



Learning part-alignment feature for person re-identification with spatial-temporal-based re-ranking method

Zhongyi Li¹ · Yi Jin¹  · Yidong Li¹ · Congyan Lang¹ · Songhe Feng¹ · Tao Wang¹

Received: 28 February 2019 / Revised: 15 August 2019 / Accepted: 22 September 2019 /

Published online: 13 November 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Person re-identification is to identify a target person in different cameras with non-overlapping views. It is a challenging task due to various viewpoints of persons, diversified illuminations, and complicated environments. In addition, body parts are usually misaligned because of the less precise bounding boxes, which play a significant role in person re-identification, so it is crucial to make them aligned for better performance. In this paper, we propose a network to learn powerful features combining global features and local-alignment features for person re-identification. For each body part, instead of fixed horizontal partition, a key points detection network is adopted to locate body parts that contain more precise and distinctive information. Besides, a novel re-ranking approach is proposed to refine the rough initial rank list by exploiting the spatial-temporal information. Unlike most existing re-ranking based methods fine-tuning the rough initial rank list only by k-nearest neighbors and their k-reverse-nearest neighbors, our method exploits spatial-temporal information which can be easily stored in the name of images, so it can be implemented in any baseline to improve the performance. Experiments on the GRID, Market-1501, and DukeMTMC-reID are conducted to prove the effectiveness of our method.

Keywords Person re-identification · Part alignment · Re-ranking

1 Introduction

As one of the most challenging problems in the computer vision field, person re-identification aims to judge whether two images from different non-overlapping cameras contain the same pedestrian. It has potential practical value in video surveillance by saving

This article belongs to the Topical Collection: *Computational Social Science as the Ultimate Web Intelligence*

Guest Editors: Xiaohui Tao, Juan D. Velasquez, Jiming Liu, and Ning Zhong

✉ Yi Jin
yjjin@bjtu.edu.cn

Extended author information available on the last page of the article.

a large number of human resources, so it received many significant efforts in the past few years. However, because of many variances between the different cameras, such as changes in person's appearance, body pose, camera angle, occlusion and illumination conditions, identifying the same pedestrian across different camera views has not been solved yet.

In order to address the problem, many efforts are made on person re-identification. With the rapid development of deep learning, some researchers start employing CNN to learn global features [2, 3, 6]. However, it is not discriminative enough just by global features to identify the same pedestrian across different camera views, because a) global features contain much useless background information, and b) the lack of useful local information.

Therefore, in recent years, some works pay more attention to local details. To focus on exploring local features, some of them cut images into fixed rigid parts, which horizontal stripes are the most common, and then learn the discriminative local features [7, 33, 35, 40]. These simple partition methods assume that the position and pose of each person are similar in the bounding box, which is usually impossible. Therefore, it is important to design a method that can alleviate the negative effect caused by body parts misalignment. Some works deal with misalignment by extracting features from patches [16, 48], stripes [1], or pose-guided region of interest (RoI) [31, 49, 53]. However, these methods still contain much useless information caused by background. Recently, a few methods combining part maps and feature maps to form part-aligned representations are proposed [32, 45]. They usually design two sub-networks, one to learn part maps and another to learn appearance maps. Then they combine these two kinds of maps together as the final appearance features. Some body-part misalignment examples are shown in Figure 1. However, it is also difficult to judge whether an image belongs to a certain person only by local features, so some researchers combine global and local features together to get a stronger feature. In [42], they first extract global feature and local feature by their Harmonious Attention CNN model, and combine these two features to get better performance. Li et al. [18] use the advantages of jointly learning local and global features to find their correlation.

In this paper, we propose a network to learn powerful features combining global features and part-alignment local features. More specifically, first a pose estimation model is employed to detect the key points, then we process the person images into several part-based images with the help of these key points to reduce the influence of background for the local features. Then these part-based images are fed into the deep network to extract local features, which will be concatenated to achieve part alignment. Finally, global features and local features are concatenated to form the final features.

Re-ranking is an effective way to improve the final performance. It can be implemented in any baseline because it doesn't need any additional training samples. Some previous

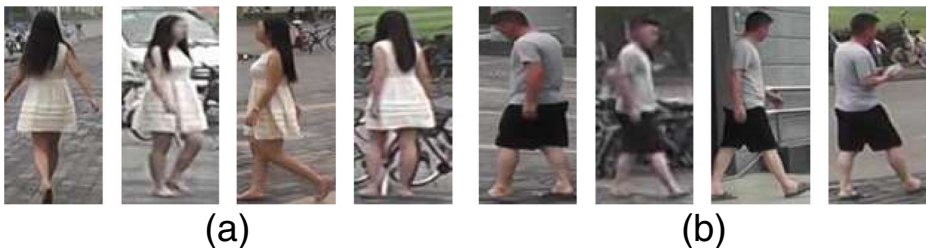


Figure 1 Some examples of misalignment caused by different poses/viewpoints in different cameras. We can see the corresponding body parts are usually not spatially aligned, which increases the difficulty of person re-identification

works conduct re-ranking step by the similarity between probe and top-ranked gallery images (e.g. k-nearest neighbors). They assume that gallery image in the top rank is likely to be a true match. However, the accuracy based on this hypothesis depends largely on the initial rank list. Once the top-k ranks of the initial rank list are all false, it cannot improve the final accuracy or even lead a worse result. Some re-ranking methods make efforts to alleviate the negative effect of the re-ranking methods aforementioned. Leng et al. [14] exploit a bidirectional ranking method to simultaneously compute both content and context similarities between bidirectional ranking lists. Garcia et al. [9] try to find the visual ambiguities in a ranking list and remove them. Zhong et al. [55] introduce a concept called k-reciprocal nearest neighbors to alleviate the bad effect of false matches. However, all existing re-ranking methods only exploit images' visual clues to refine the initial rank list, ignoring the potential spatial-temporal information. Our hypothesis is that there are different spatial-temporal models between different camera pairs, which means the probabilities of transfer time between different cameras are varied. Based on this hypothesis, we propose a spatial-temporal information based re-ranking method for person re-identification. To summarize, the contributions of this paper are as below:

