Scalable Person Re-identification: A Benchmark

Liang Zheng[†‡*], Liyue Shen[†*], Lu Tian[†*], Shengjin Wang[†], Jingdong Wang[§], Qi Tian[‡]

[†]Tsinghua University [§]Microsoft Research [‡]University of Texas at San Antonio

Abstract

This paper contributes a new high quality dataset for person re-identification, named "Market-1501". Generally, current datasets: 1) are limited in scale; 2) consist of hand-drawn bboxes, which are unavailable under realistic settings; 3) have only one ground truth and one query image for each identity (close environment). To tackle these problems, the proposed Market-1501 dataset is featured in three aspects. First, it contains over 32,000 annotated bboxes, plus a distractor set of over 500K images, making it the largest person re-id dataset to date. Second, images in Market-1501 dataset are produced using the Deformable Part Model (DPM) as pedestrian detector. Third, our dataset is collected in an open system, where each identity has multiple images under each camera.

As a minor contribution, inspired by recent advances in large-scale image search, this paper proposes an unsupervised Bag-of-Words descriptor. We view person re-identification as a special task of image search. In experiment, we show that the proposed descriptor yields competitive accuracy on VIPeR, CUHK03, and Market-1501 datasets, and is scalable on the large-scale 500k dataset.

1. Introduction

This paper considers the task of person re-identification. Given a probe image (query), our task is to search in a gallery (database) for images that contain the same person.

Our work is motivated by two aspects. First, most existing person re-identification datasets [10, 44, 4, 13, 22, 19] are flawed either in the dataset scale or data richness. Specifically, the number of identities is often confined in several hundred. This makes it infeasible to test the robustness of algorithms under large-scale data. Moreover, images of the same identity are usually captured by two cameras; each identity has one image under each camera, so the number of queries and relevant images is very limited. Furthermore, in most datasets, pedestrians are well-aligned by hand-drawn bboxes (bboxes). But in reality, when pedestrian detectors are used, the detected persons may undergo misalignment or part missing (Fig. 1). On the other hand, pedestrian detectors, while producing true positive bboxes, also yield false alarms caused by complex background or occlusion (Fig. 1). These distractors may exert non-ignorable influence on recognition accuracy. As a result, current methods may be biased toward ideal settings and their effectiveness may be impaired once the ideal dataset meets reality. To address this problem, it is important to introduce datasets that reach closer to realistic settings.

Second, local feature based approaches [11, 40, 38, 3] are proven to be effective in person re-identification. Considering the "query-search" mode, this is potentially compatible with image search based on the Bag-of-Words (BoW) model. Nevertheless, some state-of-the-art methods in person re-identification rely on brute-force feature-feature matching [39, 38]. Although good recognition rate is achieved, this line of methods suffer from low computational efficiency, which limits its potential in large-scale applications. In the BoW model, local features are quantized to *visual words* using a pretrained *codebook*. An image is thus represented by a visual word histogram weighted by TF-IDF scheme. Instead of performing exhaustive visual matching among images [39], in the BoW model, local features are aggregated into a global vector.

Considering the above two issues, this paper makes two contributions. The main contribution is the collection of a new person re-identification dataset, named the "Market-1501" (Fig. 1). It contains 1,501 identities collected by 6 cameras. We further add a distractor set composed of 500K irrelevant images. To our knowledge, Market-1501 is the largest person re-id dataset featured by 32,668+500K bboxes and 3,368 query images. It is distinguished from existing datasets in three aspects: DPM detected bboxes, the inclusion of distractor images, and multi-query, multi-ground truth per identity. This dataset thus provides a more real-

[*] Three authors contribute equally to this work.
Dataset and code are available at http://www.liangzheng.com.cn.

1550-5499/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCV.2015.133

IEEE computer society

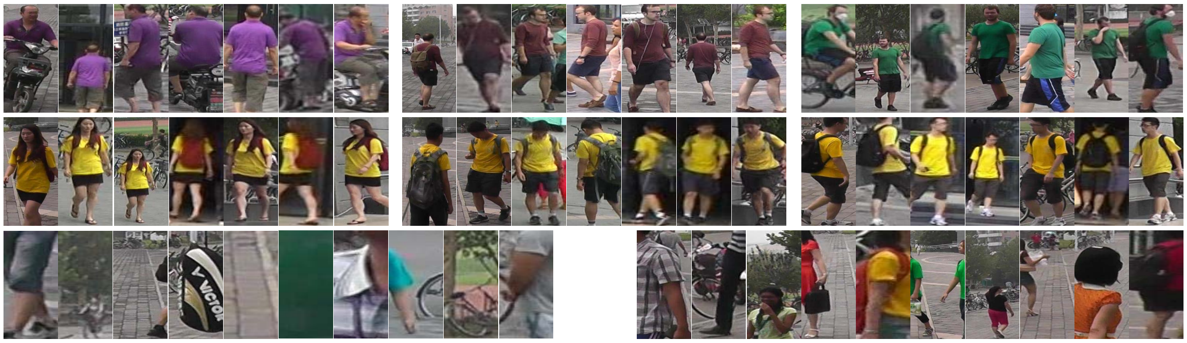


Figure 1. Sample images of the Market-1501 dataset. All images are normalized to 128×64 (**Top:**) Sample images of three identities with distinctive appearance. (**Middle:**) We show three cases where three individuals have very similar appearance. (**Bottom:**) Some samples of the distractor images (left) as well as the junk images (right) are provided.

istic benchmark. For accuracy evaluation, we propose to use mean average precision (mAP), a more comprehensive measurement compared to the commonly used Cumulated Matching Characteristics (CMC) curve [38, 39, 20].

