Person Re-identification in Crowded Scenes with Deep Learning

Yupeng Wang, Huiyuan Fu, Shuangqun Li
Beijing Key Lab of Intelligent Telecommunication Software and Multimedia
Beijing University of Posts and Telecommunications
Beijing, China
E-mail: {wyp, fhy, lsq}@bupt.edu.cn

***Abstract*—Person re-identification in crowded scenes is very important. Most images come from different surveillance video and cameras, and one person may look different in a variety of scenes, viewpoints, lighting and so on. The existing methods have limited effects in practical applications. In this paper, we propose a convolutional neural network for person re-identification in crowded scenes. The model structure of this network combines pedestrian detection and re-identification. In addition, we propose a loss function to better match the target person by calculating Pearson correlation evaluation. The experimental results show that our method is effective.**

Keywords-deep learning; person re-identification; crowded scenes; convolutional neural network; loss function

I. INTRODUCTION

Person re-identification means to match the target person in different images. These images come from different cameras, so the challenges lie in the different backgrounds, the changes of human postures, camera viewpoints, lighting, occlusions and so on. Person re-identification can be applied to surveillance video, such as cross-camera searching and tracking of criminals [1], which is very helpful for public safety. Therefore, person re-identification draws the attention of scholars and a lot of researches have been done in recent years [3][7-10][12].

Person re-identification consists of two steps: pedestrian detection and re-identification. Many researches have been done in the field of pedestrian detection. Dalal et al. [4] proposed Histogram of Oriented Gradient (HOG) feature for human detection. It can describe the edge features of human body. Zhu et al. [5] extracted Multi-scale Intrinsic Motion Structure features for pedestrian detection. These methods use hand-crafted features and linear classifiers to detect persons. Hand-crafted features may lose some important information of the original images, and the classification results are not good enough. In recent years, Deep Learning has attracted much attention in the fields of image, audio, natural language processing and so on. Yang et al. [6] proposed Convolutional Channel Features which achieved good performances in pedestrian detection. Zhang et al. [2] used Region Proposal Network (RPN) followed by boosted forests on high-resolution convolutional feature maps. As for person re-identification, the common method is part matching. Zhao et al. [7] adopted patch matching and estimated patch salience. Zheng et al. [10] proposed a PoseBox structure, which pose is estimated by affine transformations. However, these methods may introduce errors in part detection. Yi et al. [8] and Varior et al. [9] used the siamese convolutional neural network for person re-identification. Most of the existing person re-identification methods assume perfect pedestrian detections. In fact, these hand-cropped bounding boxes are unavailable in the applications of realistic scenes. Xiao et al. [3] proposed an end-to-end framework for person re-identification. Also, Yamaguchi et al. [13] used a natural language query to handle this task. But these methods are still difficult to satisfy complex realistic scenes. In these methods, the scenes are simple and contain only 1 to 5 people. The effect is not good in the scenes with many people.

In order to deal with person re-identification in crowded scenes, we propose a convolutional neural network (CNN) which combines pedestrian detection and re-identification. It can learn more effective deep features due to CNN with deeper network layers. In addition, we propose a loss function to better match the target person. The similarity of the target person and persons in crowded scene images is calculated by Pearson correlation evaluation. Finally, with fine-tuning on the weights initialization, experimental results on Large Scale Person Search dataset (PSDB) [3] show that the proposed method gains new state-of-the-art performances.

II. PERSON RE-IDENTIFICATION IN CROWDED SCENES

Our work consists of two parts: deep learning network model and matching loss function for person re-identification. In the detection phase, we construct residual network units (ResNet) [11] and RPN [14]. In order to converge better and faster, we fine-tune the network via Xavier [16] filler instead of Gaussian filler. And in the re-identification phase, we construct aggregated residual network units (ResNeXt) [15] to extract the deep features of detected persons. After that, we propose a loss function to match the target person by calculating the similarity between the target person features and detected person features.

