Pedestrian Recognition Based On Improved Semantic Segmentation Neural Network

Xiucai Guo, Yaodong Wang*
Xi'an University of Science and Technology, Shaanxi 710054, China
*Corresponding author e-mail: 1694480194@qq.com

Abstract. At present, the existing Re-ID methods can't align the target and the target in the image that should to be detected, which has been affecting the recognition accuracy. Although using semantic segmentation method can solve this problem, it brings new problems. The region which is considered as background by semantic segmentation neural network and discarded is likely to contain feature representations that are helpful for recognition. In this paper, an improved semantic segmentation method is proposed to re-divide the foreground and background regions. In order to avoid the influence of semantic segmentation errors on recognition accuracy, a new method is used to combine background and foreground regions. Experiments show that the accuracy of Re-ID is 84.52%, which is 3.33% higher than before.

1. Introduction
Re-ID refers to automatically finding other images of the pedestrian in the monitoring system according to the input target pedestrian image. When a query image is provided, all pictures with the same identity will be retrieved in the picture library through Re-ID technology [1]. The pictures in these libraries are basically taken by different cameras, and the shooting areas of some cameras are not overlapped, so Re-ID can also be considered as the problem of cross camera data association.

Due to the location and environment changes of camera deployment in real life, the pedestrian image changes in occlusion, lighting and pedestrian posture, which poses a great challenge to pedestrian recognition algorithm.

In order to achieve better recognition, it is necessary to align the target and the image to be detected. However, because the camera definition and the targets are sometimes far away from the camera, it is impossible to align the target and the object to be detected just like face recognition.

Many researches have explored the problem of alignment, such as lomo [2] Based on the assumption that the pedestrian image keeps the trunk upright, the pedestrian image is divided into several strip image blocks, and the most significant color texture histogram feature of each strip image block is used to overcome the change of perspective. However, due to the change of pedestrian posture, the upright state of pedestrian image is not always guaranteed, which leads to different pedestrian areas described by the same dimension features; and the background information will bring great interference to the description of features. M.kostinger et al. Put forward a method called kiss metric learning [3], which assumes that the characteristic difference between two samples satisfies the Gaussian distribution with the mean value of zero, transforms the objective function into the form of log likelihood ratio, and directly obtains the mapping matrix m according to the samples, thus
avoiding the complex optimization process in other measurement learning work, greatly improving the learning speed. The disadvantage of this method is that it discards some features after feature dimensionality reduction and needs to pair each image, which is too expensive. The above analysis reflects the shortcomings of the current algorithm, including the instability of feature description, and feature matching can’t fundamentally solve the problem of spatial dimension migration.

Therefore, this paper uses the method of semantic segmentation, because it can establish the accuracy and performance of the pixel level of any contour model, so it can effectively solve the problem of background interference and alignment.

2. Improved Semantic Segmentation Method

2.1. Semantic Segmentation Method Introduction

Semantic segmentation can be divided into region based image semantic segmentation and pixel based image semantic segmentation [4].

The method of image semantic segmentation based on region classification is to combine the traditional image processing algorithm with DNN. First, the original image is divided into different target candidate regions, and a series of image patches are obtained. Then, DNN is used to classify each pixel in the image block or image block. Finally, the original image is labeled according to the classification results, and the final segmentation results are obtained. Although the method of image semantic segmentation based on region classification has achieved good results, there are some problems, such as the low precision of image segmentation and the low speed of image segmentation.

The principle of image semantic segmentation method based on pixel classification is: in an image, some spatial information characteristics (such as gray level, color, contour, texture, etc.) in the target region are usually different from the surrounding background. According to these differences, semantic segmentation technology divides the image into several specific and unique regions, so as to effectively segment the foreground region out. Moreover, the image samples annotated manually and accurately can provide a lot of detail information and local features, which is conducive to improving the efficiency of network training and segmentation accuracy.

Therefore, by using the semantic segmentation method to deal with pedestrians, we can establish the model of any contour, and get the accuracy and performance of pixel level, which can effectively solve the problem of background interference and alignment.

Firstly, the application of semantic segmentation algorithm to Reid was proposed by Mahdi M. kalayeh et al. In 2018. They first put Re-ID image into the trained semantic segmentation model and perception V3 model, respectively, to get the segmented head, upper body, leg, foot and the whole five parts and the image features extracted by the common method, and then combine them to get new features, so as to improve Re-ID Accuracy [5].

2.2. Improved Semantic Segmentation Method

In the same year, Di Chen et al. Considered that part of the background is helpful for Re-ID [6], so they first use the detection method to detect people from the pictures, at this time, some background is attached to the results. Then we use the method of segmentation to separate people from the background. Then we combine people with some background to get a new image, which is called foreground.

However, although it trains part of background information as recognition information and achieves good results, there is no more discussion on this part of background information.

However, if a picture shows a woman carrying a bag, and it needs to identify two women who are almost the same, but one of them carries a similar bag, then the similarity between this woman and the woman in the previous picture is far greater than that of the other woman who does not carry a bag.

