Vehicle Re-identification Using Joint Pyramid Feature Representation Network

Xiangwei Lin[1], Huanqiang Zeng[1(✉)], Jinhui Hou[1], Jianqing Zhu[2], Jing Chen[1], and Kai-Kuang Ma[3]

[1] School of Information Science and Engineering, Huaqiao University, Xiamen 361021, China
zeng0043@hqu.edu.cn
[2] School of Engineering, Huaqiao University, Quanzhou 362021, China
[3] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore

Abstract. Vehicle re-identification (Re-ID) technology plays an important role in intelligent video surveillance systems. Due to various factors, e.g., resolution variation, viewpoint variation, illumination changes, occlusion, etc., vehicle Re-ID is a very challenging computer vision task. In order to solve this problem, a joint pyramid feature representation network (JPFRN) is proposed in this paper. Based on the consideration that various convolution blocks with different depths hold various resolution and semantic information of the vehicle image, which can help to effectively identify the vehicle, the proposed JPFRN method obtains four vehicle feature blocks with different depths by designing pyramidal feature fusion of each convolution block in a basic network. After that, a joint representation of these pyramidal features is feed into the loss function for learning discriminative features for vehicle Re-ID. We validated the proposed approach on a commonly used vehicle database i.e., VehicleID. Extensive experimental results show that the proposed method is superior to multiple state-of-the-art vehicle Re-ID methods.

Keywords: Vehicle re-identification · Joint pyramid feature representation · Deep learning

This work was supported in part by the National Natural Science Foundation of China under the grants 61871434, 61602191, and 61802136, in part by the Natural Science Foundation for Outstanding Young Scholars of Fujian Province under the grant 2019J06017, in part by the Natural Science Foundation of Fujian Province under the grant 2017J05103, in part by the Fujian-100 Talented People Program, in part by High-level Talent Innovation Program of Quanzhou City under the grant 2017G027, in part by the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University under the grants ZQN-YX403 and ZQN-PY418, and in part by the High-Level Talent Project Foundation of Huaqiao University under the grants 14BS201, 14BS204 and 16BS108, and in part by the Subsidized Project for Postgraduates Innovative Fund in Scientific Research of Huaqiao University.

B. Li et al. (Eds.): IoTaaS 2019, LNICST 316, pp. 527–536, 2020.
https://doi.org/10.1007/978-3-030-44751-9_44

1 Introduction

Similar to pedestrians, vehicle has become an important target in the intelligent video surveillance systems, because vehicle has been an indispensable part of human daily life. Some related researches on vehicles, such as vehicle classification [1,2], vehicle tracking [3,4], vehicle detection [5,6], vehicle re-identification (Re-ID) [7] draw increasing attentions from both academic and industry. Among these related tasks, vehicle Re-ID is a significant but a frontier area that aims to match all the same vehicles captured by different cameras under various viewing angles. Therefore, vehicle Re-ID can be widely used in many fields, such as intelligent transportation, urban computing, criminal tracking for public safety, to name a few.

In the practical situation, vehicle Re-ID is a very challenging task, due to the influences of many uncertain factors, such as blur, resolution variation, illumination change and viewpoint variation, which can be referred to Fig. 1. It is worthwhile of mentioning that among all the influencing factors, the large varying resolutions of vehicle images collected by different kinds of cameras under different distances is the primary factor needed to be solved in vehicle Re-ID tasks. In practice, vehicle images have different resolutions that tend to make the vehicle target appear larger or smaller, seriously affecting the accuracy of the vehicle recognition. Therefore, matching vehicle images only based on a single resolution has certain limitations. Based on this motivation, by considering the correlation between high resolution and low resolution images in the feature space of convolutional neural networks, we propose a joint pyramidal feature representation network (JPFRN), which integrates vehicle features with different resolutions and different strengths of semantic information to achieve the more rich representation of vehicle targets. Extensive experiments are conducted on large vehicle database VehicleID [7] to evaluate the performance of the proposed method. The corresponding results show that the proposed JPFRN method consistently outperforms multiple state-of-the-art related works.

The rest of this paper is organized as follows: Sect. 2 introduces the related work, Sect. 3 describes the proposed method JPFRN, Sect. 4 presents the experimental results to validate the superiority of the proposed method, Sect. 5 concludes this paper.

2 Related Work

In this section, the existing vehicle Re-ID methods will be briefly reviewed. They can be roughly classified into two categorizes: sensor/clue based method and vehicle appearance feature based method.

