Research on Pedestrian Re-Identification Using CNN Feature and Pedestrian Combination Attribute

Mengke Jiang[1(✉)], Jinlong Chen[2], and Baohua Qiang[2]

[1] Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, China
446084066@qq.com
[2] Guangxi Key Laboratory of Cryptography and Information Security, Guilin University of Electronic Technology, Guilin, China

Abstract. Aiming at the problem that the existing pedestrian recognition technology re-identification effect is not good and the traditional method has low recognition effect. A feature fusion network is proposed in this paper, which combines the CNN features extracted by ResNet with the manual annotation attributes into a unified feature space. ResNet solved the problem of network degradation and multi-convergence in multi-layer CNN training, and extracted deeper features. The attribute combination method was adopted by the artificial annotation attributes. The CNN features were constrained by the hand-crafted features because of the back propagation. Then the loss measurement function was used to optimize network identification results. In the public datasets VIPeR, PRID, and CUHK for further testing, the experimental results show that the method achieves a high cumulative matching score.

Keywords: Pedestrian re-identification · ResNet · Pedestrian attribute

1 Introduction

Re-Id, also known as pedestrian re-identification, is a technology that uses computer vision technology to judge whether there is a same target pedestrian in a camera image that does not overlap. The research on pedestrian re-identification initially focused on the manual annotation of pedestrian feature extraction, the identification of different camera angles, and the learning methods based on distance measurement [1]. However, due to the cost of manual labeling and the lack of traditional methods, this research has not obtain great progress. With the development of machine learning and deep learning, in the ImageNet competition in 2012, the Hinton team [2] first tried to integrate the convolutional neural network into the pedestrian detection technology, and achieved good experimental results. Since 2014, researchers have attempted to incorporate deep learning into the issue of pedestrian re-identification. Deep learning provides a way to solve computer vision problems without the need for excessive manual annotation of image features. The back propagation algorithm dynamically

R. Mao et al. (Eds.): ICPCSEE 2019, CCIS 1059, pp. 468–475, 2019.
https://doi.org/10.1007/978-981-15-0121-0_36

adjusts the parameters in the CNN so that the feature extraction and pairwise comparison processes are unified into a single network.

However in the real world, the appearance of a pedestrian will be largely affected by the difference of camera angle, illumination, height, etc. The manual annotation feature can solve the problem well, and the application can make the method more reliable in the pedestrian re-identification task. In order to effectively combine the artificial annotation features with the CNN features, a deep feature fusion network is proposed. The artificial annotation features are used to adjust the CNN process. The features extracted by CNN can also be used as a supplement to the manual annotation features.

Experiments on three challenging pedestrian re-identification attribute data sets (VIPeR, PRID, and CUHK) demonstrate the validity of the new features. Compared with the existing method, the recognition rate of rank-1 has been significantly improved. In summary, the manual annotation feature can improve the extraction process of CNN features and achieve a more robust image representation.

2 Related Work

Due to the outstanding performance of deep learning in pedestrian detection and recognition, the re-identification of pedestrians in recent years mainly focuses on deep learning. Many researchers at home and abroad have proposed different improved algorithms in this process. There are two main methods for pedestrian re-identification using deep learning. One is to use the end-to-end technology to extract pedestrian features using the convolutional neural network (CNN) to achieve pedestrian re-identification, such as DeepReID : deep filter pairing neural network for person re-identification; Another is to achieve pedestrian re-identification in combination with the high-level semantic features of pedestrians. For example, Li Jiali et al. [3] proposed to add multi-classification features based on human body structure detection based on deep learning features, and established a multi-feature fusion model with enhanced depth features; reference [4] not only identifies each type of pedestrian attribute separately, but also arranges and combines pedestrian attributes, and then jointly identifies multiple attributes and individual attributes; literature [5] jointly recognized pedestrian attributes and pedestrian IDs, making full use of the pedestrian's annotation information, and improving the accuracy of pedestrian recognition; paper [8] used color histogram features (RGB, HSV, YCbCr, Lab and YIQ) and texture features, combined with the features attracted by the traditional CNN, largely improved the recognition accuracy. These efforts have led to a leap-forward development in pedestrian recognition.

