Video-based Person Re-identification Based on Feature Learning of Valid Regions and Distance Fusion

Danmei Yang\textsuperscript{1a}, Meibin Qi\textsuperscript{1b}, Jingjing Wu\textsuperscript{1c} and Jianguo Jiang\textsuperscript{1d}

\textsuperscript{1}School of Computer and Information, Hefei University of Technology, Hefei, Anhui, 230009, China
\textsuperscript{a}dmyang_hfut@163.com, \textsuperscript{b}qimeibin@163.com, \textsuperscript{c}hfutwujingjing@mail.hfut.edu.cn, \textsuperscript{d}jjiang@hfut.edu.cn

Abstract. Video-based person re-identification (re-id) is a very challenging problem because of the occlusion and the changes of viewpoints, pedestrian postures and illumination. Most existing methods of video-based person re-id usually concatenates the extracted appearance and space-time features directly, which do not consider the discrepancy of different features fully. In order to deal with this problem well, we propose a simple but effective method based on feature learning of valid regions and distance fusion, which combines three distances. The first distance is local distance of valid regions calculated using the Gaussian of Gaussian (GOG) feature. Pedestrian images are divided in horizontal directions, and then regions with smaller distances are selected as valid ones to be reserved and regions with larger distances are as invalid ones to be removed due to occlusions and posture changes. The other two distances are obtained by independent metric learning using the histogram of oriented gradient 3D (HOG3D) feature and the Local Maximal Occurrence (LOMO) feature. The three distances are added as the final distance between pedestrian pair, so the matching rank of gallery is obtained. Extensive experiments are performed on the iLIDS-VID and PRID-2011 datasets, and the results prove the effectiveness of our methods.

1. Introduction

Person re-identification aims to match persons across non-overlapping camera views, which has a wide range of applications in video retrieval and surveillance. However, due to partial occlusion and the variations of camera viewpoints, poses and illumination, it is very challenging. Generally, person re-id is mainly categorized into two types: image-based person re-id and video-based person re-id.

Image-based person re-id only can use the information of single image. Literature [1] proposes an effective feature representation of local maximum occurrence. Literature [2] proposes a person re-id algorithm with multi-feature fusion and independent measure. A novel method is proposed in literature [3], which enlarges the image and then splits the image into several regions in vertical and horizontal directions. Literature [4] presents the GOG descriptor based on a hierarchical distribution of pixel features.

Different from image-based methods, video-based re-id methods can extract spatial-temporal features to match, so this paper focuses on video-based person re-id. Literature [5] formulates re-id problem as a block sparse recovery problem and solves the associated optimization problem using the alternating directions framework. Literature [6] proposes a PaMM (Pose Aware Multi-shot Match) framework. Literature [7] proposes a novel spatio-temporal feature which is able to extract statistical
motion information based on densely computed multi-direction gradients and an adaptive fusion process. The top-push distance learning (TDL) method is proposed in Literature [8] which can optimize top-rank re-id matching.

In recent years, deep learning has made great breakthroughs in the field of computer vision, and deep learning methods are widely used in person re-id, such as the Spatial and Temporal Attention Pooling Network (ASTPN) [9] and the quality network (QAN) [10]. However, deep learning algorithms require a large amount of training data and long training time, which have advantages in large datasets. For small datasets, traditional methods might work better. Therefore, this paper focuses on the traditional method, that is, the extraction of manual features and fusion approaches.

In order to make the most of image information, we divide pedestrian images into 6 horizontal regions and then remove 2 invalid regions, which can solve the occlusion problem to some extent. In addition, extracted features are concatenated directly in most of video-based person re-id algorithms, which leads to the learned metric matrix not well characterizing the characteristics of different features. So, this paper proposes a method of distance fusion. Local and global features are respectively fed into the metric learning algorithm, so we can obtain several different metric matrices.

In general, a novel method of feature learning of valid regions and distance fusion is presented to address person re-id problems. In this paper, the main contributions are as follow: 1) We propose a method of image segmentation and removing invalid regions, which can reduce the impact of invalid regions and increase robustness to occlusions. 2) We propose a method of distance fusion instead of directly concatenating the extracted various features, which can fully reflect the characteristics of different features. 3) The better recognition rates are achieved in this paper.