- We propose a new network to extract discriminative local features for person re-identification from a series of parts. By introducing a pose estimation model to detect key points, the local features are aligned automatically, which results in more distinctive visual cues. Then, combining global and local features can make the final features more discriminative and robust.
- We propose a novel re-ranking method by exploiting the spatial-temporal information. The proposed approach is full-automatic without any human feedback. All it needs is spatial-temporal information which can be easily stored in the name of images, so it can be implemented to any baseline to improve the performance.
- Extensive experiments on the GRID, Market-1501, and DukeMTMC-reID demonstrate that our approach is of effectiveness and efficiency.

The rest of our paper is organized as follows. The related work will be discussed in Section 2. In Section 3, we will talk about the detail of our proposed method. Then, Section 4 is about the experimental results. Finally, Section 5 comes to a conclusion and future work.

2 Related work

Traditional methods for person re-identification mainly focus on two aspects, (1) to exact robust and discriminative hand-crafted features [4, 19, 24, 24, 26, 27] or deep learning features [7, 15, 34, 36, 38] and (2) to learn a more robust metric [5, 11, 13, 19, 20, 28]. In addition, some works bring in more thoughts such as attributes [30], transfer learning [25], spatial-temporal information [23] and re-ranking, etc..

Regular spatial-partition based methods This kind of method usually divides every person image into several parts by fixed partition such as grid cell [2, 15] or horizontal stripe [7, 35]. Their assumption is that the position and pose of each person are similar in the bounding box. However, it is very hard to be that way in the real world. Therefore, if the image can not satisfy this assumption, the result will be unsatisfied.

Body part-aligned based methods Body part misalignment is unavoidable in person re-identification, and it becomes one of the most crucial problems to be solved. In the past few

years, body parts and key points detection have been brought into person re-identification to solve the misalignment problem. And because of the popularity of deep learning, some works [17, 41, 49] try to use deep learning techniques to achieve the goal. Some of them separate the images of body parts which are detected by pose estimator to extract a more discriminative local feature. And some of them bring in the use of attention maps [32, 45]. These methods usually design two sub-networks, one to learn part maps and another to learn appearance maps. Then they combine these two kinds of maps together as the final appearance features. Our method combines global features and part-alignment local features to form the final powerful features.

Re-ranking methods Re-ranking methods also can be divided into two categories: 1) re-ranking with human feedback and 2) re-ranking without human feedback, depending on whether the human feedback is needed during the re-ranking process. In [21], the end user needs to select one strong negative sample or additional weak negative samples as feedback to refine the initial rank list during the test stage. In [37], both similar and dissimilar samples are chosen by the end user. In [39], they propose a new incremental model which becomes stronger with human feedback. However, a re-ranking method that requires human feedback is not good enough because the human feedback could be expensive. Thus some researches pay more attention to automatic re-ranking approach without human feedback. In [9], a re-ranking method is proposed to remove the visual ambiguities in a ranking by analyzing the content and context information in the initial ranking. In [46], the author believe that if two images belong to the same person, these two images should have similar appearance not only in global view, but also in local view. In [47], two initial rank lists are obtained by different baseline methods, and then the final rank list is computed through the aggregation of similarity and dissimilarity information in the two initial rank lists. In contrast to the above re-ranking methods, our proposed method explores the importance of the potential spatial-temporal information. We believe that there are some inherent behavior patterns between different camera pairs, that is to say, it is feasible to filter the negative samples which are of high visual similarity in the top-k list with the help of the spatial-temporal models.

3 Proposed method

In this work, we aim to learn discriminative feature representations for persons and improve the performance of person re-identification. Our proposed network is shown in Figure 2. For deep network, its goal is to extract discriminative global features and part-alignment local features at the same time. For re-ranking stage, we design a novel spatial-temporal-based re-ranking method to refine the initial ranking list by exploiting spatial-temporal information.

In this section, we introduce our method as follows: the proposed network for powerful features in Section 3.1, and re-ranking method to achieve a better result in Section 3.2.

3.1 Proposed network

The proposed network consists of two components, the global feature learning component and the local feature component. In the first part, our goal is to extract discriminative global features from the whole person images. And in the second part, a 3-branch deep convolutional network is designed to obtain part-alignment local features.