As a minor contribution, inspired by the state-of-the-art image search systems, an unsupervised BoW representation is proposed. After generating a codebook on training data, each pedestrian image is represented as a visual word histogram. In this step, a number of techniques are integrated, *e.g.,* root descriptor [2], negative evidences [14], burstiness weighting [16], avgIDF [41], *etc*. Moreover, several further improvements are adopted, *i.e.,* weak geometric constraints, Gaussian Mask, multiple queries, and reranking. By simple dot product as similarity measurement, we show that the proposed BoW representation yields competitive recognition accuracy while enjoying a fast response time.

2. Related Work

For person re-identification, both supervised and unsupervised models have been extensively studied these years. In discriminative models [28, 12, 7, 20, 3], classic SVM (or the RankSVM [28, 40]) and boosting [11, 30] are popular choices. For example, Zhao *et al.* [40] learn the weights of filter responses and patch matching scores using RankSVM, and Gray *et al.* [11] perform feature selection among an ensemble of local descriptors by boosting. Recently, li *et al.* [20] propose a deep learning network to jointly optimize all pipeline steps. This line of works are beneficial in reducing the impact of multi-view variations, but require laborious annotation, especially when new cameras are added in the system. On the other hand, in unsupervised models, Farenzena *et al.* [8] make use of both symmetry and asymmetry nature of pedestrians and propose the Symmetry-Driven Accumulation of Local Features (SDALF). Ma *et al.* [25] employ the Fisher Vector to encode local features into a global vector. To exploit the salience information among pedestrian images, Zhao *et al.* [38] propose to assign higher weight to rare colors, an idea very similar to the Inverse Document Frequency (IDF) [41] in image search. This paper proposes an unsupervised method which is well-adaptable to different camera networks.

On the other hand, the field of image search has been greatly advanced since the introduction of the SIFT descriptor [24] and the BoW model. In the last decade, a myriad of methods [15, 42, 45] have been developed to improve search performance. For example, to improve matching precision, Jégou *et al.* [15] embed binary SIFT features in the inverted file. Meanwhile, refined visual matching can also be produced by index-level feature fusion [42] between complementary descriptors. Since the BoW model does not consider the spatial distribution of local features (also a problem in person re-id), another direction is to model the spatial constraints [45, 37]. Spatial coding [45] checks the geometric consistency between images by the offset map, while Zhang *et al.* [37] discover visual phrases to encode spatial information. For ranking problems, an effective reranking step typically brings about improvements. Liu *et al.* [23] design a "one shot" feedback optimization scheme which allows a user to quickly refine the search results. Zheng *et al.* [43] propose to leverage the profile of the score lists to adaptively assign weights to various features. In [29], the top-ranked images are used as queries again and final score is the weighted sum of individual scores. When multiple queries are present [1], a new query can be formed by average or max operations. This paper integrates several state-of-the-art techniques in image search, yielding a competitive person re-id system.

3. The Market-1501 Dataset

3.1. Description

In this paper, a new person re-id dataset, the "Market-1501" dataset, is introduced. During dataset collection, a total of six cameras were placed in front of a campus supermarket, including five 1280×1080 HD cameras, and one 720×576 SD camera. Overlapping exists among these cam-

Datasets	Market-1501	RAiD [5]	CUHK03 [20]	VIPeR [10]	i-LIDS [44]	OPeRID [22]	CUHK02 [19]	CAVIAR [4]
# identities	1,501	43	1,360	632	119	200	1,816	72
# BBoxes	32,668	6920	13,164	1,264	476	7,413	7,264	610
# distractors	2,793 + 500K	0	0	0	0	0	0	0
# cam. per ID	6	4	2	2	2	5	2	2
DPM or Hand	DPM	hand	DPM	hand	hand	hand	hand	hand
Evaluation	mAP	CMC	CMC	CMC	CMC	CMC	CMC	CMC

Table 1. Comparing Market-1501 with existing datasets [20, 10, 44, 22, 19, 4].

eras. This dataset contains 32,668 bboxes of 1,501 identities. Due to the open environment, images of each identity are captured by at most six cameras. We make sure that each annotated identity is captured by at least two cameras, so that cross-camera search can be performed. Overall, our dataset has the following featured properties.

First, while most existing datasets use hand-cropped bboxes, the Market-1501 dataset employs a state-of-the-art detector, *i.e.,* the Deformable Part Model (DPM) [9]. Based on the "perfect" hand-drawn bboxes, current methods do not fully consider the misalignment of pedestrian images, a problem which always exists in DPM based bboxes. As is shown in Fig. 1, misalignment and part missing are common among the detected images.

