A Deep-learning Based Approach to Vehicle Re-identification

Graduation Thesis for the Bachelor's Degree

Author: QUAN, Pengrui 1155092203

Supervisor: Prof. Wang Xiaogang

Associate Examiner: Prof. Thierry Blu

2018-4-11

DEPARTMENT OF ELECTRONIC ENGINEERING, FACULTY OF ENGINEERING, THE CHINESE UNIVERSITY OF HONG KONG

Acknowledgement

I would like to express my sincere thanks to both my advisor Prof. Hongsheng and Prof. Xiaogang who guides and instructs me patiently through the whole project and my peers who offer their helpful advice, as well as senior colleagues at Multi-media & Signal Processing Lab, Electronic Engineering Department, who contributed to frameworks of person re-identification as well as human pose estimation from which my vehicle re-identification work is inherited.

A Deep-learning Based Approach to Vehicle Re-identification

Contents

Abstract

Vehicle re-identification (ReID) [1][5][6] is an important problem and has many applications in video surveillance and intelligent transportation. The existing approaches to vehicle ReID mainly relies on the appearance information of vehicles with an oversimplified spatiotemporal information. In this project, we make use of both the appearance of vehicles and their time and geo-location relationship to improve the retrieved accuracy. We make progress on the design of deep learning frameworks and finally, extensive experiments and analysis demonstrates the performance of our framework.

1. Introduction

1.1 Vehicle Re-identification

Vehicle, which is an essential object class in modern camera surveillance, has been often attached great significance by vision community. The task of vehicle re-identification (ReID) aims at associating vehicle images through the entire camera surveillance system, where the ability of automatically tracking suspicious vehicles is extremely important in criminal investigation processes.

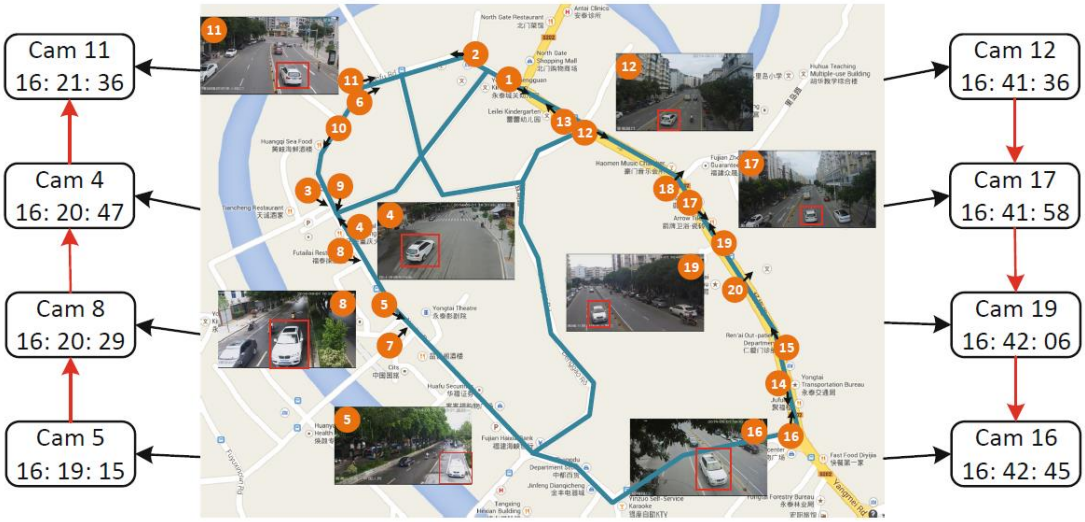

Figure 1 vehicle ReID scenario

The goal of vehicle re-identification is that given one query image of one specific vehicle, a vehicle ReID system is expected to provide all the images of the same vehicle from a large gallery database. With the vehicle ReID system, all the vehicles are expected to be tracked all the time across cameras. In this project, we will analyze the vehicle ReID problem and provide suitable solutions.

