Task-specific Neural Networks for Pose Estimation in Person Re-identification Task

Kai Lv[1], Hao Sheng[1], Yanwei Zheng[1], Zhang Xiong[1], Wei Li[1], and Wei Ke[2]

[1]State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, China
{lvkai,shenghao,zhengyw,xiongz}@buaa.edu.cn
liwei@nlsde.buaa.edu.cn
[2]Macao Polytechnic Institute, Macao, China
wke@ipm.edu.mo

Abstract. Person re-identification is a challenging task because of severe appearance changes of a person due to diverse camera viewpoints and person poses. To alleviate the impact of different poses, more and more studies have been done in pose estimation. In this work, we present three task-specific neural networks (TNN) algorithm designed to address the problem of pose estimation for re-identification in both single-shot and multi-shot matching. In order to recognize the human pose as one of the four classes(front, back, left, right), a PoseNet-A is first required to estimate the pose as class front-back or class left-right. Based on the results, we select the appropriate network(PoseNet-B1, PoseNet-B2) to obtain the final pose. According to the results, our method achieves very good results on a large data set(CUHK03-Pose). One thing that needs to be pointed out is that we build the dataset CUHK03-Pose which is based on the person re-identification dataset CUHK03.

Keywords: re-identification, pose estimation, single-shot, deep learning

1 Introduction

Person re-identification (re-id) is of great importance in security, human computer interaction and many other systems. A typical scenario of person re-id is to identify people across images that have been taken using different cameras, or across time using a single camera. However, re-id still remains a challenging problem due to illumination, pose and viewpoint changes for the people across images that have been taken using different cameras. Thereinto, the pose diversity is one of the most prominent problems.

We classify previous person re-id methods into single-shot and multi-shot matching-based methods. For single-shot matching, most of the works have focused on appearance-based techniques such as feature and metric learning for the efficient person re-id. M. Farenzena *et al.* [6] presented an appearance-based method for person re-id. It consists in the extraction of features that model three complementary aspects of the human appearance. Feature learning methods that

(a) (b) (c) (d)

Fig. 1. Sample images from the CUHK03 dataset [13]. We have roughly divided the pose into four categories: (a) front, (b) back, (c) left and (d) right.

select or weight discriminative features have been proposed in [15, 19]. T. Xiao *et al.* [18] presented a pipeline for learning deep feature representations from multiple domains with Convolutional Neural Networks (CNNs). Several methods have been applied to the re-id problem, such as KISSME [11], LMNNR [5]. Martin *et al.* [11] raised important issues on scalability and the required degree of supervision of existing Mahalanobis metric learning methods. Mert Dikmen *et al.* [5] used a metric learning framework to obtain a robust metric for large margin nearest neighbor classification with rejection.

Regarding the multi-shot matching, several person re-id methods have been proposed in recent years. Farenzena *et al.* [6] provided multi-shot matching results by comparing each possible pair of histograms between different signatures (a set of appearances) and selecting the obtained lowest distance for the final score of matching. T. Wang [16] presented a novel model to automatically select the most discriminative video fragments from noisy image sequences of people where more reliable space-time features can be extracted, whilst simultaneously to learn a video ranking function for person re-id. Y. Li *et al.* [14] also proposed an algorithm based on the Fisher criterion to learn a representative and discriminative feature sub-space from image sequences in person re-id task.

As shown in Fig. 1, it is quite challenging for person re-id due to target pose variations. However, it is reasonable to explore and utilize the pose priors of targets in every non-overlapping camera in advance. Recently, a few works [3, 17, 4] used target pose priors (pose cues) for person re-id very recently. S.Bak

et al. [3] proposed to learn a generic metric pool which consists of metrics, each one learned to match specific pairs of poses. Z. Wu *et al.* [17] built a model for human appearance as a function of pose, using training data gathered from a calibrated camera and then applied this pose prior in online re-id to make matching and identification more robust. Cho *et al.* [4] propose a novel framework for by analyzing camera viewpoints and person poses, which robustly estimates target poses and efficiently conducts multi-shot matching based on the target pose information.

However, previous works [3, 17, 4] estimated the pose only by using 3D scene information and motion of the target. Pose orientation is computed as a dot product between the viewpoint vector and the motion vector. It is viable for multi-shot matching, which enables us to extract additional cues such as the 3D position of the targets. Because of the lack of additional cues for single-shot re-id, pedestrian pose estimation is a quite difficult task.

