Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Optimizing Multi-granularity Region Similarity for Person Re-identification

CUIQUN CHEN, MEIBIN QI, NING YANG, WEI LI, AND JIANGUO JIANG
School of Computer Science and Information Engineering, Hefei University of Technology, Anhui, 230009, China

Corresponding author: Meibin Qi (e-mail: qimeibin@163.com).

This work was supported in part by the National Natural Science Foundation of China under Grant 61771180 and Grant 61876056, and Key Research and Development Project of Anhui Province, China(1704d0802183)

ABSTRACT Person re-identification(re-id) is one of the hottest research topics due to its great value in video analysis applications such as indoor security and road surveillance. It has been verified as beneficial for re-id to joint global and local features in recent literature. However, most existing methods usually extract features of the global region or divide the whole image into several parts without considering the alignment of different parts, which are not discriminating or robust to the complex scenarios. In this paper, we propose a novel method that optimizes multi-granularity similarity fusion(MGSF) based on the coarse region and the fine region. We extract the global feature representations on the coarse region and the local feature representations on the fine region. Instead of using the pose estimation method, we align the local parts by calculating the similarity between local parts, which is optimized by the refined longest path(RLP). Extensive experiments on four challenging datasets are carried out and indicate that our method has achieved state-of-the-art performances. For instance, on VIPeR dataset, our method achieves 67.25% rank-1 matching rate, outperforming state-of-the-art approaches by a large margin.

INDEX TERMS Person re-identification, coarse region, fine region, multi-granularity similarity fusion, the refined longest path.

I. INTRODUCTION

PERSON re-identification aims at associating the person across cameras and temporal periods in a non-overlapping view. Given one probe image of a pedestrian, a re-id system is expected to provide all the images of the same person from a large gallery dataset. However, the re-identification results may be inaccurate and not robust due to the appearance variation of the same person caused by large variations in illumination, viewpoint, occlusion, pose and uncontrollable camera settings. To address these difficulties, works can be divided into two categories. The first category works [1]–[9], [48] design pedestrian feature descriptors to deal with camera setting differences. It is divided into bottom layer features (color features and texture features), middle layer features (combination of bottom layer features), and deep features (features obtained by deep learning networks). The second category works [10]–[15] leverage a similarity function which maximizes the inter-class similarity and minimizes the intra-class similarity.

When it comes to extracting discriminative features, many hand-crafted-based and CNN-based(Convolutional Neural Network) methods learn global features without considering the spatial structure characteristics of pedestrian body. Due to high complexity for images captured with different cameras, it is hard for the global features to distinguish the similar inter properties or large intra variances. Some challenges are listed in Fig.1: 1) the inaccuracy of detector which introduces additional background interference and half-baked body images in Fig.1 (a-c); 2) the large variations of human poses and viewpoints make the occlusion problems, because the appendants that the pedestrians carry may appear at different viewing angles. How to distinguish occlusion regions is an important issue in re-id in Fig.1 (d-f). For solving these challenges, the combination of global features and local features has been confirmed to be a valid approach for re-id. A requisite premise of learning discriminative local features is that the body parts are accurately located [38]. Some part-based works [6], [15], [37] divide the body into several parts by empirical knowledge about the body structure without alignment. The others works [25], [41]–[43] rely on pose estimation for alignment. However, the latent datasets bias between pose estimation dataset and person re-identification

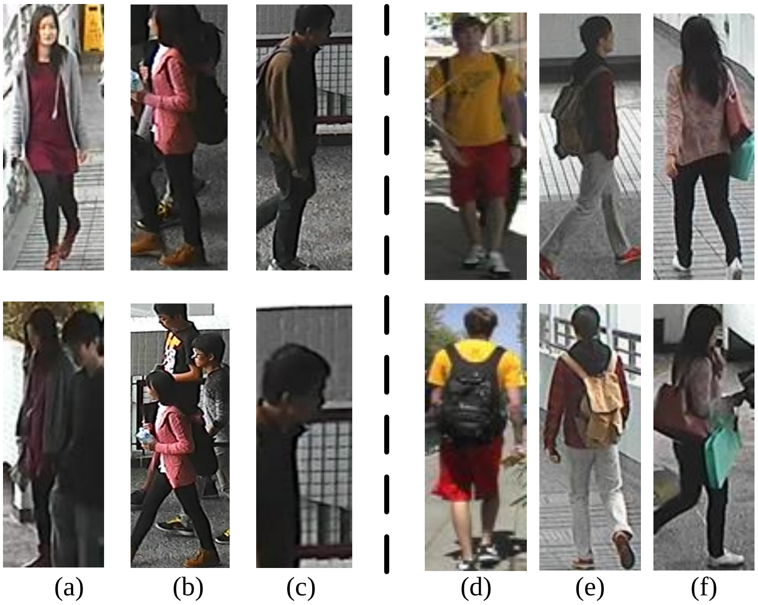

FIGURE 1: Challenges of person re-id. Images in the same column represent the same people. (a-c) Inaccurate detection. (d-f) Occlusions.

dataset remains an obstacle against ideal semantic partition, which is not conducive to person re-identification. So an effective and simple alignment method is urgently needed.

In this paper, we optimize multi-granularity similarity fusion(MGSF), which is a new approach for re-id and combines the coarse and fine region in different granularity. As shown in Fig. 2, we partition images evenly into several stripes and the various number of stripes denote the variety of granularity. The coarse region can express the overall characteristics of pedestrians and we exploit the joint learning method to learn its similarity. The fine region can extract more discriminative information and the independent learning method is used to reduce the interaction between different regions. For the fine region, we align the local parts by introducing the longest path. We consider the motivation that the corresponding parts should be as high proportion as possible in the longest path. We refine the longest path by adaptive weighting for the similarity and remove the regions with low similarities that exert a bad influence on pedestrian matching. And the alignment method is defined as the refined longest path(RLP). Then, we joint the coarse region and fine region similarity with a selective weighting strategy, which defines the final similarity between two images.

The main contributions are threefold:

(1) We propose a part-based method called MGSF, which combines the coarse region and fine region and employs a simple and effective uniform partition strategy. We find that using different similarity learning methods on coarse region and fine region can further promote performance.

(2) We propose an alignment method with the refined longest path(RLP) for fine region so that increasing robustness of fine region similarity. We also optimize the combination of regions similarity by the selective weighting strategy.

(3) We demonstrate experimentally the superior performance of our method on four person re-id datasets, i.e., VIPeR [33], GRID [34], PRID450S [35] and CUHK01 [26]. Moreover, the simple and effective combination of the coarse region and fine region is attractive to the practical person re-id system, without costly pose estimation.

The rest of the paper is organized as follows. Section II briefly reviews related works. Section III describes the details of our method. The experiments and results are shown in Section IV. We finally draw a conclusion and discuss possible future works in Section V.

