SCIENCE CHINA
Information Sciences

• LETTER •

November 2020, Vol. 63 219102:1–219102:3
https://doi.org/10.1007/s11432-019-2782-3

Mobile person re-identification with a lightweight trident CNN

Mingfu XIONG[1], Dan CHEN[2]* & Xiaoqiang LU[3]

[1]*Engineering Research Center of Hubei Province for Clothing Information, School of Mathematics and Computer Science, Wuhan Textile University, Wuhan 430200, China;*
[2]*National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China;*
[3]*Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710068, China*

Received 8 November 2019/Revised 11 December 2019/Accepted 8 January 2020/Published online 15 April 2020

Citation Xiong M F, Chen D, Lu X Q. Mobile person re-identification with a lightweight trident CNN. Sci China Inf Sci, 2020, 63(11): 219102, https://doi.org/10.1007/s11432-019-2782-3

Dear editor,
Person re-identification (re-ID) aims to recognize or associate a person and time at non-overlapping physical locations after the person had been previously observed visually elsewhere. This technology is considered increasingly important in sustaining social security for smart cities [1]. The performance of re-ID has recently been significantly enhanced through deep learning algorithms and they have now dominated the re-ID applications [2]. However, despite their great potentials in daily social security tasks, mobile re-ID applications are still rare because the optimization of such deep neural network demands huge computing resources [2,3]. Its real-time operation will certainly encounter performance bottleneck on mobile devices.

In this study, we develop a lightweight solution using a "trident" convolutional neural network (T-CNN) that comprises three branches, each conforming to the following CNN architecture: (1) the semantic network branch (s-Net) captures the "semantic" features of an image by mimicking the visual processing mechanism of humans and (2) color network (c-Net) and body network (b-Net) branches strengthen the low-level visual features (e.g., color and body parts). These features are then passed through the l_2-normalization layer to construct a coarse-to-fine descriptor together that serves as the key to mobile re-ID applications.

The proposed framework. Figure 1 shows the architecture of T-CNN, and the backbone network is CaffeNet [4], well-known for its good performance with small datasets. It comprises three branches, namely, s-Net, c-Net, and b-Net for learning different types of features defined as follows:

- The s-Net comprises seven convolutional layers and five max-pooling layers to capture the semantic features for each person, and the semantic information may be hair, clothing, trousers, shoes.
- The c-Net structure includes four convolutional layers and three max-pooling layers that are used for the low-level visual feature learning (i.e., color information).
- Information on different regions of the human body is obtained via the b-Net that is formed using five convolutional layers, three max-pooling layers, and a spatial pyramid pooling layer.

It should be noted that the outputs of the last two full connected layers of each branch are both regarded as the person descriptors.

The s-Net excavates the "semantic" information of an image inspired by the neural perception mechanism of human brain. It first extracts the convolutional feature map $X \in \mathbb{R}^{H\times W\times C}$, where $H \times W$ define the size of the map and C denotes

* Corresponding author (email: dan.chen@whu.edu.cn)

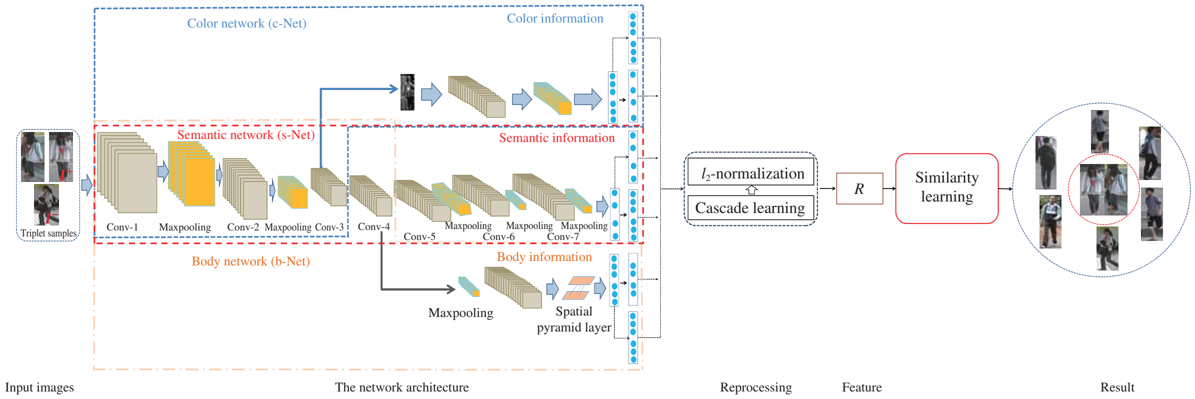

Figure 1   (Color online) The solution of T-CNN architecture for person re-identification. The input images include the triplet samples denoted by $\langle I, I^+, I^- \rangle$; the T-CNN comprises the s-Net, c-Net, and b-Net; person feature (R) is formed via cascade learning and l_2-normalization; metric learning applies to the final results of person re-identification.

the channel. Then, the body part detector [5] divides X into four horizontal parts, defined as the regions of interests (ROIs) for each person as described in (1), via the peak activation ($X_i(h,w)$), for each location (h,w) on the i-th feature channel.

$$(h_i, w_i) = \arg\min_i X_i(h,w). \tag{1}$$

A global average pooling is then performed on the feature maps of the four ROIs similar to AFN [5] to support recognition of the semantic attributes.