A. Model Structure for Person Re-identification

The model structure for person re-identification in crowded scenes is shown in Fig. 1. We construct ResNet-50 [11] as our base CNN model to extract image features. The kernel size is 7 in the first convolutional layer with 64 channels. We perform Batch Normalization (BN) layer after convolutional layers and Rectified Linear Units (ReLU) layer is performed after each BN layer. After the layer, there are three blocks. In the first block, there are three residual

 ISBN: 978-1-61208-592-0

units which include three convolutional layers with 1, 3, and 1 of kernel size respectively. In the second block, there are four residual units. In the third block, there are three residual units. These convolutional layers are different in channels. The residual units can produce convolutional feature maps with 1024 channels.

Next, we need to get the person bounding boxes. An RPN [17] with a Softmax classifier is added to get 9 anchors and predict whether each bounding box is a person or not. It selects the top 128 bounding boxes as final proposals of persons in the image.

The weights initialization method of the RPN is Gaussian filler. Weights are randomly drawn from Gaussian distributions with fixed mean (e.g., 0) and fixed standard deviation (e.g., 0.01). This is the most common initialization method in deep learning. To make the information in the network flow better, the variance of each layer's output should be equal. Xavier [16] filler makes weights a uniform distribution with the mean of 0 and the variance of *Var*, as follows

$$Var = \frac{N}{n_i + n_{i+1}}, \quad (1)$$

where n_i is the i-th input number, and n_{i+1} is the i+1-th input number, as well as the i-th output number. By default, the variance takes into account only the number of inputs (n_{i+1} = 0, N = 1). But the number of inputs and outputs is often unequal, and the variance takes into account only the number of outputs (n_i = 0, N = 1). For balanced consideration, N = 2. In our network, Xavier filler is adopted to make the effect better.

Then, the task is to match the target person in these proposals. An RoI-Pooling layer is added for each proposal. And this layer links two blocks of ResNeXt-50(32×4d) [15]. In the first block, there are three aggregated residual units, which the second convolutional layer is grouped convolutions with 32 groups. In the second block, there are three aggregated residual units which are different from units of the first block in kernel output channels. Finally, we add the matching layer to calculate the similarity of these features with the target person by our proposed loss function.

B. Matching Loss Function for Person Re-identification

To match the target person in an image, we store the features of all people in this image, and calculate the similarity of these features with the target person. If one similarity is the largest, then the corresponding person is likely to be the target. In the labeled boxes, the features of the target person box is denoted as x, where $x \in R^D$ and D is the feature dimension, that is, x is a feature vector of D dimensions. The feature of one person box in the image is denoted as y^l, where $y^l \in R^D$ and D is the feature dimension. During the forward propagation, we compute the Pearson correlation coefficient C_j^l between the target features and the box features with the j-th class ($j \in [1, L]$ and L is the number of the boxes), as follows

