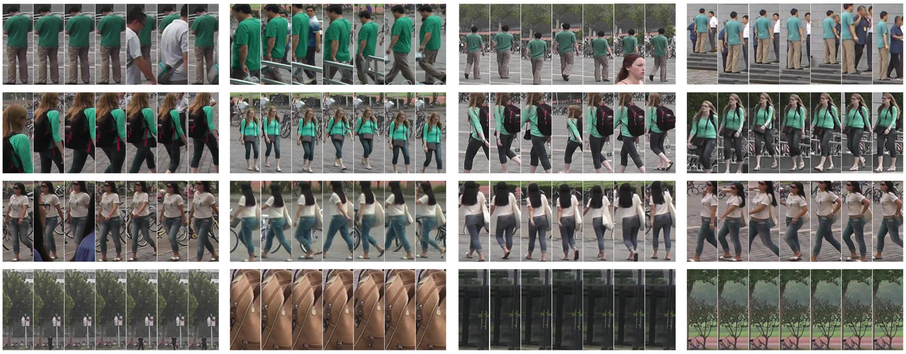


Fig. 1. Sample tracklets in MARS. The first three rows each corresponds to an identity, and tracklets in each column belong to different cameras. The last row presents four examples of false detection and tracking results. Images are shown every 6 frames

In light of the above discussions, it is of importance to (1) introduce large-scale and real-life video re-id datasets and (2) design effective methods which fully utilizes the rich visual data. To this end, this paper contributes in collecting and annotating a new person re-identification dataset, named "Motion Analysis and Re-identification Set" (MARS) (Fig. 1). Overall, MARS is featured in several aspects. First, MARS has 1,261 identities and around 20,000 video sequences, making it the largest video re-id dataset to date. Second, instead of hand-drawn bboxes, we use the DPM detector [11] and GMMCP tracker [7] for pedestrian detection and tracking, respectively. Third, MARS includes a number of distractor tracklets produced by false detection or tracking result. Finally, the multiple-query and multiple-ground truth mode will enable future research in fields such as query re-formulation and search re-ranking [45].

Apart from the extensive tests of the state-of-the-art re-id methods, this paper evaluates two important features: (1) motion features including HOG3D [18] and the gait [13] feature, and (2) the ID-disciminative Embeddings (IDE) [46], which learns a CNN descriptor in classification mode. Our results show that although motion features achieve impressive results on small datasets, they are less effective on MARS due to intensive changes in pedestrian activity. In contrast, the IDE descriptor learned on the MARS training set significantly outperforms the other competing features, and demonstrates good generalization ability on the other two video datasets after fine-tuning.

2 Related Work

Re-id dataset review. Most previous studies of person re-id are based on image datasets. The most commonly used one is VIPeR [12], which consists of 632 identities and 1,264 images. Datasets with similar scale include RAiD [6], i-LIDS [47], *etc.* Recently, two large-scale image datasets are released, *i.e.,* CUHK03 [23] and Market1501 [43]. Both datasets contain over 10k bboxes that are generated by DPM detector. The Market1501 dataset further adds 500k distractor bboxes of false detection results. Results of the large-scale datasets demonstrate that re-id accuracy drops considerably with the increase in database size, thus calling for scalable re-id methods. In video re-id (Table 1), iLIDS-VID [36] and PRID-2011 [15] contain several hundred identities and twice the number of box sequences. 3DPES [3] and ETH [32] are of similar scales. On such small datasets, the scalability of methods cannot be fully evaluated. Considering the trend of large scale in vision, the re-id community is in need of scalable video re-id datasets which reflect more practical problems.

Motion features in action recognition and person re-identification. In action recognition, an image sequence is viewed as a 3-dim space-time volume. Space-time features can be extracted based on the space-time interest points [9,21,37]. These methods generate compact representation of an image sequence using the sparse interest points, which are sensitive to variations such as viewpoint, speed, scale, *etc* [17]. An improved version associates with space-time

Table 1. Comparing MARS with datasets in videos [3,15,32,36] and images [12,23,43]

Datasets	MARS	iLIDS	PRID	3DPES	ETH	CUHK03	VIPeR	Market
#identities	1,261	300	200	200	146	1,360	632	1,501
#tracklets	20,715	600	400	1000	146	-	-	-
#BBoxes	1,067,516	43,800	40k	200k	8580	13,164	1,264	32k
#distractors	3,248	0	0	0	0	0	0	2,793
#cam./ID	6	2	2	8	1	2	2	6
Produced by	DPM+GMMCP	hand	hand	hand	hand	hand	hand	DPM
Evaluation	mAP &CMC	CMC	CMC	CMC	CMC	CMC	CMC	mAP

volume based representations [30]. Popular descriptors include HOG3D [18], 3D-SIFT [33], *etc*, which can be viewed as extensions of their corresponding 2-dim versions. In person re-id, few works focus on motion features because it is challenging to discriminate pedestrians solely by motion. Among the few, Wang *et al.* [36] employ the HOG3D descriptor with dense sampling after identifying walking periodicity. Nevertheless, [36] has only been tested on two small video datasets without further validation on large-scale settings. In this paper, we mostly follow [36] in the video description part, and show that this strategy has some flaws in dealing with practical video data.

