A Study of Community Surveillance System Improvement based on ResNet Person Re-identification

Xinyi Gu and Chengzhang Qu*
Wuhan Business University, 430056, No.816 of Dongfeng Street, Hanyang District
Wuhan, Hubei Province, China
* corresponding author.
Email: Quchengzhang@foxmail.com

Abstract. Most existing community surveillance systems rely on face recognition to identify pedestrian. The face image captured by community surveillance camera are not always clear enough. Thus, Person re-identification (ReID) imply critical applications in surveillance as it has more image information of pedestrian. In this paper, we refine a ResNet50 based reID model which only adds a Linear layer, a Batch Norm layer and a ReLU layer in front of the classifier. The refined model is simple to build on the surveillance system, and we have tested in our demo surveillance system and Market1501 data-set. The experiment result shows it can work well on real-time. On Market1501, our results are competitive that rank-1=0.875000, rank-5=0.944477, and mAP =0.706647 since its low complexity. Our demo community surveillance system shows the refined ReID model is competent and practical for identification tasks.

1. Introduction
Nowadays, surveillance cameras which capture pictorial information have been widely used for security and it makes a huge surge in video images. The definition of ReID is that a specific pedestrian image [1] (query in Figure 1) has been given. And we need to compare it with the pedestrian image in the gallery captured by the cross-camera device (1 to 10 in Figure 1), and retrieve the pedestrian image. Lots of related work use ideas like data augmentation [2] and re-ranking [3] to solve the problems, and we focus on a stronger baseline to help the next steps.

![Figure 1. An example of pedestrian images](image-url)
2. Method and Model

2.1. Deep Residual Network

ResNet50 is adopted as backbone to build our ReID model, which is a kind of deep residual network. Load the pre-trained model on ImageNet to improve efficiency [4]. Deep residual network, i.e. ResNet [5] referenced VGG19 and appended the residual block by shortcut connections [6]. ResNet solves the degradation problem caused by increasing depth, which can improve the performance of the model. The architecture of ResNet in several different depths is shown in Figure 2.

![Figure 2. The architecture of ResNet in several different depths [5]](image)

On the right of Figure 2, deeper bottleneck architectures is introduced. The three layers are 1×1, 3×3, and 1×1 convolutions. The 1×1 layers are responsible for reducing and then restoring dimensions, leaving the 3×3 layer a bottleneck with smaller input/output dimensions [5]. It makes it easier for the 3×3 convolution to get more information.

![Figure 3. A deeper residual structure](image)

2.2. The Basic Method of ReID

ReID is usually regarded as a classification problem [7]. It is a mainstream approach to solve the problem of ReID to extract the identity features and measure the distance [8], and we use it in the paper. The steps are as follows. (1) Extract features from every image in query and gallery. (2) Calculate the distances from query image to each image in gallery. (3) Sort the images by distance from small to large.

3. Design and Implementation of the ReID System

3.1. Description of the Network Structure

The ResNet50 network has five parts (Figure 2), which are conv1, conv2_x, conv3_x, conv4_x, conv5_x, where the convolution with step size of 7*7*64 and stride of 2 is input and then goes
through $3+4+6+3=16$ blocks, each block has 3 layers, and there are $16 \times 3 = 48$ layers in total. The FC layer at the end served as classification, so there are $1+48+1=50$ layers in this network. The 50 layers do not include ReLU or pooling layers, it only have convolution layers and full connection layers.

A resnet50-based model in this paper appended a linear layer between conv5_x and FC layer (classifier) to map the conv5_x’s output features from 2048 to 512 dimensions. A BN layer of BatchNorm1d and ReLU layer was added. In the full connection layer, we dropout some neurons randomly to get a more robust model. So the FC layer in the original ResNet50 model is replaced by the network whose architecture is in the order of Linear, BN, ReLU and Linear.