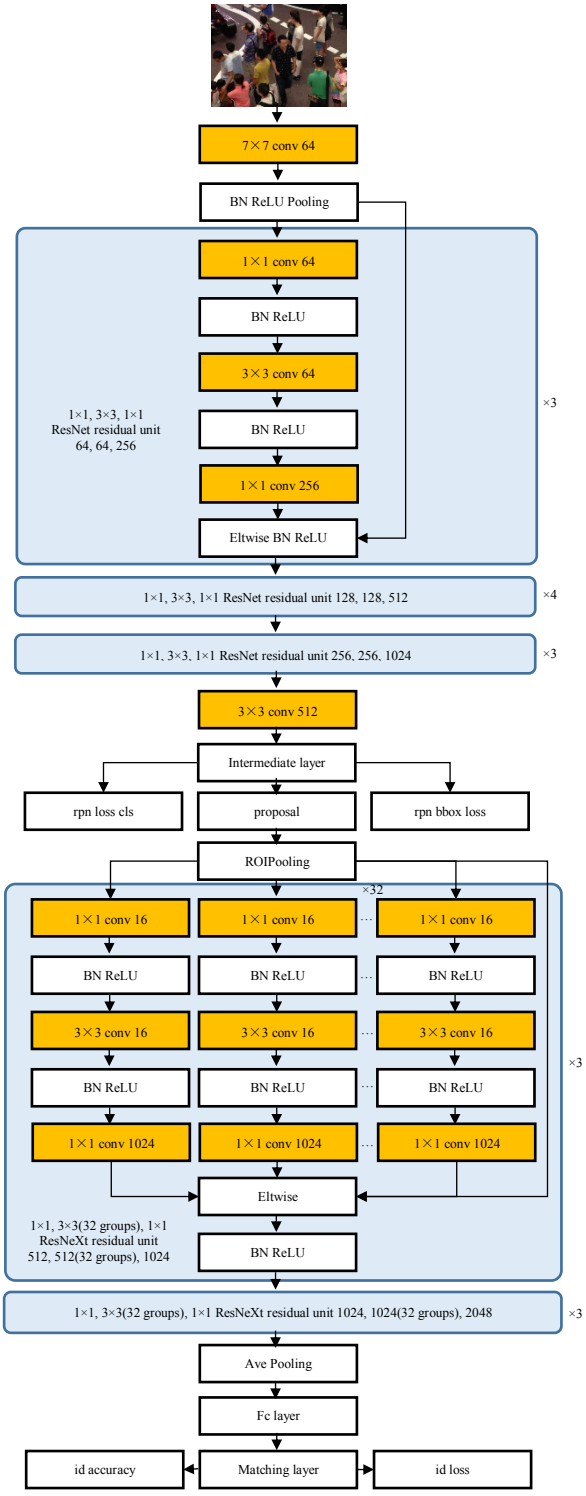

Figure 1. Our model structure.

$$C_j^l = \frac{\sum_{i=1}^{D}(x_i - \bar{x})\left(y_i^l - \overline{y^l}\right)}{\sqrt{\sum_{i=1}^{D}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{D}\left(y_i^l - \overline{y^l}\right)^2}}, \quad (2)$$

where $\bar{x}$ and $\bar{y^l}$ are the mean values of x and y^l. During the backward propagation, we update y^l_j by

$$y_j^l \leftarrow \gamma y_j^l + (1-\gamma)x\,, \tag{3}$$

where $\gamma \in [0, 1]$. Similarly, in the unlabeled boxes, the feature of one person box in the image is denoted as y^u, where $y^u \in R^D$ and D is the feature dimension. The Pearson correlation coefficient C^u_j between the target features and the box features with the j-th class ($j \in [1, U]$ and U is the number of the boxes), as follows

$$C_j^u = \frac{\sum_{i=1}^{D}(x_i-\bar{x})\left(y_i^u-\overline{y^u}\right)}{\sqrt{\sum_{i=1}^{D}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{D}\left(y_i^u-\overline{y^u}\right)^2}}\,, \tag{4}$$

where $\bar{x}$ and $\bar{y^u}$ are the mean values of x and y^u. The y^u_j is updated by

$$y_j^u \leftarrow \gamma y_j^u + (1-\gamma)x\,. \tag{5}$$

So, the probability of the feature vector x as the i-th class labeled person is

$$p_i^l = \frac{e^{\left(c_i^l/\tau\right)}}{\sum_{j=1}^{L} e^{\left(c_j^l/\tau\right)}+\sum_{k=1}^{U} e^{\left(c_k^u/\tau\right)}}\,, \tag{6}$$

where τ makes softer probability distribution. In the same way, the probability of the feature vector x as the i-th class unlabeled person is

$$p_i^u = \frac{e^{\left(c_i^u/\tau\right)}}{\sum_{j=1}^{L} e^{\left(c_j^l/\tau\right)}+\sum_{k=1}^{U} e^{\left(c_k^u/\tau\right)}}\,. \tag{7}$$

In summary, the Pearson correlation matching loss function is defined as follows:

$$\mathrm{L} = \mathrm{E}(\log p_t^l)\,, \tag{8}$$

where $t \in [1, L]$. The Pearson correlation coefficient is the cosine similarity with subtracting average. It is more applicable to data using different evaluation criteria. Then, the loss effectively compares the features of the target person with all the features of the boxes and finds the person whose feature vector is the most similar with the target one from the crowded scene images.

