DEEP PARTIAL PERSON RE-IDENTIFICATION VIA ATTENTION MODEL

Junyeong Kim and Chang D. Yoo

Korea Advanced Institute of Science and Technology
School of Electrical Engineering
291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Korea

ABSTRACT

This paper considers a novel algorithm referred to as deep partial person re-identification (DPPR) for partial person re-identification where only a part of a person is observed and full body images are available for identification. The DPPR is based on an end-to-end deep model which make use of convolutional neural network (CNN), RoI Pooling layer and attention model. The RoI Pooling layer enables the extraction of feature vector corresponding to predefined part of input image. The attention model selects a subset of CNN feature vectors. For qualitative evaluation of proposed model, data from CUHK03 are randomly cropped in constructing p-CUHK03. Experimental results show that DPPR outperforms our baseline model on p-CUHK03.

Index Terms— Partial person re-identification, Convolutional Neural Network, RoI Pooling, Attention model, DPPR

1. INTRODUCTION

Person re-identification has been actively studied for the last couple of years. It targets to re-identify probe image of a person from a gallery image set of people. This problem in general is difficult: Matching a person across different cameras is a challenging problem due to its intra-class variations of pose, viewpoint, illumination, resolution and occlusion. Due to various inherent difficulties, all previous person re-identification algorithms have drawbacks. Since existing dataset provides a perfectly aligned person with manually cropped full body images, algorithms trained on these datasets have difficulties in practical applications where severe occlusions occur. In order to overcome these limitations, Partial REID *et al.* [1] proposed new person re-identification problem referred to partial person re-identification. As shown in Fig.1, partial person re-identification aims to match partial body probe image from a set of full body gallery images.

This work was partly supported by the ICT R&D program of MSIP/IITP [2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion] and partly supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(NRF-2017R1A2B2006165).

Fig. 1. An illustration of the partial person re-identification problem. The left most column represents examples of partial body probe image and rest of the columns show examples of full body gallery images. Matching gallery images are marked with green border line.

Most person re-identification algorithms are based on person descriptor and a distance metric, and previous algorithms have been based on person descriptors using hand-crafted features. Commonly used features for person descriptor are color histograms and texture features [2], [3], [4], since the color and texture of skin and clothing remain same as view point differs. Distance metric learning focuses on how to train a metric to discriminate person identity. To this end, metric learning methods in person re-identification such as KISSME [5] and RDC [6] were proposed.

In recent years, a CNN-based deep models have had great success in person re-identification. The very first deep-learning based person re-identification algorithm [7] enabled the joint learning of feature extraction part and metric learning part in an end-to-end manner. In [7], Siamese network is utilized using CNN as a feature extractor and cosine similarity as a distance metric. In [8], CNN is used as feature extractor for triplet network and an effective Triplet generation algorithm was proposed. Classification model makes full usage of given label information by using person iden-

