TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/11 pp415–424
DOI: 10.26599/TST.2019.9010031
Volume 25, Number 3, June 2020

Person Re-Identification with Effectively Designed Parts

Yali Zhao, Yali Li*, and Shengjin Wang*

Abstract: Person re-IDentification (re-ID) is an important research topic in the computer vision community, with significance for a range of applications. Pedestrians are well-structured objects that can be partitioned, although detection errors cause slightly misaligned bounding boxes, which lead to mismatches. In this paper, we study the person re-identification performance of using variously designed pedestrian parts instead of the horizontal partitioning routine typically applied in previous hand-crafted part works, and thereby obtain more effective feature descriptors. Specifically, we benchmark the accuracy of individual part matching with discriminatively trained Convolutional Neural Network (CNN) descriptors on the Market-1501 dataset. We also investigate the complementarity among different parts using combination and ablation studies, and provide novel insights into this issue. Compared with the state-of-the-art, our method yields a competitive accuracy rate when the best part combination is used on two large-scale datasets (Market-1501 and CUHK03) and one small-scale dataset (VIPeR).

Key words: person re-IDentification (re-ID); Convolutional Neural Network (CNN); part model

1 Introduction

This paper focuses on person re-IDentification (re-ID) on a large scale, which is currently a most challenging computer vision issue. This task features two inputs, i.e., a query image and a database of pre-defined bounding boxes, and the objective is to output a ranking list sorted by the degree of similarity between the database and the query. Person re-ID plays a crucial role in a range of applications, including multi-camera pedestrian tracking[1], multi-person association[2], and group behavior analysis[3]. As indicated by previous works, the challenge for the person re-ID research community consists of misalignment, occlusion, varient illuminations, and varient scales, etc. Figure 1 shows some image samples of these four situations, which is the main problem to be solved in this paper. Due to its value for research and application, person re-ID has been receiving increasing attention from computer vision researchers.

Person re-ID is essentially a matching problem, in which the feature of the query image is matched against the gallery features. In instance retrieval[4], for example, objects may undergo severe variations in scale, position, affine transformations, etc. In this scenario, local invariant descriptors are used to match two images, in the hope that the corresponding local patches can be well aligned. Similarly, in face recognition, face landmark detection and alignment have also been extensively studied, and well-aligned faces can improve the recognition accuracy[5]. In person re-ID, it is

Fig. 1 Examples of the unsatisfactory persons. They correspond to the case of misalignment, occlusion, varient illuminations, and varient scales.

• Yali Zhao, Yali Li, and Shengjin Wang are with Tsinghua University, Beijing 100084, China. E-mail: zhaoluckydog@163.com; liyali@ocrserv.ee.tsinghua.edu.cn; wgsgj@tsinghua.edu.cn.

∗ To whom correspondence should be addressed.

Manuscript received: 2018-12-28; revised: 2019-06-22; accepted: 2019-07-17

@ The author(s) 2020. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

typically assumed that the bounding box provides roughly aligned pedestrians, especially in a statistical view. Despite the introduction of datasets[6,7] that are produced by pedestrian detectors instead of being hand-drawn, it is shown in Ref. [8] that horizontal stripes help improve the matching precision. In fact, while it is true that detection errors are introduced, pedestrians are aligned fairly well, from a global point of view. Given this consideration, it is acceptable for body parts roughly divided by regular grids to be introduced in this paper.

In previous hand-crafted part-based works, horizontal stripes have been most commonly used[6,7,9,10], based on an underlying assumption that pedestrians are more or less aligned on the vertical axis. Traditional methods[7,9,10] explicitly extract stripe features and concatenate them to form the final feature vector. There are also some discriminatively deep-learned part-based works, but these leverage external cues, such as assistance from the latest progress in landmark detection or human pose estimation[11–13], which means that they rely on external human pose estimation datasets and sophisticated pose estimators. From the point of view of efficiency and simple real-world implementation, the present study focuses on hand-crafted part models. To the best of our knowledge, this is the first work to provide insights into the combination of different categories of parts.

Compared with traditional part matching strategies, which view parts as horizontal stripes across a bounding box, this paper extends classical part partitioning to include various definitions of part regions. Specifically, apart from horizontal parts, evaluation is conducted on (1) vertical parts, by considering cross-part visual information; (2) occluded parts, by aiming to mimic the occlusion artifacts encountered in crowded scenes; and (3) partial parts, by viewing them as sub-regions of the traditional horizontal parts. In total, 18 parts are evaluated individually or in combination. Our motivation is that extensive variances in scale, illumination, and body misalignment and occlusions are likely in real-world applications. In our experiment, we find that vertical, occluded, and partial parts are effective in providing auxiliary cues to the global feature, while horizontal parts are less effective. When the most complementary parts are combined, the resulting part ensemble achieves an accuracy rate that is competitive with state-of-the-art methods. The contributions of this paper are summarized as follows.