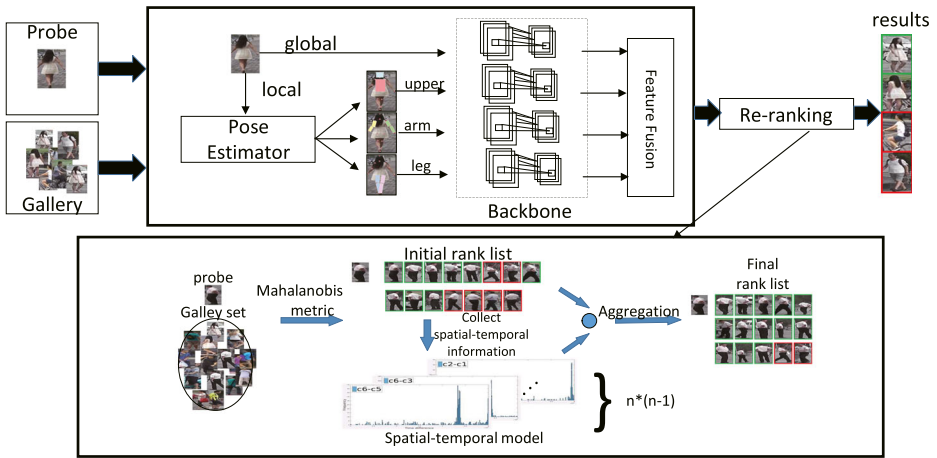


Figure 2 The framework of our proposed method

Global features learning The global feature learning component is trained with a multi-task network. One is identification task and the other is verification task. As shown in Figure 3, the global part contains two branches and they share the same parameters during training. A pair of images is fed into the global feature learning component as input. After the parameters-shared network, each image’s representation can be extracted and fed into the final fully connected layer to classify the ID of image for identification task. As for verification task, it aims to judge whether two images belong to the same ID, and it’s accomplished through a sub-module. The sub-module contains a square layer and a fully connected layer, and outputs a two-dimensional vector. Finally, after the trained model obtained, the global features f_{global} can be extracted through it.

Part-alignment local features learning As mentioned above, body part alignment is a key problem to improve the result. So in this paper, we proposed a 3-branch local network to generate local features, which can make body part aligned. As shown in Figure 4, to learn the part-alignment local features, we first employ a pose estimation model [43] to detect the key points, then we process the person images into several part-based images with the help

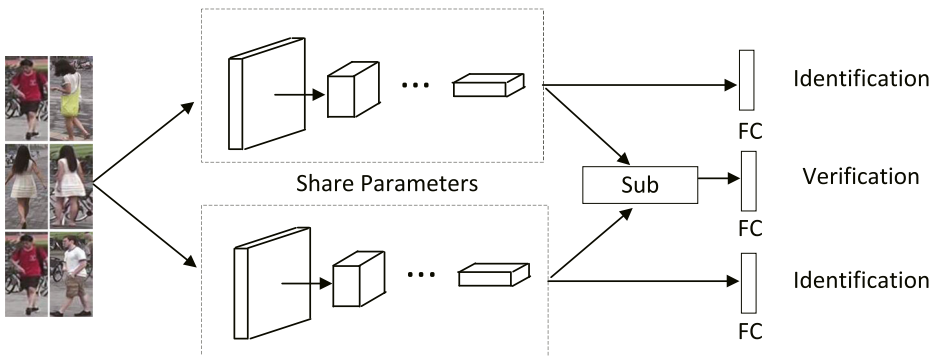


Figure 3 The structure of Global Features Learning Network

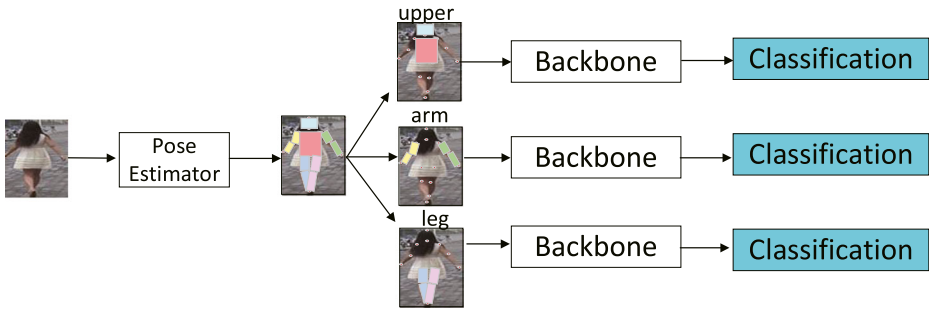


Figure 4 The structure of Local Features Learning Network

of these key points. Then these part-based images are fed into deep network to extract local features. Finally, all these local features can be concatenated and the body part alignment can be achieved automatically. In this part, three branches do not share the same parameters. During the training of each branch in local features learning network, each branch has an independent classifier based on only part of image as input, which can enforce the network to extract discriminative details of each part.

Person representation In the test stage, the final robust features are formed by concatenating global features and all local features, which can be written as below:

$$f = [\alpha \times f_{global}, \beta \times f_{upper}, \gamma \times f_{arm}, \sigma \times f_{leg}] \tag{1}$$

where $\alpha, \beta, \gamma, \sigma$ are weight parameters. f_{global} represents the global features and $f_{upper}, f_{arm}, f_{leg}$ are 3-branch local features, respectively.

Training In the global phase, the loss function can be formulated as:

$$L_g = \beta \ell_{id} + \gamma \ell_{veri} \tag{2}$$

where β, γ are weights. ℓ_{id} is the loss of identification task and ℓ_{veri} is the loss of verification task.

In the local phase, the loss function can be formulated as:

$$L_p = \lambda_1 \ell_{id}^u + \lambda_2 \ell_{id}^a + \lambda_3 \ell_{id}^l \tag{3}$$

where $\lambda_1, \lambda_2, \lambda_3$ are weights. $\ell_{id}^u, \ell_{id}^a, \ell_{id}^l$ are loss from upper branch, arm branch and leg branch, respectively.

The Softmax loss is used as the classification loss and cross-entropy loss is used as the loss of verification task.