2. The Proposed Method

2.1. Algorithm Framework

![Figure 1. Instances of valid regions and invalids regions. Regions with red wireframes are invalid regions. Regions without wireframes are valid regions. The three connected images in each part are videos of the same pedestrian in datasets. The person in a1 and a2 is the same person. The person in b1 and b2 is the same person. The person in c1 and c2 is the same person.](image)

Images in datasets of video-based person re-id are continuous sequences which are taken when people walk. When pedestrians walk on the road, they usually take something, such as backpacks, bags and so on. And due to changes of viewpoints and pedestrian postures, some images may appear some items which don’t exist in others. For example, in figure 1, the pedestrian’s backpack of a1 does
not appear in $a_2$. In addition, some items might cover target regions, which affect the recognition effect. In this paper, we call these regions invalid regions, and other regions are called valid regions correspondingly. Figure 1 shows some samples that include valid regions and invalid regions from the iLIDS-VID dataset.

In order to extract finer local information, each pedestrian image is divided into 6 horizontal regions, as shown in figure 2. Region 1 marked with red wireframes, region 2 marked with orange wireframes, region 3 marked with yellow wireframes and region 4 marked with green wireframes represent the upper body of pedestrians which accounts for 60% of the entire image. Region 5 marked with blue wireframes and region 6 marked with purple wireframes represent the lower body of the pedestrian which accounts for 40% of the entire image. The upper body contains richer information, so the upper body is divided equally into four regions to get more detailed local information. The lower body contains less effective information, so it is divided into two equal regions.

![Figure 2](image)

Figure 2. Illustration of the proposed method. $D_H$ is the distance of corresponding region 1. $D_S$ is the distance of corresponding region 2. The green box shows that the final distance between probe image and gallery image is a fusion of three distances.

Image segmentation not only can get finer features, but also can help us further remove invalid regions. For each local region, we extract GOG feature which is based on color and texture features such as RGB and scale invariant local ternary pattern [11] (SILTP). And we calculate distances of six corresponding regions by TDL. By observing images of datasets, it is found that region 1 and region 2 are generally not occluded by other items, so region 1 and region 2 are considered as valid ones to be reserved. And the occlusion of the upper body usually appears in region 3 and region 4. Therefore, comparing the distance of corresponding region 3 which is denoted as $D(x^A_3, x^B_3)$ and the distance of corresponding region 4 which is denoted as $D(x^A_4, x^B_4)$, the smaller one which is denoted as $D_L$ is regarded as the distance of valid regions and regions with larger distance is considered as invalid regions to be removed. The lower body of pedestrians is more easily occluded during walking, so the same operation of comparing distance is done in the lower body of the pedestrian image. Finally valid regions with smaller distance of the lower body are obtained, which is denoted as $D_L$. As shown in equation (1) and (2) and figure 2, $x^A_i$ ($x^B_i$) represents the extracted GOG feature of region $i$ ($i=1, 2, 3, 4, 5, 6$) from camera A(B).
\[ D_U = \min \left[ D(x^a_i, x^a_j), D(x^b_i, x^b_j) \right] \]  
\[ D_L = \min \left[ D(x^a_i, x^b_j), D(x^b_i, x^a_j) \right] \]  

For person re-id, we always hope that the top matching of gallery images is correct for a query image. It means that the distances between any unmatched gallery sample and the query should be larger than the one between any matched one and the query. However, the distance of occlusion regions is relatively larger. Local regions are measured independently to obtain the distances between the corresponding horizontal strips and then several regions with smaller distances are selected as valid regions to perform distance fusion, so that the distance between the same sample pair is smaller. As shown in equation (3), the distance of region 1 which is denoted as \( D_H \), the distance of region 2 which is denoted as \( D_S \), two distances of \( D_U \) and \( D_L \) are fused to obtain one kind of sample pair distance.

\[ D_{GOG} = D_H + D_S + D_U + D_L = D(x^a_i, x^b_j) + D(x^b_i, x^a_j) + D_U + D_L \]  

