SCIENCE CHINA
Information Sciences

• LETTER •

July 2020, Vol. 63 179203:1–179203:3
https://doi.org/10.1007/s11432-018-9633-7

Deep learning network for UAV person re-identification based on residual block

Shujian ZHANG* & Chen WEI

School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, *China*

Received 21 June 2018/Accepted 4 July 2018/Published online 27 February 2020

Citation Zhang S J, Wei C. Deep learning network for UAV person re-identification based on residual block. Sci China Inf Sci, 2020, 63(7): 179203, https://doi.org/10.1007/s11432-018-9633-7

Dear editor,

Unmanned aerial vehicles (UAVs) are well known for their flexibility and adaptability [1]. In recent years, UAVs have been commonly used as maneuverable cameras in surveillance systems. Consequently, people, as the primary monitoring targets, have become the most popular objects for detecting or tracking [2, 3]. Naturally, the task of aerial person image re-identification has gained considerable importance.

The classic person re-identification problem is a challenging task derived from the multi-target tracking task in computer vision. The main idea is to match the same person across different camera views. It is challenging because of the variety of viewpoints and illumination conditions. Person re-identification is usually performed by a combination of feature extraction and similarity calculation [4].

Some of the commonly used feature extraction methods include color histogram, Gabor feature and local binary patterns (LBP). These approaches try to modify the feature extraction part, thus eliminating the effects caused by a variety of poses and angles to some extent. In terms of similarity calculation, a large number of studies aim to find a suitable calculation metric for the re-identification task, such as Mahalanobis distance local Fisher discriminant analysis, cross camera quadratic discriminant analysis, and local adaptive decision function.

Obviously, all of the hand-crafted feature extraction and similarity calculation approaches have been developed based on people's experience in machine learning, which in consequence have inevitable limits, especially for aerial images. Compared with the images for classic person re-identification, the aerial person images, shown in Figure S1, usually have lower resolution and are obtained under harsh illumination conditions. In this scenario, the classic methods are mostly unable to obtain satisfying results. Besides, separately conducting the feature extraction and similarity calculation also has a negative impact on the global optimal solution. Schumann et al. [5] used a relevance feedback strategy to improve the feature selection process and produce more training data. However, the training process still requires a similarity calculation step, which may lead to classification failure when the target has severe appearance differences.

We develop a new deep learning network and unify the feature extraction and similarity calculation processes. Experiments are conducted on public datasets in order to demonstrate the improvements in the classic person re-identification task. Meanwhile, we modified the training process by introducing the affine transformation mechanism and aligned our network with the aerial person image re-identification problem. Visualization results are presented at the end of this study.

UAV person re-identification. Person re-identification is a challenging computer vision task. As the name implies, it involves re-

* Corresponding author (email: zhangshujian@buaa.edu.cn)

identifying the images that belong to the same person but obtained from different camera views. In UAV person re-identification, the person re-identification task is performed using aerial person images that have lower resolutions and worse deformation conditions. There are macro and micro accuracy metrics in the person re-identification task, (mean average precision) mAP [6] and (cumulative matching curve) CMC [7]. CMC metrics calculates the accuracy from the perspective of similarity ranking. When a query image is given, it ranks all the gallery images by calculating the distance between the query image and each image in the gallery. Those gallery images that have the top similarity results are considered to have the same ID. mAP metrics describes the person re-identification accuracy from a higher view. As it focuses on the average accuracy precision, it is designed for large datasets in which each query image has more than one correct image in the gallery. Both the metrics are able to reflect the accuracy of our classifier.

We regard the person re-identification problem as a classification task. In other words, our network is a classifier whose task is to decide whether two images belong to the same person. If two images are of the same person, they form a positive sample pair, otherwise, a negative sample pair. Therefore, the training sample can be given as (x_i, x_j), y_{ij}, in which x_i and x_j represent the two input images and y_{ij} is the label, which represents the relationship between the inputs. When $y_{ij} = 1$, it means that x_i and x_j belong to the same person and x_i and x_j form a positive pair [8]. Therefore, the classification loss function can be given in the Euclidean space as follows:

$$L_{\text{class}} = \sum_{i,j}^{N} [y_{ij} d + (1 - y_{ij}) \max(0, \alpha_{\text{class}} - d)],$$
$$d = \|f(x_i) - f(x_j)\|_2^2, \tag{1}$$

where $f(x)$ in (1) represents the feature map of the input image through the deep learning network, α_{class} is the threshold value for controlling the distance between two different classes, and N represents the amount of training data.

A deep learning network based on residual block. ResNet is a deep network architecture, which was first proposed by He et al. [9] in 2016 to solve the degeneration problem of deep neural networks. The degradation problem indicates that, in some cases, the optimization process can be simplified. Therefore, the basic and most important structure in ResNet is the residual block which can provide a shortcut for the input and make it possible to build a much deeper network. Through the residual block, the origin learning function $H(x)$ is recast into $F(x) + x$. As a consequence, the number of weighted parameters is decreased considerably, and according to the experiments in [9], the optimization performances are barely changed.