3.2 Proposed re-ranking method

Most existing re-ranking methods only exploit images’ visual clues to refine the initial rank list, ignoring the potential spatial-temporal information. Our hypothesis is that there are different spatial-temporal models between different camera pairs, which means the probabilities of transfer time between different cameras are varied. Based on this hypothesis, we propose a spatial-temporal information based re-ranking method for person re-identification. Specifically, given probe images, an initial rank list for each probe image is obtained through baseline. Then, some reliable gallery samples are selected for each

probe image according to the initial rank list, and these gallery samples are treated as true matches. After we get the samples, spatial-temporal models between different camera pairs can be learned from them. Finally, the final distance is calculated as the combination of the original distance and the probability according to the camera pairs. The re-ranking list can be obtained through the final distance.

Suppose there are 3 sets, a probe set P , a gallery set G and a train set T , and the amount of these 3 sets is N_p, N_g, N_t respectively. Given a probe image p_i ($i = 1, 2, \dots, N_p$) and a gallery image g_j ($j = 1, 2, \dots, N_g$), the initial distance between p_i and g_j can be computed through Euclidean distance or Mahalanobis distance,

$$d(p_i, g_j) = (x_{p_i} - x_{g_j})^T M (x_{p_i} - x_{g_j}) \tag{4}$$

where x_{p_i} and x_{g_j} are the features of probe p_i and gallery g_j respectively, and M is a positive semidefinite matrix. Then we can obtain the initial rank list $R(p_i, G) = \{g_1, g_2, \dots, g_{N_g}\}$ by sorting the original distances calculated between probe p_i and each gallery g_j in the gallery set G , where $d(p_i, g_j) < d(p_i, g_{j+1})$.

The spatial-temporal information can be collected after the initial rank list is obtained. Given an initial rank list $R(p_i, G) = \{g_1, g_2, \dots, g_{N_g}\}$, the top- k samples of the rank list, i.e., g_1, g_2, \dots, g_k , are selected and treated as true matches of probe p_i . Then, we assume a positive direction, for example, assuming the direction from the camera with a small number to the camera with a large number is positive direction. So, the spatial-temporal information is computed as follows:

$$\begin{cases} st_{c_{p_i}, c_{g_j}} = f_j - f_i, c_{p_i} < c_{g_j}, f_j > f_i \\ st_{c_{g_j}, c_{p_i}} = f_j - f_i, c_{p_i} < c_{g_j}, f_j < f_i \\ st_{c_{g_j}, c_{p_i}} = f_i - f_j, c_{p_i} > c_{g_j}, f_j < f_i \\ st_{c_{p_i}, c_{g_j}} = f_i - f_j, c_{p_i} > c_{g_j}, f_j > f_i \end{cases} \tag{5}$$

where c_{p_i}, c_{g_j} represent which camera p_i and g_j ($j = 1, 2, \dots, k$) come from respectively, and f_i, f_j represent the frame of p_i and g_j respectively. After the spatial-temporal information of every sample pairs is collected, it is sorted out into different spatial-temporal information sets $ST_{c_{p_i}, c_{g_j}}$ according to the identifiers of cameras. Cameras contain spatial information and frames contain temporal information. Finally, we sort every camera pairs' temporal information, which means there are $n * (n - 1)$ spatial-temporal models, where n is the number of cameras. Specifically, assuming that the amount of camera is 6, there are 30 time information sets except 6 time information sets consisting of the same camera. After the $N_p * k$ time differences are obtained, they are sorted into the 30 time information sets according to the camera pairs where the current image pairs come from.

Besides, different similarities between different sample pairs should make different contribution to the final result. The sample pairs with high similarities are more likely to be the true matches. Therefore, the similarities are also supposed to be collected as well when obtaining spatial-temporal information. In this paper, similarities are collected as follows:

$$\begin{cases} S_{c_{p_i}, c_{g_j}} = \sum s_{p_i, g_j}, c_{p_i} < c_{g_j}, f_j > f_i \\ S_{c_{g_j}, c_{p_i}} = \sum s_{p_i, g_j}, c_{p_i} < c_{g_j}, f_j < f_i \\ S_{c_{g_j}, c_{p_i}} = \sum s_{p_i, g_j}, c_{p_i} > c_{g_j}, f_j < f_i \\ S_{c_{p_i}, c_{g_j}} = \sum s_{p_i, g_j}, c_{p_i} > c_{g_j}, f_j > f_i \end{cases} \tag{6}$$

Table 1 The details of datasets

Datasets	ID	Box	Box/ID	Cam
QMUL GRID [22]	250	500	2	8
Market-1501 [50]	1,501	32,643	19.9	6
DukeMTMC-reID [54]	1,404	36,411	25.9	8

where $S_{c_{p_i}, c_{g_j}}$ is the sum of all samples pairs' similarities for each camera pairs. and s_{p_i, g_j} is the similarity between p_i and g_j . For example, if there are 6 cameras, 30 sums should be computed for 30 different camera pairs except 6 pairs consist of same camera.

After calculating the spatial-temporal information of each probe image and its top-k gallery images, the potential behavior patterns between different camera pairs are obtained.

