A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance

Xinchen Liu[1], Wu Liu[1(✉)], Tao Mei[2], and Huadong Ma[1]

[1] Beijing Key Lab of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China
liuwu@bupt.edu.cn

[2] Microsoft Research, Beijing 100080, China

Abstract. While re-identification (Re-Id) of persons has attracted intensive attention, vehicle, which is a significant object class in urban video surveillance, is often overlooked by vision community. Most existing methods for vehicle Re-Id only achieve limited performance, as they predominantly focus on the generic appearance of vehicle while neglecting some unique identities of vehicle (e.g., license plate). In this paper, we propose a novel deep learning-based approach to PROgressive Vehicle re-ID, called "PROVID". Our approach treats vehicle Re-Id as two specific progressive search processes: coarse-to-fine search in the feature space, and near-to-distant search in the real world surveillance environment. The first search process employs the appearance attributes of vehicle for a coarse filtering, and then exploits the Siamese Neural Network for license plate verification to accurately identify vehicles. The near-to-distant search process retrieves vehicles in a manner like human beings, by searching from near to faraway cameras and from close to distant time. Moreover, to facilitate progressive vehicle Re-Id research, we collect to-date the largest dataset named VeRi-776 from large-scale urban surveillance videos, which contains not only massive vehicles with diverse attributes and high recurrence rate, but also sufficient license plates and spatiotemporal labels. A comprehensive evaluation on the VeRi-776 shows that our approach outperforms the state-of-the-art methods by 9.28 % improvements in term of mAP.

Keywords: Vehicle re-identification · Progressive search · Deep learning · License plate verification · Spatiotemporal relation

1 Introduction

Vehicle, as a significant object class in urban video surveillance, attracts massive focuses in computer vision research field, such as detection [1], classification [2], and pose estimation [3]. However, **vehicle re-identification** (Re-Id) is still a frontier but important topic which is often neglected by researchers. The task of vehicle Re-Id is, given a probe vehicle image, to search in a database for images that contain the same vehicles captured by multiple cameras. Vehicle Re-Id has

B. Leibe et al. (Eds.): ECCV 2016, Part II, LNCS 9906, pp. 869–884, 2016.
DOI: 10.1007/978-3-319-46475-6_53

Fig. 1. (**a**) Large intra-instance differences of the same vehicles from different views (left) and subtle inter-instance differences of similar vehicles (right). (**b**) The license plates for vehicle Re-Id. (Part of the plate is covered due to privacy.)

pervasive applications in video surveillance [4], intelligent transportation [5], and urban computing [6], which can quickly discover, locate, and track the target vehicles in large-scale surveillance videos.

Different from vehicle detection, tracking or classification, vehicle Re-Id can be found as an instance-level object search problem. In the real-world vehicle Re-Id, this problem can be handled by a progressive process. For example, if the monitoring staves want to find a suspect vehicle in huge amount of surveillance videos, they will firstly filter out large numbers of vehicles by appearance features, such as colors, shapes and types, to narrow down the search space. Then, for the remaining vehicles, the license plate is utilized to accurately identify the suspects as shown in Fig. 1(b). Furthermore, the search scope is expanded from near cameras to faraway, and search period is extended from close time to distant. Therefore, the spatiotemporal information can also provide great assistance as shown in Fig. 2. The real-world practice inspires us for constructing a progressive vehicle Re-Id method, which includes two progressive search processes: (1) from-coarse-to-fine search in feature space; (2) from-near-to-distant search in the real-world spatiotemporal environment.

However, the implementation of progressive vehicle Re-Id method in real-world urban traffic surveillance still faces several significant challenges: first of all, the appearance-based approaches can hardly give optimal results due to the large intra-instance differences of the same vehicle in different cameras, and subtle inter-instance differences between different vehicles in the same views as shown in Fig. 1(a). Furthermore, traditional license plate recognition techniques may fail in unconstrained surveillance scenes due to the various illuminations, viewpoints, and resolutions as shown in Fig. 1(b). Besides, the license plate recognition is a complex multi-step process including plate detection, segmentation, shape adjustment, and character recognition as in [7,8]. How to effectively and efficiently utilize the license plate information in unconstrained traffic scenes remains great challenging. Finally, in the urban surveillance scene, it is difficult to model the patterns of vehicle's behaviors in unconstrained conditions. The