- We generate a number of different body parts by grid partitioning. These parts are different in their position, discriminative power, and in how they complement each other.
- We run in-depth experiments to produce novel insights into part complementarity through combination and ablation studies.
- We achieve competitive re-ID accuracy over three benchmarks by combining features from the complementary parts.

The remainder of this paper is organized as follows. After a brief review of related works in Section 2, the part definitions and combination strategy are presented in Section 3. We summarize and discuss the experimental results in Section 4, and draw conclusions and suggestions for further work in Sections 5 and 6, respectively.

2 Related Work

In this section, we briefly review three closely related aspects found in the literature: effective descriptor and distance metric learning combinations, part-based re-ID, and deep learning re-ID.

Effective descriptor and distance metric learning combinations. Conventionally, previous studies on person re-identification have focused on two major components: feature extraction and distance metric learning. For feature extraction, various low level descriptors of color and texture have been proposed, such as the symmetry-driven accumulation of local features[14], bag-of-words color representation[7] (which takes advantage of image retrieval techniques), the salient color names-based descriptor[15], and LOcal Maximal Occurrence (LOMO) representation[10]. Generally speaking, these hand-crafted features mainly focus on colors, and recent state-of-the-art methods introduce features learned in a Convolutional Neural Network (CNN)[6,16]. For example, Zheng et al.[17] proposed to learn an ID-discriminative embedding through a classification network, and this paper generally adopts their method to extract a discriminative feature embedding in the person space. For learning effective distance metrics for similarity measurement, the pairwise constrained component analysis proposed in Ref. [18] projects high-dimensional vectors onto a low-dimensional space, such that the projected points satisfy pre-defined constraints. Other methods include learning the

discriminative null space[19] and cross-view quadratic discriminant analysis[10].

Hand-crafted part methods. A noteworthy aspect of some of the above-mentioned methods is that geometric constraints are injected, explicitly or implicitly. In fact, the relatively rigid configuration of pedestrians is an ideal testbed for part-based recognition algorithms. Accordingly, spatial constraints have been widely adopted for person image representation and matching. For example, Farenzena et al.[14] considered the symmetric and asymmetric properties of the human body, and extracted local features from each part. Methods like Refs. [7] and [9] adopt an Spatial Pyramid Matching (SPM)-like[20] representation that separates the image into several horizontal stripes, and uses the unsupervised bag-of-words model to represent each horizontal region. In DeepReID[6], the horizontal stripes are implied in the image matching process during training, so that the learned similarity is also based on a part-wise matching strategy. Improving on Ref. [6], Ahmed et al.[16] proposed to compute the cross-input neighborhood difference features with subtraction. This method compares features from one input image to features in similar locations of the image to be matched. Similarly, part information is also integrated in Refs. [10, 21, 22], in which a bounding box is partitioned onto horizontal stripes, and Adjacency Constrained Search (ACS) is adopted for exhaustive spatial matching. Together, these previous works suggest that part integration is an effective strategy in improving re-ID accuracy, which serves as the main motivation for the present study.

Discriminatively learned part methods. Human pose estimation and landmark detection are particularly widely used techniques to discriminatively learn part features. Several recent works in re-ID use these tools for pedestrian partition and report an encouraging level of improvement. Yao et al.[23] clustered the coordinates of max activations on feature maps to locate several regions of interest. References [22, 24, 25] embed the attention mechanism[26] into the network, allowing the model to make its own decision on where to focus. Reference [27] proposed a LSTM-based attention model that can dynamically produce part attention feature by a recurrent way for localizing the discriminative local regions of the person image.

3 Proposed Approach

3.1 Feature design

We partition an image into a collection of different parts, and accordingly make use of different feature extraction methods based on the characteristics of different datasets. Specifically, for the Market-1501[7] and CUHK03[28] datasets, both the state-of-the-art hand-crafted feature extraction method (LOMO)[10] and deep neural network feature extraction are evaluated. Since the VIPeR dataset is too small to train a neural network, for VIPeR we only use the LOMO descriptor coupled with XQDA as a distance metric[10].

To extract the LOMO features, the color histogram and Scale Invariant Local Ternary Pattern (SILTP) features[10] are first extracted in 10×10 blocks with an overlap of 5 pixels. The feature representation of a row is obtained by maximizing the local occurrence of each pattern at the same horizontal location.

The large scale of the Market-1501 and CUHK03 datasets and the rich variety of the samples make it possible to take advantage of deep neural networks. We make use of two neural networks: the classic AlexNet[29] and the most discriminative ResNet-50[30]. Given an input image or part, the outputs of the "fc7" layer for AlexNet and the "fc50" layer for ResNet-50 are taken as feature vectors.

