Research on Person Reidentification Method Fusing Direction Information in Multi-camera Pedestrian Tracking Problem

Junliang Liu*, Xiangbin Shi

College of Computer Science, Shenyang Aerospace University, Shenyang, Liaoning Province, 110136, China

*Corresponding author's e-mail: 182106074003@email.sau.edu.cn

Abstract. In multi-camera pedestrian tracking, pedestrian apparent features are usually used to solve the problem of cross-camera reidentification. In order to make the pedestrian apparent features robust to the directional changes of pedestrians in different cameras. In this paper, a pedestrian feature extraction method combining directional features is proposed by using deep network Resnet50. Pedestrian feature extraction network is trained by adding pedestrian direction information. In order to improve the performance of person reidentification model, BNblock structure is added in the design, a batch normalization (BN) layer is added after the feature to get the normalized feature. It can make the triplet loss converge with the convergence of the ID loss, thereby improving the performance of the model. The proposed method was verified on DukeMTMC and Market-1501 data sets, and the results show that the pedestrian appearance features combined with direction information proposed by the pedestrian reidentification algorithm can significantly improve the reidentification accuracy. By adding BNblock into the network structure, the accuracy can be further improved.

1. Introduction

Multi-camera pedestrian tracking has excellent research value in the field of intelligent surveillance. Its purpose is to carry out pedestrian target handover in the camera network under the premise of single-camera target tracking to complete the complete tracking of pedestrian trajectories. The transfer and association of pedestrian identities between cameras are the critical points for multi-camera pedestrian tracking.

Research scholars refer to the problem of identifying the relevance of pedestrian target identity in non-overlapping view area monitoring network as the Person Re-Identification (Re-ID). There are still many challenges in the practical application of Re-ID. In actual monitoring scenes, cameras are usually affected by factors such as low resolution and unfixed viewing angles. Therefore, it is less feasible to re-identify pedestrian targets using human biological features such as human face[1] and gait[2]. At present, most of the current researches on the re-identification of pedestrian targets [3-5] regard the appearance feature of pedestrians as the main basis for Re-ID. However, there are many difficulties in re-identification based on appearance: Changes in illumination, viewing angle, scale, and camera parameters between different cameras, and pedestrians also have changes in posture. Therefore, the appearance of the same pedestrian target often shows huge differences under different cameras. The top factor impacting multi-camera tracking is the angle change of the same pedestrian in different cameras.

Aiming at the problem of pedestrian direction angle changes that often occur in cross-camera tracking, this paper starts with the design of pedestrian appearance features, uses deep neural networks...
to design a triplet loss that integrates direction information, and proposes an innovative method for describing pedestrian appearance features.

In addition, in order to further improve the performance of the Re-ID model on the basis of the above innovations, we add the BNblock structure. A batch normalization (BN) layer is added after the features (before the FC layer of the classifier) to get the normalized features. In the training phase, the triplet loss and ID loss are calculated separately. BNblock can make the network have the following characteristics: the smaller the constraint of ID loss, the easier the triplet loss will converge at the same time. In this way, the performance of the Re-ID model can be greatly improved.

2. Related Work
Re-ID has been studied in academia for many years, but it has only made a huge breakthrough in recent years with the development of deep learning. Some of the current solutions focus on the structure of the deep network model, some on the construction of training data, and some on the loss function.

For the research on the network model structure, Guanshuo Wang et al. proposed a multi-branch deep network architecture, MGN, including one branch for global feature representation and two branches for local feature representation. Although this type of research method focusing on the network model structure can obtain a good feature extraction network, it generally causes the network structure to be too large, the parameters increase, and the speed decreases.

For the research on training data construction, the popular direction is to use GANS to generate training data. Zheng[6] et al. first introduced the use of unconditional GAN to generate images. Huang[7] et al. used WGAN to assign tags to generate pictures. CycleGAN[8] can transform the style of the pedestrian data set to get a new data set. Such methods make up for the lack of data in pedestrian re-identification.