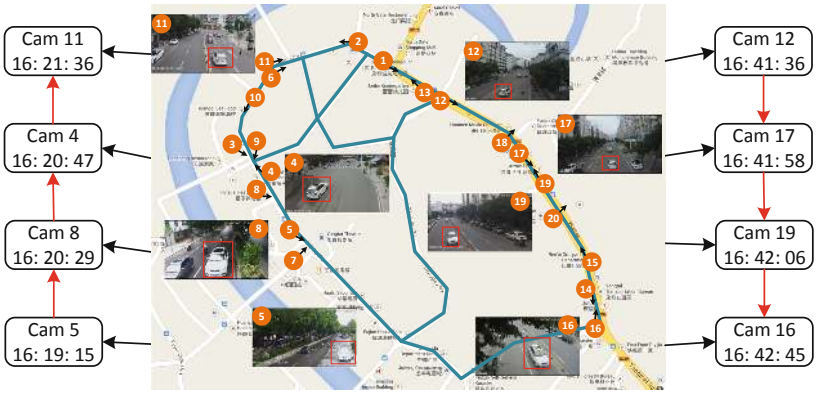


Fig. 2. The spatiotemporal information of a vehicle in the surveillance network.

traffic conditions, road maps, and weather can affect the routes of vehicles. The utilization of the spatiotemporal cues also remains challenging.

Existing methods for vehicle Re-Id mainly focus on appearance-based models [1,9]. However, these methods cannot distinguish the vehicles with similar appearance and neglect the license plate to uniquely identify a vehicle. Different from these methods, we consider both the appearance features and the license plate in a coarse-to-fine fashion. The appearance-based model firstly filters out the dissimilar vehicles, then the license plates are utilized for accurate vehicle search. Besides, most methods didn't consider the spatiotemporal information for assistance. Spatiotemporal relations have been employed in many areas such as multi-camera surveillance [10], cross-camera tracking [11], and object retrieval [12]. With spatiotemporal information in the surveillance network, we handle the search process with a from-near-to-distant principle in both the time scale and space scale.

In this paper, we propose the PROVID, a deep learning-based progressive vehicle Re-Id approach for urban surveillance, which is featured by following properties: (1) adopting the progressive approach to search for vehicles as in real-world practice; (2) an appearance attribute model learned from deep convolutional neural networks (CNNs) is exploited as a coarse vehicle filter; (3) the Siamese neural network-based license plate verification is proposed to match the license plate images; and (4) the spatiotemporal relations are explored to assist the search process. In particular, for the appearance-based coarse filtering, we adopt the fusion model of low-level and high-level features to find the similar vehicles. For license number plate, instead of accurate recognizing the characters of the license plate, we just need to verify whether two plate images belong to the same vehicle. Therefore, a Siamese neural network is trained with large numbers of plate images for license plate verification. At last, a spatiotemporal relation model is utilized to re-rank vehicles to further improve the final results of vehicle Re-Id.

To facilitate the research and validate related algorithms, we build a comprehensive vehicle Re-Id dataset named VeRi-776, which contains not only massive vehicles with diverse attributes and high recurrence rate, but also sufficient license plates and spatiotemporal labels, which can greatly facilitate the investigation of progressive vehicle Re-Id methods based on license plate and spatiotemporal information. Finally, we evaluate the PROVID on the VeRi-776 to demonstrate the effectiveness of the proposed framework, which outperforms the state-of-the-art methods by achieving 9.28 % improvements in mAP and 10.94 % in HIT@1.

2 Related Work

Vehicle Re-Id. In recent years, vehicle Re-Id is still on its early stage with a handful of related works. Feris *et al.* [1] proposed a vehicle detection and retrieval system, in which vehicles are classified into different types and colors by appearance, then indexed and searched by these attributes in the database. Recently, Liu *et al.* [9] firstly evaluated and analyzed several appearance-based models, including the texture, color, and semantic attribute, then proposed a fusion model of low-level features and high-level semantic attributes for vehicle Re-Id. However, the appearance-based approaches cannot uniquely identify a vehicle due to the similarity of vehicles and various environment factors such as illuminations, viewpoints, and occlusion. More importantly, as the unique ID of each vehicle, the license plate should be considered for accurate vehicle Re-Id.

License Plate Verification. In industry, license plate recognition has been widely used in identifying vehicles [7,8]. However, due to the high demand on the quality of plate images, existing methods can only be used in constrained conditions such as park entrances and toll gates. The license plate recognition may fail in unconstrained surveillance scenes due to the various environmental factors [1,9]. Therefore, we use the license plate verification instead of the recognition for vehicle Re-Id. In recent years, deep neural networks have achieved great success in computer vision such as object classification [13], detection [14], image understanding [15], video analysis [16], and multimedia search [17]. Among them, the Siamese Neural Network (SNN) was proposed to verify hand-write signatures by Bromley *et al.* [18]. SNN takes two weight-shared convolutional neural networks and a contrastive loss function. During training, it can simultaneously minimize the distances of similar object pairs and maximize the distances of dissimilar pairs. Chopra *et al.* [19] adopted the SNN for face verification and obtained excellent results. Zhang *et al.* [20] achieved the optimal performance in gait recognition for person identification with SNN. Therefore, we utilize SNN in license plate verification for accurate vehicle Re-Id.