Removing invalid regions can eliminate the influence of occlusions and posture changes to some extent. However, the invalid regions with larger distances are not occluded entirely. Removing all of these will lose a part of image information. Therefore, we fuse global image information as compensation. \( D_{LOMO} \) is the second kind of distance calculated by extracting the LOMO feature of the entire image between sample pair. In addition, the continuous video sequences contain rich spatiotemporal information, so HOG3D [12] feature is extracted to obtain the third kind of distance between the sample pair. For different features we use the same measure method that is TDL, so the obtained three kinds of distances are of the same order of magnitude. Therefore, three distances can add directly to reflect the role of each feature. Finally, the distance between the sample pair is shown in equation (4).

\[ D = D_{GOG} + D_{LOMO} + D_{HOG3D} \]

2.2 Feature Extraction

In datasets, all images of person videos are resized to 128×64 pixels. In region segmentation process, images are divided into 6 horizontal regions, and the size of each region of the upper body is 19×64 and the size of each region of the lower body 26×64. We extract the GOG feature from 6 divided horizontal regions respectively. The GOG feature uses directly the information of each pixel of the image. Firstly, eight kinds of position, gradient and color features are extracted for each pixel. Then, the features of each pixel in the local block are merged by the block Gaussian operation. After that, local block features are merged by the regional Gaussian operation, and finally the features of whole image with good discrimination and robustness are obtained. The GOG feature can efficiently fuse various kinds of information of image pixels to obtain distinguishing features with small dimensions, and the extracted color components in this paper adopt the GOG feature of RGB.

In order to compensate for the information loss which is caused by removing invalid regions, 26960-dimensional LOMO feature is extracted from whole image. The LOMO feature efficiently combines the HSV feature and the SILTP texture feature and can cope with viewpoint changes through the operation of the maximum value. In addition, for each person video a 1200-dimensional HOG3D feature vector is extracted to characterize spatial-temporal information.

2.3. Metric Learning

Traditional video-based person re-id algorithms are devoted to designing more robust features or learning discriminative distance metric. We pay attention to the former, making extracted features more robust. As for distance metric, we use the TDL which bases on the Mahalanobis distance, as shown in equation (5), where \( x_i \) and \( x_j \) represent two eigenvectors.

\[ D(x_i, x_j) = (x_i - x_j)M(x_i - x_j)^T \]  

TDL optimizes the Mahalanobis distance by using equation (6) to obtain the measure matrix M,
and then the distance between the sample pair is obtained.

\[ f(D) = (1 - \alpha) \sum_{i, j, k, \lambda_i = \lambda_j} D(x_i, x_j) + \alpha \sum_{i, j, k} \max \{D(x_i, x_j) - \min D(x_i, x_j) + \rho, 0\} \]  

(6)

Therefore, the distances of whole images and the distances of local regions are all calculated by TDL.

3. Experiment Result

3.1. Datasets and Settings

**Datasets**: The iLIDS-VID dataset which is taken by two disjoint cameras in an airport terminal hall contains 600 videos of 300 sampled people. Each person has one pair of videos from two camera views. In this dataset similar clothing, complex backgrounds, viewpoint changes and serious occlusion make person re-id a big challenge. The PRID-2011 dataset is recorded by two different cameras in uncrowded outdoor scenes. 200 persons are recorded in both camera views. And the 200 person samples are evaluates in experiments of this paper.

**Settings**: When extracting appearance features, this paper extracts features separately for all image sequences of the same pedestrian collected by the same camera, and then takes the mean as the feature of the sample. That means our experiments are under a single query setting. Both datasets are randomly divided into testing set and training set by half, so there are 150 and 100 individuals in the training set of iLIDS-VID and PRID-2011 respectively. The images from one of the cameras are used as the probe set and the ones from another camera are as the gallery set. We repeat the procedure 10 times under the same conditions and report the average result. In this paper Cumulated Matching Characteristics (CMC) curve is adopted to evaluate the performance of our proposed methods, which is an estimate of finding the correct match in the top n rank.