II. RETATED WORK

Feature representations Hitherto, it has been devoted significant efforts to obtain stable and distinctive features. For example, Ma et al. [1] use the covariance to describe the person image, which is robust to illumination change and background variation. Wang et.al. [46] propose to capture different low-level features from multiple channels of HSV color space. Fan et.al. [47] propose a method based on multi-feature fusion in perceptual uniform color space. These methods are using the low-level features. Liao et al. [5] propose an efficient feature representation called Local Maximal Occurrence (LOMO), using color feature HSV and texture feature SILTP to represent pedestrian appearance in a high dimension. This method locally constructs a histogram of pixel features and then takes its maximum values within horizontal strips to overcome viewpoint variations. Tetsu et al. [6] present a novel descriptor called Gaussian of Gaussian(GOG) that describes the local color information and texture structures of an image via multiple hierarchical

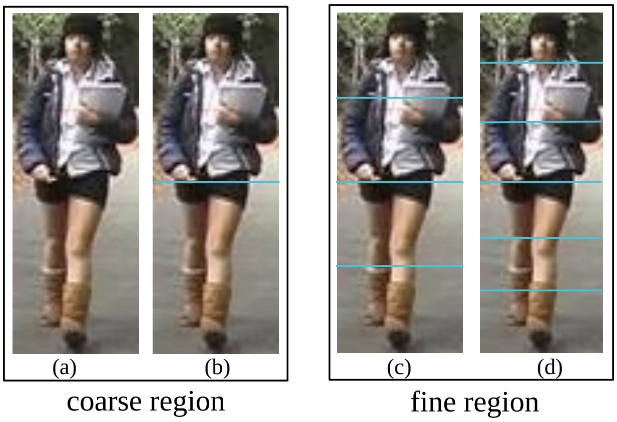

FIGURE 2: The detail of various uniform partition. (a-b) Coarse region. The coarse area is used to represent global and semi-global pedestrian feature information. (c-d) Fine region. The fine region is divided according to the spatial distribution of the pedestrian, which is used to indicate the characteristics of the body parts and the details of pedestrian images.

Gaussian distributions. Literatures [7], [25], [31], [38], [49] focus on establishing the Convolutional Neural Networks to learn feature representation of the person appearance. Tong et al. [7] present a pipeline for learning deep feature representations and the Domain Guided Dropout(DGD) algorithm to improve the feature learning procedure. Zhao et al. [25] propose a network based on human body region guided multi-stage feature decomposition and tree-structured competitive feature fusion. Cheng et al. [31] use a multi-channel CNN to learn body features from the input images. Sun et al. [38] design a Parted-based Convolutional Baseline(PCB) network, which can output a convolutional descriptor consisting of several part-level features. In this paper, we combine two hand-crafted-based and high-level features to represent pedestrian features.

Part alignment Part-based methods can be divide into two main pathways. Some works based on hand-crafted algorithms segment the images manually about the body structure without alignment. Chen et.al [15] partition images into horizontal stripes to extract color and texture features. Chu et.al [37] adopt more sophisticated subdivision horizontally and vertically to remove the bad block areas. Das et al. [40] divide images into three parts: the head, torso and legs and apply HSV histogram to capture spatial information. However, these methods still suffer from the effect of occlusion, pose variations and inaccurate detection boxes. Recently, some works align parts by pose estimation based on deep learning. Li et al. [41] detect directly key pose points and extract part features from corresponding regions. Su et al. [42] integrate a separately trained pose detection model into a part-based re-id model. Yao et al. [43] use an unsupervised method to generate a set of part boxes, and then employ the RoI pooling to produce part features. Zhao et al. [25] propose a region proposed network to locate the body region. These methods rely on sophisticated human pose estimation models which are often memory consuming and supervised.

Similarity learning Karanam et al. [44] propose a systematic evaluation of person re-id task by incorporating recent advances in metric learning. Most works learn a similarity measurement based on Mahalanobis distance. Kostinger et al. [8] propose an efficient metric computation method motivated by the log likelihood ratio test of two Gaussian distributions. Liao et al. [5]propose a subspace and metric learning method called Cross-view Quadratic Discriminate Analysis (XQDA), which learns a discriminate low dimensional subspace by cross-view quadratic discriminate analysis, and simultaneously, a QDA metric is learned on the derived subspace. Liao et al. [12] derive a logistic metric learning approach with the PSD constraint. Chen et al. [15] learn sub-similarity measurements of sub-regions and propose a unified framework that can unite local and global similarities(SCSP). In terms of deep learning, it has been derived various metric learning losses such as the contrastive loss, the triplet loss, and the triplet hard loss. Furthermore, many works combine softmax loss with metric loss to speed up the convergence [39]. In our work, we directly adopt SCSP to calculate the similarity between regions.

III. OUR APPROACH

In this section, we first introduce the framework of MGSF. Then we discuss the similarity for coarse region and fine region and explain the alignment on the fine region with more details. Finally, we fuse effectively these similarities to represent the overall similarity.

A. OVERVIEW OF THE PROPOSED FRAMEWORK

Fig. 3 illustrates the framework of our proposed method. As shown in the figure, our method not merely addresses distinctive features extraction but also robust multi-granularity similarity fusion problem. We extract two features of the image to represent image, ie.HS and R_GOG which are introduced by next part. As is shown in Fig. 3(a), we then adopt these features forming the feature map [15] which is used in similarity learning. The architecture of MGSF is shown in Fig. 3(b), it contains three branches and the change in the number of partition strips defines the multi-granularity. The Main branch contains the coarse region without any alignment process and learns the global and semi-global feature representations. The middle and lower branches which align parts through the Refined Longest Path(RLP) both represent the fine region similarity learning. We call these branches Fine-N Branch, where N refers to the number of partitions on the images, e.g the middle and lower branches in Fig. 3(b) can be named as Fine-4 and Fine-6 Branch.

We compute the similarity between any regions by SCSP which makes us to conveniently exploit the complementary strength of different local regions. The final similarity in Main branch merges the multi-granularity similarity to make the similarity more robust.

