2019

A Simple Deep Learning Architecture for City-scale Vehicle Re-identification

Eleni Kamenou
Queens University Belfast

Jesus Martinez del Rincon
Queens University Belfast

Paul Miller
Queens University Belfast

See next page for additional authors

Recommended Citation

Kamenou, E., Martinez del Rincon, J., Miller, P., & Devlin-Hill, P. (2019). A simple deep learning architecture for city-scale vehicle re-identification. *IMVIP 2019: Irish Machine Vision & Image Processing*, Technological University Dublin, Dublin, Ireland, August 28-30. doi: 10.21427/mvk7-bx33

A Simple Deep Learning Architecture for City-scale Vehicle Re-identification

Eleni Kamenou[1], Jesus Martinez del Rincon[1], Paul Miller[1], and Patricia Devlin-Hill[2]

[1]*Centre for Secure Information Technologies (CSIT), Queen's University, Belfast*
[2]*Thales, Belfast, Northern Ireland, UK*

Abstract

The task of vehicle re-identification aims to identify a vehicle across different cameras with non overlapping fields of view and it is a challenging research problem due to viewpoint orientation, scene occlusions and intrinsic inter-class similarity of the data. In this paper, we propose a simplistic approach for one-shot vehicle re-identification based on a siamese/triple convolutional architecture for feature representation. Our method involves learning a feature space in which the vehicles of the same identities are projected closer to one another compared to those with different identities. Moreover, we provide an extensive evaluation of loss functions, including a novel combination of triplet loss with classification loss, and other network parameters applied to our vehicle re-identification system. Compared to most existing state-of-the-art approaches, our approach is simpler and more straightforward for training, utilizing only identity-level annotations. The proposed method is evaluated on the large-scale CityFlow-ReID dataset.

Keywords: Re-identification, Similarity-learning, Deep-Learning, Image Processing

1 Introduction

Over recent decades, camera-based sensor networks have been installed in urban environments for traffic monitoring and surveillance purposes. In such cases, it is important to maintain vehicle identities as they move across different image sensors with non-overlapping fields of view. Therefore, wide area vehicle re-identification (ReID) has gained a lot of attention in the computer vision community [1, 2, 3, 4, 5].

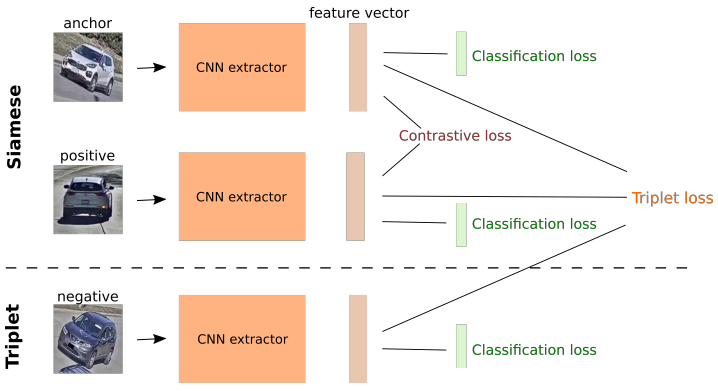

Figure 1: The structure of our proposed ReID system.

In this scenario, however, cameras are often placed far apart, resulting in large viewpoint and illumination changes across different fields of view. In addition, several inherent properties of the data, e.g., inter-class similarity, where cars of the same model and/or manufacturer can be present, render the ReID task even more challenging. Moreover, background clutter, and appearance variance of a vehicle leads to large intra-class variability.

Given the above context, this paper addresses the problem of associating different vehicle images as they transit between non-overlapping imaging sensors by introducing an image-based ReID system based on extraction of feature vectors using a convolutional architecture. In general, ReID can be considered as a retrieval problem, where, given a probe image of a vehicle we need to search over numerous gallery images to find the same vehicle as it appears in another camera. This requires a similarity learning technique to produce a

model capable of measuring a larger similarity between a pair of images showing the same vehicle (positive pair) compared to pairs of images of two different vehicles (negative pair). There are several loss functions for this purpose, including contrastive or triplet loss. In this paper we provide an extensive evaluation of these loss functions applied to vehicle ReID.

2 Related Work

With the rapid development of deep learning techniques in the computer vision community, neural network based models have also made an impact in ReID. Recently, deep features learned by convolutional neural networks (CNNs) have been utilized [2, 6, 1, 7] due to their success in extracting discriminative features from large-scale image data. From the perspective of a model, siamese models [8, 5] that use image pairs or triplets [9, 4] are commonly employed. They attempt to learn an embedding space in which vehicles of same identities are projected closer to one another, compared to the vehicles of different identities, usually by using contrastive [5] or triplet loss [4] functions. Other recent approaches [9] leverage viewpoint, vehicle type, brand or colour information to make the feature representation more robust, or to apply re-ranking for refining the retrieval results [10]. However in those cases, extra annotation is needed.