2.1 Sensor Based Method/Clue Based Method

The traditional vehicle Re-ID methods mainly relied on the sensor data or clues. In the early research phase of vehicle Re-ID works, most researchers worked

Fig. 1. The vehicle images are from the VehicleID [7] database. The images of the vehicles in each row are collected by the same vehicle, but the appearance is different under different cameras, such as blur, illumination, resolution and occlusion.

with different types of sensor data and multiple clues due to database shortages, such as the vehicle passing time [8], wireless magnetic sensors [9], and license plate, etc. Among them, the license plate number is the most important clue of vehicle and contains all the useful information about the vehicle. Therefore, the Re-ID method based on the license plate number becomes an accurate and effective solution. But in some cases, the license plate number is easily obscured, modified, or blurred. In addition, sensors-based and clue-based methods require additional hardware costs and are very sensitive to complex real-world environments. In contrast, the vehicle Re-ID method based on appearance features has more practical application scenarios.

2.2 Appearance Feature Based Method

For the appearance feature representation, the earlier approach used low-level features, such as SIFT [10], LOMO [11], BOW-CN [12]. And the recent-developed approach is to take advantage of the depth features of the image. Farenzena [13] proposed a joint representation method based on pedestrian global appearance features and local appearance features can directly employed for vehicle Re-ID. Furthermore, Liu et al. [14] proposed a coarse-to-fine vehicle Re-ID method that filters out potential matches by manual features (color, shape) and depth features, then reconstructs the rankings with license plate information and spatiotemporal information. In addition, some classic network models [15–18] are widely used as vehicle feature extractors in vehicle Re-ID tasks, such as VGGNet [17] and ResNet [18], which greatly convenient for features extraction. Furthermore, in order to improve the robustness of vehicle Re-ID, Bai et al. [19] proposed a novel depth metric learning method, triple loss function metric (the distance between features of the same class should be as small as possible). This method fully considers the inter-class similarity and intra-class variance of vehicle shape in the vehicle model. Zhang et al. [20] proposed an improved

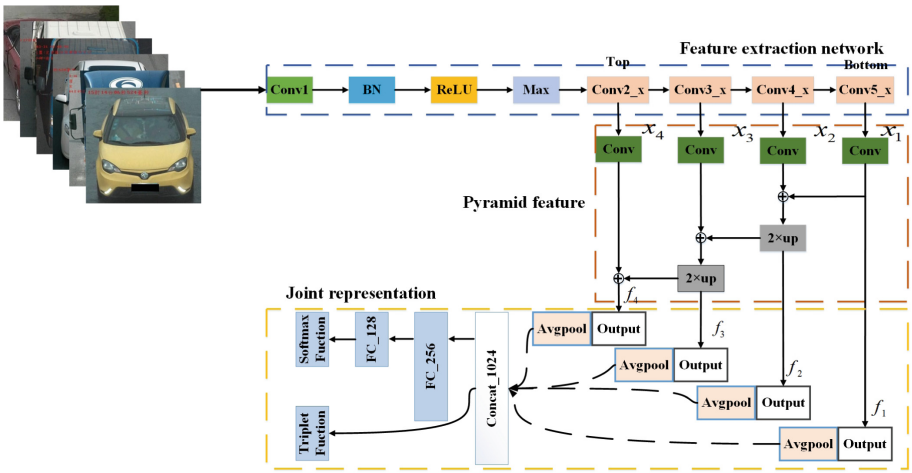

Fig. 2. The proposed network structure diagram using joint pyramid feature representation. Here, x_1, x_2, x_3, and x_4 respectively represent the output of each convolution block of ResNet, and f_1, f_2, f_3, and f_4 respectively represent the outputs of the respective reconstructed feature blocks.

training method for triple loss, the method takes three samples as a group and ensures that the distance between similar samples in the mapped space is less than the distance between different types of samples. Zhou et al. [21] focused on multi-view feature representation, proposed a viewpoint-aware attention model to select the core regions of different viewpoints, and then combined the regions of interest from multiple views by adversarial learning ideas to improve vehicle Re-ID performance. Zhou et al. [22] proposed a long-short-term memory network model to simulate the appearance transformation of different viewpoints of vehicles to increase the diversity of data under different viewpoints, so as to learn more robust and more differentiated vehicle characteristics. Through the above reviewing, we can know that the vehicle Re-ID task has made some progress. However, in the actual monitoring scenario, due to various factors such as vehicle viewpoint change, resolution change, illumination change, and occlusion, etc, the vehicle Re-ID tasks still face enormous challenges, and there is a large room for improvement.

3 Vehicle Re-ID Using Joint Pyramid Feature Representation Network

As shown in Fig. 2, the proposed joint pyramid feature representation network method consists of three parts: baseline network, feature pyramid network, and a joint representation network.