CNN in person re-id. In person re-id, the study of CNN [1,8,23,40] has only recently launched due to the small scale of re-id datasets. These works formulate person re-id as a ranking problem. Image pairs [1,23,40] or triplets [8] are defined as input to CNN, instead of single training images. Such design avoids the shortage of training images per identity by generating quadratically/cubically enlarged training sets. Then, with such input data, the network is designed to have parallel convolutional layers, max pooling layers, as well as fully connected layers to learn an optimized metric. In video-based re-id, McLaughlin *et al.* [29] propose a variant of the recurrent neural network to incorporate time flows, an idea that is later adopted by [38]. In this paper, since each pedestrian has a number of training data (from image sequences), we are capable of training a classification network [16]. In this scenario, each single image is represented by a feature vector, which will greatly accelerate online process by nearest neighbor search or ANN algorithms.

3 MARS Dataset

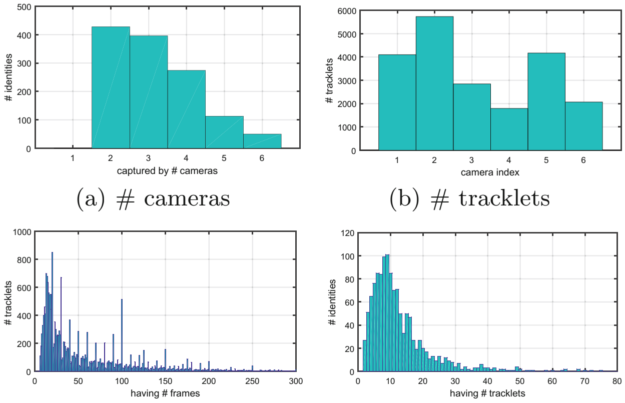
3.1 Dataset Description

In this paper, we introduce the MARS (Motion Analysis and Re-identification Set) dataset for video-based person re-identification. It is an extension of the Market-1501 dataset [43]. During collection, we placed six near-synchronized cameras in the campus of Tsinghua university. There were five 1,080 × 1920 HD cameras and one 640 × 480 SD camera. MARS consists of 1,261 different pedestrians whom are captured by at least 2 cameras.

For tracklet generation, we first use the DPM detector [11] to detect pedestrians. Then, the GMMCP tracker [7] is employed to group overlapping detection results in consecutive frames and fill in missing detection results. As output, a total of 20,715 image sequences are generated. Among them, 3,248 are distractor tracklets produced due to false detection or tracking results, which is close to practical usage. Overall, the following features are associated with MARS.

First, as shown in Table 1, compared with iLIDS-VID and PRID-2011, MARS has a much larger scale: 4 times and 30 times larger in the number of identities and total tracklets, respectively.

Second, the tracklets in MARS are generated automatically by DPM detector and GMMCP tracker, which differs substantially from existing datasets: the image sequences have high quality guaranteed by human labor. The detection/tracking error enables MARS to be more realistic than previous datasets. Moreover, in MARS, to produce "smooth" tracklets, we further apply average filtering to the bbox coordinates to reduce localization errors. As we will show in Sect. 5.3, tracklet smoothing improves the performance of motion features.



(a) # cameras

(b) # tracklets

(c) # frames

(d) # tracklets

Fig. 2. Statistics of the MARS dataset. (a): the number of identities captured by 1–6 cameras. (b): the numbers of tracklets in each camera. (c): the distribution of the number of frames in the tracklets. (d): the distribution of the number of tracklets belonging to the pedestrians

Third, in MARS, each identity has 13.2 tracklets on average. For each query, an average number of 3.7 cross-camera ground truths exist; each query has 4.2 image sequences that are captured under the same camera, and can be used as auxiliary information in addition to the query itself. As a result, MARS is an ideal test bed for algorithms exploring multiple queries or re-ranking methods [45]. Figure 2 provides more detailed statistics. For example, most IDs are captured by 2–4 cameras, and camera-2 produces the most tracklets. A large number of tracklets contain 25–50 frames, and most IDs have 5–20 tracklets.

3.2 Evaluation Protocol

In the MARS dataset, we stick to the cross-camera search mode as in previous datasets [12,23,43], *i.e.,* query and gallery are captured by different cameras. Each identity will have one probe under each camera. Since each identity may have multiple tracklets under a camera, the representative probe image is randomly selected from them, resulting in 2,009 probes. The MARS dataset is evenly divided into train and test sets, containing 631 and 630 identities, respectively, and this partition is fixed. The dataset is large, so we fix the train/test partitioning instead of repeating random partitioning for 10 or 20 times [12,23]. Then, given a query image sequence, all gallery items are assigned a similarity score. We then rank the gallery according to their similarity to the query. In our system, since a query has multiple ground truths, regular CMC curve (Cumulative Matching Characteristic, representing the expectation of the true match being found within the first n ranks) does not fully reflect the true ranking results, so we resort to both CMC and mAP as the evaluation metric [43]. The Average Precision (AP) is calculated based on the ranking result, and the mean Average Precision (mAP) is computed across all queries which is viewed as the final re-id accuracy. CMC is a pragmatic measurement focusing on retrieval precision, while mAP considers precision and recall and is useful for research purpose (Table 1).

