Multi-scale Salient Features for Person Re-identification

Yachao Dong¹ and Hongzhe Liu¹

¹ Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing, 100101, China

*Corresponding author’s e-mail: liuhongzhe@buu.edu.cn

Abstract. Person re-identification is a technology that uses computer vision to match and retrieve the same target person across cameras, which has a wide range of applications in the fields of intelligent video surveillance, intelligent security, and unmanned driving. Aiming at the problem of low accuracy caused by complex background information, occlusion and other factors in person re-identification, the extracted features are not discriminative and robust, resulting in low accuracy, a person re-identification method based on multi-scale saliency features is proposed. Firstly, the saliency detection is used to extract the distinctive feature areas with discriminative power in person, so that the features of distinguishing person images are more prominent. Secondly, these saliency features are merged with global features and the fused features are divided into different scales, and the global features are divided into three different scales to enhance the expression of features. Finally, according to the difference of global and local features, three kinds of loss functions are combined for learning. In the inference stage, the global feature and the local feature are fused into a new feature vector for metric learning. The proposed method has been used in the person re-identification public datasets Market1501 and DukeMTMC-reID to do a lot of effective verification experiments. The experimental results show that the features extracted by the proposed method have strong distinguishability and robustness.

1. Introduction

Person re-identification is the process of matching the same target person in different camera perspectives, and it plays an important role in traffic, public security and video surveillance. As shown in Figure 1, the difficulty of person re-identification lies in the difference in viewing angle and illumination produced by different cameras, as well as unfavourable factors such as occlusion and complex background, which lead to huge changes in the posture and appearance of person. These unfavourable factors will affect the accuracy of person re-identification.

![Difficult samples in person re-identification](image)

Figure 1. Difficult samples in person re-identification
The general process of the deep learning method to solve the task of person re-identification is to first use the convolutional neural network to extract the characteristics of each picture, and then measure the distance similarity between the characteristics of the checked person picture and the picture characteristics in the base library one by one. According to the distance, sort the top K pictures that are most similar to the checked person.

Sun et al. [1] proposed part-based convolutional baseline (PCB), which means that the extracted features are uniformly divided into six local features in the vertical direction and the person identity information is predicted separately. Wang et al. [2] proposed a multi-granularity network, which divides features into three branches, combines global features and multi-scale local features, and uses multiple classification losses and triplet loss joint learning. Finally, all the features are combined as the person feature representation. Deng et al. [3] used GAN (Generative Adversarial Networks) to convert the image style of the source domain to the style of the target domain while maintaining the identity of the original person, and then fine-tune the model. Li et al. [4] used GAN to transfer the person's posture, so that the model learns features that are independent of the person's posture. Jin et al. [5] adopted a plug-and-play style normalization and stylization framework to normalize different styles such as different cameras, different illuminations, and different resolutions to reduce the difference between the source domain and the target domain. The retrieval performance of these methods highly depends on the ability of domain conversion, such as the quality of images generated by GAN. Wei et al. [6] used the extracted key points of the human body to divide the picture into three parts: head, upper body and lower body to extract features.

Although these methods improved the accuracy of person re-identification to a certain extent, there is still a problem that it is difficult to extract robust features. To solve this problem, this paper proposes a person re-identification method based on multi-scale saliency features.

2. Proposed approach

The overall network structure and implementation process of the proposed method are shown in Figure 2.

![Figure 2. Framework of the proposed method model. SG-feat is the fusion of Global Feat and Saliency Map. Cut means to divide global features into multi-scale features.](image)

The overall network structure is composed of saliency detection, backbone network, dimensionality reduction layer, fully connected layer and multiple loss functions. The feature extraction backbone network uses ResNet50 [7], and uses ImageNet as the pre-training model. The difference is that this paper has made the following modifications to ResNet50: 1) Removed the average pooling layer and
full pooling layer after layer 4 in ResNet50. It has been changed to a maximum pooling layer corresponding to different branches and a fully connected layer that adapts to different scale features. 2) The Refined Layer is added after layer 3 in ResNet50, which consists of constructing the Residual Bottleneck in ResNet50. On the one hand, it is to obtain a larger feature map, and on the other hand, it is to unify the feature dimension and better integrate the Saliency Map. 3) A dimensionality reduction layer is added in front of the changed fully connected layer, which is composed of a two-dimensional convolution with a $2 \times 1$ convolution kernel, batch normalization processing, and the ReLU activation function. The purpose is to reduce feature blocks of different scales. To the same dimension 512, it not only reduces the calculation of the parameter amount, but also alleviates the over-fitting of the network.