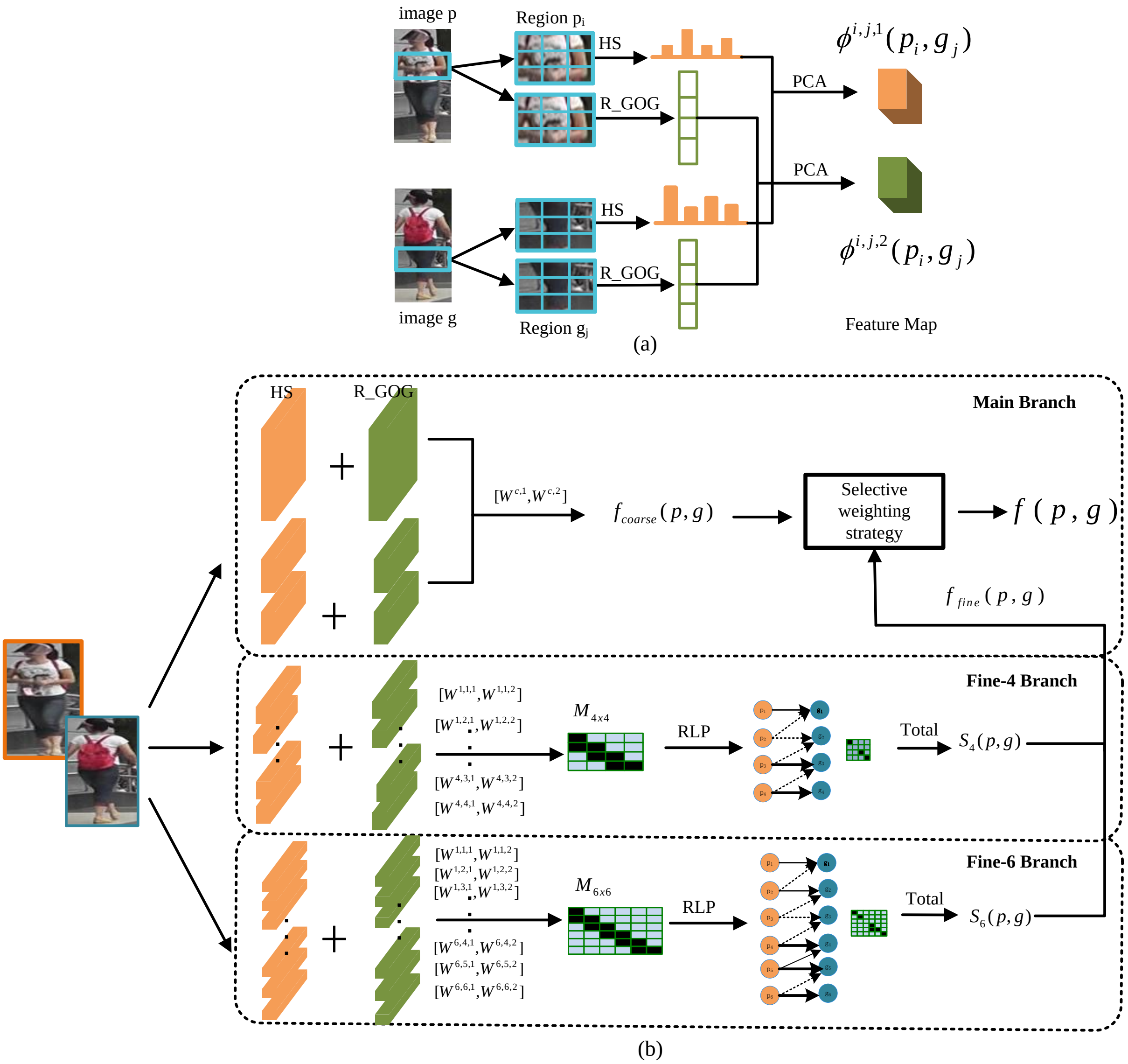

FIGURE 3: System overview. (a) The process of feature extraction of regions. (b) The framwork of MGSF. Our similarity combines the coarse region similarity and fine region similarity by the selective weighting strategy.

B. SIMILARITY MEASURE

For the problem which is shown in Fig. 1, pedestrian global features can not accurately express pedestrian information. In order to reduce the influence of mismatching and increase the robustness to occlusion, some works [15], [17] have adopted a single horizontal splitting method. This single horizontal splitting does not fully utilize the characteristics of human spatial distribution. As is shown in Fig. 2, we segment the images so that we make full use of the spatial distribution of pedestrian body. Considering that the pedestrian upper body information is more discriminating than the lower body information, the pedestrian image is divided into two regions, as shown in Fig. 2(b). Based on [18], the pedestrian carry appendants (such as backpacks, books, handbags etc.) may help to improve recognition rate, when the accessory can appear in different cameras at the same time in the same image, and vice versa. Additionally, we horizontally divide the image into 4 parts in Fig. 2(c) taking the difference of the position and size in occlusion areas into account. We further divide the image into 6 equal parts(head, chest, abdomen, thigh, keen and calf), which represent more pedestrian details, as shown in the Fig. 2(d).

In the learning stage, we adopt the fusion method of low-level features and middle-level features on a region. The feature extraction process is shown in Fig. 3(a). Due to factors such as low resolution of the pedestrian image and differences in posture, the color information becomes the most important clue to describe pedestrian characteristics, such as the color histogram and the color name descriptor [45]. The HSV color space can intuitively express the light and darkness of the object's color, which is not sensitive to changes in illumination and makes it easy to compare colors. SILTP [17] is an improved LBP description operator. It has excellent anti-interference ability for area-wide noise, especially when the detection area is covered by shadows or very dark. At the same time, the SILTP operator has a scale invariance, which makes it highly adaptable to illumination variations. We cascade the HSV and SILTP feature to form a low-level feature HS. GOG [6] models mean and covariance information of pixel features in the patch and region hierarchies. We extracted GOG feature for the coarse region and GOG_RGB(only extract RGB color feature) feature for fine region. We term the hierarchical Gaussian descriptor for this specific region as R_GOG. PCA is used to reduce the dimensions of the two features and preserve the effective features.

Given pedestrian feature descriptors p and g of image pairs, the similarity of this pedestrian image pairs is expressed as $f(p,g)$. The SCSP similarity can be formulated to:

$$f(p,g) = < \Phi_M(p,g), W_M >_F + < \Phi_B(p,g), W_B >_F \tag{1}$$

where $< . >_F$ is the Frobenius inner product, $\Phi_M(p,g) = (p-g)(p-g)^T$ and $\Phi_B(p,g) = pg^T + gp^T$. $\Phi_M(p,g)$ is connected to Mahalanobis distance and $\Phi_B(p,g)$ corresponds to bilinear similarity [3]. W_M and W_B are optimized by ADMM algorithm [15] and the details of optimization are shown in literatue [15]. $\Phi(p,g) = [\ \Phi_M(p,g) \quad \Phi_B(p,g)\ ]$, which is defined as the feature map. In this paper, we need calculate two features similarity on a region pair and fuse them by sum.

C. THE REFINED LONGEST PATH

When using the part-based mothed, the misalignment and occlusion caused by the detection error and the large variations of human pose and viewpoint become a crucial factor affecting the re-id accuracy. In this paper, we propose an effective alignment method called the refined longest path(RLP).

We argue that similarity learning should contains certain spatial constraints, which indicates that the similarity between the corresponding semantic body parts should be calculated. For example, the region that contains the head of a person should be compared with the region that contains the head rather than the region that contains the feet [15]. Based on the spatial constraints, [39] matches local parts from top to bottom to find the alignment of local features and uses the minimum total distance as the local distance. However, we find that the distance of the corresponding semantic parts and the distance of the non-corresponding semantic parts which is adverse to the pedestrian matching are included in the minimum total distance. And the occlusion regions also are not removed.

We propose the RLP as shown in Fig. 4, which contains two traits: 1) we calculate the similarity rather than the distance between two parts, so we use the dynamic programming to find the longest path(the more similar the pedestrian, the greater the similarity.) and use the sum of the longest path as the local similarity. 2) we adopt the adaptive weighting for the similarity in the RLP according to the similarity value between parts, which can increase the proportion of the corresponding parts with high similarity and discard the parts with low similarity that have a bad effect on pedestrian matching. The details are as follows.