In this work, we present three task-specific neural networks (TNN) algrithm designed to address the problem of pose estimation in single-shot re-id. Also, the method can be applied in the multi-shot matching task since only an input image is required to estimate pose. Another contribution is that we build the dataset CUHK03-Pose which is based on the re-id dataset CUHK03. The CUHK03-Pose dataset contains 13,164 images of 1,360 pedestrians and each image has a pose label.

2 Task-specific Neural Networks

In order to recognize the human pose as one of the four classes(front, back, left, right), three PoseNets (PoseNet-A, PoseNet-B1, PoseNet-B2) are required. In this section, we describe the proposed framework for the pose estimation. First, we show the overall framework of PSE for person re-identification. (Sec 2.1). Second, we introduce the two pose classes (front-back and left-right) we classified for pedestrians Sec 2.2). Third, we present the deep neural network architecture of PoseNet-A and PoseNet-B (Sec 2.3). Finally, we introduce the dataset CUHK03-Pose which is built for the pedestrain pose estimation problem (Sec 2.4).

2.1 Overall Framework

As shown in Fig. 2, we have several images in appearance pool A and we eventually classify them into four categories: front, back, left and right. The PoseNet-A of TNN is first needed to estimate the pose as front-back or left-right. The class front-back images are put into the appearance pool B1 and the class left-right images are put into the appearance pool B2. Based on the first classification, we then select the corresponding network. PoseNet-B1 is designed and trained to divide the class front-back into the class front and class back while PoseNet-B2 is to divide the class left-right into the class left and right. Finally, we yield the final estimation of which class the input images are. It should be pointed out that

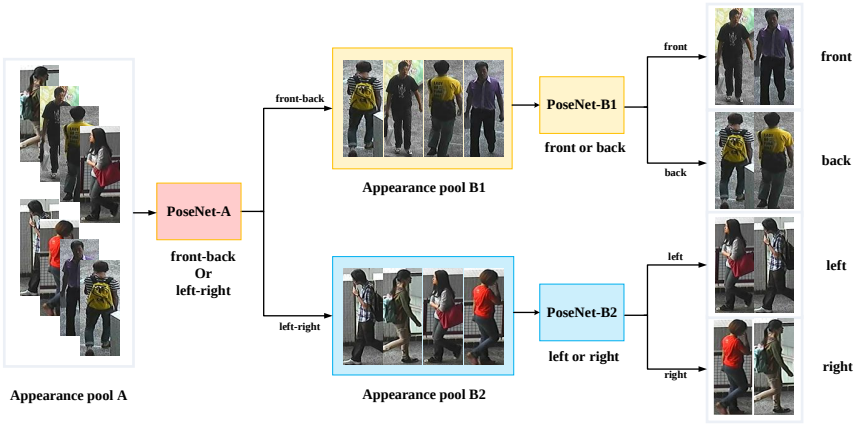

Fig. 2. Overall framework. PoseNet-A, PoseNet-B1 and PoseNet-B2 have the same architecture but different tasks. A group of images are classified into four categories using the proposed algorithm.

the three PoseNets have the same architecture but are different in classification use. 1

2.2 Front-back and Left-right

As shown in Fig. 1, pedestrian pose estimation is a problem of dividing the pose into four categories (front, back, left and right). It is quite reasonable and valid to turn the question into two binary problems. First, we estimate the pose as class front-back or class left-right using PoseNet-A. The poses are classified like this for the reason that more parts of the body are exposed to the class front and class back images. Moreover, the body silhouette of the class front and class back ones are relatively fixed. After the first binary classification, TNN can get the class front-back or class left-right and choose the corresponding PoseNet-B1 or PoseNet-B2. We regard the TNN as three experts, who have an advantage in their own field. By dividing the pose into two classes (front-back, left-right), TNN split the big task into small tasks. Thus, the experts (PoseNets) can work better in their respective fields by which the network architecture can also be designed lighter and work faster.

2.3 Network Architecture of TNN

As the three PoseNets of TNN are of the same architecture, we describe the PoseNet-A as an example in this subsection. As shown in Fig. 3, to determine whether the pose is the class front-back or class left-back, we need to design an appropriate architecture to solve the binary classification problem.