Spatiotemporal Relation. Spatiotemporal relations have been widely used in multi-camera systems [10–12]. Among them, Kettnaker *et al.* [10] proposed to assemble likely paths of objects using Bayesian estimates over cameras. Javed *et al.* [11] utilized spatiotemporal information to estimate the inter-camera correspondence for object tracking. Xu *et al.* [12] proposed a graph-based object

retrieval system in distributed camera network. However, these methods mainly focused on slow-moving objects such as persons in constrained environments like campuses. In the large-scale unconstrained traffic scene, it is difficult to model the patterns of vehicles due to the complicated traffic conditions, road maps, and weather.

3 The Proposed Method

3.1 Overview

Figure 3 shows the architecture of the proposed progressive vehicle Re-Id approach. The query contains an image of the vehicle with the camera ID and timestamp which record where and when it is captured. Given the query, the proposed method considers the task of vehicle Re-Id as progressive processes: (1) appearance-based coarse filtering: the appearance-based model is utilized to filter out most vehicles with different colors, textures, shapes, and types in the vehicle database; (2) license plate-based fine search: for remaining filtered vehicles, the license plate similarities between query and source vehicles are calculated by the Siamese neural network to find the most similar vehicles; (3) Based on the proposed from-near-to-distant principle, the spatiotemporal properties are exploited to re-rank the vehicles, which further improve the vehicle search process.

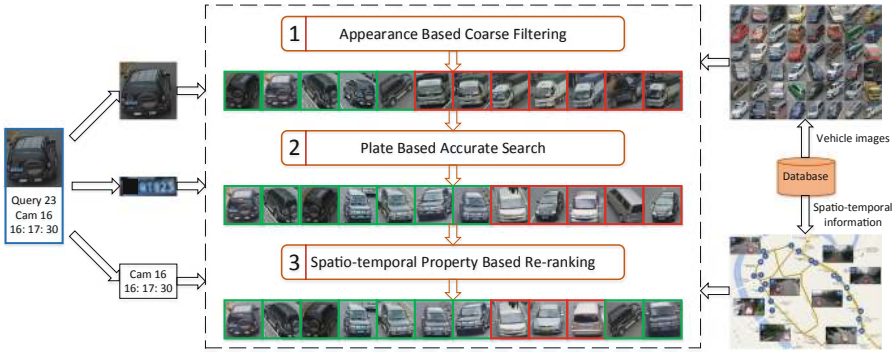


Fig. 3. The architecture of the PROVID method.

3.2 Appearance Attributes Extracted by CNN

In real-world practice, appearance features such as colors, shapes, and types are very effective to filter out the dissimilar vehicles. In addition, they are efficient to be extracted and searched in large-scale dataset. Consequently, we adopt the fusion model of texture, color and semantic attribute which have been evaluated

by Liu *et al.* [9] as the coarse filter to find the vehicles with similar appearance to the query.

The **texture feature** is represented by the conventional descriptors such as Scale-Invariant Feature Transform (SIFT) [21]. Then the descriptors are encoded by the bag-of-words (BOW) model due to its accuracy and efficiency in image retrieval [22]. The **color feature** is extracted by Color Name (CN) model [23] which is quantized by the BOW model for its excellent performance in person Re-Id [24]. The **high-level attribute** is learned by a deep convolutional neural network (CNN), i.e., the GoogLeNet [25]. This model is fine-tuned on the CompCars dataset [2] to detect the detailed attributes of vehicles, such as the number of doors, the shape of lights, the number of seats, and the model of vehicles. At last, the three types of features are integrated by the distance-level fusion.

By fusion of texture, color, and semantic attribute, the appearance-based approach can filter out most of the vehicles that have different colors, shapes, and types to the query. Therefore, the search space narrows down from the whole vehicle database to a relatively small amount of vehicles. However, appearance-based model cannot uniquely identify a vehicle due to the similarity of the vehicles and the environment factors. So we utilize the license plate, which is the unique ID of vehicles, for accurate vehicle Re-Id.

3.3 Siamese Neural Network-Based License Plate Verification

For accurate vehicle search, license plate is a significant cue because it is the unique ID for vehicle. In unconstrained surveillance scenes, the license plate may not be recognized correctly due to the view points, low illuminations, and image blurs as shown in Fig. 1(b). Besides, the license plate recognition technique is a complicated process which includes plate localization, shape adjustment, character segmentation, and character recognition. Therefore, it is not effective and efficient for the vehicle Re-Id task. Nonetheless, in vehicle Re-Id, we just need to verify whether two plates are the same instead of recognizing the characters. The Siamese neural network (SNN) introduced in [18] is applied for signature verification tasks. The main idea of the SNN is to learn a function that maps input patterns into a latent space, in which the similarity metric will be large for pairs of the same objects, and small for pairs from different ones. Therefore, it is best suited for verification scenarios where the number of classes is large, and/or samples of all the classes are not available during training. Definitely, the license plate verification is one of such scenarios.