3.2 Parts

In this paper, we design 18 different parts (as shown in Fig. 2) from an input pedestrian image that varies extensively in visual appearance, including scale, position, and shape. Our experimental results show that the re-ID performance differs from part to part (as shown in Fig. 3). During the training process, we separately train a CNN model for each part, so that

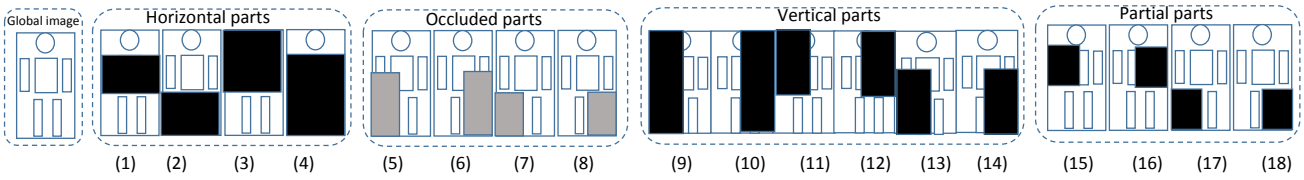

Fig. 2 A global image and 18 newly designed Parts (1) – (18). Each dashed box represents the global image or one of the four part categories, i.e., horizontal, vertical, occluded, and partial parts. The black boxes represent the corresponding parts, and the grey boxes are the occluded regions, represented by the matrix with elements of 128.

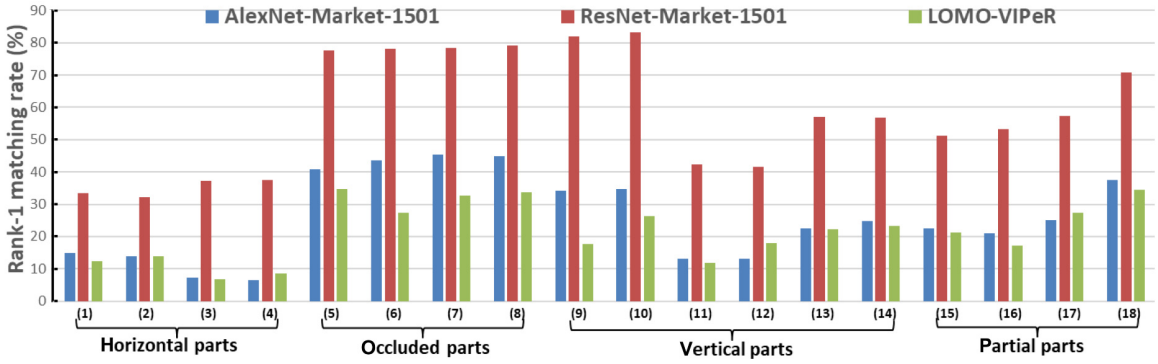

Fig. 3 Person re-identification performance of individual parts. We use CNN feature (extracted by AlexNet and ResNet-50 respectively) on Market-1501 and LOMO+XQDA on VIPeR. The index for each part is consistent with that in Fig. 2.

altogether we have 18 part models plus one model for the global image.

Roughly, as shown in Fig. 2, we divide these parts into four categories: horizontal parts, occluded parts, vertical parts, and partial parts. We define four types of horizontal parts, namely torso, legs, head+torso, and torso+legs, as Parts (1) – (4) respectively. The intuition behind this design is that the torso and legs possess the most discriminative cues, as demonstrated in Ref. [8]. To illustrate the effectiveness of the head in re-ID accuracy, Part (3) includes the head while Part (4) excludes it. We will show in the experiments that Parts (1) and (2) yield the best performance when used alone, which is consistent with Ref. [8], in which, horizontal parts are commonly used for person re-identification and work effectively when the samples are well-aligned horizontally. In the present work, however, we show that the combination of horizontal parts does not provide any benefit when testing on the Market-1501 dataset, where pedestrians are not well aligned.

The occluded parts are labelled as Parts (5) – (8). In Fig. 2, the gray boxes denote the occluded regions. Since the torso and the legs are the most important parts, the design of the occluded parts is mainly focused on these two regions. In practical usage, bags and some other carry-ons may occlude the body of the pedestrian, so the occluded part aims to mimic these situations. To the best of our knowledge, previous work on re-ID does not explicitly consider the occlusion problem, which is in fact a critical influencing factor, and this work provides the first attempt to resolve this issue. In Part (5), we occlude the left half of the torso and the leg, while Part (6) is the mirror image of Part (5). In Parts (7) and (8), the left and right legs are occluded, respectively. Figure 3 shows that the occluded model yields decent results even when some parts of the image are not informative.