3 Proposed Method

As shown in Fig 1, our ReID pipeline starts by passing every image into a CNN feature extractor. The CNN consists of three convolutional layers, each of them followed by max-pooling and non-linearity tanh layers. Following the convolutional layers, a fully connected (FC) layer produces a feature representation for every image. For simplicity, this network can be interpreted as a function, $v = C(x)$, that takes an image x as input and produces a feature vector v as output. The ReID task aims to map vehicle images with the same identity to feature vectors that are close to one another, while on the other hand, it intends to map images of different vehicle identities to feature vectors that are widely separated. In order to achieve this, three different loss functions and combinations of them are thoroughly examined and evaluated.

3.1 Loss functions

Contrastive Loss: This is a distance-based loss function calculated on pairs of feature vectors. In the simplest case the distance metric is the Euclidean distance between the two vectors v_i and v_j. The main idea of Contrastive loss is to encourage all positive pair distances to approach zero while keeping all negative pair distances to be above a certain threshold α.

$$E_{contr}(v_i, v_j) = \begin{cases} \frac{1}{2}\|v_i - v_j\|^2, & i = j. \\ \frac{1}{2}[\max(\alpha - \|v_i - v_j\|, 0)]^2, & i \neq j. \end{cases} \tag{1}$$

Triplet Loss: In this case, the feature vectors come in triplets of v_a, v_p and v_n, representing the feature representations of the anchor, positive pair and negative pair images, respectively. The dissimilarity (distance) between v_a and v_p should be less than the dissimilarity between v_a and v_n. The margin hyperparameter μ is added to the loss equation in order to define how far away the dissimilarities should be.

$$E_{trip}(v_a, v_p, v_n) = \max\big(0, (\|v_a - v_n\| - \|v_a - v_p\|) + \mu\big) \tag{2}$$

Classification Loss: Another method for obtaining a discriminative representation for an image is to utilize a traditional classification layer. Each identity is considered as a separate class and the number of classes in the last layer is equal to the number of identities in the training set. In similarity problems, once the network is

trained using some classification loss function (e.g. cross entropy), then in the testing phase the classification layer is detached and the feature representation is obtained from the new final layer of the network.

$$I(v,i) = \log\left(\frac{\exp v(i)}{\sum^{K} \exp v}\right), \text{ K= number of identities} \tag{3}$$

However in the testing phase, the identities are unknown and different from the training phase. Due to this limitation, classification losses are rarely used in ReID. Instead, we propose to use them as complement to the other losses to increase their discriminative power, in a similar fashion to additional attributes [10] but without the need for extra annotation. In the proposed combined losses, weighting factors λ, λ_1 and λ_2 are used to balance between them.

$$E_{contr-entr} = E_{contr}(v_i, v_j) + \lambda(I(v_i) + I(v_j)) \tag{4}$$

$$E_{trip-entr} = E_{trip}(v_a, v_p, v_n) + \lambda_1(I(v_a) + I(v_p)) + \lambda_2 I(v_n) \tag{5}$$

4 Experimental Results

4.1 Dataset overview and experimental setting

Our models have been trained and evaluated on the CityFlow-ReID dataset [11] which is a subset of the CityFlow tracking dataset for image-based vehicle ReID. It consists of 36,935 images with 333 vehicle identities taken by 40 cameras in total, placed across the urban environment of a city. On average, each vehicle has 84.50 image signatures from 4.55 camera views.

We separated the dataset by using 233 vehicle identities for training and the remaining 100 for evaluation. For each training epoch and for each vehicle identity an anchor image is randomly selected along with a negative pair image and a positive pair image of the same identity. This results in 699 images per epoch; 233 anchor, 233 positive sample and 233 negative sample images. It should be noted that the training set is not fixed but it is dynamically and randomly selected at the beginning of every epoch to prevent overfitting.

The distance metric based on which we measured similarity between samples in all cases is the Euclidean distance. The margins α and μ are set to 2.0 and 1.0 in equations (1) and (2) while the feature representation vector size is set to 128, which is equal to the resulting dimensionality of the CNN last layer. The network was trained for 2000 epochs using stochastic gradient descent with a learning rate of 1e-4 and with a batch size set to one. For the sake of evaluation, we use CMC curves [5] and rank-k accuracy.

4.2 Evaluation of input image size

Input image size is one of the most critical parameters that affects the learning process and subsequently determines the model accuracy. Figure 2 shows preliminary experiments indicating the differences in performance for different input image sizes while using contrastive loss function. As can be observed, best performance was obtained with an input image size of 100×100. Therefore, this value was used for the remaining experiments.