The SNN designed for plate verification contains two parallel CNNs as illustrated in Fig. 4. Each CNN is stacked with two parts: (1) two convolution layers and max-pooling layers, and (2) three full connection layers. The contrastive loss layer is connected on the top of the output layers. The network parameters are set as shown in Fig. 4. Before training, two license plate images are paired as a training sample and labeled with 1 if they belong to the same vehicle and 0 otherwise. During training, the pairwise plate images are fed into the two CNNs separately. After the forward propagation, the outputs of CNNs are combined into the contrastive loss layer to compute the loss of the model. Then through

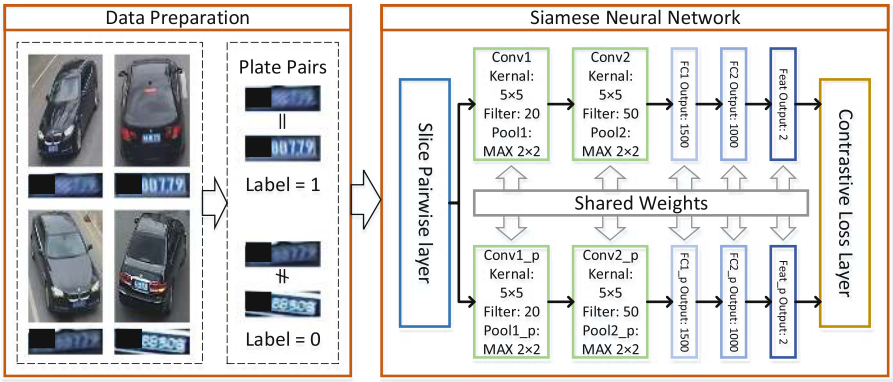


Fig. 4. The structure of the Siamese neural network for license plate verification.

the back propagation with the contrastive loss, the shared weights of the two CNNs are optimized simultaneously.

Specifically, let W be the weights of the SNN, given a pair of license plate images x_1 and x_2, we can map the data into the latent metric space as $S_W(x_1)$ and $S_W(x_2)$. Then, the energy function, $E_W(x_1, x_2)$, which measures the compatibility between x_1 and x_2, is defined as

$$E_W(x_1, x_2) = ||S_W(x_1) - S_W(x_2)||. \tag{1}$$

With the energy function, the contrastive loss can be formulated as

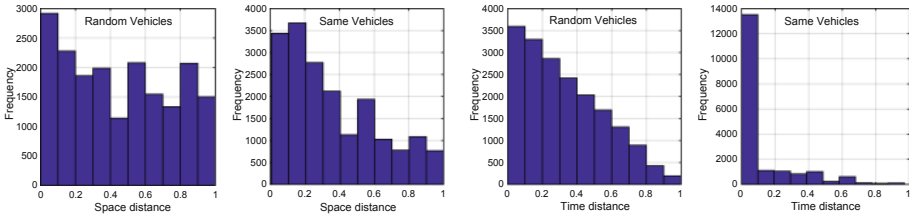
$$L(W, (x_1, x_2, y)) = (1 - y) \cdot max(m - E_W(x_1, x_2), 0) + y \cdot E_W(x_1, x_2), \tag{2}$$

where (x_1, x_2, y) is a pair of samples with the label, m is a positive margin. In the implementation, we adopt the Caffe framework [26] with the default margin value, $m = 1$. During test, we use the learned SNN to extract the 1000-D feature of the FC2 layer from the plate images. The Euclidean distance is adopted to estimate the similarity scores of two plate images.

3.4 Vehicle Re-ranking Based on Spatiotemporal Relation

As discussed in Sect. 1, in real-world practice, it is reasonable to perform vehicle search with a from-near-to-distant fashion in spatiotemporal domain. Based on this principle, we exploit the spatiotemporal relation to further improve the vehicle Re-Id.

However, in the unconstrained traffic scenarios, it is difficult to model the travel patters of vehicles and predict the spatiotemporal relations of two arbitrary vehicles. To investigate whether the spatiotemporal relation is effective for vehicle Re-Id, we analyze the space and time distances of 20,000 image pairs from the same vehicles and 20,000 pairs from randomly selected vehicles. The statistics are illustrated in Fig. 5. We obviously find that the space and time



(a) Histograms of space distances. (b) Histograms of time distances.