Our method obtains the local features $p = \{p_1, p_2..p_h\}$ and $g = \{g_1, g_2..g_h\}$ by horizontally segmenting h stripes for two images. The similarity $f(p_i, g_j)$ (p_i is the ith part of image p, g_j is the jth part of image g) between p_i and g_j is calculated as we mentioned above. A similarity matrix $M_{h\times h}$ is formed based on similarity $f(p_i,g_j)$(i and j are from 1 to h), where $f(p_i,g_j)$ is the element of the ith row and jth column. We define the refined longest path sum from $(1,1)$ to (h,h) in the matrix $M_{h\times h}$ as the local similarity between p and g. The refined longest path S is calculated through dynamic programming [39] as follows:

$$S(p_i,g_j) = \begin{cases} W(f_{i,j}) * f_{i,j} & i=1, j=1 \\ S_{i-1,j} + W(f_{i,j}) * f_{i,j} & i \neq 1, j=1 \\ S_{i,j-1} + W(f_{i,j}) * f_{i,j} & i=1, j \neq 1 \\ max(S_{i,j-1}, S_{i-1,j}) + W(f_{i,j}) * f_{i,j} & i \neq 1, j \neq 1 \end{cases} \tag{2}$$

where $S(p_i,g_j)$ is the sum of the path from $(1,1)$ to (i,j) in the matrix $M_{h\times h}$. $S(p_h,g_h)$ is the sum of the path, which is also the local similarity between two images. $W(f_{i,j})$ is the weight function that decreases to zero as the similarity $f_{i,j}$ decreases. We define the weighting function $W(f)$ as following equation:

$$W(f) = \begin{cases} exp(\alpha+\beta) & f >= q \\ 0 & f < q \end{cases} \tag{3}$$

where f represents the similarity between regions, and α, β and q are the evaluation parameters, which are determined by experiments. The weight is 0 when f is less than q, otherwise it will change with the change of f value and the monotonicity is consistent with f.

As shown in Fig. 4, there are two pairs of pedestrian images pair and image p and image g are the same people captured by different cameras. Each arrow represents an element in the refined longest path and the dotted arrows show that we discard the element in the refined longest path. We adopt the adaptive weighting for the similarity according to the similarity between local parts as eq.3. The weight rapidly declines to zero when the similarity value less than q, which is helpful to reduce the impact of mismatching and occlusion caused by the inaccurate detection and the variation in viewpoint and pose. For example, it can increase the proportion of the corresponding parts with high similarity and discard the non-corresponding parts with low similarity on pedestrian matching in the Fig. 4(a). The thicker the arrow is, the weight value is greater. Hence, the path contains as many semantic corresponding parts as possible and the local similarity is almost decided by semantic corresponding parts. As shown in Fig. 4(b), due to the change of the viewing angle, the pedestrian backpack obscures the pedestrian body, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2890664, IEEE Access

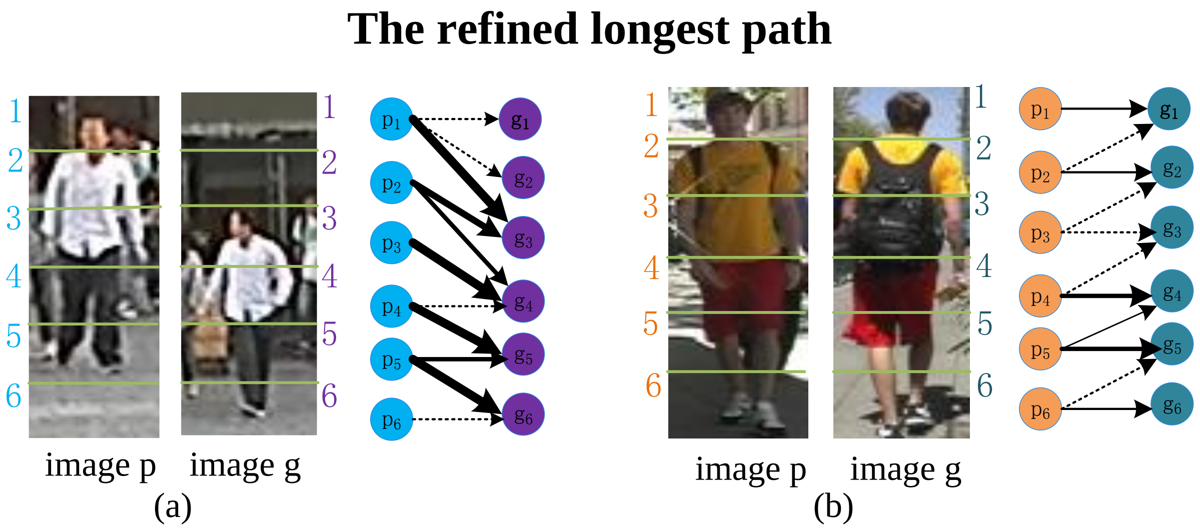

FIGURE 4: Examples of the RLP. (a) Misalignment in the RLP. (b) Occulsion in the RLP.

the RLP can reduce the impact of the occlusion region by adaptive weighting, thereby increasing the robustness to the occlusion.

D. MATCHING AND OPTIMIZING REGION SIMILARITY

Coarse region similarity For Pseudo-damaged Regions [37], we make full use of the overall information of pedestrians to compensate for the loss of Pseudo-damaged Regions that have been disabled. Joint similarity learning is used for coarse region in training dataset, which is jointly learning W shown in Main Branch of Fig. 3(b), generating the coarse region similarity:

$$f_{coarse}(p,g) = \sum_{n=1}^{2} < \Phi^{c,n}(p,g), W^{c,n} >_F \tag{4}$$

Where n is the feature descriptor($n = 1$ reoresents HS, $n = 2$ represents R_GOG), c is the partition number, and the part $\Phi^{c,n}(p,g) = [\ \Phi^{c,1}(p,g) \quad \Phi^{c,2}(p,g)\ ]^T$ is the coarse region feature map, and $W^{c,n} = [\ W^{c,1} \quad W^{c,2}\ ]$ is counted by ADMM.

Fine region similarity Independent similarity learning is performed on the fine region, which refers to train regions independently to obtain multiple similarities. The similarity between i-th part of image p and j-th part of image g is expressed as:

$$f(p_i,g_j) = \sum_{n=1}^{2} < \Phi^{i,j,n}(p_i,g_j), W^{i,j,n} >_F \tag{5}$$

Where $\Phi^{i,j,n}(p_i,g_j)$ is the feature map of region p_i and g_j. Then we align fine region by RLP to obtain the optimized fine region similarity. The fine region similarity is given by:

$$f_{fine}(p,g) = S_4(p,g) + S_6(p,g) \tag{6}$$

Here $S_4(p,g)$ and $S_6(p,g)$ indicate respectively multi-granularity segmentation in the Fine-4 and Fine-6 Branches shown in Fig. 3(b).