The c-Net extracts the low-level color information that should be consistent for each person. It yields a 1028-dimensional feature vector from each image that is calculated by subtracting the mean of the image and performing a forward propagation.

Then, the raw image is divided into three channels (RGB) to extract the features of three color components. We apply cascade learning to integrate the components to maintain the consistency of the color information for each person.

The b-Net is designed to capture the body information. Its pooling layer comprises three kinds of pooling maps (256-dimensional) organized in a "spatial pyramid" form, and the samples are set as 16, 4, and 1, corresponding to enable downsampling. Then, the hierarchical sampling features are joined together to represent the body information that includes three regions: head, torso, and legs.

The traditional CNNs fix the size of input of the full connection layer (e.g., 224 × 224 for AlexNet). However, the spatial pyramid design can adapt to inputs of various sizes and scales for body representation. The three kinds of spatial pooling operations are implemented on previously divided regions. The "multi-region" descriptors, i.e., the body features, are then obtained using downsampling feature learning.

We performed the network optimization via the triplet loss. In the training phase, the triplet images (i.e., anchor, positive, and negative samples, written as I, I^+, and I^-, respectively) are fed into the T-CNN to obtain the person features (written as R_w, R_w^+, R_w^-, and W, denoting the parameter set). The learned features of the positive samples R_w^+ should have a higher similarity with R_w. As shown in (2), the distance between R_w and R_w^+ should be shorter than that between R and R_w^-. This study only calculates the overall loss of the T-CNN for simplification. In (2), given the training set I, the triplet constraints are converted to the minimization problem of minimizing the distance between the same class and maximizing the different ones. N is the number of the training triplets, and the formula is given as

$$\begin{aligned} \mathrm{d}(W,I) = \sum_1^N & \max\left(\left\|R_W(I) - R_W(I^+)\right\|^2 \right. \\ & \left. - \left\|R_W(I) - R_W(I^-)\right\|^2, \varepsilon\right), \end{aligned} \tag{2}$$

where $\varepsilon = 1$ similar to that in hinge-loss functions.

Deep hierarchical feature learning. The T-CNN is designed to ensure the richness and integrity of the person features. To quickly converge T-CNN, the l_2-normalization layer is exploited to normalize the features:

$$y = \frac{f}{\sqrt{\sum_{p=1}^{k} f_p^2}}, \tag{3}$$

where $f = [f_1, f_2, \ldots, f_p]$ is the output of the concatenation layer with the dimension k.

In the testing phase, a triplet unit is fed forward into the trained model and exports the triplet features (f, i.e., the person descriptor). The ranking units can then be obtained according to the similarities between fs, and entries on the top can be referenced to match the correct person pairs.

Efficient person re-identification upon mobile devices. This study develops a lightweight solution toward mobile person re-identification. In this model, the efficiency of feature extraction is very important [6,7]. We reduce the number of network parameters significantly, unlike the ResNet, and still maintain the accuracy. The time complexity of optimizing T-CNN is as low as $O(n)$ for person feature extraction. It is ideal for applications on mobile devices that require a limited amount of computing resources [8, 9]. For example, on top of the MagicBook Pro series smart Tablet PC[1], this proposed model can match more than 5000 special instances in 3 s. More importantly, we perform experiments on the public datasets (Market-1501 and VIPeR), and the results demonstrate that the proposed method outperforms the current cutting-edge counterparts (e.g., the performance improvement of nearly 10% at rank@1 on Market-1501 and 2% for VIPeR in contrast to MVLDML+). Moreover, we have provided a more detailed comparison with the state-of-the-art methods in Appendix A.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61772380, 61977027), Major Project for Technological Innovation of Hubei Province (Grant No. 2019AAA044), Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) (Grant No. 2019AEA170), Foundation for Innovative Research Groups of Hubei Province (Grant No. 2017CFA007), Science and Technology Planning Project of Shenzhen (Grant No. JCYJ20170818112550194), and Open Foundation for Engineering Research Center of Hubei Province for Clothing Information (Grant No. 184084012).

Supporting information Appendix A. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Bai S, Bai X, Tian Q. Scalable person re-identification on supervised smoothed manifold. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 2530–2539

2 Hou R, Ma B, Chang H, et al. VRSTC: occlusion-free video person re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 7183–7192

3 Bai S, Tang P, Torr P H S, et al. Re-ranking via metric fusion for object retrieval and person re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 740–749

4 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012. 1097–1105

5 Yu R, Dou Z-Y, Bai S, et al. Hard-aware point-to-set deep metric for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 188–204

6 Yu R, Chao Z Z, Bai S, et al. Divide and fuse: a re-ranking approach for person re-identification. In: Proceedings of British Machine Vision Association, 2017. 1–13

7 Zhao C R, Chen K, Zang D, et al. Uncertainty-optimized deep learning model for small-scale person re-identification. Sci China Inf Sci, 2019, 62: 220102

8 Chen D, Tang Y, Zhang H, et al. Incremental factorization of big time series data with blind factor approximation. IEEE Trans Knowl Data Eng, 2019. doi: 10.1109/TKDE.2019.2931687

9 Huang T T, Xu Y C, Bai S, et al. Feature context learning for human parsing. Sci China Inf Sci, 2019, 62: 220101

[1] http://www.huawei.com/.