The vertical parts are labelled as Parts (9) – (14). The vertical parts may be useful when the samples suffer from noise on the horizontal axis, due to illumination changes, occlusions, etc. Again, we focus on the torso and the legs for the vertical parts. For Parts (9) and (10), the part is cropped to the left and right half of the pedestrian image, respectively. Parts (11) and (12) are defined as the left and right upper body, respectively, and Parts (13) and (14) are defined as the left and right lower body, respectively. Figure 3 shows that Parts (11) and (12) produce poorer results compared with the other four vertical parts. Compared with Parts (13) and (14), Parts (11) and (12) lack the information from the upper body, so this result leads us to speculate that the upper body is critical for pedestrian matching.

The four partial parts are labelled as Parts (15) – (18). The first two partial parts denote the left and right upper body, while the last two denote the left and right legs, respectively. These small parts are subsets of the horizontal parts. There are two reasons why we include these subsets: first, they specify whether the symmetrical nature of the body parts is discriminative towards an effective re-ID system; and second, the partial parts may contain discriminative cues, such as bags, that can facilitate the re-ID training process in yielding better descriptors.

This paper investigates the performance of each part category and studies their complementarity. Specifically, the impact of each part category will be accessed by cascading the extracted feature vector of the global image with that of the corresponding part category as the final representation of an input image. We further explore the complementarity of the above four categories with the help of ablation experiments. Specifically, the baseline consists in cascading the feature vectors of the global image and all 18 parts; we

then eliminate each of the features of the part categories in turn to gain insights into the configuration of these parts.

3.3 Similarity learning

We utilize the extra part information by cascading the features of the whole image and that of the corresponding parts. The similarity of a probe image and the gallery images is learnt by XQDA[10], which learns a discriminant metric by learning a discriminant low-dimensional subspace with cross-view quadratic discriminant analysis. Thus if some dimensions of the feature vector are not effective for discrimination, the elements in the corresponding columns or rows of the feature vector also tend to be less effective.

4 Experiments

In this section, we present a comprehensive evaluation of the proposed algorithm in different settings and compare the experimental results with several state-of-the-art person re-identification algorithms.

4.1 Datasets and evaluation protocol

The **VIPeR** dataset contains two views of 632 pedestrians with only one image per person in each view. It is a challenging dataset in the person re-identification arena owing to the huge variance and discrepancies. All images are normalized to 128×48. The testing protocol is to split the dataset into halves, giving 316 for training and 316 for testing.

The **Market-1501** dataset is the largest of the available person re-ID datasets, making it appropriate for deep learning. It contains 32 643 images of 1501 identities, with each identity is captured by between 2 and 6 cameras. The dataset is randomly divided into the test set of 751 identities and the training set of 750 identities. The selected 3368 queries are hand-drawn, rather than detected using the Deformable Part Model (DPM) as in the gallery. We evaluate our method under both the single-query and multi-query settings. For the multi-query evaluation, we take the mean descriptor of a person from a single camera as the query descriptor.

The **CUHK03** dataset contains 13 164 images of 1467 identities, captured by six surveillance cameras with each person appearing in only two views. Apart from the manually labeled pedestrian bounding boxes, this dataset also provides the samples detected with a pedestrian detector called DPM, which causes some misalignments and missing body parts for a more realistic setting. We use this last setting in this paper. CUHK03 originally adopts 20 random train/test splits; because this is time-consuming for deep learning we adopt the new training/testing protocol proposed in Ref. [17] instead.

Evaluation protocol. We adopt the widely-used Cumulative Match Curve (CMC)[31] technique for quantitative evaluation. As such, the top-n matching rate indicates the expectation of finding any one of the correct matched images. As in Ref. [7], the person re-identification is treated mainly as a retrieval problem, and the mean Average Precision (mAP) is also used in our experiments to evaluate the performance. Average Precision (AP) is calculated by the area under the precision-recall curve, and mAP is further calculated as the mean score of the AP of all the query images, so it essentially reflects both the precision and recall of the person re-ID process.

4.2 Evaluation

For the two large-scale datasets, we only provide figures for the results of our method on Market-1501 with AlexNet and ResNet-50 for brevity. As shown in the corresponding tables, the results are consistent with CUHK03, which is expected given that these two datasets are similar in many ways. For example, they both use DPM for detection, resulting in many misaligned samples. For VIPeR, the result with the LOMO feature is analyzed.

4.2.1 Performance of individual parts

As discussed in Section 3.2, the motivations for the designation of the 18 parts vary, and we speculate that they will exhibit distinct re-identification accuracies. We study this by observing how the performance changes with each single part, with the experimental results shown in Fig. 3. From these results, we observe that the occluded Parts (5) – (8), two vertical Parts (9) and (10), and one partial Part (18), have a superior accuracy compared with the others. In fact, these results accord with our intuition that a larger part will typically yield higher accuracy since more information is preserved. We will show that the combination of the three-part categories does improve the overall re-identification accuracy.

4.2.2 Performance gain from part combinations

We combine the feature vector of the global image with that of the four part categories separately. The performance of the part combinations is shown in Fig. 4.