Table 2. Three important features evaluated in the baseline

Features	Dim	Description
CNN	1,024	Using AlexNet [20], the three fully convolutional layers have 1,024, 1,024, and 631 blobs. Trained on MARS, fine-tuned on PRID and iLIDS. Using FC7 (after RELU) for testing.
HOG3D [18]	2,000	Motion feature. Detecting walking cycles by FEP [36]. HOG3D feature extracted from $8 \times 8 \times 6$ or $16 \times 16 \times 6$ patches and quantized using a codebook of size 2,000.
GEI [13]	2,400	Gait feature. Detecting walking cycles by FEP [36]. Pedestrian segmentation using code from [26]. Resulting maps within a cycle are resized to 80×30 and averaged to obtain the feature

4 Important Features

4.1 Motion Features

The **HOG3D** [18] feature has been shown to have competitive performance in action recognition [22]. In feature extraction, given a tracklet, we first identify walking cycles using Flow Energy Profile (FEP) proposed in [36]. For bboxes aligned in a cycle, we densely extract HOG3D feature in $8 \times 8 \times 6$ (or $16 \times$

16 × 6) space-time patches, with 50 % overlap between adjacent patches. The feature dimension of each space-time patch is 96. Since videos with different time duration have different numbers of the dense space-time tubes, we encode the local features into a Bag-of-Words (BoW) model. Specifically, a codebook of size 2,000 is trained by k-means on the training set. Then, each 96-dim descriptor is quantized to a visual word defined in the codebook. So we obtain a 2,000-dim BoW vector for an arbitrary-length video. We do not partition the image into horizontal stripes [43] because this strategy incurs larger feature dimension and in our preliminary experiment does not improve re-id accuracy.

The **Gait Energy Image (GEI)** [13] is widely applied in gait recognition. In GEI extraction, we also first find walking cycles using FEP. Then, for bboxes within a cycle, we segment each bbox into foreground (pedestrian) and background using the code released by Luo *et al.* [26]. The resulting binary images within a cycle are averaged to yield the GEI of the tracklet. In our experiment, the size of GEI is 80 × 30, which is reshaped into a column as the final vector.

After feature extraction, we learn a metric on the training set using several metric learning schemes such as Kissme [19] and XQDA [24], due to their efficiency and accuracy.

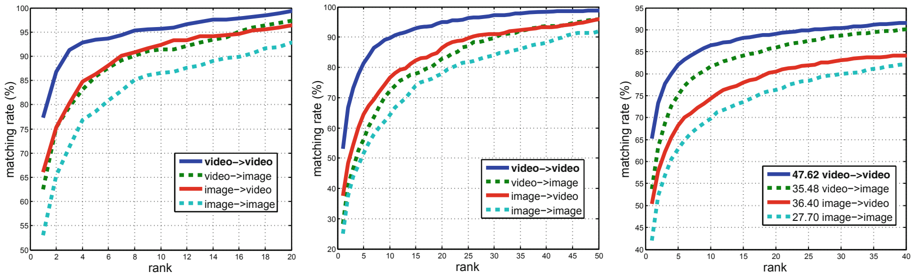
4.2 CNN Features

The Convolutional Neural Network (CNN) has achieved state-of-the-art accuracy in a number of vision tasks. In person re-identification, current CNN methods [1,8,23,40] take positive and negative image pairs (or triplets) as input to the network due to the lack of training data per identity. In this paper, we employ the ID-discriminative Embedding (IDE) [46] using CaffeNet [20] to train the re-id model in classification mode. More sophisticated networks [14,34] may yield higher re-id accuracy.

During training, images are resized to 227 × 227 pixels, and along with their IDs (label) are fed into CNN in batches. Through five convolutional layers with the same structure as the CaffeNet [20], we define two fully connected layers each with 1,024 blobs. The number of blobs in the 8th layer is equal to the number of training identities which in the case of MARS is 631. The total number of training bboxes on MARS is 518k.

In testing, since re-id is different from image classification in that the training and testing identities do not overlap, we extract probe and gallery features using the CNN model before metric learning steps. Specifically, we extract the FC7 features for all bboxes in an input tracklet. Then, max/average pooling is employed to generate a 1,024-dim vector for an tracklet of arbitrary length (A comparison between the two pooling methods can be accessed in Sect. 5). Finally, metric learning is leveraged as in image-based re-id. In Sect. 5.4, we will demonstrate that IDE descrptors learned through person classification can be effectively used in re-id.

When transferring the CNN model trained on MARS to other video re-id datasets, we fine-tune the MARS-learned CNN model on the target datasets. In experiment, we find that fixing parameters in the convolutional layers typically



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

Fig. 3. Comparison of four re-id strategies on three datasets. CNN features trained/fine-tuned on the corresponding datasets are used. We adopt XQDA for metric learning (Kissme yields very similar performance with XQDA). Numbers in the legend of (c) MARS are mAP results. Clearly, the video-to-video mode is superior

results in compromised accuracy, so in practice, all the 7 CNN layers including the convolutional and fully connected layers are fine tuned. The Last fully connected layer (FC8) is trained from scratch.