3.2. Work Procedure
We have implemented the model designed in this paper. The code structure is shown in Figure 4.

![Figure 4. Code structure](image)

prepare. py classifies person identities in the dataset; model.py contains the definition and parameter settings of the network structure; train.py includes training the model and the outputs of the training process curve. Test.py extracts the features of each image and saves them; Evaluate_gpu. py uses GPU to evaluate the model.

The model is trained on the market-1501 dataset in this paper. First, put the same identities' images into a folder, and every identity is sorted by ID.

Then define the model based on ResNet50. We defined the designed fc layer and classifier, which is informed in 3.1. Initialized the model by ImageNet pretrained model of ResNet50, and we changed the number of categories from 1000 to 751 to adapt the training dataset of Market1501. Initialized the weight of network by kaiming function. In pooling layers, we use global pooling. Because the height of pedestrian image in Market1501 is higher than its width, so the function named AdaptiveAvgpool2d is specified to improve the performance of extracting features. The shape of pedestrian images is rectangular, output sizes of the pooling layers are usually 8x4 or 16x8 which instead of the original ImageNet network of 7x7. Global pooling integrates the information of the whole image.

Train the model in 60 epochs and plot the training process image. The following 3 steps are required to train an epoch of training dataset. First, get outputs of the forward propagation. The outputs are predictive values of pedestrians identity. Pay attention to set the gradient to zero in each iteration, otherwise the gradient would accumulated. Next, calculate the loss using cross entropy loss function. Loss function is calculated as follows:

$$\text{loss}(x,\text{class}) = -\log \left( \frac{\exp(x[\text{class}])}{\sum \exp(x[j])} \right) = -x[\text{class}] + \log \sum \exp(x[j])$$

The tag of class does not participate in the direct calculation, but serves as an index of the actual labels. For backward, each epoch calculates and optimized during training to update the model’s parameters. Also, identification loss is used here, and multiple classifications are made directly with the identity label. The loss of every batch is calculated, then the average of loss and accuracy of every epoch are calculated, and loss and err of this epoch are updated and added together.

The images’ features in the folders named ‘gallery’ and ‘query’ are extracted. Just do the forward steps in training process and obtain the predictive value of dataset’s labels. Abandon the images which
is detected incorrectly or captured by the same camera as the query image. CMC and mAP are calculated ranking images by similarity.

Pay attention to separating images in folders named ‘train’ and ‘val’, which is convenient for analyse. ‘train’ contains pedestrians captured by several images, and ‘val’ contains pedestrians who only have single image. Also, evaluation uses the data in the folder named ‘train_all’, because data of multiple identities and cross-camera should be evaluated.

4. Analysis Experimental Results

4.1. The Experimental Results
The training curves of the ResNet50-based model is shown in the Figure 5.

![Figure 5. Result of training curves](image)

The evaluation of the ResNet50-based model are shown in the Table 1 below.

| Distance (m) | Market1501 |
|-------------|------------|
| Rank-1      | 0.875000   |
| Rank-5      | 0.944477   |
| Rank-10     | 0.963183   |
| mAP         | 0.706647   |

4.2. Comparison and Analysis
Training results of the model is shown in Figure 5. According to the training curves, this model can finally identify the pedestrian images in the training set of multiple cameras and multiple pedestrian images, but the training loss of a single pedestrian image can no longer be reduced to 0.5 before learning with a higher recognition degree.

The training curve in the first 6 epochs obtained from other tests of this model is shown in Figure 6. The model uses random erasing for data augmentation and re-ranking. The training environment is different and the parameters are adjusted, so the accuracy of identification is increased.

The ResNet50-based model uses the method based on representation learning, which regards a problem as a classification problem or a verification problem. The test model uses a scheme that input single images as query, extract features, calculate the Euclidean distance with others, and then sort
them by distance. All the features of pedestrians images are extracted in advance, and then we only need to compare with them. The comparison can be realized simply by matrix multiplication.