III. EXPERIMENTS

A. Dataset

We use PSDB [3], a large-scale dataset for person re-identification which covering hundreds of scenes. There are 18,184 images, 8,432 identities, and 96,143 pedestrian bounding boxes in this dataset. The training set contains 11,206 images and 5,532 query persons. The test set contains 6,978 images and 2,900 query persons.

TABLE I. PERSON RE-IDENTIFICATION METHODS

Methods	Top-1(%)	mAP(%)
DSIFT[17]+Euclidean	39.4	34.5
DSIFT[17]+KISSME[19]	53.6	47.8
LOMO+XQDA[18]	74.1	68.9
OIM[3]	78.5	75.4
Ours	**79.8**	**76.7**

TABLE II. WEIGHTS INITIALIZATION METHODS

Methods	Top-1(%)	Top-5(%)	Top-10(%)	mAP(%)
OIM (Gaussian)	78.5	90.1	92.3	75.4
OIM (Xavier)	79.0	90.0	92.8	76.0
Ours(Gaussian)	79.3	90.2	92.9	76.1
Ours(Xavier)	**79.8**	**90.3**	**92.9**	**76.7**

B. Experimental Results

For comparisons, we use some feature representations and the state-of-art method (Online Instance Matching, OIM) [3] for person re-identification on PSDB. The results are summarized in Table 1. Dense scale-invariant feature transform (DSIFT) [17] method is based on unsupervised salience learning. Euclidean distance metric and KISSME (keep it simple and straightforward metric) [19] distance metric are used for DSIFT. And Cross-view Quadratic Discriminant Analysis (XQDA) distance metric is used for Local Maximal Occurrence (LOMO) [18]. We use two kinds of evaluation metrics: cumulative matching characteristics (CMC top-K) and mean averaged precision (mAP). CMC top-K is the probability that the result hits within K times. And mAP is obtained by the average for precision results of each class of people.

The results are summarized in Table 1. Our method outperforms the others. On the one hand, the detector is also important for person re-identification. Most existing methods focus only on cropped images, so they may not be suitable for re-identification in the crowded scene images. On the other hand, different distance metric methods also have an impact on the experimental results.

In addition, we compare the different weights initialization methods with OIM [3]. As shown in Table 2, Xavier filler is superior to Gaussian filler in most cases. These results are relevant to specific network models and tasks.

The samples of our method in the experiment are shown in Fig. 2. The inputs are an image of the target person and some images with people. The outputs are the result images of bounding boxes with re-identification similarities. The person with maximum similarity is the target. Both in non-

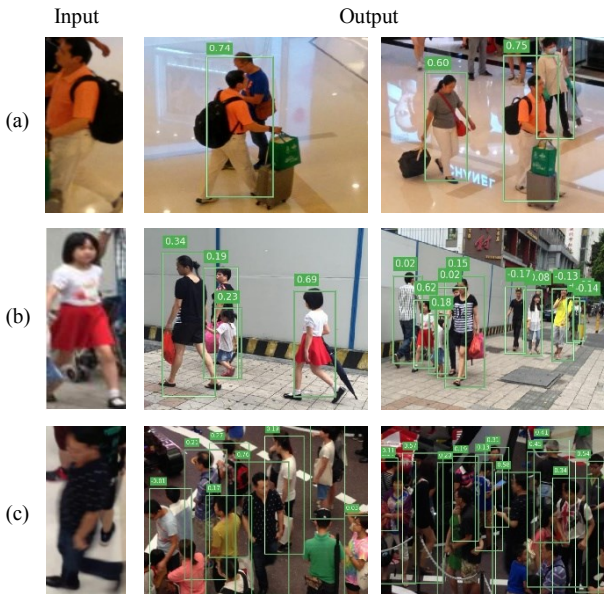

Figure 2. Samples of our method for person re-identification. (a) Non-crowded, (b) mid-crowded, (c) crowded scenes.