5 Experiments

5.1 Datasets

We use three datasets, *i.e.,* PRID-2011 [15], iLIDS-VID [36], and MARS. For the former two, we use the Cumulative Match Characteristics (CMC) curve for evaluation, which is averaged over ten train/test partitions. We use the same partition rule as [36]. For MARS, a fixed partitioning is used (our preliminary experiments show that different paritions yield stable and consistent results), and mAP and CMC are both reported.

PRID-2011 dataset contains 400 image sequences of 200 pedestrians under two cameras. Each image sequence has a length of 5 to 675 frames. Following [36], sequences with more than 21 frames from 178 persons are used. So the probe and gallery both have 89 identities.

iLIDS-VID dataset is a newly released dataset consisting of 300 identities and each has 2 image sequences, totaling 600 sequences. The length of image sequences varies from 23 to 192, with an average number of 73. This dataset is more challenging due to environment variations. The test and training set both have 150 identities.

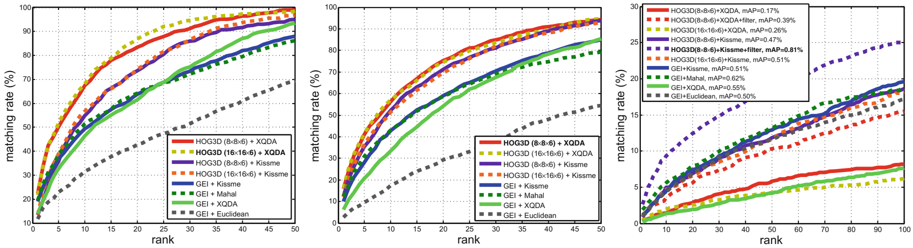
5.2 Why Do We Prefer Video-Based Re-Identification?

This paper mentioned four re-id modes in the Sect. 1. In this section, we will evaluate their performance and gain insights in the reason why video re-id should be preferred. Among the four modes, “video->video” is what we have described in this paper; in "video->image", the "image" is chosen as the first frame of a video, and this mode corresponds to the multiple-query method in image retrieval [2]; the "image->video" mode chooses the query image as the first frame as well, and corresponds to the multi-shot scenario in person re-id; finally, "image->image" is the common re-id problem in which both query and gallery images are the first frame of the tracklets. Note that we select the first frame as the representative of a tracklet only to ease experiment, and all frames roughly have similar quality ensured by the concatenation of DPM and GMMCP. For MARS, we use the CNN feature learned on its training data, initialized with the ImageNet model; for iLIDS and PRID, the fine-tuned CNN models are leveraged, initialized by ImageNet and MARS, respectively. Max pooling is used for MARS and PRID, and averge pooling for iLIDS, to aggregate bbox features. We use XQDA [24] in metric learning.

We observe from Fig. 3 that the video-to-video re-id strategy significantly outperforms the other three, while the image-to-image mode exhibits the worst performance. On MARS, for example, video-based re-id exceeds image-based re-id by +19.92 % in mAP. The video-to-image and image-to-video modes also have considerable improvment over using single images only, but are inferior to video re-id. On MARS, "image->video" outperforms "video->image" probably because the former has richer visual cues and a finer distance metric can be learned. Previous studies on person re-identification mostly focus on the image-to-image mode, while this paper argues that the generation of tracklets instead of single images will be both more natural and higher accuracy can be expected. Our results lay a groundwork for the argument: other things being equal, video re-id consistently improves re-id accuracy and should be paid more emphasis on.

5.3 Evaluation of Motion Features

As described in Sect. 4.1, we use HOG3D and Gait Energy Image (GEI) features for motion representation. Results are presented in Fig. 4 and Table 4.



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

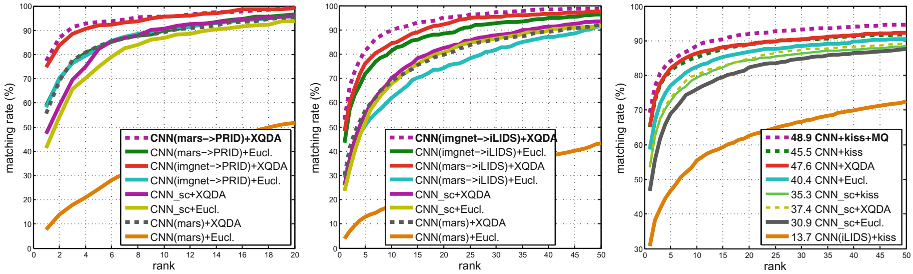
Fig. 4. CMC curves of HOG3D and GEI features on three video re-id datasets. HOG3D with 8 × 8 × 6 and 16 × 16 × 6 sampling strategies are presented. For MARS, "filter" denotes average filtering to smoothen consecutive bboxes

Performance of HOG3D and GEI. In Fig. 4 and Table 4, we observe that HOG3D and GEI feature both yield decent accuracy on iLIDS-VID and PRID-2011 datasets. Specifically, for HOG3D, the rank-1 accuracy is 21.68 % and 16.13 % on PRID-2011 and iLIDS-VID datasets, respectively; for GEI, the rank-1 accuracy is 19.00 % and 10.27 %, respectively. Our implementation is competitive with [36]: matching rate is relatively lower in low ranks, but higher in larger ranks (see Fig. 6 for clear comparison). Therefore, on the two small datasets, both features have competitive performance, and HOG3D is slightly superior.