Second, in addition to the false positive bboxes, we also provide false alarms. We notice that the CUHK03 dataset [20] also uses the DPM detector, but the bboxes in CUHK03 are relatively good ones in terms of detector. In fact, a large number of the detected bboxes would be very "bad". Considering this, for each detected bbox to be annotated, a hand-drawn ground truth bbox is provided (similar to [20]). Different from [20], for the detected and hand-drawn bboxes, the ratio of the overlapping area to the union area is calculated. In our dataset, if the area ratio is larger than 50%, the DPM bbox is marked as "good" (a routine in object detection [9]); if the ratio is smaller than 20%, the DPM bbox is marked as "distractor"; otherwise, the bbox is marked as "junk" [27], meaning that this image is of zero influence to re-id accuracy. Moreover, some obvious false alarm bboxes are also marked as "distractors". In Fig. 1, examples of "good" images are shown in the top two rows, while "distractor" and "junk" images are in the bottom row. These images undergo extensive variations in pose, resolution, *etc.*

Third, each identity may have multiple images under each camera. Therefore, during cross-camera search, there may be multiple queries and multiple ground truths for each identity. This is consistent with practical usage, especially where multiple queries can be fully exploited to obtain more discriminative information about the person of interest. In terms of performance evaluation, for a re-id system, a perfect method should be able to locate all instances of the query identity. In this sense, our dataset provides testbed for methods applied in open systems.

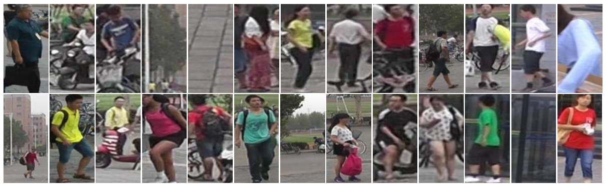


Figure 2. Sample images of the distractor dataset.

3.2. A Distractor Dataset

We emphasize that scale is a vital problem for person re-id studies. Therefore, we further augment the Market-1501 dataset with an additional distractor set. This dataset contains over 500,000 bboxes, consisting of false alarms on the background, as well as pedestrians not belonging to the 1,501 identities. Sample images are shown in Fig. 2. In the experiment, apart from the Market-1501 dataset, we will also report the results on the enlarged Market-1501 + 500K dataset.

A statistics comparison with existing datasets is shown in Table 1. Our dataset contains 1,501 identities, which is lower than CUHK02 [19]. With respect to this point, we plan to release version 2.0 to include more identities. The original dataset contains 32,668 fully annotated bboxes, making it the largest person re-id dataset to date. Since images containing a pedestrian are annotated with a hand-drawn bbox as well as an ID, this dataset can also be used for pedestrian detection. Moreover, our dataset is greatly enlarged by the 500K distractor images, and efficiency/scalability analysis can be reliably done. Compared with other benchmark datasets, Market-1501 is also featured by 6 cameras. In place of a close-system with 2 cameras only, our dataset serves as an ideal benchmark for metric learning methods, so that their generalization capacities can be evaluated for practical usages.

3.3. Evaluation Protocol

Current datasets typically use the Cumulated Matching Characteristics (CMC) curve to evaluate the performance of person re-id algorithms. The CMC curve shows the probability that a query identity appears in different-sized candidate lists. This evaluation measurement is valid only if there is only one ground truth match for a given query (see

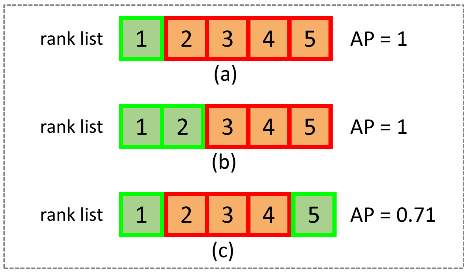


Figure 3. A toy example of the difference between AP and CMC measurements. True matches and false matches are in green and red boxes, respectively. For all three rank lists, the CMC curve remains 1. But AP = 1, 1, and 0.71, *resp.*

Fig. 3(a)). In this case, precision and recall are the same issue. However, if multiple ground truths exist, the CMC curve is biased because "recall" is not considered. For example, CMC curves of Fig. 3(b) and Fig. 3(c) both equal to 1, which fail to provide a fair comparison of the quality between the two rank lists.

For Market-1501 dataset, there are on average 14.8 cross-camera ground truths for each query. Therefore, we use mean average precision (mAP) to evaluate the overall performance. For each query, we calculate the area under the Precision-Recall curve, which is known as average precision (AP). Then, the mean value of APs of all queries, *i.e.,* mAP, is calculated, which considers both precision and recall of an algorithm, thus providing a more comprehensive evaluation. When average precision (AP) is used, rank lists in Fig. 3(b) and Fig. 3(c) are effectively distinguished.

Our dataset is randomly divided into training and testing sets, containing 750 and 751 identities, respectively. During testing, for each identity, we select one query image in each camera. Note that, the selected queries are hand-drawn, instead of DPM-detected as in the gallery. The reason is that in reality, it is very convenient to interactively draw a bbox, which can yield higher recognition accuracy [20]. The search process is performed in a cross-camera mode, *i.e.,* relevant images captured in the same camera as the query are viewed as "junk". In this scenario, an identity has at most 6 queries, and there are 3368 query images in total. Queries of two sample identities are shown in Fig. 4.

4. Our Method

4.1. The Bag-of-Words Model

For three reasons, we adopt the Bag-of-Words (BoW) model. First, it well accommodates local features, which are indicated as effective by previous works [25, 38]. Second, it enables fast global feature matching, instead of exhaustive feature-feature matching [40, 39, 3]. Third, by quantizing similar local descriptors to the same visual word, the BoW model achieves some invariance to illumination, view, *etc*. We describe the individual steps below.