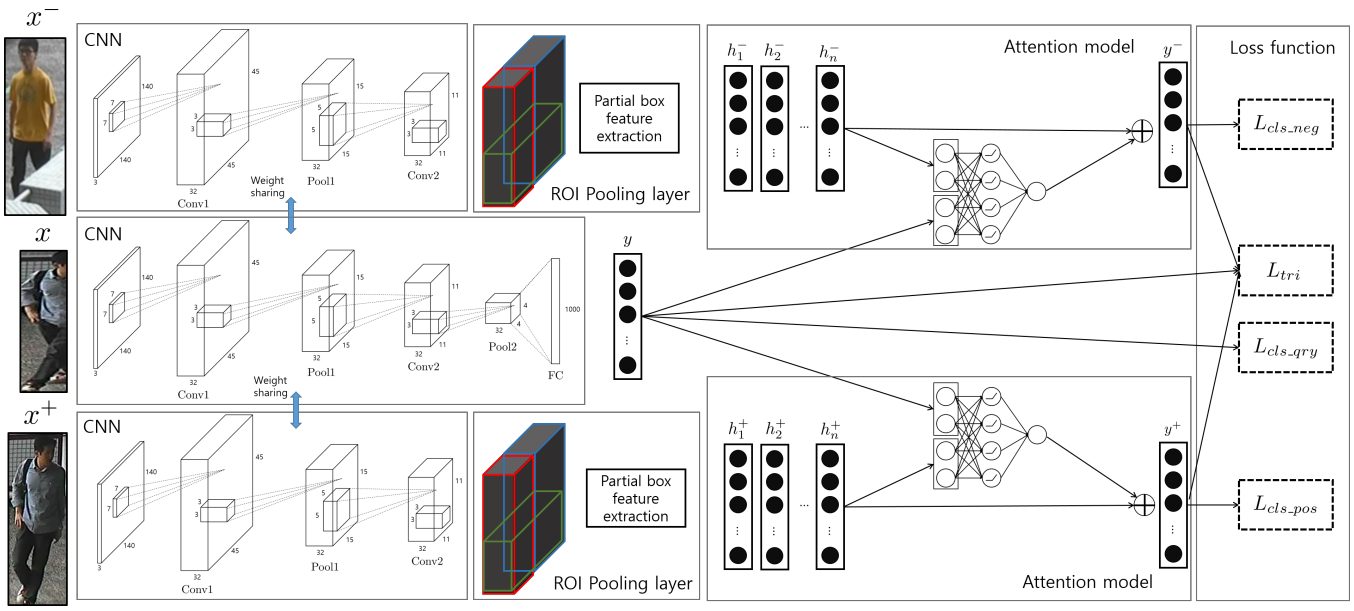

Fig. 2. overall framework of proposed DPPR. It consists CNN, RoI Pooling layer, Attention model and multi-task loss functions.

tity information. In previous works including [9], it is shown that classification task could learn discriminative feature thus greatly help the person re-identification. In [10], object detection framework Faster R-CNN [11] is used to jointly learn pedestrian detection and person re-identification.

In this paper, end-to-end deep model for partial person re-identification is proposed. The proposed algorithm will be referred as Deep Partial Person Re-identification (DPPR) for the rest of this paper. The main contributions of this paper are as follows. 1) To our knowledge, we are the first to apply deep learning on Partial Person Re-identification. 2) A novel framework for matching partial image to full image is proposed by utilizing CNN, RoI Pooling and Attention model. The RoI Pooling layer enables the extraction of feature vector corresponding to predefined part of input image. The attention model selectively focuses on the subset of CNN feature vectors.

2. PROPOSED ALGORITHM

2.1. The overall framework

As shown in Fig.2, the proposed framework consists of four main components: CNN, RoI Pooling, Attention model and multi-task loss functions. DPPR make use of CNN under the triplet framework. Triplet network is composed of three weight-sharing CNNs and uses triplet examples for training. Three input images $X = [x, x^+, x^-]$, form triplet example, where positive image x^+ and query image x are from the same person, while negative image x^- is from a different person. For partial person re-identification problem, query image corresponds to partial body probe image, and positive image, negative image are whole body gallery images.

To extract corresponding feature vectors $Y = [y, y^+, y^-]$, CNN, RoI Pooling layer and Attention model is utilized. Weight sharing CNNs can map raw input image into learned feature space. From the positive image and the negative image, we extract the feature vector corresponding to the portion of where the partial body probe image corresponds to the entire probe image. We utilize RoI Pooling layer and attention model for this. Then three classification loss functions and one triplet loss function are deployed to train the network.

2.2. CNN architecture

The CNN structure considered is inspired by CNN model proposed in [12]. As shown in Fig.3, it consists two convolution layers, two pooling layers and one fully-connected layer. First convolution layer (Conv1) consists of 32 feature maps with 7x7 sized kernel and stride of 3. Then MAX pooling layer

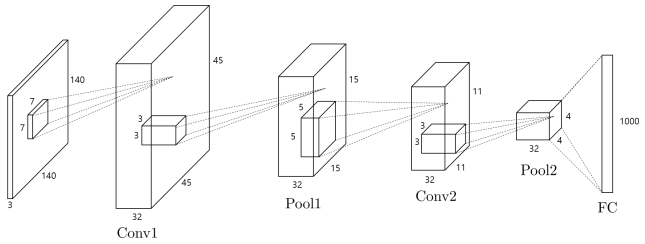

Fig. 3. CNN architecture of DPPR. The CNN architecture consists two convolution layer denoted as Conv1, Conv2 and two max pooling layer denoted as Pool1, Pool2, and fully-connected layer denoted as FC.

Fig. 4. Examples of predefined 13 partial boxes.

(Pool1) with 3x3 sized kernel and stride of 3 is used. Second convolution layer (Conv2) has 32 feature maps with 5x5 sized kernel and stride of 1. Then again, MAX pooling layer (Pool2) with 3x3 sized kernel and stride of 3 is used. Finally, fully-connected layer (FC) generates an output of 1000 dimension feature vector for each image input. For query image, above CNN is used to extract 1000 dimension feature vector. But for positive image and negative image, Pool2 layer is substituted to RoI Pooling layer which will be described in consecutive subsection.

2.3. RoI Pooling

Partial person re-identification uses partial body images as probe images. There is a large difference between the features extracted from partial body image and full body image and it will be inappropriate to use full body feature for partial person re-identification. In DPPR, we predefined 13 partial boxes for full body gallery image. As shown in Fig.4, 13 partial boxes includes a full image, left, right, up, down half of images, four horizontally equally spaced translation of left side partial image and four vertically equally spaced translation of upside partial image.