On the MARS dataset, however, the performance of HOG3D and GEI both drops considerably. The rank-1 accuracy of HOG3D is 2.61 % with Kissme, and mAP is 0.81 %; for GEI, rank-1 accuracy and mAP are 1.18 % and 0.40 %, respectively. For the two features, both precision (rank-1) and recall (mAP) are low on MARS. **The reason why motion features have poor performance on MARS is two-fold.** On one hand, a larger dataset will inevitably contain many pedestrians sharing similar motion feature with the probe, and it is challenging to discriminate different persons based on motions. On the other hand, since MARS is captured by six cameras, motion of the same identity may undergo significant variations due to pedestrian pose change (see Fig. 7 for visual results), so the motion-based system may miss the same pedestrian under motion variations. For example, a walking person with frontal and side views will have large intra-class variability, let alone considering persons standing still with hardly any motion at all.

Impact of tracklet smoothing. In MARS, consecutive bboxes in the tracklets may not be smooth due to detection errors. To correct this, we employ the average filtering to smoothen bboxes within each tracklet. In our experiment, we use a window of size 5, and across the 5 bboxes, compute the average coordinates of the upper left point as well as the frame width and height, which is taken as the smoothed bbox of the frame in the middle. Features are then extracted using the smoothed tracklets. In Fig. 4(c), we find that, the smoothing strategy yields some marginal improvement for HOG3D feature (mAP increases from 0.47 % to 0.81 %). For GEI, the segmentation method [26] already corrects this artifact, so the improvement is limited (not shown, because it overlaps with other lines).

In summary, on PRID-2011 and iLIDS-VID datasets, motion features such as HOG3D and GEI are effective for two reasons: both datasets have relatively small scales; image sequences in both datasets do not undergo significant variances. On the MARS dataset, our observation goes that motion features have much lower accuracy. In comparison with PRID-2011 and iLIDS-VID datasets, MARS has much more tracklets, and the tracklets have more intensive variations in viewpoint, pose, *etc.* Intra-class variance is large, while inter-class variance can be small, making it challenging for effective usage of motion features. Figure 7 presents re-id examples in which motion feature fails.



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

Fig. 5. CMC curves on three video re-id datasets with the IDE feature. "(mars->PRID)" represents CNN model pre-trained on MARS and fine-tuned on PRID. "sc" means training from scratch. "(mars)" and "(iLIDS)" indicate that the model is trained on MARS or iLIDS and then directly transferred to the target set. Kissme and XQDA are used as distance metric; otherwise, Euclidean (Eucl.) distance is used. "MQ" denotes multiple queries

Table 3. Method comparisons on three datasets. "Self" means training IDE on the target dataset set. "MARS" denotes directly transferring MARS-learned model to the target dataset. "Self pretrained on MARS" stands for fine-tuning IDE on a MARS-initialized model. "avg" means using average pooling, or otherwise, max pooling. All models are first initialized with the ImageNet model. Red and blue numbers indicate average pooling compromises and improves accuracy, respectively

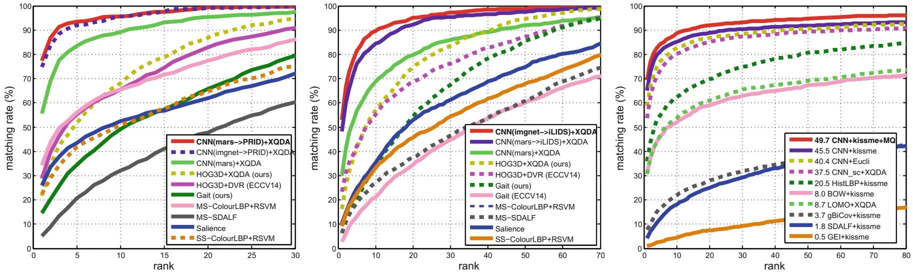
Train. Set	Metric	*PRID-2011*				*iLIDS-VID*				*MARS*			
		1	5	10	20	1	5	10	20	1	5	20	mAP
Self	Eucl.	58.2	82.7	90.6	98.2	40.5	70.0	78.9	84.7	58.7	77.1	86.8	40.4
	avg.	-1.1	-0.7	-0.4	-0.2	+3.1	+1.9	+2.5	+2.7	+1.3	+0.8	+1.1	+2.0
	Kiss.	66.3	88.5	93.9	98.2	47.6	76.1	86.1	92.5	65.0	81.1	88.9	45.6
	avg.	+0.2	-1.6	-1.1	-0.4	+1.2	-0.5	+1.1	+0.1	-0.2	-0.1	-0.5	+1.6
	XQDA	74.8	92.1	95.7	99.1	51.3	79.1	87.2	94.3	65.3	82.0	89.0	47.6
	avg.	-2.0	-1.8	-0.2	-0.4	+1.7	+2.3	+2.5	+0.8	-0.7	-0.6	+0.1	-0.1
MARS	Eucl.	7.6	24.6	39.0	51.8	2.9	9.9	14.7	23.0	-	-	-	-
	avg.	-0.3	-1.1	-1.2	-1.1	+1.0	+3.1	+2.8	+3.9				
	XQDA	55.5	83.6	89.3	95.4	24.3	56.3	66.9	79.3	-	-	-	-
	avg.	-2.1	-3.3	-0.2	-0.2	+5.7	+4.1	+2.0	+0.4				
Self pretrained on MARS	Eucl.	58.9	93.5	95.7	99.3	25.9	44.4	57.2	71.5	-	-	-	-
	avg.	+2.2	+2.5	+2.5	+0.4	-0.9	-1.9	+0.9	-0.3				
	XQDA	77.3	93.5	95.7	99.3	47.1	76.7	85.6	93.3	-	-	-	-
	avg.	-3.6	-0.9	-1.1	-0.9	+1.1	-0.4	-1.2	-1.1				