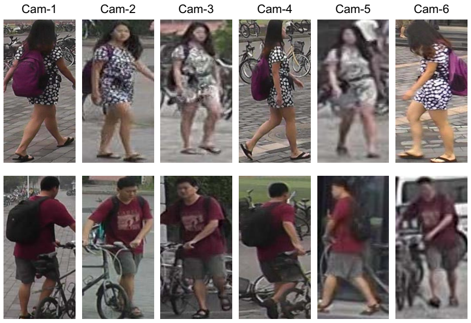


Figure 4. Sample query images. In Market-1501 dataset, queries are hand-drawn bboxes. Each identity has at most 6 queries, one for each camera.

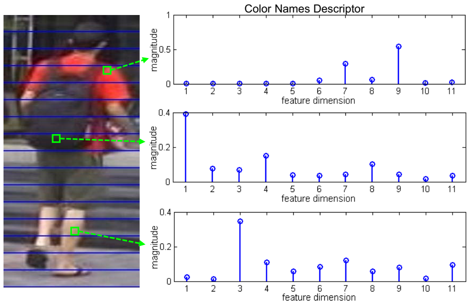


Figure 5. Local feature extraction. We compute the mean CN vector for each 4×4 patch. Local features are quantized, and pooled in a histogram for each horizontal stripe.

Feature Extraction. We employ the Color Names (CN) descriptor [32]. Given a pedestrian image normalized to 128×64 pixels, patches of size 4×4 are densely sampled. The sampling step is 4, so there is no overlapping between patches. For each patch, CN descriptors of all pixels are calculated, and are subsequently ℓ_1 normalized followed by $\sqrt{(\cdot)}$ operator [2]. The mean vector is taken as the descriptor of this patch (see Fig. 5).

Codebook. For Market-1501, we generate a codebook on its training set. For other datasets, the codebook is trained on the independent TUD-Brussels dataset [35]. Standard k-means is used, so codebook size is k.

Quantization. Given a local descriptor, we employ Multiple Assignment (MA) [15] to find its near neighbors under Euclidean distance in the codebook. We set MA = 10, so a feature is represented by the indices of 10 visual words.

TF-IDF. The visual word histogram is weighted by TF-IDF scheme. TF encodes the number of occurrences of a visual word, and IDF is calculated as $\log \frac{N}{n_i}$, where N is the number of images in the gallery, and n_i is the number of images containing visual word i. In this paper, we use the avgIDF [41] variant in place of the standard IDF.

Burstiness. Burstiness refers to the phenomenon where a query feature finds multiple matches in a test image [16]. For CN descriptor, burstiness could be more prevalent due to its low discriminative power compared with SIFT. Therefore, all terms in the histogram are divided by $\sqrt{tf}$.

Negative Evidence. Following [14], we calculate the mean feature vector in the training set. Then, the mean vector is subtracted from all test features. So the zero entries in the feature vector are also taken into account with dot product.

Similarity Function. Given a query image Q and a gallery image G, we calculate the dot product between their feature vectors. Note that, after normalized by l_2-norm, dot product is equivalent to Euclidean distance. In large-scale experiments, Euclidean distance is employed for Approximate Nearest Neighbor algorithm [33].

4.2. Improvements

Weak Geometric Constraints. In person re-id, popular approaches on encoding geometric constraint include "Adjacency Constrained Search" (ACS) [38, 39]. This method is effective in incorporating spatial constraints, but suffers from high computational cost. Inspired by Spatial Pyramid Matching [18], we integrate ACS into the BoW model. As illustrated in Fig. 5, an input image is partitioned into M horizontal stripes. Then, for stripe m, the visual word histogram is represented as $\boldsymbol{d}^m = (d_1^m, d_2^m, ..., d_k^m)^T$, where k is the codebook size. Consequently, the feature vector for the input image is denoted as $\boldsymbol{f} = (\boldsymbol{d}^1, \boldsymbol{d}^2, ..., \boldsymbol{d}^M)^T$, *i.e.* the concatenation of vectors from all stripes. When matching two images, dot product sums up the similarity at all corresponding stripes. Therefore, we avoid the expensive computation on patch distances for each query feature.

Background Suppression. The negative impact of background distraction has been studied extensively [8, 38, 39]. In one solution, Farenzena *et al.* [8] propose to separate the foreground pedestrian from background by segmentation.

Since the process of generating a mask for each image is both time-consuming and unstable, this paper proposes a simple solution by exerting a 2-D Gaussian template on the image. Specifically, the Gaussian function takes on the form of $N(\mu_x, \sigma_x, \mu_y, \sigma_y)$, where μ_x, μ_y are horizontal and vertical Gaussian mean values, and σ_x, σ_y are horizontal and vertical Gaussian standard variances, respectively. We set (μ_x, μ_y) to the image center, and set $(\sigma_x, \sigma_y) = (1, 1)$ for all experiments. This method assumes that the person lies in the center of an image, and is surrounded by background.

Multiple Queries. The usage of multiple queries is shown to yield superior results in image search [1] and re-id [8]. Because intra-class variation is taken into account, the algorithm is more robust to pedestrian variations.

When each identity has multiple query images in a single camera, instead of a multi-multi matching strategy [8], we merge them into a single query for speed consideration. Here, we employ two pooling strategies, *i.e.,* average and max pooling. In average pooling, the feature vectors of multiple queries are pooled into one by averaged sum; in max pooling, the final feature vector takes the maximum value in each dimension from all queries.