Combining feature extraction and image pair classification into a single CNN network, the comparison and symmetrical structure of paired pictures are widely used, but pairwise comparison demanded to form lots of pairs for each probe image and perform deep convolution on these pairs [6]. Inspired by [6], a feature fusion network is proposed that extracts depth features from a single image without using pairs of inputs. The artificially labeled pedestrian attribute features are combined and merged with the depth features to be unified into a single network, and then the loss measurement

function proposed in this paper is used to optimize the network recognition results. The test results show that the proposed algorithm is superior to the current mainstream pedestrian recognition algorithm in public data sets, especially the first accuracy rate (rank-1).

3 Model

3.1 Network Structure

We use the fine-tuned feature fusion network to learn new features. The network structure consists of two parts, as shown in Fig. 1. The first part uses the ResNet network to extract features from pedestrian sample images. ResNet can solve the problem of network degradation and multi-convergence in multi-layer CNN training, and can extract deeper features. The second part deals with the hand-crafted feature of the same picture. Finally, the two sub-networks are combined to produce a complete image description. In the second part of the study, the results of the first part will be adjusted.

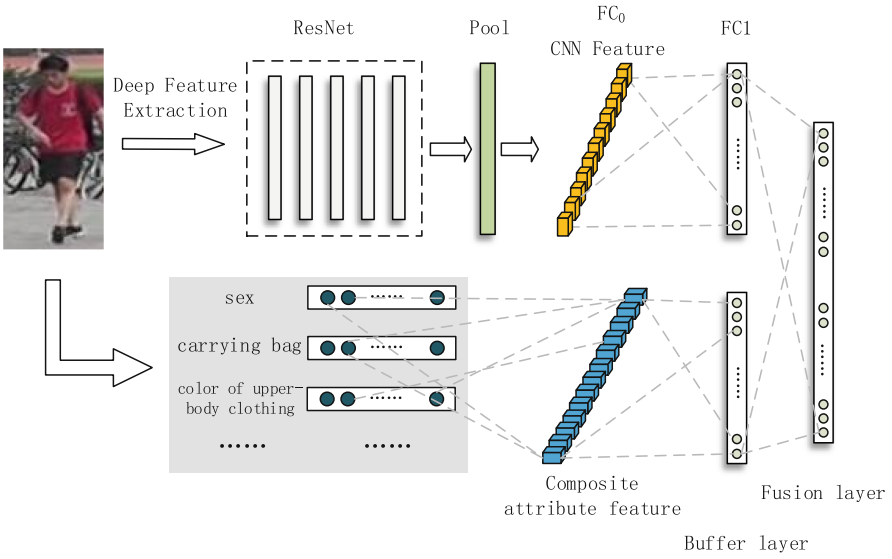

Fig. 1. Fusion feature network structure model

3.2 CNN Features

The extraction task of image features is using ResNet, the upper part of Fig. 1, which uses the ResNet-50 network, that is a residual network with a depth of 50 layers, and the network is mainly composed of convolutional layer, pooling layer and residuals. In the process of deep learning development, the researchers found that as the number of

network layers deepens, the network will undergo gradient dispersion and gradient explosion during training, as a result, the network can't be able to converge during training. Differ from the general CNN, ResNet has a unique residual block structure, which avoid the network degradation and the convergence problem during training without introducing additional parameters and computational complexity by learning the residual function and realizes the identity mapping.

In our framework, by using back propagation, the parameters of the entire CNN network are affected by the hand-crafted features. Our goal is to combine features into a unified feature space. A feature fusion deep neural network is proposed to adjust CNN features by using hand-crafted features, so that CNN can extract complementary features and have a more complete feature representation. The buffer layer can be used as a bridge between the extracted CNN feature and the composite attribute feature to reduce the huge difference between different features and ensure the convergence of the fusion layer. If the input of the fusion layer is:

$$X = [\text{Composite\_Attribute\_Features}, \quad \text{CNN\_Features}], \tag{1}$$

Then the output of this layer is calculated by the following formula:

$$Z_{Fusion}(x) = h\left(W_{Fusion}^{T} x + b_{Fusion}\right), \tag{2}$$

Where h(·) indicates the activation function.