Fig. 5. Statistics of spatiotemporal information.

distances of the same vehicles are relatively smaller than those of the randomly selected vehicles. From this observation, we make a general assumption that: two images have higher possibility to be the same vehicle if they have small space or time distance, and lower possibility to be the same vehicle if they have large space or time distance. With this assumption, for each query image i and test image j, the spatiotemporal similarity $ST(i,j)$ is defined as:

$$ST(i,j) = \frac{|T_i - T_j|}{T_{max}} \times \frac{\delta(C_i, C_j)}{D_{max}} \tag{3}$$

where T_i and T_j are the timestamps of query image i and test image j, T_{max} is the maximal time difference between all query images and test tracks. $\delta(C_i, C_j)$ is the length of the shortest path between camera C_i and C_j, D_{max} is the maximal length between all cameras. The shortest path between two cameras is obtained from the Google Map and stored in a matrix as shown in Fig. 6. At last, either a post-fusion strategy or a re-ranking strategy can be adopt for the combination of the spatiotemporal information with the appearance and plate features.

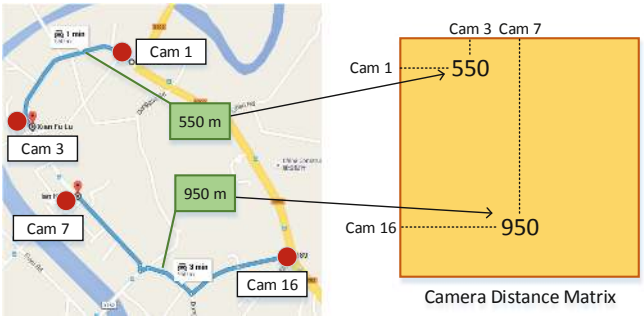


Fig. 6. The real-world camera distance matrix obtained from the Google map.

4 Experiments

4.1 Dataset

To well investigate the spatiotemporal relation and evaluate the proposed progressive vehicle Re-Id approach, we build the VeRi-776 dataset from the VeRi dataset in [9]. The VeRi dataset has three featured properties. First, it contains about 40,000 images of 619 vehicles captured by 20 surveillance cameras. In addition, the images are captured in a real-world unconstrained traffic scene and labeled with varied attributes, e.g. BBoxes, types, colors, and brands. Furthermore, each vehicle is captured by 2 $\sim$ 18 cameras in different viewpoints, illuminations, and occlusions, which provides high recurrence rate for vehicle Re-Id. Finally, we extend the VeRi dataset with (1) data volume expansion, (2) license plate labels, and (3) spatiotemporal information.[1]

Data Volume Expansion. With the video frames provided by Liu *et al.* [9], we add over 20 % new vehicles into the VeRi dataset. The new vehicles are also labeled with BBoxes, types, colors, brands, and cross-camera relations as in [9]. This makes the dataset contain over 50,000 vehicle images, about 9000 tracks, and 776 vehicles, which further improves the scalability for vehicle Re-Id.

License Plate Annotation. The most important contribution of the new VeRi-776 dataset is the annotation of license plates. Before annotating, we divide the dataset into the testing set of 200 vehicles and 11,579 images, and the training set of 576 vehicles and 37,781 images. For the testing set, we pick out one image from each track as the query and obtain 1,678 queries. Then, for each query image and test image, we annotate the BBox of license plate if the plate can be detected by the annotators. For the quality of annotation, each image is annotated by at least three human annotators with the majority vote. At last, we obtain 999 plate images from the query images, 4,825 plate images from the test images, and 7,647 plate images from the train images. About 50 % of the query and test images can utilize the license plate to improve the vehicle Re-Id.

Spatiotemporal Relation Annotation. We annotate the spatiotemporal relation for tracks of all vehicles. The track is the trajectory of a vehicle captured by one camera at the same time, the images belonging to one track are clustered together. For each track, we firstly label the ID (from 1 to 20) of the camera which captures the track. Then we use the timestamp of the first captured image in the track as its timestamp. Furthermore, to facilitate the computation of the space distances used in the spatiotemporal relation-based re-ranking, we obtain the length of the shortest path between each pair of the 20 cameras in the surveillance network via Google Map as shown in Fig. 6.

4.2 Experimental Settings

The VeRi-776 dataset is divided into two subsets for training and testing as in Sect. 4.1. The training set has 576 vehicles with 37,781 images and the testing set

[1] The latest dataset can be obtained at https://github.com/VehicleReId/VeRidataset.

has 200 vehicles with 11,579 images. In the evaluation, the cross-camera search is performed, which means we use one image of a vehicle from one camera to search for tracks of the same vehicle in other cameras. Moreover, in [9], the vehicle Re-Id is in an image-to-image manner, which means using a query image to search for the target images as in person Re-Id [24]. Different from [9], we conduct the vehicle Re-Id in an image-to-track fashion, in which the query is an image, while the target units are tracks of vehicles. The similarity between a query image and a test track is denoted by the maximum of the similarities between the query image and all images of the track. In real-world practice, we just need to find the track in one camera to capture the target vehicle. Therefore, the image-to-track search is more reasonable in the practical scenario. For the image-to-track search, we have 1,678 query images and 2,021 testing tracks.