For the research on the loss function, Hermans and Mischuk [9] proposed a batch-hard triplet loss with difficult sample mining. In their method, the most difficult positive samples and negative samples are considered, which increases the robustness of features, but the impact of changes in the direction of pedestrians has not been significantly improved.

3. Appearance Features Integrated with Directional Information
In this paper, the Resnet50 deep network is used to extract the pedestrian appearance features. In the process of feature extraction network training, the loss function plays an important role in the quality of the final trained network model. Hermans proposed a triplet loss, as in Equation 1.

$$\left[ \sum_{x_p \in P(a)} \omega_p d(x_a, x_p) - \sum_{x_n \in N(a)} \omega_n d(x_a, x_n) + \alpha \right]_+$$

Figure 1 shows the network after learning through the triplet loss will make the distance between the anchor point and the positive sample very small, and the distance from the negative sample is very large.
Although the triple loss is very effective, it also has disadvantages: First of all, the data distribution is not necessarily uniform due to the selection of triplet, so the model training process is very unstable and converges slowly, and need to constantly adjust the parameters. Secondly, the triplet loss is easier to overfit than the classification loss. Therefore, many scholars combine it with the Softmax function (classification loss).

However, for multi-camera tracking, the appearance change caused by pedestrian direction changes is the most likely problem. The above methods do not focus on the influence of direction changes on appearance characteristics. Due to the influence of the camera installation position, pedestrians appear on the front in some cameras, and on the back or side in other cameras, which will greatly affect the accuracy of the information on the appearance of pedestrians. In order to make up for this defect, we design a triplet loss function that integrates pedestrian direction information, as shown in Equation 2.

$$\frac{1}{N} \sum_{i=1}^{N} \left[ d(x_i^a, x_i^p) - d(x_i^a, x_i^n) + \alpha \right] + \gamma \sum_{i=1}^{N} \left[ d(x_i^a, x_i^{pd}) - d(x_i^a, x_i^{ns}) + \beta \right]$$

In the above equation, $x_i^a$ represents the anchor of picture $i$, $x_i^p$ represents the positive sample of picture $i$, and $x_i^n$ represents the negative sample of picture $i$. $x_i^{pd}$ indicates that the picture $i$ is a positive sample in a different direction, and $x_i^{ns}$ indicates that the picture $i$ is a different kind of negative sample in the same direction. $d(,)$ represents the Euclidean distance. $\alpha$ and $\beta$ are the minimum distances separating positive and negative sample pairs. According to the empirical value, set $\alpha=1$, $\beta=0.01$, $|a| = \max\{0,a\}$. A batch is N pictures. "$\gamma$" is a proportional relationship between the two parts of the loss function, here is 0.03. The purpose of the loss function is to make the distance between the same person in different directions become closer, and the distance between different people in the same direction becomes longer. The training method is shown in Figure 3.

After the detection frames are extracted, these detection frames are divided into four types: front, back, left, and right. Use the pedestrian joint point information (pose estimation) obtained by OpenPose to classify the detection frame, then use the ResNet50 deep network to train on the RAP data set [10] labeled with four directions to obtain a trained direction classification model. This model will be used to generate the method classification of the input OpenPose in Figure 3. For each anchor in each batch, randomly select its negative samples, positive samples in different directions, and negative samples in the same direction to form different triplets for training.

The directional loss function consists of two parts. One part comes from the triple constraint between the anchor, the positive and negative samples; The other part is the triplet loss of the same person in different directions, which distinguishes the distance from the anchor to the same direction of the positive sample and different directions of the positive sample. The features used by these triple loss functions are still extracted by the ResNet50. The total loss function is composed of w1 times the Softmax loss and w2 times the direction loss function, w1 =1, w2 =0.5. All training branches share parameters, and the Softmax loss is only applied to the branch where the anchor is located. Finally,
according to the experiment, it can be concluded that the appearance characteristics of pedestrians fused with direction information can significantly improve the re-identification of pedestrians.

