A G2G Similarity Guided Pedestrian Re-identification Algorithm

Guangcai Wang, Shang Gao and Di Fan*

College of Electronic Information Engineering, Shandong university of science and technology, Qingdao, Shandong, 266590, China

E-mail: fandi_93@126.com

Abstract. Pedestrian re-identification aims to settle the matching problem for a given target pedestrian under the multi-camera with non-overlapping visual field, so as to achieve the goal of pedestrian retrieval. In this paper, the similarity between the gallery images (G2G similarity) will be used to guide and refine the similarity between query images and gallery images (P2G similarity). It is also introduced into the training process and playing a supervisory role. To fully utilize details of the images, the image features are horizontally overlapped into groups, and the similarities between each group of the query images and the gallery images are calculated respectively. By learning the weights of grouping features in the training process, the importance of key parts can be automatically perceived. The results on the Market-1501 and CUHK03 datasets prove the effectiveness of proposed algorithm.

1. Introduction

As one of the important means of public safety management, monitoring video is extremely significant for the management of urban order, the inquiry of lost pedestrian and the case investigation. Pedestrian re-identification is a technique of pedestrian matching under a multi-camera network with non-overlapping visual field using computer vision technology. In other words, given a monitor pedestrian image, it retrieves the images of the pedestrian across devices.

The research hotspot of pedestrian re-identification can be summarized as two main directions. One for feature-based representation method and the other is measurement-based learning method. Feature-based representation method studies to learn more robust features, such as LOCAL Maximal Occurrence Representation (LOMO)[1], Feature Fusion Net (FFN)[2] and Symmetry-Driven Accumulation of Local Features (SDALF)[3]. However, because of the limit of manual feature description, it is difficult to adapt to large data tasks in complex scenarios. In the meantime, a lot of researches have done for feature measurement. Some methods with the verifying loss, based on the Siamese convolutional neural network, such as contrast loss[4,5], triplet loss[6], quadruplet loss[7], are used to make the distance with the same ID pedestrians closer, and those with different ID farther.

However, in complex scenes, due to the angle of shooting, illumination, and human body posture, the retrieval results based on the similarities ranking between the probe and the gallery, have large error and poor robustness. In response to these problem, some scholars have considered using the similarity information of the gallery to re-rank the retrieval results[8-11], and re-estimate the similarity between the probe image and gallery images, to ensure the correct results ranked higher. The supervised smoothed manifold (SSM)[8] uses the underlying manifold structure to refine the similarity between the probe and gallery images. The CNN embedding method[9] which combines coarse-grained and fine-grained human pose information, introduces the concept of extended cross-neighbour distance to re-rank of retrieval results. The k-reciprocal[10] neighbours of each probe image are encoded
into a single vector to re-rank the retrieval results under Jaccard distance. Bai proposed the Unified Ensemble Diffusion (UED)\cite{11} algorithm to learn the weights of the measurement in the re-ranking stage by introducing the replicator equation.

The existing methods mainly deal with the re-ranking process separately. In the training stage, the similarity between the gallery is not fully utilized to make the network learn more features. This paper makes the utmost of the similarity of the gallery, optimizes the similarity between the gallery images and the probe through the refined module, which brings rich supervisory information to the training process. In addition, by horizontally overlapping the image features, the image detail information is fully utilized to refine the similarity, and the weights for groups of similarities fusion are all integrated into the training process, refining P2G similarity accuracy and optimizing the re-identification retrieval results.

2. The whole idea of the algorithm

For a probe image (P) and some gallery images (G), we can measure the similarity score (P2G similarity) between the probe image and the gallery images. Then, according to the P2G similarity score, the gallery images are ranked to get the final result. The algorithm we proposed, which guided by G2G similarity information, introduces G2G similarity information to optimize P2G similarity, and in the way of horizontal feature grouping to automatically perceive the importance of key location features, to achieve the re-identification in higher precision. As is shown in figure 1.

![Figure 1. The optimization of P2G similarity guided by G2G similarity.](image)

The initial P2G and G2G similarity scores are obtained by the Siamese network based on the ResNet-50 backbone, and then optimized by the refined module. In the meantime, in order to make the utmost of the detail information in images, the features are horizontally overlapping grouped and the partial block similarities are measured and refined, to get optimized local P2G similarity score. Finally, the partial similarity scores are fully integrated through the fusion module, and the key parts are automatically perceived, to get more accurate P2G similarity score and improve the precision of re-identification.