To extract feature vectors of corresponding partial boxes, RoI Pooling layer is utilized. Feature vector from each partial box represents the corresponding part of full body gallery person image. In details, Pool2 layer in CNN is substituted to a RoI Pooling layer for positive image and negative image. Since size of input image is 140x140 and feature map size of Conv2 layer is 11x11, we used a spatial ratio of 0.08 for RoI Pooling layer. then it generates 4x4 sized feature map just like Pool2 layer. RoI Pooling layer and successive FC layer outputs 13 feature vectors for each positive image and negative image, which is denoted positive RoI feature $h^+ = [h_1^+, \cdots, h_{13}^+]$, and negative RoI feature $h^- = [h_1^-, \cdots, h_{13}^-]$.

2.4. Attention model

Given positive RoI feature and negative RoI feature, Attention model [13],[14] is utilized to selectively focus on a small subset of predefined boxes and extract one feature vector out of 13 RoI features. Each attention weight represents how much the corresponding feature (or location of a partial box) is relevant to re-identify partial body probe image from a set of gallery images. For example, if partial probe image contains upper part of a body, large weights for partial boxes including upper body is desired. To accomplish this, we use a neural network conditioned on query feature y, since y contains the information of which part of query image belongs to a person.

Given the RoI features h^+ and h^-, calculate weighted sum of RoI features to generate positive feature and negative feature:

$$y^{\{+,-\}} = \sum_{i=1}^{13} \alpha_i^{\{+,-\}} h_i^{\{+,-\}}, \tag{1}$$

where α_i^+ and α_i^- are attention weights for positive, negative image corresponding to i-th partial box.

We use feedforward neural network conditioned on query feature y to calculate unnormalized score s_i^+ and s_i^-:

$$s_i^{\{+,-\}} = W^T \tanh(U h_i^{\{+,-\}} + V y + b), \tag{2}$$

where W, U, V, b are Attention model parameters learned during training.

After the unnormalized scores s_i^+ and s_i^- are calculated for all predefined boxes, we normalize them using softmax function:

$$\alpha_i^{\{+,-\}} = \frac{\exp(s_i^{\{+,-\}})}{\sum_{j=1}^{n} \exp(s_j^{\{+,-\}})}. \tag{3}$$

2.5. Loss function

Triplet loss function is utilized to pull feature vectors of query image and positive image and push feature vectors of query image and negative image. Given triplet example $X = [x, x^+, x^-]$ and extract feature vectors, $Y = [y, y^+, y^-]$, the similarity between input images are measured by the euclidean distance of feature vectors. The distance between query feature and positive feature is defined as $d^+ = ||y - y^+||_2$ and the distance between query image and negative image is also defined as $d^- = ||y - y^-||_2$. While training, triplet loss function constrains d^+ to be smaller than d^- by predefined margin τ. Following the constraint, triplet loss function is designed as follows:

$$L_{tri} = \max(0, \tau - d^- + d^+). \tag{4}$$

Along with triplet loss function, we also use three cross-entropy loss functions. For each feature vectors, one more fully-connected layer to produce logits, $\hat{t}, \hat{t}^+, \hat{t}^-$, and softmax output function are used to classify corresponding person identity. Following equations are utilized cross-entropy loss functions:

$$L_{cls\_qry} = -\sum_{c=1}^{C} t_c \log \hat{t}_c + (1 - t_c) \log(1 - \hat{t}_c), \tag{5}$$

$$L_{cls\_pos} = -\sum_{c=1}^{C} t_c \log \hat{t}_c^+ + (1 - t_c) \log(1 - \hat{t}_c^+), \tag{6}$$

$$L_{cls\_neg} = -\sum_{c=1}^{C} t_c^- \log \hat{t}_c^- + (1 - t_c^-) \log(1 - \hat{t}_c^-), \tag{7}$$

where C is the number of total train person identities, t is the label of query image, positive image, and t^- is the label of negative image.

The total loss is the sum of above four loss functions:

$$L = L_{tri} + L_{cls\_qry} + L_{cls\_pos} + L_{cls\_neg}. \quad (8)$$

In the proposed framework, we performed multi-task learning of classification and metric learning. Label information used in metric learning does not use full annotation of provided dataset label. Triplet loss function only requires sameness information of input images as label. Thus, we supplemented three classification loss functions to make full usage label information.

3. EXPERIMENTS

3.1. Datasets

The Partial REID [1] consists 600 images of 60 people. In Partial REID [1], 42 person identities are used as test set. The Partial REID dataset is not suited for training deep model since training set is smaller than test set. Hence, we constructed simulated dataset partial CUHK03 (p-CUHK03) based on CUHK03 [15] and used for evaluation.