4. BNblock

When training the Re-ID model, the commonly used method is to combine ID loss and triple loss for training, but in fact ID loss and triple loss are optimized for different purposes during the training process, as shown in Figure 4 (a, b). After the features are mapped to the hyperspace, the ID loss is aimed at optimizing the angle between the features. The smaller the angle of the same pedestrian and the larger angle of different pedestrians, the better. The triple loss is mainly aimed at optimizing the distance between features. The distance of the same pedestrian is as close as possible, and the distance of different pedestrians is as far away as possible. The triple loss is dominated by Euclidean distance, and the ID loss is dominated by cosine distance.

As shown in Figure 4(a), ID loss constructs a hyperspace, and uses the hyperplane to divide it into multiple subspaces, and each category embeds a different subspace. Because the cosine has a range, and the Euclidean distance has no range, in the inference stage, the cosine distance is more suitable for the ID loss optimization model than the Euclidean distance. But the disadvantage of this method is that the cohesion within the class is not strong. On the other hand, as shown in Figure 4(b): In Euclidean space, the advantage of triplet loss is that it enhances the cohesion intra-class and the separability inter-class. The disadvantage is that because the triplet loss cannot provide the global optimal constraint, the inter-class distance is sometimes smaller than the intra-class distance.

The current mainstream method is to combine ID loss and triplet loss to train the model, as shown in Figure 4(c). This method allows the model to better distinguish different types of features. It still has shortcomings: for the image features embedded in the space, the model trained by ID loss is mainly measured by the cosine distance in the inference stage, and the triple loss is mainly to optimize the Euclidean distance. If we use both losses at the same time, their goals may be inconsistent. During the training process, one loss may decrease, while the other loss may oscillate or even increase.

In order to overcome the above problems, we designed a structure called BNblock, adding the loss of BNblock can make the feature vectors distributed on the same hypersphere as much as possible, as shown in Figure 4(d), so that when the angles are similar, the distance will be similar, and the purpose of triple loss and ID loss training will be more consistent. As shown in Figure 2, BNblock adds a batch normalization (BN) layer after the feature (before the classifier FC layer). The feature before the BN layer is denoted as $f_i$. We let $f_i$ pass through a BN layer to acquire the normalized feature $f_i^\prime$. In the training stage, $f_i$ and $f_i^\prime$ are used to compute triplet loss and ID loss, respectively. Normalization balances each dimension of $f_i$. The features are gaussianly distributed near the surface of the hypersphere. This distribution makes the ID loss easier to converge. In addition, BNblock reduces the constraint of the ID loss on $f_i$. Less constraint from ID loss leads to triplet loss easier to converge at the same time. Thirdly, normalization keeps the compact distribution of features that belong to one same person. Because this method makes the purpose of training more consistent, both cosine distance and Euclidean distance can be applied in the inference stage, and the effect is the same.
In the inference stage, we choose $f_i$ to do the task of re-identification. The cosine distance metric has better performance than the Euclidean distance metric. The following experimental results show that BNblock can significantly improve the performance of the ReID model.

5. Experiments

We use the recognized pedestrian re-identification data sets Market1501 [11] and Duke-ReID [12] in our experiments. To evaluate the effective of using deep learning features of fusion direction information and BNblock on pedestrian re-identification, we use fair pedestrian re-identification evaluation indicators Rank-1 and mAP evaluates this algorithm.

In the experimental phase, we use an open-source pedestrian re-recognition method open-reid as the benchmark. This method uses ResNet50 as the backbone network. In the training phase, this method allows each picture to be horizontally flipped with a 50% probability, using triple loss and cross entropy loss, the Adam method is used to optimize the model. The initial learning rate is set to 0.00035, and the 40th and 70th epochs are reduced by 0.1 respectively. There are 120 epochs in total.