In this paper, we will do further research on the objects that are divided into backgrounds in semantic segmentation.
Let the whole effective feature area as M. The traditional segmentation method takes bicycle, motorcycle, handbag and backpack as the background, but this new method thinks that they are useful and let them as y. The part of pure pedestrian is x, and the part of pedestrian surrounding background is z.

\[ M = x + cy + \beta z \] (1)

Because the accuracy of semantic segmentation method is not high enough, it can not completely depend on the segmented image. Compared with the high-resolution image used in semantic segmentation, the resolution of pedestrian recognition is much lower. Therefore, it is necessary to eliminate the negative effects of semantic segmentation errors through the superposition of regional features. The effect of the new semantic segmentation method is shown in Figure 1.

![Figure 1. Improved semantic segmentation method.](image)

2.3. Semantic segmentation network
Because DFN [6] network has higher mIoU accuracy and lower GPU resource consumption, this paper uses it for experiments.

The image is converted to 512 × 512 size by image preprocessing, and then the global feature representation is obtained by the encoder part composed of ResNet [7], and then the decoding operation is carried out layer by layer, finally the predicted image is obtained.

The whole network is divided into two parts: a backbone composed of ResNet101 and a decoder network for combining high-level features and low-level features and recovering image size. After that, the collected and marked training set and test set pictures will be sent to the neural network for training. In order to reduce over fitting, instead of using the way of saving the network after the whole training, the test set will be used to test the current neural network after each training cycle, and the test accuracy will be saved at the same time. If the test result is the historical maximum, the neural network and network will save parameters. When the set number of training is reached, the training and test are finished.

2.4. Re-ID network
For the wrong part of semantic segmentation, global representation learning can be added to the network to correct it. That is to say, in the process of segmentation, the pedestrian and background are separated first, and the output obtained is input into two networks. One is to continue the more detailed segmentation, and then the target and data set are matched one by one to learn the effective part of the representation; the other is to extract the characteristics of the segmented pedestrian, learn the effective global representation, and then use the fusion network to extract the effective global representation. Finally these two parts are combined in proportion.
3. Experimental Process And Results

3.1. Dataset introduction
CCP dataset [8] and Pascal dataset [9] are used as training sets of semantic segmentation network. In order to achieve segmentation effect, the marks of these two datasets have been modified to some extent. Moreover, since the image clarity of CCP dataset and Pascal dataset is much higher than that of maket-1501 [10], which will be used for Re-ID, Gaussian blur is used to simulate in the image preprocessing stage to picture’s definition lower. The picture is as follows

![Figure 3. The left figure is taken from CCP dataset, and the right figure is taken from Pascal dataset.](image)

Using maket-1501 as the data source of Re-ID, the market-1501 dataset was collected on the campus of Tsinghua University, photographed in summer, constructed and published in 2015. It includes 1501 pedestrians captured by six cameras, including five high-definition cameras and one low-definition camera. As shown in the figure below:

![Figure 4. Market-1501 dataset.](image)

3.2. Experimental Result
In order to better analyze the training process and results, after each training cycle, the training results of this training cycle are saved in tensorboard.

In order to verify the effectiveness of the improved algorithm, it is compared with pedestrian recognition network and neural network model using pre optimization semantic segmentation network.
In order to ensure the consistency of training conditions, the parameters of all neural networks are set the same.

![Graph showing accuracy rate of Re-ID before and after optimization]

Figure 5. Accuracy rate of Re-ID before optimization rank-1 and after optimization rank-1

In the same case, before using the optimization algorithm, the accuracy of Re-ID is 81.8%, after using the optimization algorithm, the accuracy of Re-ID is 84.52%, 3.33% higher than before. So the improved algorithm can improve the accuracy of Re-ID.

4. Conclusion
In view of the problem that the area which is considered to be the background to be abandoned in the current Re-ID neural network based on the semantic segmentation method is likely to contain the feature representation which is helpful for recognition, this paper proposes to re divide the foreground and background areas by improving the semantic segmentation method, and use a new way to combine the background and the foreground area. This method can effectively improve the recognition accuracy. The test results show that this method can improve the accuracy of pedestrian recognition. After using the optimization algorithm, the accuracy of Re-ID is 84.52%, 3.33% higher than before.

References
[1] Zheng, Liang, Yi Yang, and Alexander G. Hauptmann. "Person re-identification: Past, present
and future." arXiv preprint arXiv:1610.02984 (2016).

[2] Liao, Shengcai, et al. "Person re-identification by local maximal occurrence representation and metric learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[3] Koestinger, Martin, et al. "Large scale metric learning from equivalence constraints." 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012.

[4] Tian Xuan, Wang Liang, and Ding Qi. "An overview of image semantic segmentation based on deep learning." Journal of software 2 (2019): 14

[5] Kalayeh, Mahdi M., et al. "Human semantic parsing for person re-identification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[6] Yu, Changqian, et al. "Learning a discriminative feature network for semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[7] Chen, Di, et al. "Person search via a mask-guided two-stream cnn model." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[8] Yang, Wei, Ping Luo, and Liang Lin. "Clothing co-parsing by joint image segmentation and labeling." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[9] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303–338, 2010.

[10] Kalayeh, Mahdi M., et al. "Human semantic parsing for person re-identification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.