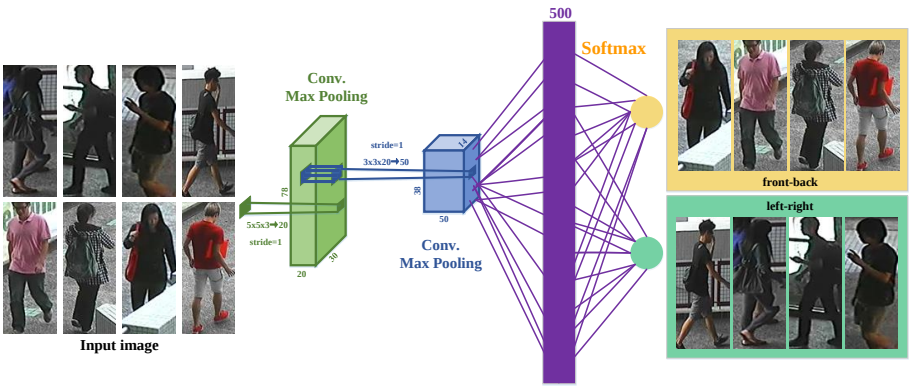

Fig. 3. Proposed architecture. An input image is passed through the network. PoseNet-A's outputs means class front-back or class left-right, while the PoseNet-B1 (PoseNet-B2) network means front (left) or back (right). The number and size of convolutional filters that must be learned are shown. For example, in the first convolution layer, $5 \times 5 \times 3 \rightarrow 20$ indicates that there are 20 convolutional features in the layer, each with a kernel size of $5 \times 5 \times 3$.

In the deep learning literature, convolutional features have proven to provide representations that are useful for a variety of classification tasks. The first layer is convolution layers, which we use to compute higher-order features. In the first convolution layer, we pass an input RGB image of size $64 \times 160 \times 3$ through 20 learned filters of size $5 \times 5 \times 3$. The resulting feature maps are passed through a max-pooling kernel that halves the width and height of features. Then the features are passed through another convolutional layer that uses 50 learned filters of size $3 \times 3 \times 50$, followed by a max-pooling layer that again decreases the width and height of the feature map by a factor of 2. At the end of the two layers, each input is represented by 50 features maps of size 14×38. We then apply a fully connected layer. The resultant feature vector of size 500 is passed through a ReLu nonlinearity. These 500 outputs are then passed to another fully connected layer containing 2 softmax units, which represent the probability that the image is class front-back or class left-back.

2.4 CUHK03-Pose Dataset

In this paper, a pose estimation dataset, the "CUHK03-Pose" dataset, is proposed. The CUHK03-Pose dataset is based on the large dataset CUHK03 [13] which includes 13,164 images of 1360 pedestrians. The whole dataset captured with ten surveillance cameras. Each identity is observed by two disjoint camera views and has an average of 4.8 images with pose labels in each view. Overall, there are two reasons to utilize CUHK03 dataset.

First, a large amount of images are required for deep learning task. There are many specialized datasets for person identification such as Market-1501 [20],

Datasets	CUHK03[13]	CUHK01[12]	VIPeR [7]	PRID[8]	3DPeS[2]	i-LIDS[21]
cameras	10	2	2	2	8	2
identities	1360	971	632	934	193	119
images	13,164	1,552	1264	1,134	1,012	476

Table 1. Details of existiong datasets [13, 12, 7, 8, 2, 21].

CUHK03 [13], CUHK01 [12], VIPeR [7], PRID2011 [8], 3DPeS [2], i-LIDS [21] and Shinpuhkan [10]. As we can see in Table 1, the identities and images of CUHK03 dataset are larger than the ohter four datasets.

Second, when observing the CUHK03 dataset, we found the dataset has two characteristics: there exist no more than 10 images per person and the poses of image 1-5 are always left-right while the rest are front-back. Then, it is reasonable to label the images as front-back and left-right respectively. Finally, each image of the CUHK03 dataset is labeled as front, back, left or right. At the same time, because there are also very few obvious false labels, we have made a manual correction. To the best of our knowledge, this is the first attempt to build a large dataset for pedestrian pose estimation. Sample images are shown in Fig. 4