Reranking. When viewing person re-id as a ranking problem, a natural idea consists in the usage of reranking algorithms. In this paper, we use a simple reranking method which picks top-T ranked images of the initial rank list as queries to search the gallery again. Specifically, given an initial sorted rank list by query Q, image R_i which is the i^{th} image in the list is used as query. The similarity score of a gallery image G when using R_i as query is denoted as $S(R_i, G)$. We assign a weight $1/(i+1), i = 1, ..., T$ to each top-i ranked query, where T is the number of expanded queries. Then, the final score of the gallery image G to query Q is determined as,

$$\hat{S}(Q, G) = S(Q, G) + \sum_{i=1}^{T} \frac{1}{i+1} S(R_i, G), \qquad (1)$$

where $\hat{S}(Q, G)$ is the weighted sum of similarity scores obtained by the original and expanded queries, and the weight gets smaller as the expanded query is located away from the top. This method departs from [29] in that Eq. 1 employs the similarity value while [29] uses the reverse rank.

5. Experiments

5.1. Datasets

VIPeR dataset [10] is composed of 632 identities, and each has two images captured from two different cameras. All images are normalized to 128×48 pixels. VIPeR is randomly divided into two equal halves, one for training, and the other for testing. Each half contains 316 identities. For each identity, we take an image from one camera as query, and perform cross-camera search.

CUHK03 dataset [20] contains 13,164 DPM bboxes of 1,467 identities. Each identity is observed by two cameras and has 4.8 images in average for each view. Following the protocol in [20], for the test set, we randomly select 100 persons. For each person, all the images are taken as query in turns, and a cross-camera search is performed. The test process is repeated 20 times. We report both the CMC scores and mAP for VIPeR and CUHK03 datasets.

5.2. Important Parameters

Codebook size k. In our experiment, codebooks of various sizes are constructed, and mAP on Market-1501 dataset is presented in Table 2. We set $k = 350$ where the peak value is achieved.

Number of stripes M. Table 3 presents the performance of different numbers of stripes. As the stripe number increases, a finer partition of the pedestrian image leads to a more discriminative representation. So the recognition accuracy increases, but recall may drop for a large M. As a trade-off

Methods	*Market-1501*		*VIPeR*			*CUHK03*	
	r = 1	mAP	r = 1	r = 20	mAP	r = 1	mAP
BoW	9.04	3.26	7.82	39.34	11.44	11.47	11.49
BoW + Geo	21.23	8.46	15.47	51.49	19.85	16.13	15.12
BoW + Geo + Gauss	34.38	14.10	21.74	60.85	26.55	18.89	17.42
BoW + Geo + Gauss + MultiQ_avg	41.21	17.63	-	-	-	22.35	20.48
BoW + Geo + Gauss + MultiQ_max	42.64	18.68	-	-	-	22.95	20.33
BoW + Geo + Gauss + MultiQ_max + Rerank	42.64	19.47	-	-	-	22.95	22.70

Table 5. Results (rank-1, rank-20 matching rate, and mean Average Precision (mAP)) on three datasets by combining different methods, *i.e.*, the BoW model (BoW), Weak Geometric Constraints (Geo), Background Suppression (Gauss), Multiple Queries by average (MultiQ_avg) and max pooling (MultiQ_max), and reranking (Rerank). Note that, here we use the Color Names descriptor for BoW.

k	100	200	350	500
mAP (%)	13.31	14.01	14.10	13.82
r=1 (%)	32.20	34.24	34.38	34.14

Table 2. Impact of codebook size on Market-1501. We report results obtained by "BoW + Geo + Gauss".

M	1	4	8	16	32
mAP (%)	5.23	11.01	13.26	14.10	13.79
r=1 (%)	14.36	27.53	32.50	34.38	34.58

Table 3. Impact of number of horizontal stripes on Market-1501. We report results obtained by "BoW + Geo + Gauss".

T	0	1	2	3	4	5
mAP (%)	18.68	19.47	19.20	19.16	19.10	19.04

Table 4. Impact of number of expanded queries on Market-1501. $T = 0$ corresponds to "BoW + Geo + Gauss + MultiQ_max".

between speed and accuracy, we choose to split an image into 16 stripes in our experiment.

Number of expanded queries T. Table 4 summarizes the results obtained by different numbers of expanded queries. We find that the best performance is achieved when T = 1. When T increases, mAP drops slowly, which validates the robustness to T. The performance of reranking highly depends on the quality of the initial list, and a larger T would introduce more noise. In the following, we set T to 1.

5.3. Evaluation

BoW model and its improvements. We present results obtained by BoW, geometric constraints (Geo), Gaussian mask (Gauss), multiple queries (MultiQ), and reranking (Rerank) in Table 5 and Fig. 6.

First, the baseline BoW vector produces a relatively low accuracy: rank-1 accuracy = 9.04%, 10.56%, and 5.35% on Market-1501, VIPeR, and CUHK03 datasets, respectively.

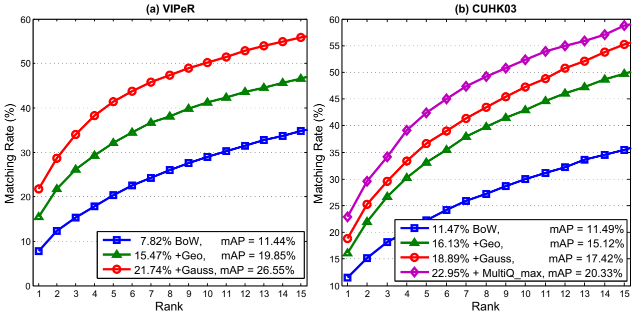


Figure 6. Performance of different method combinations on VIPeR and CUHK03 datasets.