Fig. 3. The vehicle images are from the VehicleID [7] database. The images are collected by the same vehicle, but the resolution is different under different cameras.

3.1 Baseline Deep Learning Network

In this work, we use the ResNet-50 [18] architecture as a feature extractor in our experiments. ResNet-50 [18] is a particularly popular network because it is closely monitored during the training process. Among each residual block, the output of each residual block contains the output of the previous block, so it combines the low-level and high-level features of the input image. Specifically, the training images are resized into $256 \times 256 \times 3$ to adapt the baseline model. We replace the fully-connected layer in the original network with two other fully-connected layers and a classification layer. The first fully-connected layer has 1024 units, and the second fully-connected layer has 128 units. The classification layer has K neurons to predict the K classes, where K is the number of classes in the training set.

3.2 Pyramid Feature Extraction

The human visual system has different perceptions of different resolution images, that is, the human eyes can acquire more image features for high-resolution images, and less for low-resolution images. As shown in Fig. 3, these vehicle images are of the same vehicle, but their resolutions are different due to diversifying capture distances, so the features obtained by the visual system are different. Ultimately, the visual system recognizes the target by combining the different scale features acquired. Similarly, in computer vision, after extracting features from the basic network, the image at the bottom of the network has lower resolution, the implicit semantic information is richer, while the image at the top of the network has higher resolution, but its semantic information is relatively less. Therefore, we give full consideration to the characteristics of the bottom convolutional block and the top convolutional block, propose a pyramidal feature reconstruct method, which can combine vehicle image features with different resolution and semantic information to achieve unified representation and make the task of vehicle Re-ID more robust. The specific network as shown in Fig. 2, the bottom layer, convolutional block $Conv5_x$, first goes through a dimensionality

reduction operation and then adds with convolutional block $Conv4_x$ after an up-sampling operation. It is worth noting that $Conv4_x$, the convolutional block, also goes through a descending dimension operation. Repeatedly, the reconstructed image block is continuously up-sample and added to the last block, so as to fuse the semantic information and resolution information of all convolutional blocks. This feature pyramid can be represented as:

$$F_n = \sigma(W_n * x_n + b_n) \qquad n = 1, 2, 3, 4 \tag{1}$$

$$f_1 = F_1 \tag{2}$$

$$f_n = F_n + upsample(f_{n-1}) \qquad n = 2, 3, 4 \tag{3}$$

where F_n refers to dimensionality reduction of convolution block. The W and b represent weight and bias respectively. Upsample() represents up-sampling operation of the target (no up-sampling for features with the same resolution), and f_n refers to feature blocks at all levels after pyramid reconstruction.

3.3 Joint Representation Network

In the previous section, we proposed a method to reconstruct vehicle features from the bottom layer of the network to the top layer through up-sampling. However, there are still limitations of single resolution vehicle images after reconstruction for complex identification tasks. Since the resolution of the vehicle images collected by different cameras is different, and vehicle size of the vehicle is also inconsistent. For this problem, we consider the correlation between the reconstructed convolutional blocks, and combine the four feature blocks with different resolutions after reconstruction to realize the joint representation. Therefore, our model can contain multiple resolution vehicle features and identify vehicle images with different target sizes at different resolutions. As shown in Fig. 2, the reconstructed fused feature blocks f_1, f_2, f_3, f_4 are respectively aggregated into a (36×1024) feature vector for joint representation after Avgpooling layer. One branch connects the triplet loss function, and the other connects the Softmax loss function through two full connection layers.

4 Experiment and Analysis

4.1 Databases and Evaluation Index

In order to verify the superiority of the proposed joint pyramid feature representation network, we compare it with multiple state-of-the-art methods. The corresponding database and evaluation indexes used in our experiment are described as follows.

The VehicleID database contains 221763 vehicle images collected from multiple non-overlapping cameras in 26267 vehicles, each with its own front and back viewpoint images. And in the database, all vehicles are marked with their

Table 1. Network parameters for JPFRN.