Existing deep re-identification networks for person re-identification adopt Deviance Loss or Maximum Mean Discrepancy as loss function [6]. Our goal is to effectively extract the depth features of each image, rather than comparing the image pairs through deep neural networks. So we use the softmax loss function, a more discriminative feature representation will get a lower softmax value.

3.3 Hand-Crafted Features

The recognition of pedestrian images by semantic attributes, such as gender, wearing, backpack color, etc., can be used as auxiliary information to improve pedestrian recognition accuracy. There are several advantages to using manual annotation of attribute features: First, because most people have similar appearances (such as clothing color, backpack, hair color, etc.) it is more difficult to manually mark the same pedestrian in a camera with low pixels. In contrast, the labeling of pedestrian attributes is simpler and more accurate. Second, the number of classification of pedestrian attributes is less than that of different pedestrians, since different pedestrians will have the same attributes.

The training data set contains N pedestrian images, and a pedestrian image has multiple attribute annotations (such as gender male, upperbody black), and we group these attributes. According to the paper [4], each image is marked by G attribute groups, e.g., Gender, Age, and every attribute group has a different number of attributes, denoted as K(g), e.g., group gender has male, female, group upperbody clothing has sweater, t-shirt, suit and so on. We assume that each attribute group can only have

one attribute value. For example, a pedestrian's upperbody color is black and white. We only select one color, that is, one attribute value.

Since the categories of each attribute group are inconsistent, we define a weighted cross entropy loss function. The loss of output node j is calculated as follows:

$$P_{(y=j)} = \frac{e^{a_j}}{\sum_{j=1}^{T} e^{a_k}} \tag{3}$$

Where T represents the number of categories. The cross entropy loss function is as follows:

$$\mathrm{L} = -\frac{1}{N^g} \sum_{j=1}^{N} \sum_{k=1}^{K^g} \frac{y_j log P_j}{N_{k(i)}^g} \tag{4}$$

Where N represents the number of pedestrian pictures, N^g represents the number of pedestrian images in the g-th attribute group, $N_{k(i)}^g$ is the number of training samples of the k-th attribute in the g-th attribute group, and P_j is calculated by the formula 3 inferred.

4 Experiment

4.1 Dataset

This paper uses the PETA (PEdesTrian Attribute) data set, which is the largest data set currently open for pedestrian attribute recognition tasks. The data set contains 8705 pedestrians for a total of 19,000 images (resolutions from 17*39 to 169*365). Each pedestrian is marked with 61 binary value and 4 multi-category attributes (binary values such as whether they are under 15 years old, and multiple categories such as upperbody colors can have multiple coexistences). In fact, the PETA data set is a collection of multiple smaller pedestrian re-identification data sets that are labeled by attributes. The partial data sets included are shown in the following Table 1:

Table 1. Partial data set of PETA.

Datasets	Images	Resolution
3DPeS	1012	From 31*100 to 236*178
CAVIAR4REID	1220	From 17*39 to 72*141
CUHK	4563	80*160
GRID	1275	From 29*67 to 169*365
PRID	1134	64*128
VIPeR	1264	48*128

For all manual annotation properties, we divided then into 8 groups, each group contains a different number of attribute characteristics, as shown in Table 2. Among them, some attribute will be removed if the number of them is less than 10. In addition, if an attribute group has two attribute values at the same time (for example, upperbody has black and white), and we randomly select one value as the attribute label.