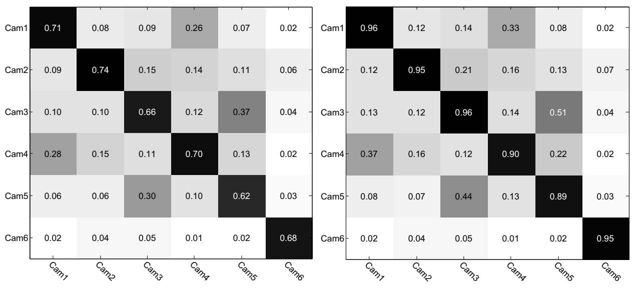
Second, when we integrate geometric constraints by stripe matching, we observe consistent improvement in accuracy. On Market-1501 dataset, for example, mAP increases from 3.26% to 8.46% (+5.20%), and an even larger improvement can be seen from rank-1 accuracy, from 9.04% to 21.23% (+12.19%).

Third, it is clear that the Gaussian mask works well on all three datasets. We observe +5.64% in mAP on Market-1501 dataset. Therefore, the prior that pedestrian is roughly located in the center of the image is statistically sound.

Then, we test multiple queries on CUHK03 and Market-1501 datasets, where each query identity has multiple bboxes. Results suggest that the usage of multiple queries further improves recognition accuracy. The improvement is more prominent on Market-1501 dataset, where the query images take on more diverse appearance (see Fig. 4). Moreover, multi-query by max pooling is slightly superior to average pooling, probably because max pooling gives more weights to the rare but salient features and improves recall.

Finally, we observe from Table 4 and Table 5 that reranking generates higher mAP. Nevertheless, one recurrent problem with reranking is the sensitivity to the quality of initial rank list. On Market-1501 and CUHK03 datasets, since a majority of queries DO NOT have a top-1 match, the improvement in mAP is relatively small.

Results between camera pairs. To further understand the Market-1501 dataset, we provide the re-id results between all camera pairs in Fig. 7. We use the "BoW+Geo+Gauss"



(a) Pairwise mAP (b) Pairwise rank-1 accuracy

Figure 7. Re-id performance between camera pairs on Market-1501: (a) mAP and (b) rank-1 accuracy. Cameras on the vertical and horizontal axis are probe and gallery, respectively. The cross-camera average mAP and average rank-1 accuracy are 10.51% and 13.72%, respectively.

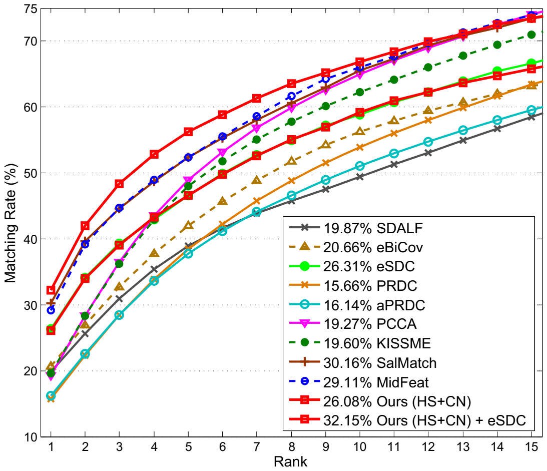


Figure 8. Comparison with the state-of-the-arts on VIPeR. We combine HS and CN features, and the eSDC method.

representation. It is easy to tell that re-id within the same camera yields the highest accuracy. On the other hand, as expected, performance among different camera pairs varies a lot. For camera pairs 1-4 and 3-5, the BoW descriptor generates relatively good performance, mainly because the two camera pairs share more overlap. Moreover, camera 6 is a 720×576 SD camera, and captures distinct background with other HD cameras, so re-id accuracy between camera 6 and others are quite low. Similarly low result can be observed between camera pairs 5-1 and 5-2. We also compute the cross-camera average mAP and average rank-1 accuracy: 10.51% and 13.72%, respectively. We weight mAPs between different camera pairs according to their number of queries, and do not calculate the results on the diagonals. Compared with the "BoW+Geo+Gauss" line in Table 5, both measurements are much lower than pooling images in all cameras as gallery. This indicates that re-id between camera pairs is very challenging on our dataset.

Methods	*CU03*
	r = 1
SDALF [8]	4.87
ITML [6]	5.14
LMNN [34]	6.25
eSDC [39]	7.68
KISSME [17]	11.70
FPNN [20]	19.89
BoW	18.89
BoW (MultiQ)	22.95
BoW (+HS)	**24.33**

Methods	*Market-1501*	
	r = 1	mAP
gBiCov [26]	8.28	2.23
HistLBP [36]	9.62	2.72
LOMO [21]	26.07	7.75
BoW	34.38	14.10
+LMNN [34]	34.00	15.66
+ITML [6]	38.21	17.05
+KISSME [17]	39.61	17.73
BoW (MultiQ)	42.64	19.47
BoW (+HS)	**47.25**	**21.88**

Table 6. Method comparison on CUHK03 and Market-1501.

Stage	SDALF [8]	SDC [8]	Ours
Feat. Extraction (s)	2.92	0.76	**0.62**
Search (s)	2644.80	437.97	**0.98**

Table 7. Average query time of different steps on Market-1501 dataset. For fair comparison, Matlab implementation is used.