The original ResNet contains 34 layers including a convolutional layer, a fully-connected layer, and 16 residual blocks. As for the person re-identification task, the 34-layer ResNet cannot achieve the required level of accuracy. Comparison experiments are described later in this study. A deeper network does not necessarily indicate a better final performance. Owing to the peculiarity of the person re-identification task, the datasets for person re-identification are much smaller compared with the general classification datasets. When a new ID is added, the manual labeling cost increases exponentially. Therefore, almost all the public person re-identification datasets fail to support huge deep learning networks. Hence, we build a 62-layer network based on the residual blocks. The network structure is shown in Figure S2. Every three squares represent a residual block, and cov denotes convolutional layers. Owing to the existence of the residual block, one specific layer has a certain possibility of becoming a simple connection instead of a series of weighted parameters. Therefore, the duplication of layers is necessary. The solid line indicates that the connection between two residual blocks is a simple connection, similar to the structure mentioned above. The dashed line represents another connection between convolutional layers with different dimensions. In that case, we need to add a 1×1 convolutional layer in order to fit the dimension based on

$$Y = F(X, \boldsymbol{W}_i) + \boldsymbol{W}_s X, \tag{2}$$

where $\boldsymbol{W}_s$ is a linear projection matrix for dimension matching and $\boldsymbol{W}_i$ is a set of weighted parameters in the residual blocks.

Experiment result. We chose three well-known public person re-identification datasets as our experiment target and ResNet34 as our benchmark. The small dataset is VIPeR and the two larger datasets are Market1501 and CUHK03. The accuracy results are shown in Table 1.

From the table, it can be seen that our network works better on the larger datasets. In both Market1501 and CUHK03, our network significantly outperforms ResNet34 by 6% and 4.4%. However, in terms of VIPeR, our network fails to improve the final performance. The main reason for this is the variation in the size of the datasets. Because each person in the training dataset represents a

Table 1 ResNet34 accuracy

Dataset	CMC rank 1 (%)	CMC rank 5 (%)	mAP (%)
VIPeR	72.1	74.8	62.7
Market1501	75.2	82.5	66.9
CUHK03	79.3	85.3	69.9
VIPeR[Ours]	71.9	73.9	62.1
Market1501[Ours]	81.2	89.6	70.9
CUHK03[Ours]	83.7	89.8	72.9

class, the number of images of a person is equal to the number of training data for such a class. As for VIPeR, each person has only two training images, and such a small number of training samples is far from enough. The deep learning network structure proposed in this study requires a certain number of training samples, so the performance on VIPeR is unsatisfactory. Above all, our proposed network is feasible and has its own advantages in person re-identification tasks.

When we transfer the deep learning network trained on classic person re-identification datasets to aerial images, the differences in viewpoints have a crucial impact on the final accuracy performance. The aerial person images always have a higher viewpoint compared with the classic person re-identification images. Therefore, in order to align the network with the aerial viewpoint better, we introduce the affine transformation in image preprocessing. Figure S3 shows a couple of qualitative results of the UAV person re-identification experiments based on aerial images. The blue image is the query image and the images that follow are the gallery images ranked according to the similarity distance. Green and red indicate the correct and wrong re-identification results. As shown in the qualitative result, we can acquire satisfying results using the proposed method. When the target persons have similar main clothes color, the classifier may make a mistake. The last case in the qualitative result shows the worst performance when the illumination condition is severe.

Conclusion. The proposed network achieves considerable improvement in the classic person re-identification task. Our network was successfully applied to the novel problem of aerial person image re-identification. Admittedly, the accuracies of both tasks still have room for improvement. Our work provides an instructive idea and takes a step in this new area.

Supporting information Figure S1 aerial person images, Figure S2 network structure, and Figure S3 qualitative results. 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 Peng K M, Lin F, Chen B M. Online schedule for autonomy of multiple unmanned aerial vehicles. Sci China Inf Sci, 2017, 60: 072203
2 Avola D, Foresti G L, Martinel N, et al. Aerial video surveillance system for small-scale UAV environment monitoring. In: Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance, Lecce, 2017
3 Duan H B, Li H, Luo Q N, et al. A binocular vision-based UAVs autonomous aerial refueling platform. Sci China Inf Sci, 2016, 59: 053201
4 Roth P M, Wohlhart P, Hirzer M, et al. Large scale metric learning from equivalence constraints. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012. 2288–2295
5 Schumann A, Schuchert T. Person re-identification in UAV videos using relevance feedback. Inter Soc Opt Eng, 2015, 9407: 8
6 Zheng L, Shen L Y, Tian L, et al. Scalable person re-identification: a benchmark. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), Santiago, 2015. 1116–1124
7 Hirzer M, Roth P M, Bischof H. Person re-identification by efficient impostor-based metric learning. In: Proceedings of Advanced Video and Signal Based Surveillance (AVSS), Beijing, 2012. 203–208
8 Hadsell R, Chopra S. Dimensionality reduction by learning an invariant mapping. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, New York, 2006. 1735–1742
9 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016