crowded and crowded scenes, our method can well identify the target person with the largest similarity.

IV. Conclusion and Future Work

In this paper, we propose a deep learning framework for end-to-end person re-identification. We adopt one CNN to link the two tasks of pedestrian detection and person re-identification. And ResNeXt residual units are constructed in the model to extract features more effectively. We also propose a Pearson Correlation Matching loss function to match the target person. Compared with existing methods, the performance of our method is improved with fine-tuning through experiments on PSDB dataset.

In the future, we plan to study person re-identification for surveillance video. The application of video is necessary for industry and society. On the one hand, many images make up video frame sequences. That is, the methods of images can be extended to video. On the other hand, there are new and available features due to the continuity of video frames. We will also combine person re-identification with Natural Language Processing to achieve a wider range of applications in the future.

Acknowledgment

The research reported in this paper is supported by the National Natural Science Foundation of China under Grant No.61402048; The NSFC-Guangdong Joint Found under No.U1501254.

References

[1] X. Wang, “Intelligent multi-camera video surveillance: A review,” Pattern Recognition Letters, vol. 34, no. 1. pp. 3–19, 2013.

[2] L. Zhang, L. Lin, X. Liang, and K. He, “Is faster R-CNN doing well for pedestrian detection?,” in European Conference on Computer Vision. Springer International Publishing, vol. 9906 LNCS, pp. 443–457, 2016.

[3] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, “Joint Detection and Identification Feature Learning for Person Search,” ArXiv e-prints, 2017.

[4] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. I, pp. 886–893, 2005.

[5] J. Zhu, et al., “Pedestrian detection in low-resolution imagery by learning Multi-scale Intrinsic Motion Structures (MIMS),” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3510–3517, 2014.

[6] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Convolutional channel features,” in Proceedings of the IEEE International Conference on Computer Vision, vol. 11–18–Dece, pp. 82–90, 2015.

[7] R. Zhao, W. Ouyang, and X. Wang, “Person re-identification by salience matching,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 2528–2535, 2013.

[8] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Deep metric learning for person re-identification,” in Proceedings - International Conference on Pattern Recognition, pp. 34–39, 2014.

[9] R. R. Varior, M. Haloi, and G. Wang, “Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification,” ECCV16, pp. 1–18, 2016.

[10] L. Zheng, Y. Huang, H. Lu, and Y. Yang, “Pose Invariant Embedding for Deep Person Re-identification,” ArXiv e-prints, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[12] W. S. Zheng, et al., “Partial person re-identification,” in Proceedings of the IEEE International Conference on Computer Vision, vol. 11–18–Dece, pp. 4678–4686, 2015.

[13] M. Yamaguchi, K. Saito, Y. Ushiku, and T. Harada, “Spatio-temporal Person Retrieval via Natural Language Queries,” ArXiv e-prints, 2017.

[14] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2016.

[15] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, “Aggregated Residual Transformations for Deep Neural Networks,” CoRR, vol. abs/1611.0, 2016.

[16] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” Proc. 13th Int. Conf. Artif. Intell. Stat., vol. 9, pp. 249–256, 2010.

[17] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3586–3593, 2013.

[18] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by Local Maximal Occurrence representation and metric learning,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2197–2206, 2015.

[19] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, “Large scale metric learning from equivalence constraints,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2288–2295, 2012.

 ISBN: 978-1-61208-592-0