3. P2G optimization under the guidance of G2G similarity

The overall network consists paired similarity calculation module, refined module and fusion module. At first, a pair of images are sent into the paired similarity calculation module to get the image features and a similarity score between them. In the paired similarity calculation module, we adopt the backbone of ResNet-50 for the Siamese network. The original global pooling layer and the fully connected layer are removed to obtain the images features size as 7*7*2048. In order to make the utmost of the detailed features of each part of the image, the feature maps are horizontally overlapped divided into 3 groups of upper, middle and lower features size as 3*7*2048.

The grouped features of two images are subtracted and squared by element, and then normalized. To get the similarity score, after the batch normalization, we deal with the features by a fully connected layer. We can get the initial P2G similarities between a probe and \( n_g \) gallery images \( y_{k}^{(0)} \in \mathbb{R}^{n_g} \) and G2G similarities \( S_k \in \mathbb{R}^{n_g \times n_g} \), for \( k = 1, 2, 3 \), which are estimated by the paired similarity calculation module. And then we can get 3 sets of refined P2G similarities by the refined module. After 3 groups of features similarities fusion module, the final P2G similarities is optimized.
### 3.1. P2G similarity update algorithm

From the paired similarity calculation module, we can get all P2G similarities and G2G similarities. Suppose that the similarity vector between the probe image and all gallery images is $\bar{Y} \in \mathbb{R}^n$, and $S \in \mathbb{R}^{n \times n}$ represents the normalized G2G similarity score matrix. In order to prevent the self-enhancement of the similarity score during the iterative process, $S(i,i) = 0$.

At the beginning, the similarity score between the probe image and the $j$th gallery image is $y_j$. For the probe image, the score $y_j$ after passing through the $k_{th}$ gallery image is:

$$y_j = S_{jk} y_k \quad (1)$$

In order to distinguish the influence of sample difficulty on similarity, we introduce adaptive weight $\tilde{a} = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T$ to balance the different roles of different positive(p) and negative(n) samples in similarity updating:

$$\alpha_k = \begin{cases} e^{y_k} / \sum_{i \in p} e^{y_i}, & k \in p \\ e^{-y_k} / \sum_{i \in n} e^{-y_i}, & k \in n \end{cases} \quad (2)$$

As shown in figure 2, from the probe image to the $j$th gallery image, the score $y_j$ passes through all intermediate gallery images and becomes:

$$y_j^{(1)} = \sum_{k=1}^{n} S_{jk} \alpha_k y_k^{(0)} \quad (3)$$

![Diagram](image)

Figure 2. The process of similarity iteration.

Then, after $t+1$ times of iteration, the similarity score for the probe is combined into a new vector:

$$\bar{Y}^{(t+1)} = S(\tilde{a} \circ \bar{Y}^{(t)}) \quad (4)$$

When the similarity score $S(i,j)$ becomes larger, the influence on $y_j$ will be greater. In order to ensure that the refined P2G similarity is not too far from the initial P2G similarity, a weighting factor $\lambda \in [0,1]$ is added to balance the updated P2G similarity and the initial P2G similarity:

$$\bar{Y}^{(t+1)} = (1 - \lambda) S(\tilde{a} \circ \bar{Y}^{(t)}) + \lambda (\tilde{a} \circ \bar{Y}^{(0)}) \quad (5)$$

Extend $t$ to infinity, and finally get the similarity score between the probe and gallery images:

$$\lim_{t \to \infty} \bar{Y}^{(t)} = \left[ \frac{1}{\lambda} E - \frac{(1-\lambda)}{\lambda} S \right]^{-1} (\tilde{a} \circ \bar{Y}^{(0)}) \quad (6)$$

Where $E$ donates the identity matrix. Then Least Mean Square method is adopted to optimize the result from the inverse operation of the matrix and get more accuracy score:

$$H = \lim_{t \to \infty} \bar{Y}^{(t)} + 2\varepsilon \left[ \frac{1}{\lambda} E - \frac{(1-\lambda)}{\lambda} S \right]^{-1} e \quad (7)$$

In which $H$ represents the final similarity score vector, $\varepsilon = 0.5[trace(A^T A)]^{-1}$ represents the step size of update, and $e = \left[ \frac{1}{\lambda} E - \frac{(1-\lambda)}{\lambda} S \right] \lim_{t \to \infty} \bar{Y}^{(t)} - (\tilde{a} \circ \bar{Y}^{(0)})$ represents the error of similarity score.