Similarity fusion Part-based feature representations can focus on the discriminative pedestrian details. Coarse region features contain the overall information of the pedestrian, which can complement to local features. Due to viewpoint and pose variations or detection errors which cause the misalignment and occulsion, the contents of the fine region of different images of the same person might large different. It is significant to effectively combine the coarse region similarity and fine region similarity .In this paper, we combine the coarse region similarity and the fine region similarity to generate the final similarity result according to the following equation:

$$f(p,g) = \begin{cases} (f_{coarse}(p,g) + f_{fine}(p,g))/2 & t < \delta \\ (1-\lambda) f_{coarse}(p,g) + \lambda f_{fine}(p,g) & otherwise \end{cases} \tag{7}$$

where $t = |f_{coarse}(p,g) - f_{fine}(p,g)|$. The tradeoff parameter δ and $\lambda(\lambda > 0.5)$ are selected via cross validation. When t is smaller than the parameter δ, it indicates that the gap between the fine region similarity and the coarse region similarity is relatively small and the influence caused by the occlusion of the pedestrian image or the change of the posture is weak. And the mean of two similarities is taken as the final similarity. When t is larger than the parameter δ, it indicates that the similarities of the coarse region and fine region are very different, and the image is greatly affected by the occlusion and pose variation, thus weighting and fusing two similarities are used to obtain the final similarity.

IV. EXPERIMENTS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2890664, IEEE Access

A. DATASETS AND SETTING

Datasets: four widely used datasets are selected for experiments, including VIPeR,GRID,PRID450S and CUHK01.

VIPeR The VIPeR dataset contains 632 persons ,with a total of 1264 images. Each person contains two images captured by camera A and camera B, respectively. In the experiments, 316 persons are randomly selected as the training set, remaining 316 persons as the test set.

GRID The GRID dataset is a very challenging dataset for re-id because of the poor image quality. Due to color changes, lighting changes, low image resolution and differences in pedestrian posture, the dataset can hardly achieve high pedestrian matching rates. The GRID dataset contains 250 persons with 1275 images and there are 775 images that do not belong to any one of the 250 people to augment the dataset. For each experiment, 125 persons are randomly selected as the training set, and the remaining 125 persons and 775 images are used as the test set. Namely, 125 probe images and 900 gallery images are included in one test.

PRID450S The PRID450S dataset is an extension of the PRID2011 dataset, which contains 450 persons from two different static monitoring cameras. In each experiment, 225 people are randomly selected as the training set and 225 individuals are left as the test set.

CUHK01 There are 971 persons contained in the CUHK01 dataset. All persons are captured from two cameras and each camera takes two pictures of the same person. Camera A captures the front and back angles of the pedestrians. Camera B shoots the left and right sides of the pedestrians. According to the literatures [26-27], 485 people are randomly selected as the training set and 486 are left as the test set.

Settings The dimensions of PCA reduction in the four datasets are 120,90,90 and 160, respectively. The tradeoff parameter δ and λ are selected via cross validation, which is set to 0.2 and 0.7 in our experiments. The parameter α, β and q are set to 0.5, 0.1 and 0.2 by experiments. The results are evaluated by the cumulative match characteristic(CMC) curves, which can well reflect the statistics of the ranks of true matches. In all experiments, we randomly choose images from each dataset for training and testing and the procedure is repeated 10 times and the average CMC is computed for final result.

B. PERFORMANCE COMPARISON

1) EXPERIMENTS ON VIPeR

The comparison results of the VIPeR dataset are shown in Fig. 5(a) and Table 1. We have selected the classic re-id algorithms in recent five years for comparison. The recognition rates of the algorithms Spindle [25], SSM [24], SCSP [15], and NFST [23] are all over 50% and the algorithms S-CIR [19], IDLA [20], SSDAL [22], and Spindle [25] are based on deep learning. From Table 1, we can see that the proposed method obtains a better performance than other methods with rank-1 reaching 67.25%, which is 13.45% higher than Spindle. It shows that our method is optimal on the VIPeR dataset and is superior to some methods based on deep learning.

2) EXPERIMENTS ON GRID

Similar to the performance on VIPeR, our method on GRID dataset significantly outperforms the previous state-of-the-art, achieving 31.70% rank-1 matching rate which is shown in Fig. 5(b) and Table 2. The results on VIPeR and GRID show that our method has obvious advantages on small datasets and achieves high recognition rates.

3) EXPERIMENTS ON PRID450S

Our method obtains superior results again as shown in Fig. 5(c) and Table 3. Our method achieves the best rank-1 and rank-5 matching rates, but rank-10 and rank-20 matching rates are worse than those of GOG [6]. The reason is that GOG adopts XQDA [5] as the distance measurement, while XQDA is non-linear, which is applicable to the dataset.

4) EXPERIMENTS ON CUHK01

Fig. 5(d) and Table 4 present the matching rates of various methods on the CUHK01 dataset and our method reaches better performance than most of the handcrafted features algorithms, which is 10% higher than GOG [6]. As a matter of fact, we achieve the second best matching rate when comparing with Spindle [25] which is based on deep learning. It is indispensable to underline that Spindle is very sensitive to the number of the training samples and very computationally intensive. Diversely, our method keeps less computation and smaller memory requirements because we need the less training parameters.

C. EMPIRICAL ANALYSIS OF THE PROPOSED METHOD

The VIPeR dataset is the most challenging for person re-id. Therefore, we perform experimental analysis of the proposed method on the VIPeR with 316 gallery images.

TABLE 1: Matching rates(%) of different methods on VIPeR

Methods	r=1	r=5	r=10	r=20
IDLA [20]	34.81	54.30	76.50	87.60
S-CIR [19]	35.76	67.00	83.00	–
Polymap [16]	36.77	70.35	83.70	91.74
SCNCD [21]	37.80	68.05	81.20	90.04
LOMO [5]	40.00	68.13	80.51	91.08
SSDAL [22]	43.50	71.80	81.50	89.00
GOG_{fusion} [6]	49.72	79.70	88.67	94.53
NFST [23]	51.17	82.09	90.51	95.92
SCSP [15]	53.54	82.59	91.49	96.65
SSM [24]	53.73	–	91.49	96.08
Spindle [25]	53.80	74.10	83.20	92.10
Ours	**67.25**	**90.37**	**95.89**	**97.44**

TABLE 2: Matching rates(%) of different methods on GRID

Methods	r=1	r=5	r=10	r=20
Polymap [16]	16.30	35.80	46.00	57.60
LOMO [5]	16.56	–	41.84	52.40
MLAPG [12]	16.64	–	41.20	52.96
SCSP [15]	24.24	44.56	54.08	65.20
GOG_{fusion} [6]	24.80	47.00	58.40	68.88
SSM [24]	27.20	–	61.12	70.56
Ours	**31.70**	**52.40**	**61.32**	**72.10**

TABLE 3: Matching rates(%) of different methods on PRID450S

Methods	r=1	r=5	r=10	r=20
NFST [23]	40.9	64.70	73.20	81.00
SCNCD [21]	41.60	68.90	79.40	87.80
Semantic [28]	43.10	70.5	78.20	86.20
TMA [29]	54.20	73.80	83.10	90.20
SCSP [15]	61.60	85.60	92.00	96.60
Spindle [25]	67.00	89.00	89.00	92.00
GOG_{fusion} [6]	68.47	88.80	**94.50**	**97.80**
Ours	**70.19**	**89.70**	94.02	97.52

TABLE 4: Matching rates(%) of different methods on CUHK01

Methods	r=1	r=5	r=10	r=20
eSDC [30]	19.7	32.40	–	50.20
Semantic [28]	32.70	51.2	–	76.30
MLF [31]	34.30	55.10	–	75.00
CSBT [32]	51.20	76.30	–	91.80
LOMO [5]	63.10	–	90.80	94.90
SCSP [15]	66.70	84.96	89.92	94.36
GOG_{fusion} [6]	67.30	86.90	91.80	95.90
Spindle [25]	**79.90**	**94.40**	**97.10**	**98.60**
Ours	77.10	92.50	95.96	98.32

TABLE 5: Performance comparison on VIPeR without key descriptors.