The results suggest that adding the horizontal parts benefits VIPeR but compromises the accuracy on

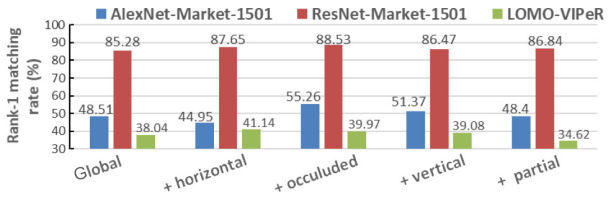

Fig. 4 Impact of adding each part category to the global image. Four combinations are presented, i.e., "global image + horizontal parts", "global image + occluded parts", "global image + vertical parts", and "global image + partial parts".

Market-1501 and CUHK03, both with the simple AlexNet and the complex ResNet-50. This may be due to the fact that the pedestrians in Market-1501 and CUHK03 are detected with DPM, resulting in many misaligned samples, while VIPeR is a hand-labeled dataset in which the pedestrians are well aligned. This generates noisy samples in Market-1501 and CUHK03, but not in VIPeR, so we attribute the lower effectiveness of the fixed horizontal partition to the severe misalignments of body parts. As shown in Fig. 1, the corresponding part regions may contain totally distinguished parts.

The occluded parts bring benefits in all of the three experiments. As each part of the feature is associated with one local region, once some region is occluded, the features of other regions still work. Such a mechanism implies that our method is potentially robust to occlusion. This is straightforward since all three datasets are common in regard to occlusions. Figure 5a presents a visual example in which the occluded parts help refine the re-ID accuracy when combined with the global image.

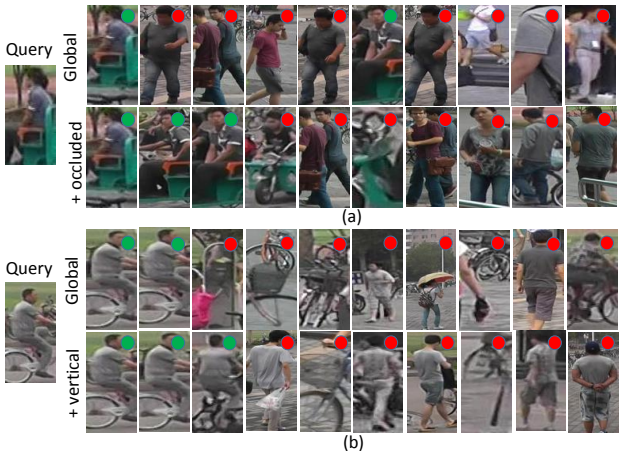

Fig. 5 An example of the effectiveness of combining, (a) occluded parts and (b) vertical parts with the global image on Market-1501 with ResNet-50. Green discs represent the true matches of the query, and red ones represent the false matches.

The vertical parts also improve re-ID performance in all three cases. This can be attributed to the extra cross-part information brought by such partitioning. For example, the vertical Part (13) in Fig. 2 better models the cross-part structure of a human torso and legs. Figure 5b is a visual example in which the vertical parts boost the performance. Apart from an increase in AP, the false matches are much more similar to the query after adding the extra vertical parts. In this case, the additional parts better model the structure of human parts even though there is significant deformation for this identity. The vertical parts also perform better when there is some illumination change in the horizontal direction, because when the left or right half of a person suffers from the illumination, the other half can still model the global structure. Since there is more illumination changes in the horizontal direction than in the vertical direction for person re-identification tasks, this kind of part always refines the overall performance.

The impact of partial parts varies with the CNN model and dataset. On the one hand, for the VIPeR dataset, and for Market-1501 and CUHK03 when using AlexNet, the small size of these parts means less information and thus leads to deteriorated performance. On the other hand, the depth of ResNet-50 leads to smaller receptive fields when compared with AlexNet, making it better at matching the small parts and leading to improved accuracy. One interesting observation is that, due to their complementarity, partial parts with a lower performance in Fig. 3 also help when combined with the global image.

4.2.3 Part complementarity analysis

In this section, ablation experiments are conducted to provide insight into part complementarity. The results are shown in Fig. 6.

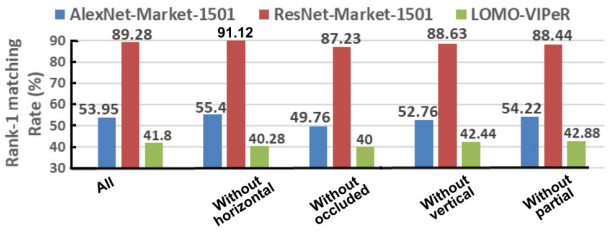

Fig. 6 Ablation study on Market-1501 with AlexNet and ResNet-50 and on VIPeR with LOMO+XQDA. "All" denotes the global image combined with all the four categories of parts defined in Fig. 2. We find that (1) the horizontal parts benefit VIPeR instead of Market-1501, (2) the occluded and vertical parts are necessary for both datasets, and (3) the partial parts vary with them.