The results of evaluation show that $\text{Rank-5} = 94.5\%$, which indicates that this model has a good performance on Market1501. The mAP is about 70%, and the baseline can be increased more, which also indicates the limitation of CMC evaluation, and the actual performance is not as high as about 90%. There’s other baseline\cite{9} whose network is similar to our model, which contains the newly full connection layers, too. The evaluations of this baseline are shown in Table 2. It shows that we get a stronger baseline on ResNet50, which makes it easier to have next steps. The ResNet50-based model in this paper is a baseline which doesn’t use tricks like re-ranking and random erasing. Other improved model\cite{10} which adopt ResNet50 as CNN backbone has a better performance, shown in Table 2, indicating that there is still large space to develop.

| Table 2. Evaluations(%) of other ResNet50-based model |
|------------------------------------------------------|
| Rank-1 | Rank-5 | Rank-10 | mAP   |
|---------|--------|---------|-------|
| Baseline\cite{9} | 80.16  | 92.03   | 94.98 | 57.82 |
| PCB\cite{10}      | 92.3   | 97.2    | 98.2  | 77.4  |
| PCB+RPP\cite{10}  | 93.8   | 97.5    | 98.5  | 81.6  |

5. A Demo of Community Surveillance System
A demo of community surveillance system is implemented. Every user in the system can use the function of retrieving pedestrian. Permissions are given to the admins to check the record and manage the users. The functional structure diagram is shown in Figure 7. The system has been implemented as designed. When users upload an query image, the interface is shown in Figure 8.

Figure 7. The functional structure diagram of community surveillance system
Figure 8. The interface of community surveillance system
6. Summary and Prospect

ReID is a hot issue in the field of intelligent monitoring nowadays. With the development of deep learning in recent years, a large number of re-ID methods based on deep learning have been proposed. In this paper, we first introduce the Network’s structure of Deep Residual Network(ResNet50) and its theory. Second, we describe our working process and model combined with ReID system. An ReID model based on the ResNet50 which appends a Linear layer, a BatchNorm layer, and a reLU layer in front of the classifier is put forward in the third part. The implementing process of the ReID system is described in detail next. Also, we show the experimental results, and do compare and analyze work. On Market1501, we obtain rank-1 = 0.875000, rank-5 = 0.944477, rank-10 = 0.963183 and mAP = 0.706647. The CMC value indicates that the hit rate of this model in the top five similarity degrees has exceeded 94%. In conclusion, the model implemented in this paper performs well in the Market1501 data set. Finally, a simple demo of community surveillance system is implemented.

Loss function and image processing techniques used in this paper are common methods. Loss function which is more suitable for specific situation can be researched in the future, and the pre-trained model can be improved according to specific scene. The re-ID system in this paper was trained and tested on Market1501 dataset. The person in this dataset is most asian people who wear summer clothes. Therefore, we can try to re-experiment and improve the model on the open data set with different people with various clothes. Moreover, our community surveillance system need more work on reducing the computational time complexity, as saving computer resources can make the system more suitable and practical.

7. Acknowledgements

Supported by Hubei Provincial Natural Science Foundation of China, No.2018CFB310; Supported by Wuhan Business University Doctoral research project, No.2017KB006.

8. References

[1] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1116-1124.
[2] Hauptmann A G, Yang Y, Zheng L. Person Re-identification: Past, Present and Future[J]. 2016.
[3] Zhong Z, Zheng L, Cao D, et al. Re-ranking person re-identification with k-reciprocal encoding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1318-1327.
[4] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252.
[5] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778. doi: 10.1109/CVPR.2016.90.
[6] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Springer, Cham, 2016: 630-645.
[7] Chen H, Wang Y, Shi Y, et al. Deep Transfer Learning for Person Re-Identification[C]. ieee international conference on multimedia big data, 2018: 1-5.
[8] Xiao T, Li H, Ouyang W, et al. Learning deep feature representations with domain guided dropout for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1249-1258.
[9] Lin Y, Zheng L, Zheng Z, et al. Improving person re-identification by attribute and identity learning[J]. Pattern Recognition, 2019: 151-161.
[10] Zheng, Ruochen et al. “Camera Style and Identity Disentangling Network for Person Re-identification.” BMVC (2019).