Methods	r=1	r=5	r=10	r=20
HS+MGSF	55.45	81.87	91.20	96.27
R_GOG+MGSF	62.50	88.74	91.14	96.71
GOG_{fusion}+MGSF	62.76	89.0	91.25	96.92
Ours	67.25	91.11	96.30	98.61

1) DEPENDENCY ON FEATURE DESCRIPTORS

Our MGSF is used to optimize multi-granularity region similarity. In order to describe regions preferably, we extract the HS feature and R_GOG feature. We evaluate the performance of MGSF without the HS feature or R_GOG feature to analyze the importance of features in our method. The results in Table 5 show that the recognition rate would descend without the HS feature or R_GOG feature and the lack of R_GOG would lead to further descend compared with the absence of HS feature. These experimental results show that our MGSF depends on features. When we use the GOG_{fusion} as the only feature, the accuracy is high 0.26% than R_GOG. However, the GOG_{fusion} need calculate four color space feature which is complicated and time-consuming. So we adopt the R_GOG as the feature. Furthermore, our MGSF result with GOG_{fusion} feature is better than GOG_{fusion} [6]. The rank-1 rate is 62.76% , which manifest a good improvement of 13.04% over GOG_{fusion} [6]. This result further explains that adopting the MGSF is conducive to improving the accuracy.

2) EFFECTIVENESS OF ALIGNMENT

The Main Branch focuses on main body part and Fine Branches focus on head, limbs, waist and feet. Furthermore, the Fine Branches with alignment help our framework to pay attention to useful detail on images and to distinguish the similar intra properties and large inter variances. We verify availability of the alignment method with three similar methods: MGSF without using the alignment(MGSF without alignment), MGSF uses the long path(LP) without weighting as the alignment method (MGSF with LP) and MGSF uses the refined long path(RLP) as the alignment method(MGSF with RLP). The result is shown in Table 6. Compared with MGSF no alignment , the MGSF with LP promotes 4.56% rank1 accuracy and MGSF with RLP promotes 6% rank1 accuracy, which indicates that our refining is effective.

3) EFFECTIVENNESS OF THE PARTS SEGMENTATION

We study the effect of multi-granularity segmentation by performing separate experiments for each segmentation method based on R_GOG feature. The experimental results are shown in Fig. 6. In comparison with the global similarity and similarity of the two stripes, it is concluded that the local similarity is more effective than the global similarity. However, the recognition rate does not always increase with number of parts, which means that the effective areas of the pedestrian image are destroyed and the alignment process becomes difficult. When number of parts is eight, the accuracy drops dramatically whether it's aligned or not. As a result, we choose to divide four parts and six parts as fine region in order to express more detail information of pedestrians.

4) EFFECT OF SIMILARITY FUSION STRATEGY

Intuitively, integrating the multi-granularity similarities will generally improve the performance, but how to effectively take their complementary strengths still remains an open problem. The proposed different similarity learning mode goes beyond the single learning mode. To verify this, joint learning and independent learning are used for both the coarse region and fine region. As shown in Table 7, in the fine region without alignment, the independent learning manner is higher 2% than the joint learning manner.

The effective fusion on the multi-granularity similarities is crucial. We contrast the following schemes to validate our method of similarity fusion strategy. In Table 8, we first find that the accuracy is enhanced about 1% when considering the combination of the coarse region similarity and fine region similarity by sum. Then, the selective weighting strategy can achieve the high performance and the rank1 matching rate is 67.25%, which indicates that the proposed fusion strategy is valid.

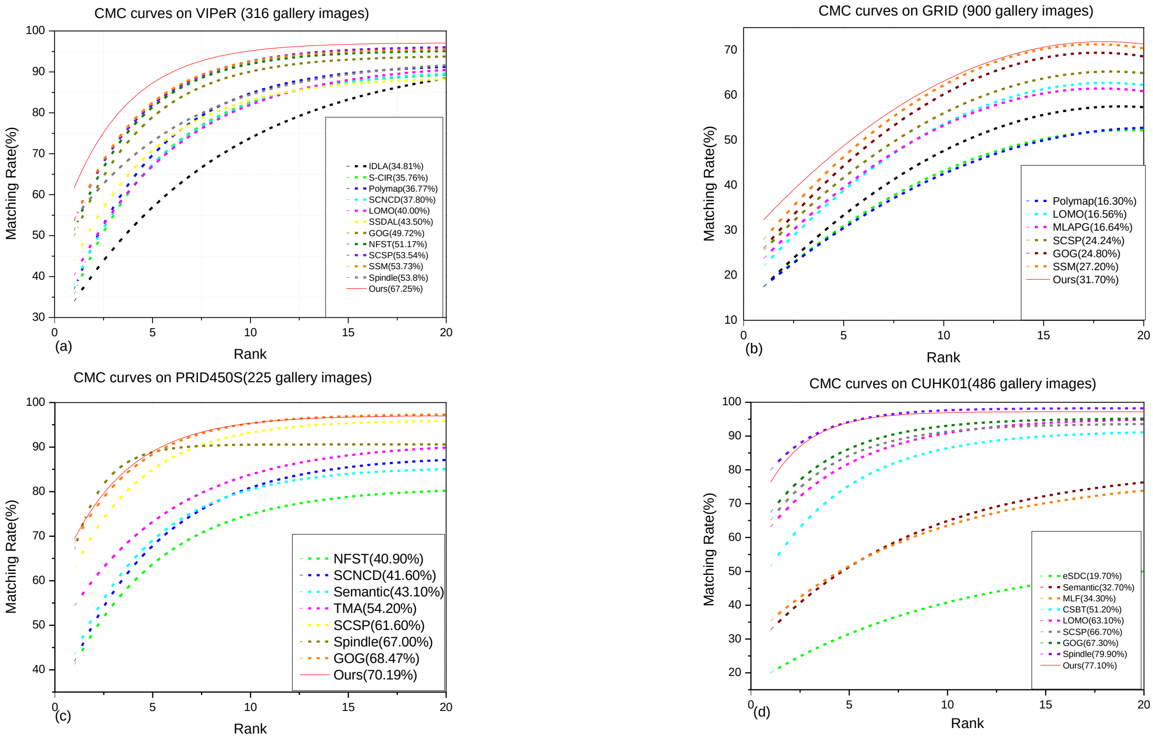

FIGURE 5: CMC curves for method comparison on (a)VIPeR dataset, (b)GRID dataset, (c)PRID450S dataset and (d)CUHK01 dataset.