5.4 Evaluation of the CNN Feature

Training from scratch vs. fine-tuning on ImageNet. For the three datasets, CNN modesl are either trained from scratch or fine-tuned on ImageNet-pretrained models. In Fig. 5, we observe that fine-tuning on the ImageNet model consistently outperforms training from scratch. On MARS, fine-tuning brins about +9.5 % and 10.2 % improvement when Euclidean or XQDA distances are used, respectively. Situation on iLIDS and PRID is simila. In the following experiments, we always employ ImageNet-initilized models, if not specified.

Comparison with motion features. In Fig. 6 and Table 4, direct comparisons are made available between the two feature types. On PRID and iLIDS, "CNN+XQDA" exceeds "HOG3D+XQDA" by 55.6 % and 46.9 % in rank-1 accuracy, respectively. On MARS, the performance gap is 65.7 % and 48.5 % in rank-1 and mAP, respectively. On all the three datasets, CNN outperforms the motion features by a large margin, validating the effectiveness of appearance models.

Generalization ability of MARS. We conduct two experiments to study the relationship between MARS and the two small datasets. First, we directly transfer the CNN model trained on MARS to iLIDS and PRID. In Figs. 5(a) and (b), and Table 3, we directly extract features with the MARS-trained model for iLIDS and PRID. We find that re-id accuracy with Euclidean distance is pretty low. This is expected because the data distribution of MARS is different from that of iLIDS and PRID. Metric learning then improves accuracy to a large extent. Our second experiment is fine-tuning a CNN model on the target set using MARS-pretrained models (Self pretrained on MARS). On both dataset, fine-tuning yields improvement over direct model transfer. On PRID, we achieve rank-1 accuracy = 77.3, which is higher than fine-tuning from ImageNet. On iLIDS, fine-tuning from MARS is lower than ImageNet by ∼4 % in rank-1 accuracy. This demonstrates that data distribution of PRID is close to MARS, while iLIDS seems to be more different. We note that MARS was captured in summer while the other two depicts scenes in colder seasons, and that PRID and MARS are both outdoor datasets, while iLIDS is an indoor one.



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

Fig. 6. Comparison with state-of-the-art methods on three video re-id datasets. Numbers before each method name in (c) denote the mAP(%). "(ours)" and "ECCV14" denote result implemented by ourselves and borrowed from [36], respectively

Max pooling vs. avg pooling. In this paper, bbox features within a tracklet are combined into a fixed-length vector using max/average pooling. Now we compare the performance of the two pooling methods. Results are summarized in Table 3, in which max pooling is used if not specified. We observe that max pooling is generally better on PRID and MARS, while for iLIDS, average pooling seems to be superior. One possible explanation is that max pooling helps to find local salient features, which is desirable under large illuminations changes like in iLIDS. Other pooling options are worth exploiting, such as the $\mathcal{L}_p-$norm pooling [44], the fisher vector encoding [31], *etc.*

Multiple queries (MultiQ). In MARS, multiple tracklets for the same ID exist within the same camera as mentioned in Sect. 3.1. They contain rich information as the pedestrian may have varying poses within each tracklet. Following [46], we re-formulate each probe by max-pooling the tracklets within the same camera. Results are presented in Fig. 6(c) and Table 4. With multiple queries, we improve the rank-1 accuracy from 65.0 % to 68.3 % (+3.3 %).