For the VeRi-776 dataset, there are multiple ground truths for each query. Therefore, we adopt mean average precision (mAP) which considers both precision and recall to evaluate the overall performance for vehicle Re-Id. For each query image, we calculate the average precision (AP) as

$$AP = \frac{\sum_{k=1}^{n} P(k) \times gt(k)}{N_{gt}} \tag{4}$$

where n is the number of test tracks, N_{gt} is the number of ground truths, $P(k)$ is the precision at cut-off k in the result lists, and $gt(k)$ is an indicator function equaling 1 if the kth result is correct. The mAP is computed over all queries as

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

where Q is the number of queries. Besides, we also adopt the Cumulative Matching Characteristic (CMC) curve, HIT@1, and HIT@5 which are widely used in person Re-Id [24].

4.3 Evaluation of Plate Verification

To evaluate the Siamese neural network-based plate verification, we compare it with the conventional handcraft features, SIFT [21]. We combine the plate features with the FACT model by post-fusion to test their performance for the appearance-based model. The settings of the two models are as follows:

(1) **FACT + Plate-SIFT**. This method adopts the conventional SIFT as the local descriptor. Then the descriptors of the plate is quantized with the bag-of-words (BOW) model. Before testing, we train a codebook of the BOW model with the training plates of VeRi-776, the size of codebook is 1000. In the test stage, each plate is represented as an 1000-D BOW feature.
(2) **FACT + Plate-SNN**. Before performing the search, we firstly train the Siamese neural network for license plate verification. With the 7,647 training plate images, we randomly pick out about 50,000 pairs of plates belonging

to the same vehicles as the positive samples and 50,000 pairs of plates of different vehicles as the negative samples. We adopt the Caffe [9] to implement the SNN as in Sect. 3.3 and train the SNN with the Stochastic Gradient Descent solver. We use the model of 60,000 iterations to extract the 1000-D FC2-layer output as the representation of license plates.

For both of the two models, the image-to-track search is performed, the similarity is calculated by the Euclidean distance. The weights for post-fusion is 0.4 for the FACT and 0.6 for the Plate-SNN. The mAP, HIT@1, and HIT@5 are used to evaluate the performance. Table 1 shows the search results which demonstrate that the deep learned model is much better than the SIFT feature. The results demonstrate that traditional handcraft features are not robust to the various illuminations, viewpoints, and resolutions in unconstrained surveillance scenes. While the SNN model which is trained on large amount of plate pairs can map input patterns into a latent space, in which the similarity metric is larger for pairs of the same objects, and lower for pairs from different ones. The abundant training license plate samples guarantee the robustness of the learned model.

Table 1. Comparison of different models for plate verification.

Methods	mAP	HIT@1	HIT@5
FACT [9] + Plate-SIFT	18.49	50.95	73.48
FACT [9] + Plate-SNN	**25.88**	**61.08**	**77.41**

4.4 Evaluation of Vehicle Re-Id Methods

To validate the effectiveness of progressive vehicle Re-Id, we compare eight methods on the built VeRi-776 dataset:

(1) **BOW-CN** [24]. This is the Bag-of-Words with Color Name descriptor which is one of the state-of-the-art appearance features for person Re-Id. It is also adopted as the color feature for vehicle re-id as in [9].
(2) **LOMO** [27]. This is the state-of-the-art texture features for person Re-Id which can effectively overcome the various illumination in real-world surveillance environment.
(3) **GoogLeNet** [2]. This method utilizes the GoogLeNet model [25] which is fine-turned on the CompCars [2]. We adopt it as a feature extractor to obtain the high-level semantic attributes of the appearance.
(4) **FACT** [9]. We adopt the FACT [9] to estimate the appearance similarities between the query images and the test tracks. The FACT considers all of the colors, textures, shapes, and semantic attributes for appearance-based filtering.

(5) **Plate-SNN**. This scheme only uses the license plate similarities between the query and tracks to search for the nearest target in test tracks. The features are calculated by the SNN model trained as the **Plate-SNN** in Sect. 4.3.
(6) **FACT + Plate-SNN**. We firstly use the FACT as the coarse vehicle filter. Then we adopt the post-fusion strategy which combines the similarities of FACT model and Plate-SNN model as fine search. The weights used in summation are 0.4 and 0.6 respectively for the FACT and the Plate-SNN due to their individual performances in vehicle Re-Id.
(7) **FACT + Plate-REC**. In this scheme, we adopt a commercial plate recognition system (Plate-REC) to replace the Place-SNN as the fine search.
(8) **FACT + Plate-SNN + STR**. This scheme integrates the similarities of the FACT, Plate-SNN, and spatiotemporal relations (STR). The spatiotemporal similarity between the query and test is calculated by Eq. 3. Then, the similarity matrixes of the FACT + Plate-SNN and STR are both normalized to $(0, 1)$. At last, the two matrixes are summed with different weights. The weights are 0.8 and 0.2 respectively due to their individual performances. By this means, the appearance, license plate, and spatiotemporal relations are combined together for the progressive vehicle search.