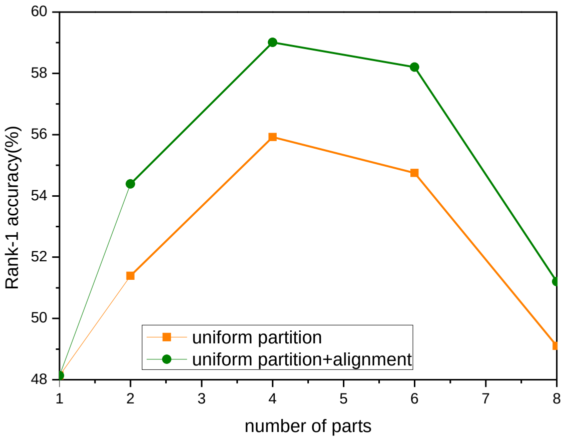

FIGURE 6: CMC curves for the number of horizontal stripes.

TABLE 6: Effectiveness of using the RLP alignmnet method. We fuse the coarse region simailarity and fine region similarity by sum in all experiments.

Methods	r=1	r=5	r=10	r=20
MGSF without alignment	60.00	87.34	93.80	97.44
MGSF with LP	64.56	89.15	95.09	98.29
MGSF with RLP	66.71	90.91	95.60	98.45

TABLE 7: Comparison of multi-mode learning. Fine region(N) represents that fine region without alignment.

Methods	Coarse region				Fine region(N)			
	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
Joint learning	58.42	85.28	92.75	97.75	56.01	84.08	92.22	97.37
Independent learning	57.75	85.54	93.39	97.82	58.77	86.68	93.29	97.22

TABLE 8: Comparison of similarity fusion strategy. Fine region(Y) represents that fine region with alignment.

Methods	r=1	r=5	r=10	r=20
Coarse region	58.42	85.28	92.75	97.75
Fine region(Y)	65.63	90.89	95.85	98.45
Sum	66.71	90.91	95.60	98.45
MGSF	67.25	91.11	96.30	98.61

5) ROBUSTNESS TO THE OCCLUSION

In this paper, the similarity of each image pair is the combination of the coarse region and fine region. As a result, when some region is occluded, the similarity measures in other regions still work. This matching mechanism is potentially robust to occlusion. To verify this, we conduct the following experiments to compare the performance of LOMO [5], GOG [6], NFST [23] with our method on the manual setting of occlusion areas.

In the experiment, we modify the probe images by randomly assigning to occlusion areas to the images during the test stage. Fig. 7 shows the results of the experiment. All

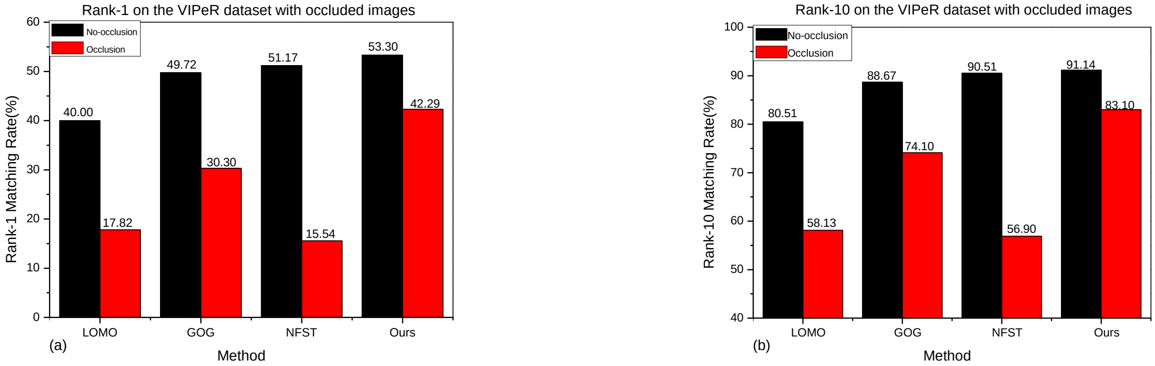

FIGURE 7: Methods comparison on occluded images of VIPeR dataset. (a) Rank-1 Matching Rate. (b) Rank-10 Matching Rate. All the experiments are trained and tested with R_GOG feature.

the four methods decline sharply due to the occlusion in Fig. 7(a), the corresponding rank-1 of several methods decreases by 22.18%, 19.42%, 35.63% and 11.01% respectively. In particular, our method is least affected by occlusion, which shows that it has good robustness against occlusion. The rank-10 has the same change rule as rank1 in the Fig. 7(b).

V. CONCLUSION

In this paper, we propose a multi-granularity similarity fusion(MGSF) method and design a novel multi-branch re-id framework. Each branch learns region features with different granularity partition. Our method introduces part locating operations through the RLP, which can not only align parts but also increase robustness to occlusion. For any pedestrian image pair, our method calculates the coarse region similarity and fine region similarity to generate the final similarity with a selective weighting strategy. Experiments on four datasets illustrate that our method reaches the comparable performance versus the state-of-the-art methods. In the future, we hope to extend our approach by applying deep learning and choosing more precise function $W(f)$, which is expected to improve performance ulteriorly.

REFERENCES

[1] B.Ma, Y.Su, and F.Jurie, "Bicov: a novel image representation for person re-identification and face verification," in proc. Brit. Mach. Vis. Conf, 2012.

[2] R.Zhao, W.Ouyang, and X.Wang, "Unsupervised salience learning for person re-identification," in proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jan. 2012.

[3] M.Farenzena, L.Bazzani, A.Perina, V.Murino, and M.Cristani, "Person re-identification by symmetry-driven accumulation of local features. In Computer Vision and Pattern Recognition", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jan. 2010.

[4] D.S.Cheng, M.Cristani, M.Stoppa, L.Bazzani, and V.Murino, "Custom pictorial structures for re-identification", in proc. Brit. Mac. Vis. Conf, Jan. 2011.

[5] S.Liao, Y.Hu, X.Zhu, and S.Z.Li, "Person re-identification by local maximal occurrence representation and metric learning", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015.

[6] T.Matsukawa, T.Okabe, E.Suzuki, and Y.Sato, "Hierarchical gaussian descriptor for person re-identification", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1363-1372.

[7] T.Xiao, H.Li, W.Ouyang, and X.Wang, "Learning deep feature representations with domain guided dropout for person re-identification", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp.1249-1258.

[8] W. Li, R. Zhao, T. Xiao, and X.Wang. Deepreid: Deep filter pairing neural network for person re-identification. In proc. IEEE Conf. Comput. Vis. Pattern Recognit., Feb. 2014, pp. 152-159.

[9] 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 proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun.2016, pp.1335-1344.

[10] M.Kostinger, M.Hirzer, P.Wohlhart, P.M.Roth, and H.Bischof, "Large scale metric learning from equivalence constraints", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jan. 2012, pp. 2288-2295.

[11] Z.Li, S.Chang, F.Liang, T.S. Huang, L.Cao, and J.R.Smith, "Learning locally-adaptive decision functions for person verification", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp.3610-3617.