The refined module can be integrated into the network as a neural layer, and trained simultaneously with the network. Therefore, the update process of the similarity information from G2G to P2G is realized, which not only refines the P2G similarity, but also the similarity gradient information of the
gallery images is propagated to the front by the S matrix in the backward propagation. And this provides more supervision from the gallery image for the model.

3.2. Groups feature similarities fusion module
Because the global features of the image cannot exhaustively represent the pedestrian’s details, the network may miss many details of the target image. Therefore, we adopt horizontal overlapping grouping to divide the visual features of the image into three group. Each group of features is used to calculate and optimize the similarity score separately to strengthen the influence of detail information on the similarity measurement.

Considering each group of the similarity has its own advantages, but not comprehensive. After the similarity optimization of each group, we introduce a fusion module, fusing the three sets of similarity scores and training the weights through back-propagation. In this way, the image detail feature information is more fully utilized, and the groups of similarities fusion module is shown in figure 3.

![Figure 3. Groups of similarities fusion module.](image)

4. Experiments and results analysis

4.1. Experimental data set and conditions
In order to investigating the availability of the proposed algorithm in pedestrian re-identification, we conducted experiments and comparative analysis on the Market-1501 dataset and CUHK03 dataset.

The CUHK03[^12] dataset was gathered from two different cameras in the Chinese University of Hong Kong. There are 14,097 images for 1,467 people, taken from two different cameras. The Market-1501[^13] dataset was gathered from six different cameras in Tsinghua University. It contains a total of 32,668 images for 1501 different people. There are 12,936 images of 751 people in the training set, and 19,732 images of 750 people in the test set.

The experiment is based on the RTX 2070 graphics card server, under the Anaconda environment, using the Pytorch open source framework. In the experiment, the paired similarity calculation module was based on the ResNet-50 backbone, which was pre-trained on the ImageNet dataset. We reshaped the input images into size of 256x128. We randomly crop and horizontal flip to expand the dataset. In experiment, λ is set to 0.05. We adapt the stochastic gradient descent method to achieve end-to-end training for pedestrian re-identification. The batch size is 32, and 8 pedestrians for each batch. For each pedestrian, one image was selected for the probe image randomly, and the other three instants are gallery images. Thus, 24 gallery images are shared for each probe. In actual training, the initial learning rate was set as $10^{-6}$. The learning rate was adjusted to 0.1 times of the previous one after every 100 iterations, and the training is terminated after 1000 iterations. In the test stage, for the purpose of evaluation for the performance of the model in learning features, query images and gallery images are put into the trained paired similarity calculation module to estimate the similarity score of P2G, and then retrieval pedestrian images are sorted by comparing the similarity scores.

4.2. Analysis of experimental results
In the paper, we use Mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) as indicators for evaluating the performance of the model.

We use rank-1, rank-5, and rank-10 scores to represent the cumulative matching characteristic curve CMC, to reflect the ranking accuracy. According to the sorted P2G similarity score, we can get the number of tests with the right label at rank $1$, $N_1$, and the total number of test samples $N_t$. The
Rank-1 rate means the ratio of $N_1$ to $N_t$. The Rank-5 rate indicates the percentage of the top five items that match the best. If the degree of similarity matching is ranked from large to small, the correct label can be matched in the head of the ranking results, indicating that the model performs better.

The mAP (mean average precision) means the average value of the AP (average precision) in all query images, which reflects the precision for pedestrian retrieval results.

As is shown in Table 1 and Table 2, we compare the accuracy in this method with other algorithms on the Market1501 dataset and CUHK03 dataset to show the performants of them.

**Table 1. Accuracy comparison of different algorithms on Market-1501 dataset**

| Models    | mAP  | Rank-1 | Rank-5 | Rank-10 |
|-----------|------|--------|--------|---------|
| LSTM      | 35.3 | 61.6   |        |         |
| Spindle Net | -    | 76.9   | 91.5   | 94.6    |
| TriHard   | 69.1 | 84.9   | 94.2   |         |
| GLAD      | 73.9 | 89.9   | -      | -       |
| K-reciprocal | 63.6 | 77.1   | -      | -       |
| AlignedReID | 79.3 | 91.8   | -      | -       |
| SSM       | 68.8 | 82.2   | -      | -       |
| PSE+ECN   | 84.0 | 90.3   | -      | -       |
| ours      | 82.5 | 92.3   | 97.1   | 98.3    |