In the test stage, spatial-temporal probabilities whether a probe image and each gallery image in the gallery set G belong to the same person should be calculated through the above behavior patterns. In order to make the probability more confident, an interval is set based on the time difference between probe image and gallery image. Thus the probability is computed as follows:

$$probability(p_i, g_j) = s_{p_i, g_j} * Num_{p_i, g_j} / S_{c_{p_i}, c_{g_j}} \quad (7)$$

$$Num_{p_i, g_j} = Index(t + \Delta) - Index(t - \Delta) \quad (8)$$

where t means the time difference between probe image p_i and gallery image g_j , Δ is an interval, $Index(x)$ means the index of x in the $ST_{c_{p_i}, c_{g_j}}$, $S_{c_{p_i}, c_{g_j}}$ means the whole weight between camera c_{p_i} and camera c_{g_j} and Num_{p_i, g_j} means the number of spatial-temporal information within the interval $[t - \Delta, t + \Delta]$.

Considering the spatial-temporal information should be treated as auxiliary information and it can be complementary to the appearance representations, we jointly aggregate the initial distance and spatial-temporal probability to revise the initial ranking list, and the final distance d_{final} is defined as

$$d_{final}(p_i, g_j) = d_{initial}(p_i, g_j) / probability(p_i, g_j) \quad (9)$$

Besides, not all pairs are supposed to calculate their spatial-temporal information, because this could lead to some negative samples that are dissimilar in visual representation jumping into the top-k list due to their high spatial-temporal probabilities. Thus it's necessary to set a distance limitation to judge whether spatial-temporal information should be considered in the final distance, and we set the spatial-temporal probability to a minimum value if it is ignored. Finally, the final rank list can be obtained by sorting the final distance.

4 Experiments

4.1 Datasets and settings

Datasets Although spatial-temporal information can be easily recorded in the name of each image, there are only 3 datasets containing spatial-temporal information. They are QMUL GRID [22], Market-1501 [50], and DukeMTMC-reID [54] respectively. Thus experiments are conducted on these 3 datasets. The overview of datasets is shown in Table 1.

QMUL GRID Loy et al.[22] is the first re-ID benchmark dataset which contains spatial-temporal information in the image name. There are 500 pedestrian images containing 250

persons. Each person has a pair of images from different cameras. The number of cameras is 8 and the background is a busy underground station.

Market-1501 Zheng et al. [50] is a large-scale image-based dataset containing spatial-temporal information in the image name. There are 32,668 pedestrian images containing 1,501 persons from 6 different cameras. These images are captured by Deformable Part Model (DPM) [8]. In the experiments, the standard training and evaluation protocols in [50] where 751 identities are used for training and the remaining 750 for testing is used.

DukeMTMC-reID Zheng et al. [54] is a subset of the DukeMTMC for image-based re-identification, whose format is the same as Market-1501 dataset. There are 36,411 pedestrian images containing 1,404 persons from 8 different cameras. In the experiments, the standard training and evaluation protocols where 702 identities are used for training and the remaining 702 for testing is used.

Evaluation metrics For small-scale dataset QMUL GRID, we only use Cumulative Match Characteristic (CMC) curve to evaluate the performance of Re-ID methods, because there is only one gallery image for each identity. For large-scale dataset Market-1501 and DukeMTMC-reID, we use both Cumulative Match Characteristic (CMC) curve and mean average precision (mAP) to evaluate the performance.

Feature representations Except the features extracted from our proposed network, we also employ some other features to demonstrate the effect of our proposed re-ranking method. The Local Maximal Occurrence (LOMO) [19] features are used to represent the person visual appearance for all 3 datasets. It is hand-crafted feature and robust to view changes and illumination variations. In addition, we also employ the ID-discriminative Embedding (IDE) feature proposed in [52] for Market-1501, whose model is based on ResNet-50 [10]. Finally, we employ a baseline deep feature proposed in [51] for Market-1501 and DukeMTMC-reID.

Implementation details We implement the proposed person re-identification model based on a improved Resnet50 framework. The stochastic gradient descent is used to optimize the networks. The initial learning rate, weight decay, and the momentum are set to 2×10^{-4} , 5×10^{-4} , and 0.9, respectively.

4.2 Experiments on QMUL GRID

We first conduct our experiments on the small-scale dataset QMUL GRID. Considering there are only 250 pairs of images, we just use the hand-crafted LOMO feature and use XQDA as the metric method. We set the interval Δ to 10, the max_interval to 30 and m to 50, where the max_interval is the threshold of frame difference and m is the number of distances used to compute a similarity threshold. Then we calculated the spatial-temporal probability threshold as follows: First, we calculated the mean value of all query images top-50 initial distances and set spatial-temporal probability threshold to $1e-7$. Then we calculated the mean value of all spatial-temporal probabilities that greater than $1e-7$ as the spatial-temporal probability threshold. The result on QMUL GRID is shown in Table 2.

It clearly shows that there is a huge margin between baseline and our method. Our method gains 54.08%, 61.12% and 45.52% improvement in rank-1, rank-5 and rank-20 respectively for LOMO + XQDA. In addition, no matter what metric we use, our method always can improve rank-1 significantly. The reason why there is a so huge margin could

Table 2 The results of our approach on the QMUL GRID

Method	Rank 1	Rank 5	Rank 20
LOMO + Euc	15.20	25.92	36.40
LOMO + Euc + OL-MANS [56]	20.88	–	45.12
LOMO + Euc + our re-rank	52.24	66.00	71.12
LOMO + XQDA	16.56	33.84	52.40
LOMO + XQDA + OL-MANS	29.20	–	50.96
LOMO + XQDA + our re-rank	70.64	94.96	97.92

be that the behavior patterns between different camera pairs is rather simple, thus the spatial-temporal information plays a dominant role.