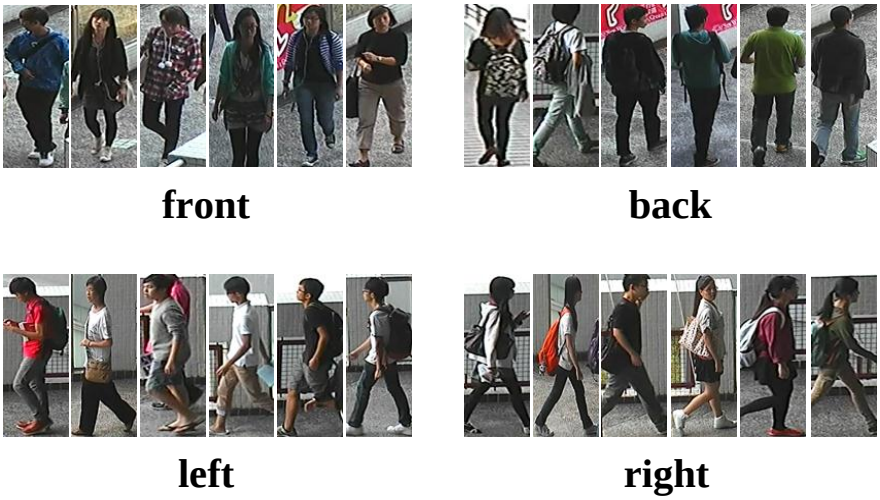

Fig. 4. Sample images of the "CUHK03-Pose" dataset.

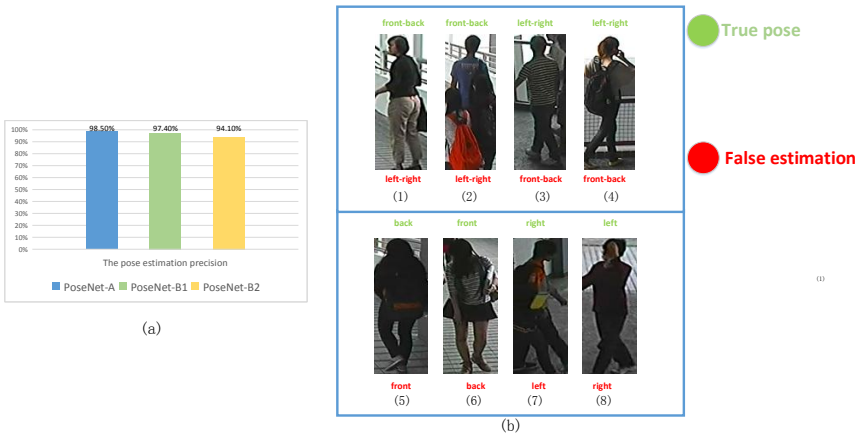

Fig. 5. Experimental results. (a) shows the precision of the three PoseNet. (b) have listed several wrong estimates. The true pose is listed upon every picture with green color and the words under each picture with red color mean a false estimation.

3 Experiments

3.1 Data Preparation

We regard the pedestrian pose estimation problem as binary classification. All the images of CUHK03-Pose are labeled as front, back, left, right. However, the training data of PoseNet-A consist of images labeled as positive (front-back) and negative (left-right). Indeed, the front-back images are the union of class front images and class back images while the left-right images are the union of class left images and class right images. When training the PoseNet-B1 or PoseNet-B2, we only use the front-back or left-right images to achieve the specific network.

In order to train the network in caffe [9], training data are randomly divided into mini-batches. The model performs forward propagation on the current mini-batch and computes the output and loss. Backpropagation is then used to compute the gradients on this batch, and network weights are updated.

To reduce overfitting, we artificially augment the data by performing random 2D translation, as also done in [1]. For an original image of size $W \times H$, we sample 5 images around the image center, with translation drawn from a uniform distribution in the range $[-0.05H, 0.05H] \times [-0.05W, 0.05W]$.

We start with a base learning rate of $\gamma^{(0)} = 0.01$ and gradually decrease it as the training progresses using the "inv" policy: $\gamma^{(i)} = \gamma^{(0)}(1 + \gamma \cdot i)^{-p}$ where $\gamma = 10^{-4}, p = 0.75$. The momentum and weight decay in our experiment are $u = 0.9$ and $\lambda = 0.0001$.