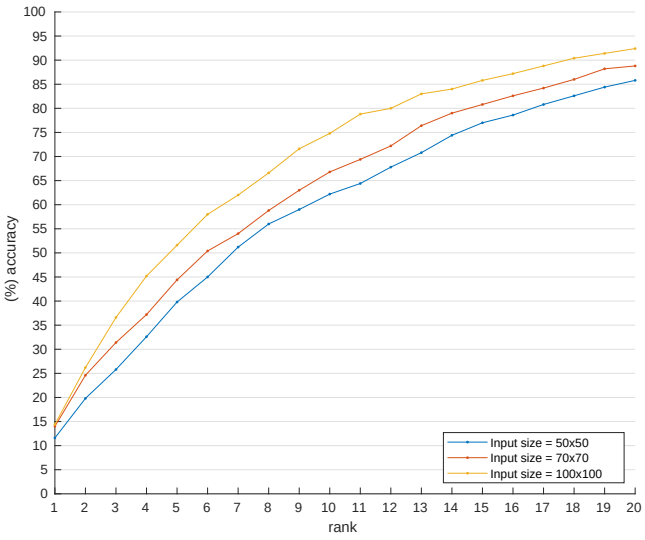

Figure 2: (%) rank-k accuracy for different input image sizes.

4.3 Evaluation of loss functions

In this experiment, the different loss functions, as well as their combinations, are compared. The values of the weight variables were set empirically using grid search to $\lambda = 0.05$, $\lambda_1 = 0.05$ and $\lambda_2 = 0.1$

Loss Function	rank 1	rank 2	rank 5	rank 10	rank 20
Contrastive	14.4	26.2	**51.6**	**74.8**	**92.4**
Contrastive + Classification	15.0	24.0	43.6	68.0	90.6
Triplet	17.2	28.4	48.8	69.8	88.0
Triplet + Classification	**18.8**	**29.0**	48.0	74.4	92.0

Table 1: (%) rank-K accuracy for different loss functions

5 Discussion and Conclusions

Our solution for image-based vehicle ReID is based on a triple convolutional architecture and similarity learning techniques. As for the loss functions used for training our model, combining triplet with classification loss resulted in the highest accuracy for rank-1 and rank-2. It can also be observed that adding classification loss, slightly improved the performance in both siamese and triple cases. Furthermore, increasing the image size provides better visual information as input for the model and thus leads to better performance. This simple architecture with our proposed classification-enhanced triplet loss has been shown to be capable of performing ReID in a city-scale scenario.

References

[1] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *CVPR 2018*, pages 6036–6046. IEEE Computer Society, 2018.

[2] Ejaz Ahmed, Michael J. Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *CVPR 2015*, pages 3908–3916. IEEE Computer Society, 2015.

[3] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. Dense 3d-convolutional neural network for person re-identification in videos. *TOMCCAP*, 15(1s):8:1–8:19, 2019.

[4] Cairong Zhao, Kang Chen, Zhihua Wei, Yipeng Chen, Duoqian Miao, and Wei Wang. Multilevel triplet deep learning model for person re-identification. *Pattern Recognition Letters*, 117:161–168, 2019.

[5] Niall McLaughlin, Jesús Martínez del Rincón, and Paul C. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR 2016*, pages 1325–1334. IEEE Computer Society, 2016.

[6] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *TOMCCAP*, 14(4):83:1–83:18, 2018.

[7] Hao Chen, Benoit Lagadec, and Francois Bremond. Partition and reunion: A two-branch neural network for vehicle re-identification. In *CVPR Workshops*, 2019.

[8] Jakub Spanhel, Vojtech Bartl, Roman Juranek, and Adam Herout. Vehicle re-identifiation and multi-camera tracking in challenging city-scale environment. In *CVPR Workshops*, June 2019.

[9] Xiao Tan, Zhigang Wang, Minyue Jiang, Xipeng Yang, Jian Wang, Yuan Gao, Xiangbo Su, Xiaoqing Ye, Yuchen Yuan, Dongliang He, Shilei Wen, and Errui Ding. Multi-camera vehicle tracking and re-identification based on visual and spatial-temporal features. In *CVPR Workshops*, June 2019.

[10] Tsung-Wei Huang, Jiarui Cai, Hao Yang, Hung-Min Hsu, and Jenq-Neng Hwang. Multi-view vehicle re-identification using temporal attention model and metadata re-ranking. In *CVPR Workshops*, 2019.

[11] Milind Naphade, Zheng Tang, Ming-Ching Chang, David C. Anastasiu, Anuj Sharma, Rama Chellappa, Shuo Wang, Pranamesh Chakraborty, Tingting Huang, Jenq-Neng Hwang, and Siwei Lyu. The 2019 ai city challenge. In *CVPR Workshops*, 2019.