Comparison with the state-of-the-arts. We compare our results with the state-of-the-art methods in Fig. 8 and Table 6. On VIPeR (Fig. 8), our approach is superior to two unsupervised methods, *i.e.,* eSDC [39], SDALF [8]. Specifically, we achieve a rank-1 identification rate of 26.08% when two features are used, *i.e.,* Color Names (CN) and HS Histogram (HS). When eSDC [39] is further integrated, the matching rate increases to 32.15%.

On CUHK03, our method without multiple-query significantly outperforms almost all presented approaches. Compared with FPNN [20] which builds a deep learning architecture, our accuracy is slightly lower by 1.00%. But when multiple-query and HS feature are integrated, rank-1 matching rate exceeds [20] by +4.44% on CUHK03 dataset.

On Market-1501, we compare with state-of-the-art descriptors including HistLBP [36], gBiCov [26], and LOMO [21]. The proposed BoW descriptor clearly outperforms these competing methods. We then apply various metric learning methods [34, 6, 17] on BoW (after PCA to 100-dim). Instead of pairwise training (can be expensive under large camera networks [31]), we take all positive and negative pairs in 6 cameras as training samples. We observe that metric learning brings decent improvement.

Some sample results on Market-1501 dataset are provided in Fig. 9. Apart from the mAP increase with the method evolution, another finding which should be noticed is that the distractors detected by DPM on complex background or body parts severely affect re-identification accuracy. Previous works typically focus on "good" bounding boxes with person only, and rarely study the detector errors.

Large-scale experiments. First, on Market-1501, we compare our method with SDALF [8] and SDC [39] in two aspects, *i.e.,* feature extraction and search time. We use the Matlab implementation by the authors and for fair comparison, run our algorithm in Matlab too. Evaluation is per-

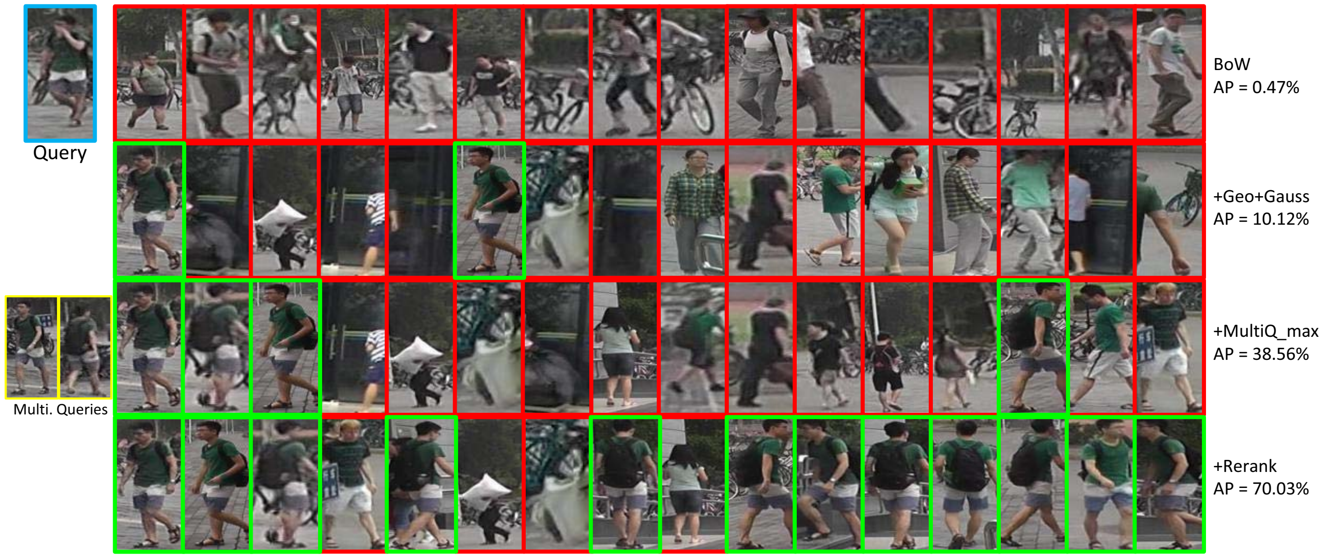
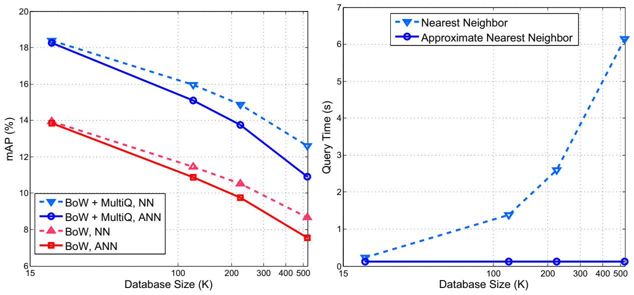


Figure 9. Sample results on Market-1501 dataset. Four rows correspond to four configurations, *i.e.,* "BoW", "BoW + Geo + Gauss", "BoW + Geo + Gauss + MultiQ", and "BoW + Geo + Gauss + MultiQ + Rerank". The original query is in blue bbox, and the added multiple queries are in yellow. Images with the same identity as the query is in green box, otherwise red.