Name	Channels	Conv window	Stride	Output size
Conv1	64	7×7	2	$64 \times 128 \times 128$
Maxpooling	64	3×3	2	$64 \times 64 \times 64$
$Conv2_x$	256	–	1	$256 \times 64 \times 64$
$Conv2_{x-}$	256	1×1	1	$256 \times 64 \times 64$
$Conv3_x$	512	–	2	$512 \times 32 \times 32$
$Conv3_{x-}$	256	1×1	1	$256 \times 32 \times 32$
$Conv4_x$	1024	–	2	$1024 \times 16 \times 16$
$Conv4_{x-}$	256	1×1	1	$256 \times 16 \times 16$
$Conv5_x$	1024	–	2	$2048 \times 16 \times 16$
$Conv5_{x-}$	256	1×1	1	$256 \times 16 \times 16$
$Pyramid_{f1}$	256	1×1	1	$256 \times 16 \times 16$
$Pyramid_{f1}$	256	1×1	1	$256 \times 16 \times 16$
$Pyramid_{f1}$	256	1×1	1	$256 \times 32 \times 32$
$Pyramid_{f1}$	256	1×1	1	$256 \times 64 \times 64$

matching ID information, most of which have color and vehicle type information. In addition, the database is divided into two parts: training subset and testing subset. The training subset contains 110178 vehicle images collected by 13134 vehicles. The testing subset can be divided into three sub-tests, test-800, test-1600, test-2400. The test-800 contains 6532 vehicle images collected by 800 vehicles, the test-1600 contains 11395 vehicle images collected by 1600 vehicles, and the test-2400 contains 17638 vehicle images collected by 2400 vehicles.

In the research of vehicle Re-ID, average precision (mAP) and cumulative matching characteristic (CMC) are generally adopted. Where the cumulative matching feature curve represents the probability of query target appearing in candidate sets of different size, that is, the curve represents the probability of finding the matching target in the candidate sets with the first k ranks. In other words, assume that there are N query targets in total and conduct N queries, $k = (k_1, k_2..., k_n)$ represents the ranking result of matching targets in each query, and set k as the size of the candidate set. The calculation formula of CMC is as follow:

$$CMC@k = \frac{\sum_{q=1}^{Q} gt(1,k)}{Q} \tag{4}$$

The average precision evaluates the overall performance of Re-ID and calculates the average precision of each query image. The formula is:

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{5}$$

Q is the number of vehicle images to be queried, and $AP(q)$ is the accuracy of each query vehicle image. In addition to CMC and mAP, Rank-N also adopted as auxiliary means in vehicle Re-ID tasks. Rank-N can represent the real performance of vehicle Re-ID. For example, Rank-1 means the first target to be matched in the return list. In particular, the greater the number of vehicle types in the candidate set, the lower the probability of finding an accurate match.

4.2 Training Configuration

In our experiments, the software tools are PyTorch, CUDA10.0, CUDNN V7.6.0. The hardware device is a workstation equipped with Inter(R) Xeon(R) CPU E5-2643 v4 @ 3.40 GHZ, two NVIDIA GeForce 2080Ti and 256 GB of memory. We adopted the following training setting: the size of all images in the database was set to 256×256, and each image was randomly flipped horizontally with a probability of flipping 0.5. The ResNet pertaining model is used in the training to initialize the parameters of the network. The number of small batches is 18, the initial learning rate is 0.0003, and the number of training sessions is 50000 and the learning rate begins to decline after 25000 training sessions. Network parameters are shown in Table 1.

4.3 Experimental Evaluation

The experimental results on VehicleID are shown in Table 2, where the performance of the proposed method is compared with multiple existing methods,

Table 2. The performance comparison on VehicleID.

Method	Test-800			Test-1600			Test-2400		
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
DRDL [7]	N/A	48.91	66.71	N/A	46.36	64.38	N/A	40.97	60.02
FACT [24]	N/A	49.53	67.96	N/A	44.63	64.19	N/A	39.91	60.49
Zhu [23]	76.54	72.32	92.48	74.63	70.66	88.90	68.41	64.14	83.37
JPFRN	**79.91**	**74.68**	**94.93**	**74.32**	**68.75**	**90.40**	**70.38**	**63.34**	**86.68**

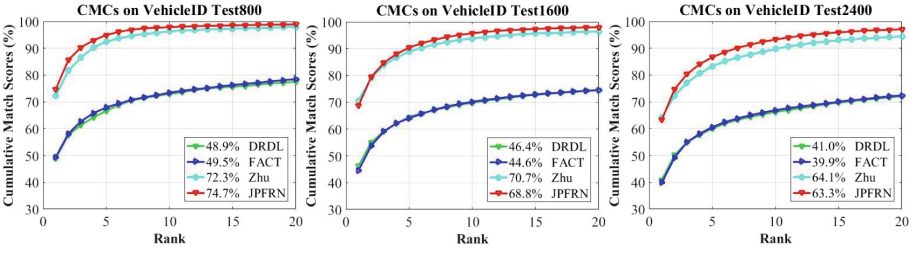

Fig. 4. The CMC curve comparisons of the proposed methods and state-of-the-art methods on VehicleID-Test800, VehicleID-Test1600, and VehicleID-Test2400 dataset, respectively.

including Zhu [23], DRDL [7], and FACT [24]. In order to ensure the stability of the experimental results, we performed ten tests on the model and took the average of ten test results as the final result. Obviously, the proposed method, including mAP, Rank-1, and Rank-5 in test-800, Rank-5 in test-1600, mAP and Rank-5 in test-2400, are slightly better than Zhu's [23] method, and far better than that of DRDL [7] and FACT [24]. Moreover, the CMC curves of various methods are further shown in the Fig. 4. One can see that the proposed method has achieved better results than other methods under comparison.