Comparing all of the results reported in Fig. 6, re-identification performance improves for Market-1501 and CUHK03 but not for VIPeR when eliminating the horizontal parts. This arises from the fact that VIPeR is well aligned, unlike Market-1501 and CUHK03. This is consistent with the observation in Section 4.2.2, and further illustrates the effectiveness of adding other part categories. Moreover, we see from Fig. 6 that re-ID accuracy drops without the occluded parts. Occlusion parts prove their effectiveness both in the combination and ablation studies, so the conclusion about the occluded parts given in Section 4.2.2 also holds for the complementarity of parts. A visual example of these two findings is shown in Fig. 7; in which there are three true matches in the gallery for the queried identity. For the vertical and partial parts, the results of the ablation experiments are not consistent and further investigation will be needed in future work.

4.2.4 Comparison with the state-of-the-arts

Based on the above analysis, we evaluate our method with the best performing part configuration: vertical + partial + occluded parts + global image. For VIPeR, we adopt the LOMO+XQDA method as shown in Table 1. For Market-1501 and CUHK03, we use ResNet-50 as the feature extractor for its superior performance compared with AlexNet. We present the comparison results in Tables 2 and 3.

For the VIPeR dataset, we compare among hand-crafted feature methods. The results in Table 1 demonstrate the competitive performance of the proposed method compared with the single models at rank-1, -5, -10, and -20. In this table, the compared methods are divided into two groups. The models in the top section use only one kind of feature and one model, similar to our method. The models in the bottom section use either multiple models or multiple features, thus this group generally outperforms the top one. As can be seen, the rank-1 matching rate of out method is 42.88%, which outperforms the next best method by 0.5%.

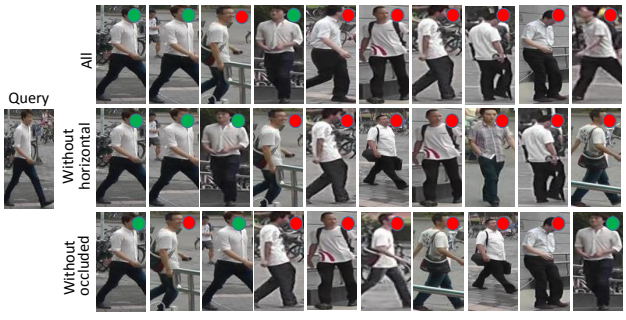

Fig. 7 An example of the part ablation study. "All" presents the global image together combined with all the four part categories. The following two rows are results of cutting off the horizontal or occluded parts from "all", respectively. Green and red discs represent true and false matches, respectively.

Table 1 Top matching rank on VIPeR. (%)

Method	Rank-1	Rank-5	Rank-10	Rank-20
LADF[32]	29.34	61.04	75.98	88.10
MtMCML[33]	28.83	59.34	75.82	88.51
Mid-level[22]	29.11	52.34	65.95	79.87
SalMatch[34]	30.16	52.31	65.54	79.15
SCNCD[33]	37.80	68.50	81.20	90.40
Deep feature[35]	40.50	60.80	70.40	84.40
LOMO+XQDA[10]	40.00	67.40	80.51	91.08
LOMO+KISSME[36]	34.81	60.44	77.22	86.71
LOMO+LDNs[19]	42.28	71.46	82.94	92.06
mFilter+LADF[22]	43.39	73.04	84.87	93.70
Metric ensembles[37]	45.90	77.50	88.90	95.80
EDFP[38]	51.06	81.01	91.39	96.90
LDNS[19]	51.17	82.09	90.51	95.92
Ours	**42.88**	**72.22**	82.82	**92.78**

Table 2 Comparison of the proposed method with state-of-the-art methods on Market-1501. (%)

Method	Single query		Multi query	
	Rank-1	mAP	Rank-1	mAP
BoW+kissme[7]	36.40	19.70	44.40	28.30
XQDA+LOMO[10]	43.79	22.22	54.13	28.41
kLFDA+LOMO[39]	51.37	24.43	52.67	27.36
IDE(baseline)[27]	78.20	59.70	85.30	68.50
LDNS+LOMO[19]	55.43	29.87	67.96	41.89
LDNS+fusion[19]	61.02	35.68	71.56	46.03
SVDNet[40]	73.10	54.30	82.30	62.10
PAN[41]	75.33	55.20	82.80	63.40
MultiRegion[42]	57.10	35.20	66.40	41.20
MultiScale[43]	79.10	65.60	88.90	73.10
MultiLoss[44]	76.30	55.20	83.90	64.40
Ours	**82.80**	**65.89**	**91.12**	**74.10**

Table 3 Comparison with prior art on CUHK03. (%)