4.3 Experiments on market-1501

We follow the standard training and evaluation protocols in [50] where 751 identities are used for training and the remaining 750 for testing. And we employ LOMO [19], IDE [52] and baseline deep feature [51] to evaluate the performance. We set the interval to 700, the max_interval to 40000 and m to 50. The spatial-temporal probability threshold is calculated the same way in the QMUL GRID part. The result is shown in Table 3.

As we can see from Table 3, our re-ranking method consistently improves the rank-1 accuracy and mAP over all features and metrics. Especially when the performance of

Table 3 The results of our re-ranking method on the Market-1501

Method	Rank 1	mAP
LOMO + Euc	27.14	8.01
LOMO + Euc+ krnn [55]	30.34	14.39
LOMO + Euc + our re-rank	40.20	11.22
LOMO + XQDA	43.56	21.93
LOMO + XQDA + krnn	48.34	32.21
LOMO + XQDA + OL-MANS [56]	51.87	–
LOMO + XQDA + our re-rank	64.82	24.14
IDE + Euc	78.92	55.03
IDE + Euc + krnn	81.09	71.15
IDE + Euc + our re-rank	85.21	60.04
IDE + XQDA	77.76	56.11
IDE + XQDA + krnn	80.23	70.20
IDE + XQDA + our re-rank	87.35	62.35
Deep feature + Euc	88.93	73.47
Deep feature + Euc + krnn	91.48	87.94
Deep feature + Euc + our re-rank	92.25	75.49
Our net + Euc	91.06	76.88
Our net + Euc + krnn	92.75	89.06
Our net + Euc + our re-rank	93.68	79.72

Table 4 The results of our deep network on the Market-1501

Features	Rank 1	mAP
Global	87.78	74.37
Upper-L	51.70	30.29
Arm-L	17.01	9.14
Leg-L	48.37	30.65
All-L	82.24	61.91
All	91.06	76.88

baseline is not so good, our method always makes a great improvement by a huge margin. Even though the performance of baseline is good enough such as Deep Feature, our method still can improve 3.32% and 2.02% in rank-1 and mAP respectively. Comparing with other re-ranking methods, our method outperforms in rank-1 accuracy. It’s reasonable that the improvement of mAP is less than the other re-ranking method, because the aim of our method mainly focuses on improving the top-k accuracy, thus we only consider the gallery images of high visual similarity, i.e., the top-m candidates of the initial rank list. Therefore, these true matches of low visual similarity are ignored by our re-ranking method, leading the benefit for mAP is not significant enough.

Considering our features are fused by several components, thus we compare different components’ effectiveness. ”Global” means f_{global} extracted by global network. ”Upper-L”, ”Arm-L” and ”Leg-L” denotes f_{upper} , f_{arm} and f_{leg} respectively. ”All-L” means all local features are fused together. The result is shown in Table 4. It shows that every part is effective and we get the best performance by fusing every part, which proves that they are complementary (Figure 5).

Table 5 shows the comparison of our best result with some other state-of-the-art methods on the Market-1501 dataset. We can see our best result can be competitive with these state-of-the-art methods.

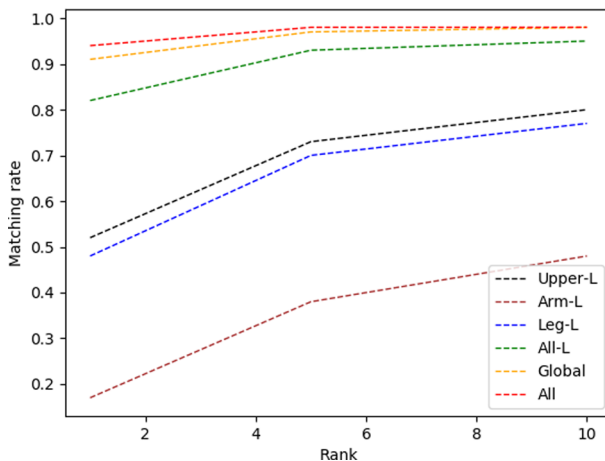


Figure 5 The CMC curve of feature fusion on Market-1501

Table 5 Comparison with state-of-the-art methods on the Market-1501

Method	Rank 1	mAP
MGCAM [29]	83.8	74.3
AACN [44]	85.9	66.9
HA-CNN [42]	91.2	75.7
SPReID [12]	92.5	81.3
PCB + RPP [33]	93.8	81.6
MGN [40]	95.7	86.9
Our net	91.1	76.9
Our net + Our re-rank	93.7	79.7

4.4 Experiments on DukeMTMC-reID

Finally, we conduct some experiments on the latest large-scale dataset DukeMTMC-reID. We follow the standard training and evaluation protocols in [8] where 702 IDs as the training set and the remaining 702 IDs as the testing set. We set the interval to 700, the max_interval to 40000 and m to 50. The spatial-temporal probability threshold is calculated the same way in the QMUL GRID part. The LOMO and baseline deep feature are employed to evaluate the performance. The result is shown in Table 6. And the result of different part fusion is shown in Table 7.

It's clear to see from table that our re-ranking method gains at least 7% improvement in rank-1 accuracy, no matter what feature or metric is used. This indicates that spatial-temporal information is a key clue for person re-identification (Figure 6).

Table 8 shows the comparison of our best result with some other state-of-the-art methods on the DukeMTMC-reID dataset. We can see our proposed method + re-ranking can achieve competitive performance.