5 Conclusion

In this paper, a joint pyramid feature representation network is designed for vehicle Re-ID. In the proposed method, each convolution block in base network has vehicle features with different resolutions and intensity semantic information. By combining the bottom-layer convolutional block with the previous level of convolutional block, we obtain the pyramid feature block that fusing the bottom-layer high intensity semantic information and the top-layer high resolution. These pyramid vehicle features are then jointly represented by concatenating the feature blocks at each level. Since the proposed JPFRN method effectively resists the adverse effects of resolution variations in vehicle images, the performance of vehicle Re-ID is improved. Experimental results shown that the proposed method is obviously superior to multiple recently-developed vehicle Re-ID methods.

References

1. Yang, L., Luo, P., Loy, C.C., Tang, X., Huang, T.: A largescale car dataset for finegrained categorization and verification. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3973–3981 (2015)
2. Sochor, J., Herout, A., Boxcars, J.H., Huang, T.: 3D boxes as CNN input for improved finegrained vehicle recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3006–3015 (2016)
3. Matei, B.C., Sawhney, H.S., Samarasekera, S.: Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3465–3472 (2011)
4. Guo, J.-M., Hsia, C.-H., Wong, K., Wu, J.-Y., Wu, Y.-T., Wan, N.-J.: Nighttime vehicle lamp detection and tracking with adaptive mask training. IEEE Trans. Veh. Technol. **65**(6), 4023–4032 (2016)
5. Chen, X., Xiang, S., Liu, C.L., Pan, C.H.: Vehicle detection in satellite images by hybrid deep convolutional neural networks. IEEE Geosci. Remote Sens. Lett. **11**(10), 1797–1801 (2017)
6. Fan, Q.F., Brown, L., Smith, J.: A closer look at Faster R-CNN for vehicle detection. In: IEEE Intelligent Vehicles Symposium (IV), pp. 124–129 (2016)
7. Liu, H., Tian, Y., Wang, Y., Pang, L., Huang, T.: Deep relative distance learning: tell the difference between similar vehicles. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2167–2175 (2016)

8. Lin, W.H., Tong, D.: Vehicle re-identification with dynamic time windows for vehicle passage time estimation. IEEE Trans. Intell. Transp. Syst. **12**(4), 1057–1063 (2011)
9. Kwong, K., Kavaler, R., Rajagopal, R., Varaiya, P.: Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. Transp. Res. Part C Emerg. Technol. **17**(6), 586–606 (2009)
10. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3586–3593 (2013)
11. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
12. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: European Conference on Computer Vision, pp. 1116–1124 (2016)
13. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry driven accumulation of local features. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2360–2367 (2010)
14. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 869–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_53
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems, pp. 1097–1105 (2012)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for largescale image recognition. arXiv:1409.1556 (2014)
17. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
19. Bai, Y., Lou, Y., Gao, F., Wang, S., Wu, Y., Duan, L.: Group-sensitive triplet embedding for vehicle reidentification. IEEE Trans. Multimedia **20**(9), 2385–2399 (2018)
20. Zhang, Y., Liu, D., Zha, Z.J.: Improving triplet wise training of convolutional neural network for vehicle re-identification. In: IEEE International Conference on Multimedia and Expo, pp. 1386–1391 (2017)
21. Zhou, Y., Shao, L.: Viewpoint-aware attentive multi-view inference for vehicle re-identification. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 6489–6498 (2018)
22. Zhou, Y., Liu, L., Shao, L.: Vehicle re-identification by deep hidden multi-view inference. IEEE Trans. Image Process. **27**(7), 3275–3287 (2018)
23. Zhu, J., et al.: Vehicle re-identification using quadruple directional deep learning features. IEEE Trans. Intell. Transp. Syst. **21**(1), 410–420 (2020)
24. Liu, X., Liu, W., Ma, H., Fu, H.: Large-scale vehicle re-identification in urban surveillance videos. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2016)