Table 2 illustrates the mAP, HIT@1, and HIT@ of the above models. The CMC curves are plotted in Fig. 7. From the results, we can find that:

(1) For the appearances based models, the BOW-CN, LOMO, GoogLeNet, and FACT have competitive performances which are all not very good for vehicle Re-Id. The FACT is better than GoogLeNet, because the GoogLeNet model only considers the semantic attributes, while the FACT also combines color and texture features. This demonstrates that the low-level features as well as high level features are both effective for appearance-based filtering. In addition, the appearance-based model can only find the vehicles that have similar appearance to the query but cannot accurately identify the vehicles.

(2) The progressive combination of the appearance-based model and Plate-SNN model achieves 7.39 % improvement in mAP and 10.13 % in HIT@1 for

Table 2. Comparison of different methods on VeRi-776 dataset.

Methods	mAP	HIT@1	HIT@5
BOW-CN [24]	12.20	33.91	53.69
LOMO [27]	9.64	25.33	46.48
GoogLeNet [2]	17.04	49.82	71.16
FACT [9]	18.49	50.95	73.48
Plate-SNN	15.74	36.29	46.60
FACT + Plate-REC	18.62	51.19	73.60
FACT + Plate-SNN	25.88	61.08	77.41
FACT + Plate-SNN + STR	**27.77**	**61.44**	**78.78**

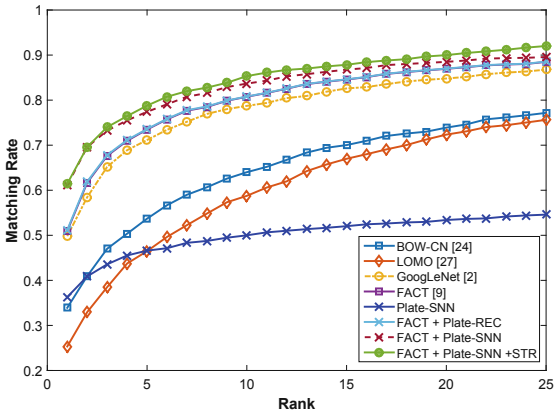


Fig. 7. The CMC curves of different methods.

vehicle Re-Id compared with the FACT model. The results validate the effectiveness of our progressive search with appearance-based coarse filtering and plate-based accurate search. The appearance-based filter can filter out most of the dissimilar vehicles, especially the vehicles have similar license plates to the query. Then, for the remaining vehicles with similar appearance to the query, the plate-based method can search for the vehicles of which the license plates are also similar to avoid the mistaken matching. The plate recognition methods could fail due to the various illuminations, occlusions, and resolutions. So the Plate-REC method only achieve marginal improvement.

(3) The FACT + Plate-SNN + STR model obtains further improvements compared with the former methods. This demonstrates that with the from-near-to-distant principle, the progressive search in the spatial and temporal domains also improves the vehicle Re-Id. In total, the proposed PROVID method achieves 9.28 % improvements in mAP, 10.94 % in HIT@1, and 5.3 % in HIT@5 compared with the state-of-the-art appearance-based model. The results validate the effectiveness of our progressive vehicle search framework and the indispensability of each feature for accurate vehicle Re-Id. More importantly, we also evaluate the speed of progressive method (157 ms/query), which reduces 87.84 % time cost than the strategy without progressive fusion (1,292 ms/query). This demonstrates that the progressive search can dramatically improve the instant-level search accuracy and speed in real-world space.

Figure 8 shows several examples of the PROVID method on VeRi-776 dataset. Sample (a) and (b) illustrate the significant effect of the license plate-based Re-Id. The appearance-based filter find the similar vehicles, but the targets are not ranked in the front of the results. Then, with the license plate-based method, the vehicles are correctly searched. Sample (c) and (d) show that the license plate-based search fails due to the severe blur and distortion of the plates. However, in these samples, the target vehicles are searched with the assistance of the spatiotemporal relations. Sample (e) perfectly shows the effectiveness of the

Fig. 8. Examples of the PROVID on VeRi-776 dataset with the top-5 results. The true positive is in green box, otherwise red. In each example, the three rows are the results of the FACT, FACT + Plate-SNN, and FACT + Plate-SNN + STR. (Best seen in color.) (Color figure online)

PROVID method. In this sample, the vehicles have similar appearance are firstly found, then the license plate-based model achieves accurate search. At last, the spatiotemporal relations guarantee the target vehicles are ranked in the top position. Sample (f) is a failure case of the proposed method. The vehicles that have different colors to the query are not filtered out due to the illuminations, so the proposed method do not distinguish the yellow cars and white cars. Besides, without the license plate in the query, the unique ID cannot be utilized to accurately search the vehicles. Therefore, the spatiotemporal relation also fails in such an uncertain situation. To overcome these problems, we need to exploit an appearance-based model which is more robust to the environment factors such as illuminations and occlusions. Furthermore, we will further utilize the license plate such as integrating the plate recognition and verification in an end-to-end multi-task deep neural network.