Table 2. The group of the attribute

Group	Attributes	Number
Gender	Male, Female	2
Age	Young, Teenager, Adult, Old	4
Hair	Long, Short	2
Upperbody Color	Black, White, Red, Purple, Yellow, Gray, Blue, Green	8
Upperbody Clothing	Long Sleeve, Short Sleeve, no Sleeve	3
Downbody Color	Gray, Black, White, Pink, Purple, Yellow, Blue, Green, Brown	9
Downbody Clothing	Dress, Pants	2
Other	Bag, Hat, Handbag, Backpack	4

4.2 Setup

We use three pedestrian re-identification databases to evaluate the fine-tuned CNN features and implement our approach using the keras framework. We resize all training images into 256*128 pixels and add a pad with 10 pixels, then randomly crop 256*128 sub-windows. For test time, we resize all the input images to 256*128 pixels. The CNN parameters are all derived from the pre-trained model ResNet-50, we start the last fully connected layer from random weights. The batch size is set to 256, the initial learning rate is set to $\gamma = 0.0001$, and every 20,000 iterations is reduced as $\gamma_{new} = 0.1 * \gamma$.

The method proposed in this paper was tested on three data sets and compared with other methods, the experimental results are shown in Table 3. The results of rank-1 on the VIPeR, PRID, and CUHK datasets were 45.23%, 50.22%, and 49.20, separately. It can be seen that compared to the methods listed in the table, the accuracy tested in the three datasets used the method proposed in this paper has improved a lot, which proves the effectiveness of our method.

Table 3. Performance comparison with different methods

Methods	VIPeR				PRID				CUHK			
	r = 1	r = 5	r = 10	r = 20	r = 1	r = 5	r = 10	r = 20	r = 1	r = 5	r = 10	r = 20
Ours	**45.23**	**73.88**	**85.23**	**92.34**	**50.22**	**68.89**	**83.23**	**90.20**	**49.20**	**73.34**	**86.34**	**95.31**
Deep Feature Learning [7]	40.50	60.80	70.40	84.40	–	–	–	–	–	–	–	–
Ahmed's Deep Re-id [8]	–	–	–	–	34.81	63.72	76.24	81.90	47.53	72.10	80.53	88.49
KISSME [9]	24.75	53.48	67.44	80.92	36.31	65.11	75.42	83.69	14.02	32.20	44.44	56.61
L2 - norm	10.89	22.37	32.24	45.19	11.33	24.50	32.22	43.89	5.63	16.00	22.89	30.63
L1 - norm	12.15	26.01	32.09	34.72	25.50	25.33	51.73	53.07	10.80	15.51	37.57	35.57

5 Conclusion

The paper proposed a network structure that combines the processed manual annotation attribute features with CNN extraction features, and then uses the loss measurement function proposed in this paper to optimize the recognition results. The CNN extraction feature is based on the ResNet network, and combine the hand-crafted attribute features. The combine the two sub-network into a single network structure. According to the back propagation, the attribute feature can supplement the CNN features extracted by the network, and obtain a more complete feature representation to achieve more accurate pedestrian recognition results. Tested on three challenging public data sets (VIPeR, PRID, CUHK), the experimental results prove the effectiveness of the proposed method. Subsequent work will further study the pedestrian attributes and hope to further improve the accuracy of pedestrian re-identification.

References

1. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multipleshot person re-identification by HPE signature. In: 20th International Conference on Pattren Recognition (ICPR) 2010, pp. 1413–1416. IEEE (2010)
2. Krizhevsky, A., Sutskerver, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing System, pp. 1097–1105 (2012)
3. Li, J., Guo, J.: Research on pedestrian re-recognition algorithm based on enhanced depth feature fusion. Inf. Technol. **42**(320(07)), 23–27 (2018)
4. Matsukawa, T., Suzuki, E.: Person re-identification using CNN features learned from combination of attributes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2428–2433 (2017)

5. Roy, A., Sural, S., Majumdar, A.K.: Minimum user requirement in role based access control with separation of duty constraints. In: Proceedings of the 12th International Conference on Intelligent Systems Design and Applications, Washington D C, USA, pp. 386–391. IEEE Press (2013)
6. Wu, S., Chen, Y., Li, X.: An enhanced deep feature representation for person re-identification. In: WACV (2016)
7. Ding, S., Lin, L., Wang, G., et al.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recognit. **48**(10), 2993–3003 (2015)
8. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: IEEE CVPR (2015)
9. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: IEEE CVPR (2012)