3.2. Performance Comparison

**Results of iLIDS-VID**: We compare our approach with the state-of-the-art on the iLIDS-VID dataset. As shown in table 1, where Class A represents traditional methods and Class B represents deep learning methods, it can be seen that our approach can get better recognition rates than those of compared methods. For instance, our method improves the Rank-1 matching rate by 21.2% compared to the baseline method TDL. Moreover, our method also can beat the deep learning approaches, including ASTPN[9], QAN[10] and so on.

| Type   | Method                  | iLIDS-VID   | PRID-2011   |
|--------|-------------------------|-------------|-------------|
|        |                         | Rank1 | Rank5 | Rank10 | Rank20 | Rank1 | Rank5 | Rank10 | Rank20 |
| Class A| Ours                    | 77.5  | 93.4  | 96.5  | 98.9  | 84.3  | 93.7  | 96.8  | 98.1  |
|        | SRID[5]                 | 24.9  | 44.5  | 55.6  | 66.2  | 35.1  | 59.4  | 69.8  | 79.7  |
|        | PaMM[6]                 | 30.3  | 56.3  | 70.3  | 82.7  | 45.0  | 72.0  | 85.0  | 92.5  |
|        | CS-FAST3D+RMLLC[7]      | 31.2  | 60.3  | 76.4  | 88.6  | 54.7  | 66.7  | 78.1  |        |
|        | TDL[8]                  | 56.3  | 87.6  | 95.6  | 98.3  | 56.7  | 80.0  | 87.6  | 93.6  |
|        | SIIDL[13]               | 48.7  | 81.1  | 89.2  | 97.3  | 64.1  | 87.3  | 89.9  | 92.0  |
|        | STFV3D+KISSME[14]       | 44.3  | 71.4  | 83.7  | 91.7  | 42.1  | 71.9  | 84.4  | 91.6  |
|        | STFV3D[15]              | 37.0  | 64.3  | 77.0  | 86.9  | 42.1  | 71.9  | 84.4  | 91.6  |
|        | ASTPN[9]                | 62.0  | 86.0  | 94.0  | 98.0  | 77.0  | 95.0  | 99.0  | 99.0  |
|        | QAN[10]                 | 68.0  | 86.8  | 95.4  | 97.4  | 90.3  | 98.2  | 99.3  | 100   |
| Class B| RNN-CNN[16]             | 58.0  | 84.0  | 91.0  | 96.0  | 70.0  | 90.0  | 95.0  | 97.0  |
|        | CNN+XQDA[17]            | 53.0  | 81.4  | 89.7  | 95.1  | 77.3  | 93.5  | 95.7  | 99.3  |
|        | CNN+SRM+TAM[18]         | 55.2  | 86.5  | -     | 97.0  | 79.4  | -     | 94.4  | -     |

**Results of PRID-2011**: Compared with the existing traditional person re-id algorithms based on feature extraction and metric learning, our algorithm still improves greatly on PRID-2011. From table...
1, at present, the highest recognition rate of traditional methods is SIIDL, and Rank1 reaches 76.7%. The Rank1 of this method is as high as 84.3%, which is 7.6% higher than SIIDL algorithm. And it is higher than all current traditional person re-id algorithms on PRID-2011. Besides, our method outperforms some deep learning methods, such as ASTPN, literature [17] and literature [18].

3.3. Effect of Major Components
To further illustrate the role of the different components of the method, comparative experiments are performed on PRID-2011 dataset.

Effect of Distance Fusion: In order to prove the effectiveness of distance fusion, we compare with the CM1 method which directly concatenates the extracted features. The results of the CM1 method are shown in table 2. It is found that our method works much better. Distance fusion can more fully reflect the performance of different features and achieve complementary functions between features.

| Method | Rank1  | Rank5  | Rank10 | Rank20 |
|--------|--------|--------|--------|--------|
| Ours   | 84.3   | 93.7   | 96.8   | 98.1   |
| CM1    | 45.2   | 80.1   | 90.6   | 95.1   |
| CM2    | 81.8   | 90.6   | 95.6   | 97.5   |
| CM3    | 82.1   | 92.5   | 95.9   | 98.0   |

Effect of removing invalid regions: In order to verify the effectiveness of removing the invalid regions after region segmentation, we conduct a comparative test which is named CM2 with all 6 horizontal regions. As shown in table 2, Rank1 of CM2 is 81.8%, which is decreased by 2.5%. It can be seen that invalid regions contain some interference information which causes certain interference to person re-id.

Effect of non-uniform segmentation: We perform a comparative experiment named CM3 to prove the effectiveness of our segmentation method. Firstly, we keep the images divided into 6 regions evenly. Then, we remove an invalid region from the upper body and the lower body respectively. As shown in table 2, it can be observed that our segmentation method has better performance than CM3.