2.1. Extract saliency map based on saliency detection

In person re-identification, the complex background global features are not enough to judge that two people are the same person, because the global features do not effectively highlight the distinguishing local feature blocks. Besides, since most existing feature dicing methods regard each feature block as the same weight, important feature blocks should be weighted. As shown in Figure 3, since two persons are occluding each other, saliency detection can be used to extract the key person areas. Some existing feature block methods treat the weight of each block equally, but after the saliency feature map, the 2, 3, and 4 feature blocks can be adaptively given greater weights.

![Figure 3. Framework of the proposed method model](image)

The current saliency detection algorithm is becoming more and more mature. We used F3Net [8] in the person re-identification datasets, which performs well. The difference is that the person re-identification datasets have low picture pixels and the effect is not obvious. This paper pre-processes the pictures and uses bilinear interpolation to up-sample them to adapt to the person re-identification datasets. Besides, we do not directly input the saliency feature map into the network for calculation, but converts it into a binarized grayscale image, and then normalizes it to be stored in the form of a feature vector. Finally, these feature calculation graphs are multiplied and fused with the middle layer layer-3 of ResNet50 by vector matrix multiplication. In this way, there will not be too much abstract information and the original feature map is also adaptively weighted.

2.2. Joint learning with multiple loss functions

The focus of multi-scale features lies in the loss function, and appropriate loss functions need to be applied according to the features of different scales. For this reason, this paper combines multiple loss functions and applies them to multiple scale feature blocks to train the data set and optimize the network model. General person re-identification loss functions include cross-entropy loss, triplet loss,
and so on. In cross-entropy loss, we used person ID as the classification category. The effect of cross-entropy loss is to increase the distance between classes. Which is defined as:

\[
L_{\text{cross}} = \sum_{i=1}^{K} -q_i \log(p_i) \begin{cases} 
q_i = 0, y \neq i \\
q_i = 1, y = i 
\end{cases}
\]

(1)

Where K is the number of person IDs, y is the truth label, and p_i is the probability value of the model predicted category i. However, since the IDs appearing in the test set do not appear in the training set, not all newly added IDs in the training classification network can be roughly judged by 0 or 1, and this learning method is likely to cause over-fitting. Therefore, this paper adds the idea of label smoothing, and changes the in formula (3-1) to:

\[
q_i = \begin{cases} 
1 - \frac{K-1}{K} \delta, y = i \\
0, y \neq i 
\end{cases}
\]

(2)

Where \(\delta\) is a set constant, which is set to 0.1 to prevent the trained model from trusting the training set too much. Therefore, the ID loss used is combined with formulas (1) and (2) to obtain the following formula (3):

\[
L_{ID} = \sum_{i=1}^{K} -q_i \log(p_i) \begin{cases} 
q_i = \delta / K, y \neq i \\
q_i = 1 - \frac{K-1}{K} \delta, y = i 
\end{cases}
\]

(3)

The role of Triplet Loss not only increases the distance between classes, but also narrows the distance within classes, which is defined as:

\[
L_{trip} = \sum_{j=1}^{B} \sum_{i=1}^{N} \left[ \alpha + \max_{p=1..P} \|f_p^{(i)} - f_p^{(j)}\|_2 - \min_{j\neq i} \|f_p^{(i)} - f_p^{(j)}\|_2 \right] + \sum_{j=1}^{B} \sum_{i=1}^{N} \left[ \|f_j - c_{y_j} \|_2^2 \right]
\]

(4)

Among them, \(f_i^a\), \(f_i^p\), \(f_i^n\) represent the anchor, and the positive and negative samples corresponding to the anchor. P represents the number of IDs in each training batch, N represents the number of pictures with the same ID, \([\varphi]_+\) represents \(\max(\varphi, 0)\), and the margin of Triplet Loss is set to 1.2 here. To make up for the shortcomings of Triplet Loss, we introduce a central loss function, which is defined as:

\[
L_{center} = \frac{1}{2} \sum_{j=1}^{B} \|f_j - c_{y_j} \|_2^2
\]

(5)