Table 6 The results of our re-ranking method on the DukeMTMC-reID

Method	Rank 1	mAP
LOMO + Euc	12.16	4.67
LOMO + Euc+ krnn [55]	14.23	7.80
LOMO + Euc + ours	32.72	9.56
LOMO + XQDA	31.10	17.15
LOMO + XQDA + krnn	37.93	28.53
LOMO + XQDA + OL-MANS [56]	51.87	–
LOMO + XQDA + ours	66.92	31.48
Deep feature + Euc	65.04	45.56
Deep feature + Euc + krnn	72.62	66.69
Deep feature + Euc + ours	81.19	55.98
Deep feature + XQDA	70.56	50.90
Deep feature + XQDA + krnn	76.97	70.75
Deep feature + XQDA + ours	80.83	57.33
Our net + Euc	81.91	68.10
Our net + Euc + krnn	86.45	82.76
Our net + Euc + our re-rank	88.32	76.77

Table 7 The results of our deep network on the DukeMTMC-reID

Features	Rank 1	mAP
Global	79.76	64.27
Upper-L	46.72	26.94
Arm-L	15.39	7.14
Leg-L	45.38	24.61
All-L	70.65	50.06
All	81.91	68.10

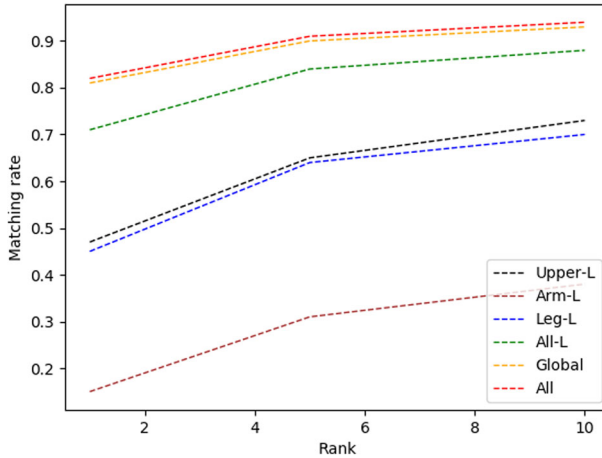


Figure 6 The CMC curve of feature fusion on DukeMTMC-reID

Table 8 Comparison with state-of-the-art methods on the DukeMTMC-reID

Method	Rank 1	mAP
AACN [44]	76.8	59.3
HA-CNN [42]	80.5	63.8
PCB + RPP [33]	83.3	69.2
SPReID [12]	84.4	71.0
MGN [40]	88.7	78.4
Our net	81.9	68.1
Our net + Our re-rank	88.3	76.8

Table 9 The results from DukeMTMC-reID to Market-1501

Method	Rank 1	mAP
Deep feature + Euc	47.65	22.15
Deep feature + Euc+ krnn [55]	51.60	32.92
Deep feature + Euc+ ours	62.29	26.36
Our net + Euc	52.36	25.77
Our net + Euc+ krnn	56.16	34.92
Our net + Euc+ Our re-rank	67.22	31.18

Table 10 The results from Market-1501 to DukeMTMC-reID

Method	Rank 1	mAP
Deep feature + Euc	29.08	15.52
Deep feature + Euc + krnn [55]	35.05	25.45
Deep feature + Euc + ours	58.53	27.40
Our net + Euc	35.01	18.46
Our net + Euc + krnn	40.33	27.72
Our net + Euc + Our re-rank	61.42	29.16

4.5 Experiments on cross-dataset

It's practical to exploit an existed model trained on a labeled dataset in a totally new environment. Therefore, two more experiments are conducted on cross dataset. One is training model on DukeMTMC-reID, and testing on Market-1501. The other is training model on Market-1501, and testing on DukeMTMC-reID. The results are shown in Tables 9 and 10 respectively.

Apparently, our re-ranking method can improve the performance a large margin even if the model was trained on another labeled dataset, which means our method is more generalized and can be applied to any new environment with a pretrained basic model.

5 Conclusion

In this paper, an effective framework for person re-identification is proposed, which includes a new deep network and a re-ranking method. In the new network, in addition to extracting global features, we also design a multi-branch network to extract features from a series of local regions effectively which can alleviate the disadvantage of misalignment. And in the re-ranking phase, we proposed a novel re-ranking method that exploits spatial-temporal information to obtain a better performance. In our further studies, we will focus on the network and try to design a more robust and generalized network.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No.61972030), the Fundamental Research Funds for Central Universities (No. 2018JBM017) and the Hebei Province Key Research and Development Projects(18210305D).

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
2. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. *Computer Vision & Pattern Recognition* (2015)
3. Almazan, J., Gajic, B., Murray, N., Larlus, D.: Re-id done right: Towards good practices for person re-identification. arXiv:1801.05339 (2018)
4. Chen, D., Yuan, Z., Chen, B., Zheng, N.: Similarity learning with spatial constraints for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1268–1277 (2016)
5. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: A deep quadruplet network for person re-identification. In: *IEEE Conference on Computer Vision & Pattern Recognition* (2017)