(a) mAP vs database size (b) Query time vs database size

Figure 10. mAP (a) and query time (b) in Market-1501+500K dataset. Dashed lines are obtained by exact NN search, while solid lines represent ANN search.

formed on a server with 2.59 GHz CPU and 256 GB memory, and efficiency results are shown in Table 7. We report the total timing by HS (we extract a 20-dim HS histogram and generate another BoW vector for fusion with CN) and CN features for our method. Compared with SDC, we achieve an efficiency gain of over two orders of magnitude. For S-DALF, three features are involved, *i.e.,* MSCR, wHSV, and RHSP. The feature extraction time is 0.09s, 0.03s, 2.79s, respectively; the search time is 2643.94s, 0.66s, and 0.20s, respectively. Therefore, our method is faster than SDALF by three orders of magnitude.

Then, we experiment on the Market-1501+500K dataset. Images in the 500K dataset are treated as outliers. For efficiency, we use the Approximate Nearest Neighbor (ANN) algorithm proposed in [33]. During index construction, we build 4 kd-trees, and store 50 neighbors for each datum in the knn-graph. The number of neighbors returned is 1000 for both NN and ANN (so mAP of NN is slightly lower than reported in Table 5).

Re-id performance on the large-scale dataset is presented in Fig. 10. As the database gets larger, accuracy drops. On Market-1501+500K dataset, when ANN is used, an mAP of 10.92% is achieved for "BoW + MultiQ_max". Compared with result on the original dataset, a relative drop of 69.7% is observed. As a result, database size has a significantly negative effect on performance, which has been rarely discussed in literature. Moreover, although ANN marginally decreases re-id accuracy, the benefit it brought obvious. With ANN, query time is 127.5ms on the 500K dataset, a 50x speedup compared with the NN case.

6. Conclusion

This paper firstly introduces a large-scale re-id dataset, Market-1501 (+500k), which reaches closer to realistic settings. Then, a BoW descriptor is proposed in the attempt to bridge the gap between person re-id and image search. The new dataset will enable research possibilities in multiple directions, *e.g.,* deep learning, large-scale metric learning, multiple query techniques, search reranking, *etc.* In the future, current test data will be treated as validation set, and new test IDs will be annotated and presented in a coming person re-id challenge.

This work was supported by the National High Technology Research and Development Program of China (863 program) under Grant No. 2012AA011004 and the National Science and Technology Support Program under Grant No. 2013BAK02B04. This work was supported in part to Dr. Qi Tian by ARO grants W911NF-15-1-0290, W911NF-12-1-0057, Faculty Research Awards by NEC Laboratories of America, and National Science Foundation of China (NSFC) 61429201.

References

[1] R. Arandjelovic and A. Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, 2012. 2, 5

[2] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012. 2, 4

[3] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, 2015. 1, 2, 4

[4] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, volume 2, page 6, 2011. 1, 3

[5] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury. Consistent re-identification in a camera network. In *ECCV*. 2014. 3

[6] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216. ACM, 2007. 7

[7] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*. 2011. 2

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367. IEEE, 2010. 2, 5, 7

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *T-PAMI*, 32(9):1627–1645, 2010. 3

[10] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, volume 3, 2007. 1, 3, 5

[11] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 2008. 1, 2

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. Springer, 2011. 2

[13] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*. 2012. 1

[14] H. Jégou and O. Chum. Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening. In *ECCV*. 2012. 2, 5

[15] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317. Springer, 2008. 2, 4

[16] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, pages 1169–1176, 2009. 2, 4

[17] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012. 7

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 5

[19] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601, 2013. 1, 3

[20] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 2, 3, 4, 5, 7

[21] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 7

[22] S. Liao, Z. Mo, Y. Hu, and S. Z. Li. Open-set person re-identification. *arXiv preprint arXiv:1408.0872*, 2014. 1, 3

[23] C. Liu, C. C. Loy, S. Gong, and G. Wang. Pop: Person re-identification post-rank optimisation. In *ICCV*, 2013. 2

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[25] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV Workshops and Demonstrations*, pages 413–422. Springer, 2012. 2, 4

[26] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 32(6):379–390, 2014. 7

[27] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, pages 1–8, 2007. 3

[28] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 1, page 5, 2010. 2

[29] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *CVPR*, 2012. 2, 5

[30] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015. 2

[31] C. Su, f. Yang, S. Zhang, Q. Tian, L. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV*, 2015. 7

[32] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *TIP*, 18(7):1512–1523, 2009. 4

[33] J. Wang and S. Li. Query-driven iterated neighborhood graph search for large scale indexing. In *ACM MM*, 2012. 5, 8

[34] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2005. 7

[35] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, pages 794–801. IEEE, 2009. 4

[36] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*. 2014. 7

[37] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *ACM MM*, 2009. 2

[38] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 1, 2, 4, 5

[39] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 1, 2, 4, 5, 7

[40] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 1, 2, 4

[41] L. Zheng, S. Wang, Z. Liu, and Q. Tian. Lp-norm idf for large scale image search. In *CVPR*, 2013. 2, 4

[42] L. Zheng, S. Wang, Z. Liu, and Q. Tian. Packing and padding: Coupled multi-index for accurate image retrieval. In *CVPR*, 2014. 2

[43] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query-adaptive late fusion for image search and person re-identification. In *CVPR*, 2015. 2

[44] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, volume 2, page 6, 2009. 1, 3

[45] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian. Spatial coding for large scale partial-duplicate web image search. In *ACM MM*, 2010. 2