5 Conclusions

In this paper, we propose a deep learning-based progressive vehicle Re-Id approach, which employs the deep CNN to extract the appearance attributes as the coarse filter, and Siamese neural network-based license plate verification as the fine search. Furthermore, the spatiotemporal relations of vehicle in real-world urban surveillance is investigated and combined into the proposed method. To facilitate the research, we build one of the largest vehicle Re-Id dataset from urban surveillance videos with diverse vehicle attributes, sufficient license plates, and accurate spatiotemporal information.

Acknowledgements. This work is supported by the National High Technology Research and Development Program of China (No. 2014AA015101), the National Natural Science Foundation of China under Grant No. 61332005, the Funds for Creative Research Groups of China under Grant No. 61421061, the Cosponsored Project of Beijing Committee of Education, the Beijing Training Project for the Leading Talents in S&T (ljrc 201502), and the Fundamental Research Funds for the Central Universities (No. 2016RC43).

References

1. Feris, R.S., Siddiquie, B., Petterson, J., Zhai, Y., Datta, A., Brown, L.M., Pankanti, S.: Large-scale vehicle detection, indexing, and search in urban surveillance videos. IEEE Trans. Multimedia **14**(1), 28–42 (2012)
2. Yang, L., Luo, P., Loy, C.C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3973–3981 (2015)
3. Matei, B.C., Sawhney, H.S., Samarasekera, S.: Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3465–3472 (2011)
4. Valera, M., Velastin, S.A.: Intelligent distributed surveillance systems: a review. IEE Proc. - Vis. Image Sign. Process. **152**(2), 192–204 (2005)
5. Zhang, J., Wang, F.Y., Wang, K., Lin, W.H., Xu, X., Chen, C.: Data-driven intelligent transportation systems: a survey. IEEE Trans. Intell. Transp. Syst. **12**(4), 1624–1639 (2011)
6. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. **5**(38), 1–55 (2014)
7. Du, S., Ibrahim, M., Shehata, M., Badawy, W.: Automatic license plate recognition (ALPR): a state-of-the-art review. IEEE Trans. Circuits Syst. Video Technol. **23**(2), 311–325 (2013)
8. Wen, Y., Lu, Y., Yan, J., Zhou, Z., Von Deneen, K.M., Shi, P.: An algorithm for license plate recognition applied to intelligent transportation system. IEEE Trans. Intell. Transp. Syst. **12**(3), 830–845 (2011)
9. Liu, X.C., Liu, W., Ma, H.D., Fu, H.Y.: Large-scale vehicle re-identification in urban surveillance videos. In: IEEE International Conference on Multimedia and Expo (2016, Accepted and to appear)
10. Kettnaker, V., Zabih, R.: Bayesian multi-camera surveillance. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 253–259 (1999)

11. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Comput. Vis. Image Underst. **109**(2), 146–162 (2008)
12. Xu, J., Jagadeesh, V., Ni, Z., Sunderrajan, S., Manjunath, B.: Graph-based topic-focused retrieval in distributed camera network. IEEE Trans. Multimedia **15**(8), 2046–2057 (2013)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
14. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
15. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: a deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems, pp. 2121–2129 (2013)
16. Liu, W., Mei, T., Zhang, Y., Che, C., Luo, J.: Multi-task deep visual-semantic embedding for video thumbnail selection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3707–3715 (2015)
17. Mei, T., Rui, Y., Li, S., Tian, Q.: Multimedia search reranking: a literature survey. ACM Comput. Surv. **46**(3), 38 (2014)
18. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a siamese time delay neural network. Int. J. Pattern Recogn. Artif. Intell. **7**(04), 669–688 (1993)
19. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 539–546 (2005)
20. Zhang, C., Liu, W., Ma, H.D., Fu, H.Y.: Siamese neural network based gait recognition for human identification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2832–2836 (2016)
21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
22. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision, pp. 1470–1477 (2003)
23. Van De Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Trans. Image Process. **18**(7), 1512–1523 (2009)
24. 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)
25. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
26. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia, pp. 675–678 (2014)
27. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition Proceedings, pp. 2197–2206 (2015)