For the effectiveness experiment of the proposed pedestrian re-identification algorithm that integrates pedestrian direction information (DI), we compare the benchmark and multiple pedestrian re-identification algorithms on the Duke-ReID and Market1501 data sets. Then add the loss function of the fusion pedestrian direction to the benchmark, and perform ablation experiments with the BNblock module. The experimental results are shown in the table 1.

| method | Market1501 |  | Duke-ReID |
|--------|------------|  |------------|
|        | Rank-1 | mAP | Rank-1 | mAP |
| TriHard | 82.9 | 66.6 | 73.2 | 54.6 |
| Baseline | 87.7 | 74.0 | 79.7 | 63.7 |
| + DI | 88.1 | 76.5 | 81.4 | 64.5 |
| + BNblock | 89.2 | 75.6 | 80.7 | 63.9 |
| + DI & BNblock | **89.8** | **79.9** | **82.2** | **68.4** |

It can be seen that Baseline's Rank-1 in Duck-ReID is 79.7%, mAP is 63.7%, and in Market1501, Rank-1 is 87.7%, mAP is 74.0%. After adding either the direction information or the BNblock module respectively, there is an improvement. And after adding both at the same time, there is a significant improvement. It has the same improvement effect for TriHard. TriHard is an improved triplet loss function, which has better generalization ability and ability to seek global solution than the binary loss function. It can be seen that the features of fusion direction information and the use of BNblock can improve the accuracy of pedestrian re-recognition.

Acknowledgments

Thanks for the help of my classmates and teachers. Thanks to the National Natural Science Foundation of China (61170185, 61602320), the Liaoning Provincial Doctoral Startup Fund Project (201601172) and the Liaoning Provincial Department of Education's scientific research general project (L201607) for funding.

References

[1] Lu, J., Liong, V.E., Zhou, X., et al. (2015) Learning compact binary face descriptor for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(10): 2041-2056.
[2] Zhang, Y., Pan, G., Jia, K., et al. (2015) Accelerometer based gait recognition by sparse representation of signature points with clusters. IEEE Transactions on Cybernetics, 45(9):1864-1875.
[3] Zheng, W.S., Gong, S., Xiang, T. (2013) Reidentification by Relative Distance Comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(3): 653-668.
[4] Farenzena, M., Bazzani, L., Perina, A., et al. (2010) Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, California, USA. pp. 2360-2367.

[5] Yang, Y., Yang, J., Yan, J., et al. (2014) Salient color names for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV). Zurich, Switzerland. pp. 536-551.

[6] Zheng, Z.D., Liang, Z., and Yi, Y. (2017) Unlabeled samples generated by Gan improve the person re-identification Baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy. pp. 3754-3762

[7] Arjovsky, M., Chintala, S., and Bottou, L. (2017) Wasserstein GAN. In: ICML. International Convention Centre, Sydney Australia. p. 3.

[8] Zhu, J.Y., Park, T., Isola, P., Efros, A. (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. Venice, Italy. pp. 3-11.

[9] Mishchuk, A., Mishkin, D., Radenovic, F., and Matas, J. (2017) Working hard to know your neighbor’s margins: Local descriptor learning loss. In: Neural Information Processing Systems. Long Beach, USA. pp. 4829–4840.

[10] Kwangjin, Y., Song, Y., and Jeon, Moongu. (2019) Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views. IET Image Processing, 12(7): 1175-1184.

[11] Zheng, L., Shen, Liyue., Tian, L., Wang, S.J., Wang, J.D., Tian Q. (2015) Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. pp.1116-1124.

[12] Ristani E., Solera F., Zou R., Cucchiara R., Tomasi C. (2016) Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In: Hua G., Jégou H. (eds) Computer Vision – ECCV 2016 Workshops. Amsterdam, The Netherlands. pp 17-35.