Method	Rank-1	mAP
IDE(baseline)[27]	43.82	38.94
BoW+kissme[7]	6.45	6.42
XQDA+LOMO[10]	12.81	11.57
PAN[41]	36.36	34.12
SVDNet[40]	41.56	37.38
MultiScale[43]	40.76	37.19
Ours	**55.88**	**50.89**

For the Market-1501 and CUHK03 datasets, we apply XQDA to the learned CNN feature to further improve performance. We use three types of methods for comparison: hand-crafted methods, deep learning methods with global features, and deep learning methods with part features. As can be seen in Tables 2 and 3, our method outperforms the state-of-the-art for both single and multiple queries on all evaluation metrics on Market-1501. Notably, we achieve a rank-1 accuracy of 91.12% and mAP of 74.1% on this dataset. On CUHK03, our method achieves a rank-1 accuracy of 55.8% and mAP of 50.89%. Both exceed their corresponding baselines by a large margin, which indicates that imposing part information improves the performance of person re-identification on large-scale datasets.

4.2.5 Robustness of the method

In order to give greater insight of the robustness of the proposed method, especially from the viewpoint of different images, we use the random erasing method[1] to imitate training samples of low quality. Random erasing is a new data augmentation method and is parameter learning free and easy to implement. For each input, either a global image or the newly designed parts, we randomly choose a rectangular region of an arbitrary size, and assign random values to the pixels within that region. After this operation, the final rank-1 matching rate is 42.23% on VIPeR, 82.36% on Market-1501, and 55.12% on CUHK03, which is only a slight decline compared with the performance using the original training samples (42.88%, 82.80%, and 55.88% on the three datasets, respectively). We attribute this result to the complementarity of the global image and the newly designed parts, which contain much detailed and discriminative information. This result indicates the robustness of the proposed method for images of different qualities.

4.2.6 Computational complexity of the method

The proposed model can be implemented with any CNN-based recognition backbones. Without loss of generality, we evaluate the computational complexity of the algorithms on ResNet-50. The evaluation is performed on a server with GTX 1080 GPU (8 GB memory), 2.60 GHz CPU, and 128 GB memory. The results show a feature extraction time of 31.3 ms, a search time of 0.82 ms, and a sorting time of 0.63 ms on Market-1501, which suggests that the proposed method is highly efficient.

5 Conclusion

In this paper, we propose the use of effectively designed part models for person re-identification. While previous hand-crafted part works typically use horizontal parts, we design a total of 18 different parts, categorized into horizontal, occluded, vertical, and partial. We benchmark the performance of the individual parts and study the complementarity among them through combination and ablation studies. We find from our experiments that various part types exert a different influence on the datasets but, in general, considering different parts in re-identification brings a consistent improvement over baseline approaches. We demonstrate that our method achieves competitive accuracy rates when compared with the state-of-the-art on the VIPeR, CUHK03, and Market-1501 datasets.

6 Further Work

We suggest further work on part-level features for pedestrian image description, since it offers fine-grained information. Rather than using external cues, such as estimation or directly locating parts, we suggest directly learning discriminative part features for person retrieval, as it is a much more general and robust approach. Since the parts of a human body have a regular spatial distribution and are sequentially aligned, a promising research direction is to incorporate Long Short-Term Memory (LSTM) into the part sequence.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 61771288 and 61701277) and the State Key Development Program of the 13th Five-Year Plan (No. 2017YFC0821601).

References

[1] T. D' Orazio and G. Cicirelli, People re-identification and tracking from multiple cameras: A review, in *Proc. of 2012 19th IEEE International Conference on Image Processing*, Phoenix, AZ, USA, 2012, pp. 1601–1604.

[2] B. Wang, G. Wang, K. Luk Chan, and L. Wang, Tracklet association with online target-specific metric learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1234–1241.

[3] W. S. Zheng, S. Gong, and T. Xiang, Group association: Assisting re-identification by visual context, in *Proc. of European Conference on Computer Vision*, London, UK, 2014, pp. 183–201.

[4] L. Zheng, Y. Yang, and Q. Tian, SIFT meets CNN: A

decade survey of instance retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1224–1244, 2017.

[5] X. Jin and X. Tan, Face alignment in-the-wild: A survey, *Journal of Computer Vision and Image Understanding*, vol. 162, pp. 1–22, 2017.

[6] W. Li, R. Zhao, T. Xiao, and X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 152–159.

[7] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, Scalable person re-identification: A benchmark, in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1116–1124.

[8] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 1335–1344.

[9] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, Salient color names for person re-identification, in *Proc. of European Conference on Computer Vision*, Zurich, Switzerland, 2014, pp. 536–551.

[10] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 2197–2206.

[11] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, Convolutional pose machines, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4724–4732.

[12] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, Deepercut: A deeper, stronger, and faster multi-person pose estimation model, in *Proc. of European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 34–50.