The CUHK03 dataset includes 13,164 images of 1360 person identities. Each person identity contains 10 full body images captured by six different surveillance cameras and every images are obtained in campus environment. In general, 1160 person identities are used as training set, 100 person identities are used as validation set and 100 person identities are used as test set. To construct p-CUHK03, we selected five images with same viewpoint for each person identities. We generated 10 partial body probe images out of selected two images and the rest three images are used as full body gallery image. Partial images are randomly generated with full-to-partial spatial area ratio p. In this experiment, three values of p (0.3, 0.5, 0.7) is used to consider various sized occlusion scenarios and analyze the effect of occlusion size.

3.2. Evaluation

For quantitative evaluation of proposed algorithm, we adopted cumulative match curve (CMC) metric. We evaluated for Single-shot case (N = 1) where only one gallery image per person identity is used and Multi-shot case (N = 3) where three gallery image per person identity is used.

We compare DPPR to DPPR without Attention model (DPPR w.o. attn). Without using Attention model, average of positive RoI feature and negative RoI feature are used as positive feature and negative feature for DPPR w.o. attn.

The following tables show the experimental result of our proposed DPPR. We conducted experiment on our simulated dataset p-CUHK03 with $p = 0.3, 0.5, 0.7$.

Table 1. Experimental evaluations on p-CUHK03, $p = 0.3$

Methods	N=1			N=3		
	r=1	r=5	r=10	r=1	r=5	r=10
DPPR w.o. attn	16.3	40.3	55.1	21.1	47.4	61.3
DPPR	**22**	**44.3**	**57.6**	**25.9**	**54.6**	**65.5**

Table 2. Experimental evaluations on p-CUHK03, $p = 0.5$

Methods	N=1			N=3		
	r=1	r=5	r=10	r=1	r=5	r=10
DPPR w.o. attn	36.2	65.1	76.6	49.3	**76.3**	84.2
DPPR	**41.8**	**66.7**	**76.9**	**53.1**	76.2	**85.1**

Table 3. Experimental evaluations on p-CUHK03, $p = 0.7$

Methods	N=1			N=3		
	r=1	r=5	r=10	r=1	r=5	r=10
DPPR w.o. attn	57.4	80.1	87.5	75.3	91	95.9
DPPR	**60.4**	**83.9**	**89.4**	**75.4**	**92.9**	**97**

The experimental results clearly shows that DPPR outperforms our baseline model, DPPR w.o.attn. As shown in Table.1, DPPR achieves 5.7% higher matching rate at rank-1. DPPR shows 22% matching rate when only 30% part of person is given as probe image. Table.2 shows that DPPR achieves 5.6% higher matching rate at rank-1. DPPR shows 41.8% matching rate when half of person is given as probe image. Table.3 shows that DPPR achieves 3% higher matching rate at rank-1. DPPR achieves 60.4% matching rate when 70% part of person is given as probe image. The gap between DPPR and DPPR w.o. attn is bigger when p is small. Since small p indicates small probe image, attention model plays more important role in extracting matching partial box.

4. CONCLUSION

This paper introduces a novel algorithm referred to as deep partial person re-identification (DPPR) for partial person re-identification where only a part of a person is observed and full body images are available for identification. The DPPR is based on an end-to-end deep model which make use of convolutional neural network (CNN), RoI Pooling layer and attention model is proposed. The RoI Pooling layer enables the extraction of feature vector corresponding to predefined part of input image. The attention model selects a subset of CNN feature vectors. For qualitative evaluation of proposed model, data from CUHK03 are randomly cropped in constructing p-CUHK03. Experimental results shows that DPPR outperforms our baseline model on p-CUHK03.

5. REFERENCES

[1] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jian-Huang Lai, and Shaogang Gong, "Partial person re-identification," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 4678–4686.

[2] Loris Bazzani, Marco Cristani, and Vittorio Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Comput. Vis. Image Underst.*, vol. 117, no. 2, pp. 130–144, Feb. 2013.

[3] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, USA, June 2013.

[4] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, December 2013.

[5] Martin Hirzer, "Large scale metric learning from equivalence constraints," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2012, CVPR '12, pp. 2288–2295, IEEE Computer Society.

[6] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang, "Reidentification by relative distance comparison.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, 2013.

[7] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li, "Deep metric learning for person re-identification," in *ICPR*, 2014.

[8] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recogn.*, vol. 48, no. 10, pp. 2993–3003, Oct. 2015.

[9] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2014, CVPR '14, pp. 1891–1898, IEEE Computer Society.

[10] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang, "End-to-end deep learning for person search," *CoRR*, vol. abs/1604.01850, 2016.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 91–99. Curran Associates, Inc., 2015.

[12] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[14] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville, "Describing videos by exploiting temporal structure," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015.

[15] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.