5.5 Comparison with State-of-the-arts

In Table 4 and Fig. 6, we compare our results with the state-of-the-art methods. On PRID-2011 and iLIDS-VID datasets, five descriptors are compared, *i.e.,* HOG3D [18], color, color+LBP, SDALF [10], Salience [42], and BoW [43]. Three metric learning methods, *i.e.,* DVR [36], XQDA [24], and Kissme [19] are evaluated. We observe that the CNN descriptor is superior to these methods, obtaining rank-1 accuracy = 77.3 % and 53.0 % on PRID and iLIDS, respectively. Comparing with recent video re-id works, the best known rank-1 accuracy is 70 % and 58 % on PRID and iLIDS, respectively, both reported in [29]. So this paper sets a new state of the art on PRID, and is 5 % lower on iLIDS. On the MARS dataset, results of another set of features are presented, *i.e.,* HistLBP [39], gBiCov [27], LOMO [24], BoW [43], and SDALF [10]. We report rank-1 accuracy = 68.3 % and mAP = 49.3 % on MARS.

Table 4. Results of the state-of-the-art methods on the 3 datasets. Accuracy is presented by mAP and precision in rank 1, 5, and 20. We use average pooling for iLIDS, and max pooling for PRID and MARS. Except for iLIDS, we use the MARS-pretrained CNN model. ImageNet initialization is always employed. Best results are in blue

Methods	*PRID-2011*			*iLIDS-VID*		
Rank R	1	5	20	1	5	20
HOG3D+DVR	28.9	55.3	82.8	23.3	42.4	68.4
Color+DVR	41.8	63.8	88.3	32.7	56.5	77.4
ColorLBP+DVR	37.6	63.9	89.4	34.5	56.7	77.5
SDALF+DVR	31.6	58.0	85.3	26.7	49.3	71.6
Salience+DVR	41.7	64.5	88.8	30.9	54.4	77.1
BoW+XQDA	31.8	58.5	81.9	14.0	32.2	59.5
GEI+Kiss.	19.0	36.8	63.9	10.3	30.5	61.5
HOG3D+XQDA	21.7	51.7	87.0	16.1	41.6	74.5
CNN+Kiss	69.9	90.6	98.2	48.8	75.6	92.6
CNN+XQDA	77.3	93.5	99.3	53.0	81.4	95.1

Methods	*MARS*			
Rank R	1	5	20	mAP
HistLBP+XQ.	18.6	33.0	45.9	8.0
gBiCov+XQ.	9.2	19.8	33.5	3.7
LOMO+XQ.	30.7	46.6	60.9	16.4
BoW+Kissme	30.6	46.2	59.2	15.5
SDALF+DVR	4.1	12.3	25.1	1.8
HOG3D+Kiss.	2.6	6.4	12.4	0.8
GEI+Kiss.	1.2	2.8	7.4	0.4
CNN+XQDA	65.3	82.0	89.0	47.6
CNN+Kiss.	65.0	81.1	88.9	45.6
+MQ	68.3	82.6	89.4	49.3

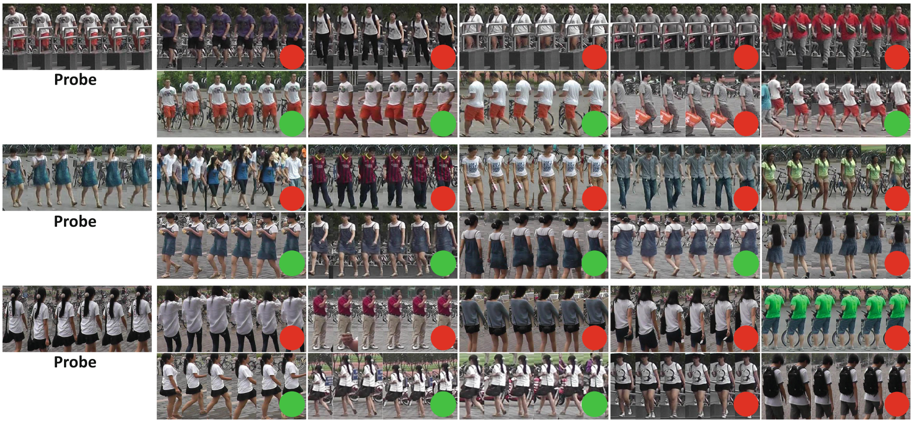


Fig. 7. Sample re-id results of three probes. For each probe, the first and the second row display ranking results obtained by HOG3D and CNN features, respectively. Green and red discs denote the same and different person with the probe, respectively

In Fig. 7, three sample re-id results are shown. Our observation is that a large number of pedestrians share similar motion, a phenomenon that is less-studied on small datasets. The usage of motion features therefore tends to find pedestrians with similar activities. In contrast, by replacing motion features with the CNN feature, We find the CNN embedding is effective in dealing with image variances and yields superior results to motion features.

6 Conclusions

This paper advocates using video tracklets in person re-identification. Attempts are made in constructing a realistic video re-id dataset, named "MARS". This dataset is four times larger than previous video re-id datasets, and is collected with automatic detector and tracker. Moreover, MARS dataset is featured by multi-query, multi-ground truth, and over 3,000 distractor tracklets produced by false detection and tracking results. These characteristics make MARS an ideal test bed for practical re-id algorithms. We employ two motion features as well as the Convolutional Neural Networks to learn a discriminative embedding in the person subspace. Our experiments reveal that motion features that were previously proved successful on small datasets turn out to be less effective under realistic settings with complex background, occlusion, and various poses. Instead, given the large amount of training data in video datasets, the learned CNN feature outperforms motion features and a number of state-of-the-art image descriptors to a large margin, and has good generalization ability on other video datasets.