[13] A. Newell, K. Yang, and J. Deng, Stacked hourglass networks for human pose estimation, in *Proc. of European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 483–499.

[14] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in *Proc. of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010, pp. 2360–2367.

[15] J. Van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, Learning color names for real-world applications, *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1512–1523, 2009.

[16] E. Ahmed, M. Jones, and T. K. Marks, An improved deep learning architecture for person re-identification，in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 3908–3916.

[17] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, Person re-identification in the wild, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 1367–1376.

[18] M. Hirzer, P. M. Roth, M. Kostinger, and H. Bischof, Relaxed pairwise learned metric for person re-identification, in *Proc. of European Conference on Computer Vision*, Heidelberg, Germany, 2012, pp. 780–793.

[19] L. Zhang, T. Xiang, and S. Gong, Learning a discriminative null space for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 1239–1248.

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

[21] R. Zhao, W. Ouyang, and X. Wang, Unsupervised salience learning for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013, pp. 3586–3593.

[22] R. Zhao, W. Ouyang, and X. Wang, Learning mid-level filters for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 144–151.

[23] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, Deep representation learning with part loss for person re-identification, *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2860–2871, 2019.

[24] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, and X. Wang, Hydraplus-net: Attentive deep features for pedestrian analysis, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 350–359.

[25] L. Zhao, X. Li, Y. Zhuang, and J. Wang, Deeply-learned part-aligned representations for person re-identification, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 3219–3228.

[26] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, and C. L. Zitnick, From captions to visual concepts and back, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1473–1482.

[27] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, End-to-end comparative attention networks for person re-identification, *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[28] Z. Zheng, L. Zheng, and Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 3754–3762.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Proc. of Advances in Neural Information Processing Systems*, Portland, OR, USA, 2012, pp. 1097–1105.

[30] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.

[31] 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 (PETS)*, Las Vegas, NV, USA, 2007, pp. 1–7.

[32] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, Learning locally-adaptive decision functions for person verification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013, pp. 3610–3617.

[33] L. Ma, X. Yang, and D. Tao, Person re-identification over camera networks using multi-task distance metric learning, *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3656–3670, 2014.

[34] R. Zhao, W. Ouyang, and X. Wang, Person re-identification by salience matching, in *Proceedings of the IEEE International Conference on Computer Vision*, Sydney, Australia, 2013, pp. 2528–2535.

[35] S. Ding, L. Lin, G. Wang, and H. Chao, Deep feature learning with relative distance comparison for person re-identification, *IEEE Transactions on Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[36] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, Large scale metric learning from equivalence constraints, in *Proc. of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 2012, pp. 2288–2295.

[37] B. McFee and G. R.Lanckriet, Metric learning to rank, in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010, pp. 775–782.

[38] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You, and W. S. Zheng, An enhanced deep feature representation for person re-identification, in *Proc. of 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, 2016, pp. 1–8.

[39] F. Xiong, M. Gou, O. Camps, and M. Sznaier, Person re-identification using kernel-based metric learning methods, in *Proc. of European Conference on Computer Vision*, Zurich, Switzerland, 2014, pp. 1–16.

[40] Y. Sun, L. Zheng, W. Deng, and S. Wang, Svdnet for pedestrian retrieval, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 3800–3808.

[41] Z. Zheng, L. Zheng, and Y. Yang, Pedestrian alignment network for large-scale person re-identification, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 2860–2871, 2018.

[42] E. Ustinova, Y. Ganin, and V. Lempitsky, Multi-region bilinear convolutional neural networks for person re-identification, in *Proc. of 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, Italy, 2017, pp. 1–6.

[43] Y. Chen, X. Zhu, and S. Gong, Person re-identification by deep learning multi-scale representations, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 2590–2600.

[44] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, Random erasing data augmentation，arXiv preprint arXiv: 1708. 04896, 2017.

Yali Zhao received the BEng degree from Sichuan University, China, in 2014. She is a PhD student in the Department of Electronic Engineering in Tsinghua University. Her research interests include image retrieval, classification, and person re-identification.

Yali Li received the BEng degree with Excellent Graduates Award from Nanjing University, China, in 2007 and the PhD degree from Tsinghua University, Beijing, China, in 2013. Currently she is a research assistant in Department of Electronic Engineering, Tsinghua University. Her research interests include image processing, pattern recognition, computer vision, and video analysis, etc.

Shengjin Wang received the BEng degree from Tsinghua University, China, and the PhD degree from the Tokyo Institute of Technology, Tokyo, Japan, in 1985 and 1997, respectively. From 1997 to 2003, he was a member of the research staff with the Internet System Research Laboratories, NEC Corporation, Japan. Since 2003, he has been a professor with the Department of Electronic Engineering, Tsinghua University, where he is currently the director of the Research Institute of Image and Graphics. His current research interests include image processing, computer vision, video surveillance, and pattern recognition.