**Table 2. Accuracy comparison of different algorithms on the CUHK03 dataset**

| Models    | mAP  | Rank-1 | Rank-5 | Rank-10 |
|-----------|------|--------|--------|---------|
| LSTM      | -    | 57.3   | 80.1   | 88.3    |
| Spindle Net | -    | 88.5   | 97.8   | 98.6    |
| Quadruplet | -    | 74.4   | 96.6   | 98.9    |
| GLAD      | -    | 82.2   | 95/8   | 97.6    |
| K-reciprocal | 67.6 | 61.6   | -      | -       |
| AlignedReID | -    | 92.4   | 98.9   | 99.5    |
| SSM       | -    | 76.6   | 94.6   | 98.0    |
| ours      | 92.0 | 94.3   | 98.7   | 99.3    |

As is shown in Table 1 and Table 2, the method we proposed performs better than others on the evaluation indicators. Compared with local or global features from images, realizing pedestrian re-identification by classification loss or metric loss, such as LSTM, AlignedReID, Spindle Net, GLAD shown in the table, our method has obvious advantages in mAP and Rank-1 retrieval accuracy. Similarly, the accuracy of Rank-1 is significantly better than that of the simple similarity measure methods, like the TriHard and the Quadruplet. Compared with the re-rank methods SSM and K-reciprocal for retrieval results, there is also a visibly improvement in the mAP and Rank-1.

Figure 4. Examples of experiment result.
The test result of the model on the Market-1501 is shown in figure 4. In the figure the first column is the query image, and the retrieved images are sorted according to their similarity score from high to low in the right of the figure. The retrieval image number of the same person as the query image, is marked with green, otherwise marked with red. It is shown in the figure that the retrieval results of the algorithm in this paper are reliable and efficient.

5. Conclusion
In this paper, we have proposed a pedestrian re-identification algorithm with G2G similarity guidance to update P2G similarity. It makes the utmost of the similarities between gallery images (G2G similarity) information, optimizes the similarity score between probe images and the gallery images (P2G similarity), and obtains more accurate similarity results, to improve pedestrian re-identification and retrieval results. At the same time, we divide the image features into 3 horizontal overlapping groups to make the utmost of the details of the image features, which are introduced into the training process to train the weights and refine the similarity scores. The experiment results show that the method we proposed is obviously superior to the re-identification methods that only takes P2G similarity into consideration, which proves the availability of the proposed algorithm in this paper. Next, we are considering to combine the similarity of horizontal overlapping group features with that of global features, to improve the accuracy of the model.

References
[1] Liao S, Hu Y, Zhu X. (2015) Person re-identification by Local Maximal Occurrence representation and metric learning. In: computer vision and pattern recognition. pp. 2197-2206.
[2] Wu S, Chen Y, Li X, et al. An enhanced deep feature representation for person re-identification[J]. workshop on applications of computer vision, 2016: 1-8.
[3] Farenzena M, Bazzani L, Perina A. (2010) Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2360-2367.
[4] Varior R R, Shuai B, Lu J. (2016) A Siamese Long Short-Term Memory Architecture for Human Re-identification. In: European conference on computer vision. pp. 135-153.
[5] Wang Y, Chen Z, Wu F. (2018) Person re-identification with cascaded pairwise convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1470-1478.
[6] Hermans A, Beyer L, Leibe B, et al. In Defence of the Triplet Loss for Person Re-Identification[J]. arXiv: Computer Vision and Pattern Recognition, 2017.
[7] Chen W, Chen X, Zhang J. (2017) Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification. In: computer vision and pattern recognition. pp.1320-1329.
[8] S. Bai, X. Bai, and Q. Tian. (2017) Scalable person re-identification on supervised smoothed manifold. In CVPR.
[9] Saquib Sarfraz M, Schumann A, Eberle A. (2018) A pose-sensitive embedding for person re-identification with expanded cross neighbourhood re-ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 420-429.
[10] Zhong Z, Zheng L, Cao D. (2017) Re-ranking Person Re-identification with k-Reciprocal Encoding. In: computer vision and pattern recognition. pp. 3652-3661.
[11] Bai S, Tang P, Tor P H S. (2019) Re-Ranking via Metric Fusion for Object Retrieval and Person Re-Identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 740-749.
[12] W. Li, R. Zhao, T. Xiao, and X. Wang. (2014) Deepreid: Deep filter pairing neural network for person re-identification. In CVPR. pp. 152–159.
[13] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. (2015) Scalable person re-identification: A benchmark. In ICCV. pp. 1116–1124.