6. Chen, W., Chen, X., Zhang, J., Huang, K.: A multi-task deep network for person re-identification. In: *Thirty-First AAAI Conference on Artificial Intelligence* (2017)
7. De, C., Gong, Y., Zhou, S., Wang, J., Zheng, N.: 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*, pp. 1335–1344 (2016)
8. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1627–1645 (2010)
9. Garcia, J., Martinel, N., Micheloni, C., Gardel, A.: Person re-identification ranking optimisation by discriminant context information analysis. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1305–1313 (2015)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
11. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification (2017)
12. Kalayeh, M.M., Basaran, E., Gokmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification (2018)
13. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2288–2295. IEEE (2012)
14. Leng, Q., Hu, R., Liang, C., Wang, Y., Chen, J.: Bidirectional ranking for person re-identification. In: *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE (2013)
15. Li, W., Rui, Z., Tong, X., Wang, X.G.: Deepreid: Deep filter pairing neural network for person re-identification. *Computer Vision & Pattern Recognition* (2014)
16. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159 (2014)
17. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification (2017)
18. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. [arXiv:1705.04724](https://arxiv.org/abs/1705.04724) (2017)
19. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197–2206 (2015)
20. Liao, S., Li, S.Z.: Efficient psd constrained asymmetric metric learning for person re-identification. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3685–3693 (2015)
21. Liu, C., Loy, C.C., Gong, S., Wang, G.: Pop: Person re-identification post-rank optimisation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 441–448 (2013)
22. Loy, C.C., Xiang, T., Gong, S.: Multi-camera activity correlation analysis (2009)
23. Lv, J., Chen, W., Li, Q., Yang, C.: Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7948–7956 (2018)
24. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical gaussian descriptor for person re-identification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1363–1372 (2016)
25. Peng, P., Xiang, T., Wang, Y., Pontil, M., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. *Computer Vision & Pattern Recognition* (2016)
26. Rui, Z., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. *Computer Vision & Pattern Recognition* (2013)
27. Rui, Z., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. *Computer Vision & Pattern Recognition* (2014)
28. Shi, H., Yang, Y., Zhu, X., Liao, S., Zhen, L., Zheng, W., Li, S.Z.: Embedding deep metric for person re-identification: A study against large variations (2016)
29. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1179–1188 (2018)
30. Su, C., Zhang, S., Xing, J., Wen, G., Qi, T.: Deep attributes driven multi-camera person re-identification. In: *European Conference on Computer Vision* (2016)
31. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Qi, T.: Pose-driven deep convolutional model for person re-identification. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3960–3969 (2017)

32. Suh, Y., Wang, J., Tang, S., Mei, T., Lee, K.M.: Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 402–419 (2018)
33. Sun, Y., Zheng, L., Yi, Y., Qi, T., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 480–496 (2018)
34. Tong, X., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. *Computer Vision & Pattern Recognition* (2016)
35. Viorio, R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: European Conference on Computer Vision, pp. 135–153. Springer (2016)
36. Viorio, R.R., Haloi, M., Gang, W.: Gated siamese convolutional neural network architecture for human re-identification. In: European Conference on Computer Vision (2016)
37. Wang, Z., Hu, R., Liang, C., Leng, Q., Sun, K.: Region-based interactive ranking optimization for person re-identification. In: Pacific Rim Conference on Multimedia, pp. 1–10. Springer (2014)
38. Wang, F., Zuo, W., Liang, L., Zhang, D., Lei, Z.: Joint learning of single-image and cross-image representations for person re-identification. *Computer Vision & Pattern Recognition* (2016)
39. Wang, H., Gong, S., Zhu, X., Xiang, T.: Human-in-the-loop person re-identification. In: European Conference on Computer Vision, pp. 405–422. Springer (2016)
40. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. arXiv:1804.01438 (2018)
41. Wei, L., Zhang, S., Yao, H., Wen, G., Qi, T., Wei, L., Zhang, S., Yao, H., Wen, G., Qi, T.: Glad: Global-local-alignment descriptor for pedestrian retrieval (2017)
42. Wei, L., Zhu, X., Gong, S.: Harmonious attention network for person re-identification (2018)
43. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 466–481 (2018)
44. Xu, J., Rui, Z., Feng, Z., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification (2018)
45. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2119–2128 (2018)
46. Ye, M., Chen, J., Leng, Q., Liang, C., Wang, Z., Sun, K.: Coupled-view based ranking optimization for person re-identification. In: International Conference on Multimedia Modeling, pp. 105–117. Springer (2015)
47. Ye, M., Liang, C., Yu, Y., Wang, Z., Leng, Q., Xiao, C., Chen, J., Hu, R.: Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Trans. Multimed.* **18**(12), 2553–2566 (2016)
48. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by saliency matching. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2528–2535 (2013)
49. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1077–1085 (2017)
50. Zheng, L., Shen, L., Lu, T., Wang, S., Wang, J., Qi, T.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)
51. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv:1610.02984 (2016)
52. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q., et al.: Person re-identification in the wild. *CVPR* **1, 2** (2017)
53. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose invariant embedding for deep person re-identification. arXiv:1701.07732 (2017)
54. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. arXiv:1701.07717. 3 (2017)
55. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3652–3661. IEEE (2017)
56. Zhou, J., Yu, P., Tang, W., Wu, Y.: Efficient online local metric adaptation via negative samples for person reidentification. In: The IEEE International Conference on Computer Vision (ICCV), vol. 2, p. 7 (2017)

Affiliations

Zhongyi Li¹ · Yi Jin¹  · Yidong Li¹ · Congyan Lang¹ · Songhe Feng¹ · Tao Wang¹

Zhongyi Li
16120398@bjtu.edu.cn

Yidong Li
yqli@bjtu.edu.cn

Congyan Lang
cylang@bjtu.edu.cn

Songhe Feng
shfeng@bjtu.edu.cn

Tao Wang
twang@bjtu.edu.cn

¹ Department of Computer Science, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing, 100044, People's Republic of China