3.2 Evaluation

Results of three PoseNets. We evaluate TNN in CUHK03-Pose by calculating the precision of each PoseNet. As shown in Fig. 5 (a), the three PoseNets get a good performance, although there exist some false pose estimations. The precisions of the three networks ranges from 94.1% to 98.5%. However, our network have some shortcomings as we can see in Fig. 5 (b). The true pose is listed upon every picture with green color and the words under each picture with red color mean a false estimation. Through image (5) (6), we can draw conclusions that we may get false estimation when the light is too dark. Also, image (2) (7) shows that the pose estimation results are easy to be affected by occlusion.
Overall Results. In this paper, TNN divides the pose of pedestrians into front-back and left-right at first. To evaluate the effectiveness of our two-stage approach, a classifier that directly decomposed into the four classes is trained. This classifier has the same network architecture and parameters but with four outputs. In our experiment, the overall precision of this four-class classifier is 90.1%. However, the final precision of TNN can be achieved by calculating the three PoseNets' precision. The overall precision of TNN is 94.6% whcih has a 4% improvement than the four-class classifier.

4 Conclusion

Previous works estimate pedestrian pose by using 3D scene information and motion of the target in multi-shot matching which is not viable in single-shot scenario. To estimate pose in both single-shot and multi-shot matching, we solve the pedestrian pose estimation problem by using three PoseNets of TNN. In order to recognize the human pose as one of the four classes(front, back, left, right), a PoseNet-A is first required to estimate the pose as class front-back or class left-right. Based on the results obtained, we then select the appropriate network(PoseNet-B1, PoseNet-B2) to obtain the final pose. Experiments show that our method is very effective and also have a high precision. In this paper, although the algorithm performs well at CUHK03-Pose dataset, there are a lot of works to do to apply this algorithm to other datasets. Moreover, we need to use the pose cues for person re-identification works. In future, we plan to explore the potential of using pose cues to solve the re-identification problem.

Acknowledgement

This study is partially supported by the National Natural Science Foundation of China(No.61472019), the National Science Technology Pillar Program (No.2015BAF14B01), the Macao Science and Technology Development Fund (No.138/2016/A3), the Programme of Introducing Talents of Discipline to Universities, the Open Fund of the State Key Laboratory of Software Development Environment under grant SKLSDE-2017ZX-09 and HAWKEYE Group.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Computer Vision and Pattern Recognition. pp. 3908–3916 (2015)
2. Baltieri, D., Vezzani, R., Cucchiara, R.: 3dpes: 3d people dataset for surveillance and forensics. In: International ACM Workshop on Multimedia Access To 3d Human Objects. pp. 59–64 (2011)
3. Bk, S., Martins, F., Bremond, F.: Person re-identification by pose priors. Proceedings of SPIE - The International Society for Optical Engineering 9399, 93990H–93990H–6 (2015)
4. Cho, Y.J., Yoon, K.J.: Improving person re-identification via pose-aware multi-shot matching. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1354–1362 (2016)
5. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian recognition with a learned metric. In: Computer Vision - ACCV 2010 - Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers. pp. 501–512 (2010)
6. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Computer Vision and Pattern Recognition. pp. 2360–2367 (2010)
7. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking (2007)
8. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. Lecture Notes in Computer Science 6688(12), 91–102 (2011)
9. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
10. Kawanishi, Y., Yang, W., Mukunoki, M., Minoh, M.: Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In: The Korea-Japan Joint Workshop on Frontiers of Computer Vision, Fcv (2014)
11. Kstinger, M., Hirzer, M., Wohlhart, P., Roth, P.M.: Large scale metric learning from equivalence constraints pp. 2288–2295 (2012)
12. Li, W., Wang, X.: Locally aligned feature transforms across views. In: Computer Vision and Pattern Recognition. pp. 3594–3601 (2013)
13. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 152–159 (2014)
14. Li, Y., Wu, Z., Karanam, S., Radke, R.J.: Multi-shot human re-identification using adaptive fisher discriminant analysis. In: British Machine Vision Conference. pp. 73.1–73.12 (2015)
15. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: what features are important? In: International Conference on Computer Vision. pp. 391–401 (2012)
16. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: European Conference on Computer Vision. pp. 688–703 (2014)
17. Wu, Z., Li, Y., Radke, R.J.: Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. IEEE Transactions on Pattern Analysis Machine Intelligence 37(5), 1095–1108 (2015)

18. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification (2016)
19. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: Computer Vision and Pattern Recognition. pp. 144–151 (2014)
20. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE International Conference on Computer Vision. pp. 1116–1124 (2015)
21. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings (2009)