Multiple research directions are made possible with MARS. For example, it is important to design view-invariant motion features that can deal with view changes in real-life datasets. Since each tracklet has multiple frames, another feasible topic is video pooling which aims to find discriminative information within video frames. Moreover, when classic CNNs can be trained on the rich visual data, a number of CNN variants can be explored such as those utilizing human parts. While this paper finds it less effective to use motion features, it is interesting to exploit the temporal cues in addition to appearance models [29,38]. Our preliminary results have revealed some moderate improvement using LSTM, and further experiment is needed to extend the temporal models. Finally, since there exists a number of tracking datasets, it remains unknown how to transfer these data to the target domains.

Acknowledgement. We thank Zhun Zhong and Linghui Li for providing some benchmarking results. This work was supported by Initiative Scientific Research Program of Ministry of Education under Grant No. 20141081253. This work was supported in part to Dr. Qi Tian by ARO grants W911NF-15-1-0290, Faculty Research Gift Awards by NEC Laboratories of America and Blippar, and National Science Foundation of China (NSFC) 61429201.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR (2015)
2. Arandjelovic, R., Zisserman, A.: Multiple queries for large scale specific object retrieval. In: BMVC (2012)
3. Baltieri, D., Vezzani, R., Cucchiara, R.: 3dpes: 3d people dataset for surveillance and forensics. In: ACM Workshop on Human Gesture and Behavior Understanding (2011)
4. Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., Lucey, P.: A database for person re-identification in multi-camera surveillance networks. In: DICTA (2012)
5. Chen, D., Yuan, Z., Hua, G., Zheng, N., Wang, J.: Similarity learning on an explicit polynomial kernel feature map for person re-identification. In: CVPR (2015)
6. Das, A., Chakraborty, A., Roy-Chowdhury, A.K.: Consistent re-identification in a camera network. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 330–345. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10605-2_22
7. Dehghan, A., Assari, S.M., Shah, M.: Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In: CVPR (2015)
8. Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recogn. **48**(10), 2993–3003 (2015)
9. 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 (2005)
10. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Pattern Anal. Mach. Intell. IEEE Trans. **32**(9), 1627–1645 (2010)
12. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, vol. 3 (2007)

13. Han, J., Bhanu, B.: Individual recognition using gait energy image. Pattern Anal. Mach. Intell. IEEE Trans. **28**(2), 316–322 (2006)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Image Analysis, pp. 91–102 (2011)
16. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM Multimedia, pp. 675–678 (2014)
17. Ke, Y., Sukthankar, R., Hebert, M.: Volumetric features for video event detection. Int. J. Comput. Vis. **88**(3), 339–362 (2010)
18. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC (2008)
19. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR, pp. 2288–2295 (2012)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
21. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. **64**(2–3), 107–123 (2005)
22. Li, W., Wang, X.: Locally aligned feature transforms across views. In: CVPR (2013)
23. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR, pp. 152–159 (2014)
24. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
25. Liu, K., Ma, B., Zhang, W., Huang, R.: A spatio-temporal appearance representation for video-based pedestrian re-identification. In: CVPR, pp. 3810–3818 (2015)
26. Luo, P., Wang, X., Tang, X.: Pedestrian parsing via deep decompositional network. In: ICCV (2013)
27. Ma, B., Su, Y., Jurie, F.: Covariance descriptor based on bio-inspired features for person re-identification and face verification. IVC **32**(6), 379–390 (2014)
28. Martinel, N., Micheloni, C., Piciarelli, C.: Distributed signature fusion for person re-identification. In: ICDSC (2012)
29. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: CVPR (2016)
30. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
31. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. IJCV **105**(3), 222–245 (2013)
32. Schwartz, W.R., Davis, L.S.: Learning discriminative appearance-based models using partial least squares. In: SIBGRAPI (2009)
33. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: ACM Multimedia (2007)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint (2014). arXiv:1409.1556
35. Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L.S., Gao, W.: Multi-task learning with low rank attribute embedding for person re-identification. In: CVPR (2015)
36. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 688–703. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_45

37. Willems, G., Tuytelaars, T., Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88688-4_48
38. Wu, L., Shen, C., Hengel, A.V.D.: Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. arXiv preprint (2016). arXiv:1606.01609
39. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 1–16. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_1
40. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: ICPR, pp. 34–39 (2014)
41. You, J., Wu, A., Li, X., Zheng, W.S.: Top-push video-based person re-identification. In: CVPR (2016)
42. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR (2013)
43. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: CVPR (2015)
44. Zheng, L., Wang, S., Liu, Z., Tian, Q.: Lp-norm idf for large scale image search. In: CVPR (2013)
45. Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., Tian, Q.: Query-adaptive late fusion for image search and person re-identification. In: CVPR (2015)
46. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Tian, Q.: Person re-identification in the wild. arXiv preprint (2016). arXiv:1604.02531
47. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: BMVC, vol. 2, p. 6 (2009)