Where B is the minimum number of images in batches, \(f_j\) is the feature vector of the j-th image, \(y_j\) is the label of the j-th image, and \(c_{y_j}\) is the \(y_j\)-th class center of the depth feature. It can be seen that it can effectively describe the intra-class changes. Increase compactness within the class. The overall loss function is defined as:

\[
L_{total} = L_{ID} + L_{trip} + L_{center}
\]

(6)

3. Experiments

In order to verify the effectiveness of the method proposed in this paper, experiments were conducted on two public person re-identification data sets, Market1501 [9] and DukeMTMC-reID [10]. The experimental evaluation standard uses mAP (mean Average Precision), Rank-1, Rank-5, and Rank-10 as comparative indicators. The experimental results are shown in Table 2 and Table 3.
Table 2. Performance comparison with the existing methods on DukeMTMC-reID dataset (%).

| Methods         | mAP | Rank-1 |
|-----------------|-----|--------|
| PSE+ECN [11]    | 62.0| 79.8   |
| HA-CNN [12]     | 63.8| 80.5   |
| Mancs [13]      | 71.8| 84.9   |
| HPM [14]        | 74.3| 86.6   |
| Pyramid [15]    | 79.0| 89.0   |
| Ours            | 85.1| 89.1   |

From the Table 2 and Table 3, we can see that the method proposed in this article has superiority compared with the existing methods. In order to show more intuitively that the proposed method still has a good re-recognition effect in complex scenarios, this article visually displays part of the query results on the Market-1501 and DukeMTMC-reID data sets, as shown in Figure 4 and Figure 5. It can be seen from the result graph that the method proposed in this paper can re-identify the same person under complex conditions.

Figure 4. Visualization of the results of the proposed method on Market1501. The solid line box is the correct search, the dashed line box is the wrong search.

Figure 5. Visualization of the results of the proposed method on Market1501. The solid line box is the correct search, the dashed line box is the wrong search.
4. Conclusion
This paper starts from the problem of the inability to extract robust features due to complex background and occlusion, and proposes a person re-identification method based on multi-scale saliency features. And through the comparison with the existing methods and the analysis of the visual experiment results, the effectiveness of the method proposed in this paper is proved.

Acknowledgments
Thank you very much for my supervisor's guidance and help on my paper. I am very grateful to the National Natural Science Foundation of China (Grant No. 61871039, 61906017, 61802019), the Beijing Municipal Commission of Education Project (No.KM202111417001, KM201911417001), and the Beijing Union University Graduate Research and Innovation Funding Project (Grant No. YZ2020K001) for funding my paper.

References
[1] Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European conference on computer vision (ECCV). pp. 480-496.
[2] Wang, G., Yuan, Y., Chen, X., Li, J., & Zhou, X. (2018, October). Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM international conference on Multimedia. pp. 274-282.
[3] Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., & Jiao, J. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 994-1003.
[4] Li, Y. J., Lin, C. S., Lin, Y. B., & Wang, Y. C. F. (2019). Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7919-7929.
[5] Jin, X., Lan, C., Zeng, W., Chen, Z., & Zhang, L. (2020). Style normalization and restitution for generalizable person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3143-3152.
[6] Wei, L., Zhang, S., Yao, H., Gao, W., & Tian, Q. (2017, October). Glad: Global-local-alignment descriptor for pedestrian retrieval. In Proceedings of the 25th ACM international conference on Multimedia. pp. 420-428.
[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778.
[8] Wei, J., Wang, S., & Huang, Q. (2020, April). F³Net: Fusion, Feedback and Focus for Salient Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34, No. 07, pp. 12321-12328.
[9] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In Proceedings of the IEEE international conference on computer vision. pp. 1116-1124.
[10] Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 152-159.
[11] Sarfraz, M. S., Schumann, A., Eberle, A., & Stiefelhagen, R. (2018). A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 420-429.
[12] Li, W., Zhu, X., & Gong, S. (2018). Harmonious attention network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2285-2294.
[13] Wang, C., Zhang, Q., Huang, C., Liu, W., & Wang, X. (2018). Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV). pp. 365-381.

[14] Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., ... & Huang, T. (2019, July). Horizontal pyramid matching for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, No. 01, pp. 8295-8302.

[15] Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., ... & Ji, R. (2019). Pyramidal person re-identification via multi-loss dynamic training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8514-8522.