[12] S.Liao and S.Z.Li, "Efficient psd constrained asymmetric metric learning for person re-identification", in proc. IEEE Int. Conf. Comput. Vis., 2015, pp.3685-3693. 3, 6, 7

[13] 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 proc. IEEE Conf. Comput. Vis. Pattern Recognit.,Jun. 2016, pp. 1335-1344.

[14] L.Ma, X.Yang, and D.Tao, "Person re-identification over camera networks using multi-task distance metric learning", in proc. IEEE Trans On Image process, vol. 23, no. 8, pp. 3656-3670, Jan. 2014.

[15] D.Chen, Z.Yuan, B.Chen, and N.Zheng, "Similarity learning with spatial constraints for person re-identification", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 1268-1277.

[16] Chen D, Yuan Z, Hua G, et al, "Similarity learning on an explicit polynomial kernel feature map for person re-identification", in proc.IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp.1565-1573.

[17] Qi M B, Hu L F, Jiang J G, " Person re-identification based on multi-features fusion and independent metric learning", in proc. Journal of Image and Graphics, vol. 21, no. 8, pp.1464-1472, 2016.

[18] Y.Yang, J.Yang, J.Yan, S.Liao, D.Yi, and S.Z.Li, "Salient color names for person re-identification", in proc. Eur. Conf. Comput. Vis., Jan. 2014.

[19] Wang F, Zuo W, Lin L, Zhang D, Zhang L, " Joint learning of single-image and cross-image representations for person re-identification", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016

[20] Ahmed E, Jones M, Marks TK, "An improved deep learning architecture for person reidentification", in proc. IEEE Conf. Comput. Vis. Pattern Recognit. , Jun. 2015, pp. 3908-3916.

[21] Yang Y, Yang J, Yan J, Liao S, Yi D, Li SZ (2014) ,"Salient color names for person re-identification", in proc. Eur. Conf. Comput. Vis., 2014, pp.536-551.

[22] Su C, Zhang S, Xing J, Gao W, Tian Q," Deep attributes driven multi-camera person reidentification", in proc. Eur. Conf. Comput. Vis. , 2016, pp.475-491.

[23] L.Zhang, T.Xiang, and S.Gong, "Learning a discriminative null space for person re-identification", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp.1239-1248.

[24] Song Bai, Xiang Bai, Qi Tian, "Scalable Person Re-identification on Supervised Smoothed Manifold", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp.3356-3365.

[25] H. Zhao, M.Tian, S.Sun, J.Shao, J.Yan, S.Yi, X.Wang,and X.Tang, " Spindle net: Person re-identification with human body region guided feature decomposition and fusion", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp.907-915.

[26] W.Li and X.Wang, "Locally aligned feature transforms across views", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp.3594-3601.

[27] R.Zhao, W.Ouyang, and X.Wang, "Learning mid-level filters for person re-identification", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp.144-151.

[28] Z.Shi, T.M.Hospedales, and T.Xiang, "Transferring a semantic representation for person re-identification and search", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp.4184-4193.

[29] N.Martinel, A.Das, C.Micheloni, and A.K.RoyChowdhury, "Temporal model adaptation for person re-identification", in proc. Eur. Conf. Comput. Vis. , 2016, pp.858-877.

[30] R.Zhao, W.Ouyang, and X.Wang, "Unsupervised salience learning for person re-identification", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp.3586-3593.

[31] 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 proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp.1335-1344.

[32] J.Chen, Y.Wang, J.Qin, L.Liu, and L.Shao, "Fast person reidentification via cross-camera semantic binary transformation", in proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp.5330-5339.

[33] D.Gray and H.Tao," Viewpoint invariant pedestrian recognition with an ensemble of localized features", in proc. Eur. Conf. Comput. Vis., 2008, pp. 262-275.

[34] C.C.Loy, T.Xiang, and S.Gong, "Time-delayed correlation analysis for multi-camera activity understanding", in proc. Int Jou. Comput Vis, vol. 90, no. 1, pp. 106-129, 2010.

[35] W.Li, R.Zhao, and X.Wang, "Human reidentification with transferred metric learning", in proc. Asi. Conf. Comput. Vis. , 2012, pp. 31-44.

[36] P.M.Roth, M.Hirzer, M.Kostinger, C.Beleznai, and H.Bischof, " Mahalanobis distance learning for person reidentification", pp. 247-267, 2014.

[37] Chu H, Qi M, Liu H, Jiang J. "Local region partition for person re-identification", in proc. Multimedia Tools & Applications, pp. 1-17, 2017.

[38] Y. Sun, L Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling", arXiv preprint arXiv:1711.09349, 2017.

[39] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification", arXiv preprint arXiv:1711.08184, 2017.

[40] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent Re-identification in a Camera Network", in proc. Springer International Publishing, 2014.

[41] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for pedestrian retrieval", in proc. ACM Multimedia, 2017.

[42] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Posedriven deep convolutional model for person re-identification", in proc. ICCV, 2017.

[43] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian, "Deep representation learning with part loss for person re-identification", arXiv preprint arXiv:1707.00798, 2017.

[44] Karanam and Gou et al, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets", in proc. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016.

[45] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification", in proc. European Conference on Computer Vision, 2014.

[46] Wang X, Zhao C, Miao D, et al, "Fusion of multiple channel features for person re-identification", in proc. Neurocomputing, 2016, pp. 125-136.

[47] Fan C, Chen Y, Cao L. "Person Re-identification Based on Fusing Appearance Features in Perceptual Color Space", in proc. China Academic Conference on Printing & Packaging and Media Technology, 2016, pp. 273-282.

[48] Li. T , Chang. H , Wang. M, Ni. B, Hong. R, Yan. S. "Crowded Scene Analysis: A Survey", in proc. IEEE Transactions on Circuits and Systems for Video Technology, vol.25, no.3, pp. 367-386, 2015.

[49] Hu. Y, Chang. H, Nian. F, Wang. Y, and Li. T, "Dense crowd counting from still images with convolutional neural networks", in proc. Journal of Visual Communication & Image Representation, vol.38(C), pp. 530-539.

CUIQUN CHEN received the B.E. degree in Fuyang Normal College, where she is currently taking successive postgraduate and doctoral programs of study for Ph.D. degree. Her research interests include digital image analysis and processing, computer vision and machine learning.

MEIBIN QI is currently a Professor in the School of Computer and Information at Hefei University of Technology. He received the B.E. degree in radio technology from Chongqing University in 1991, the M.E. and Ph.D. degrees in signal and information processing from Hefei University of Technology in 2001 and 2007. His research interests include pattern recognition, video coding, video surveillance and the application of DSP technology.

NING YANG received the B.E. degree in Hefei University of Technology, where he is currently a M. S. candidate at the Hefei University of Technology. His research interests include digital image analysis and processing, computer vision.

WEI LI received the B.E. degree in Fuyang Normal College, where he is currently a M. S. candidate at the Hefei University of Technology. His research interests include digital image analysis and processing, computer vision and deep learning.

JIANGUO JIANG is currently a professor in the School of Computer and Information at Hefei University of Technology. He received the B.E. degree in radio technology and M.E. degree in signal and information processing from Hefei University of Technology in 1982 and 1989. His research interests include digital image analysis and processing, distributed intelligent system and the application of DSP technology.

• • •