4. Conclusion
In this paper, a simple but effective method based on feature learning of valid regions and distance fusion is proposed for the video-based person re-id of traditional methods. We calculate the distances between the corresponding horizontal regions, select regions with smaller distance as the valid ones and regions with larger distance as the invalid ones. The effectiveness of our methods lies in removing the invalid regions due to occlusions or changes of views. Moreover, we use the distance fusion method to fuse the appearance and the space-time features, which could fully use the information of pedestrian videos. Experiments on iLIDS-VID and PRID-2011 datasets can demonstrate that our methods can significantly improve the recognition rate of person re-id.

Acknowledgments
This work is supported by the National Natural Science Foundation of China Grant 61771180 and 61876056 and Key Research and Development Project of Anhui Province, China 1704d0802183.

Reference
[1] Liao S, Hu Y, Zhu X, et al. 2015 Person re-identification by local maximal occurrence representation and metric learning IEEE Conference on Computer Vision and Pattern Recognition (USA: Boston) pp 2197-2206
[2] Qi M, Hu L, Jiang J, et al. 2016 Person re-identification based on multi-features fusion and independent metric learning Journal of Image and Graphics 21(11): 1464-1472(J)
[3] Chu H, Qi M, Liu H, et al. Local region partition for person re-identification Multimedia Tools & Applications 2017(7): 1-17 (J)

[4] Matsukawa T, Okabe T, Suzuki E, et al. 2016 Hierarchical gaussian descriptor for person re-identification IEEE Conference on Computer Vision and Pattern Recognition (Las Vegas: NV) pp 1363-1372

[5] Karanam S, Li Y and Radke R 2015 Sparse re-id: Block sparsity for person re-identification IEEE Conference on Computer Vision and Pattern Recognition Workshops (USA: Boston) pp 33-40

[6] Cho Y and Yoon K 2016 Improving person re-identification via pose-aware multi-shot matching IEEE Computer Vision and Pattern Recognition (USA: Las Vegas) pp 1354-1362

[7] Liu Z, Chen J and Wang Y 2016 A fast adaptive spatio-temporal 3d feature for video-based person re-identification IEEE International Conference on Image Processing (USA: Phoenix) pp 4294–4298

[8] You J, Wu A C, Li X, et al 2016 Top-push video-based person re-identification IEEE Computer Vision and Pattern Recognition (USA: Las Vegas) pp 1345-1353

[9] Xu S, Cheng Y, Gu K, et al 2017 Jointly attentive spatial-temporal pooling networks for video-based person re-Identification IEEE International Conference on Computer Vision (Italy: Venice) pp 4743-4752.(J)

[10] Liu Y, Yan J and Ouyang W 2017 Quality aware network for set to set recognition IEEE Conference on Computer Vision and Pattern Recognition (USA: Hawaii) pp 4694-4703.

[11] Liao S, Zhao G, Kellokumpu V, et al. 2010 Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes IEEE Computer Vision and Pattern Recognition (USA: San Francisco) pp 1301-1306

[12] Klasar A, Marszalek M, Schmid C 2008 A spatiotemporal descriptor based on 3D-gradients British Machine Vision Conference (UK: Lodon) pp 99.1-99.10.

[13] Zhu X, Jing X, Wu F, et al. 2016 Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics International Joint Conference on Artificial Intelligence (USA: New York) pp 3552-3558.

[14] Zhang W, Ma B, Liu K, et al. 2017 Video-based pedestrian re-identification by adaptive spatio-temporal appearance model. IEEE Transactions on Image Processing 26(4) pp 2042–2054. [J ]

[15] Liu K, Ma B P, Zhang W, et al 2015 A spatio-temporal appearance representation for video-based pedestrian re-identification IEEE International Conference on Computer Vision. (Chile: Santiago) pp 3810-3818

[16] McLaughlin N, Rincon J and Miller P 2016 Recurrent convolutional network for video-based person re-identification IEEE Computer Vision and Pattern Recognition (USA: Las Vegas) pp 1325-1334

[17] Zheng L, Bie Z, Sun Y, et al. 2016 MARS: A video benchmark for large-scale person re-identification the European Conference on Computer Vision (Netherlands: Amsterdam) pp 868-884

[18] Zhou Z, Huang Y, Wang W, et al 2017 See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification IEEE Conference on Computer Vision and Pattern Recognition (USA: Hawaii) 6776-6785