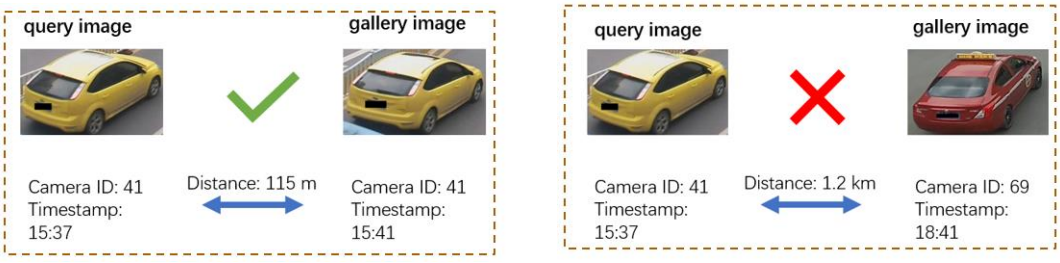

Figure 2 Vehicle ReID Task

Similar to person ReID, most of the existing vehicle re-identification research is manipulated using the appearance information of vehicles. However, due to enormous differences in

lightening conditions and various qualities in resolution of cameras, merely depending on the appearance of vehicles might not be reliable and qualified enough under the current surveillance system. Hence, we will incorporate additional spatiotemporal information of vehicles to propose a feasible approach to the ReID problem.

1.2 Recent Work

Xinchen Liu [5] propose an open vehicle dataset VeRi-776, which consists of massive vehicle images capture by urban surveillance camera, as well as the spatiotemporal information of the images being capture by the cameras.

Yantao Shen [1] propose a two-stage framework that incorporates complex spatiotemporal information, such as posterior probability of the same class and Markov Random Fields for effectively regularizing the re-identification results

Zhongdao Wang [8] design an orientation invariant feature embedding module and regularize it with spatiotemporal modeling. By adopting landmark regressor as the orientation feature framework, they can propose an orientation invariant feature aggregation network.

2. Appearance-based Vehicle ReID

2.1 Baseline Model with Single Convolutional Neural Network

In the first place, to tackle the vehicle ReID task, the similarity score that describe the confidence within a pair of vehicle is needed. Based on the visual information of vehicles, a Convolutional Neural Network (CNN) [2] can be utilized to characterize their similarity. By finetuning a pre-trained ResNet-50 [3] model on the provided training vehicle dataset with Cross-Entropy Loss, where:

$$Cross - Enropy\ Loss(x, class) = -log(\frac{e^{x[class]}}{\sum_{class'} e^{x[class']}}) \quad (1)$$

Then a vehicle classifier of the training set can be obtained. We remove the last three fully connected layer that acts as a linear classifier for high dimensional vectors, a feature extractor of vehicles can be constructed. The feature extractor produces a vector of $R^{1024\times1}$ that characterizes the identity of vehicles.

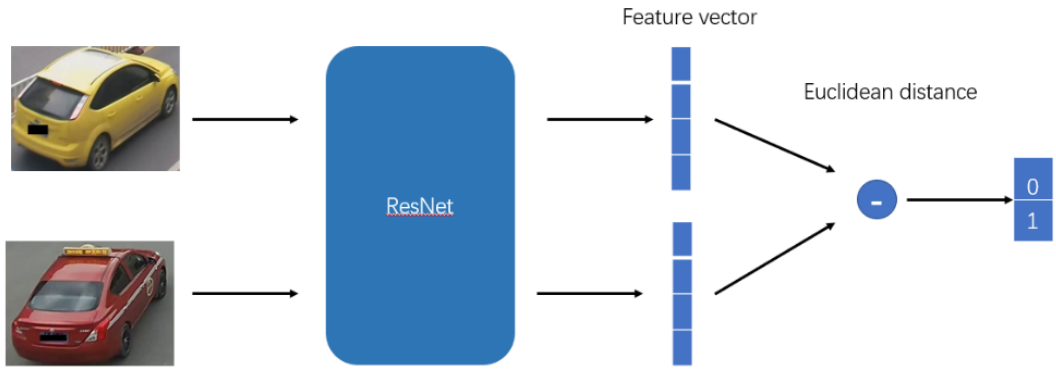

Figure 3 baseline model with single convolutional neural network

Therefore, by evaluating the distance in Euclidean Space, we can compute the similarity score between different vehicles and hence, in the validation and test process, the system can search the gallery and retrieves images of similar vehicles.

2.2 Baseline Model with Siamese Neural Network

Inspired by the idea that the Siamese Neural Network (SNN) can be utilized to verify hand-writing signatures [4], we designed a pairwise verification CNN model to augment the ReID accuracy.

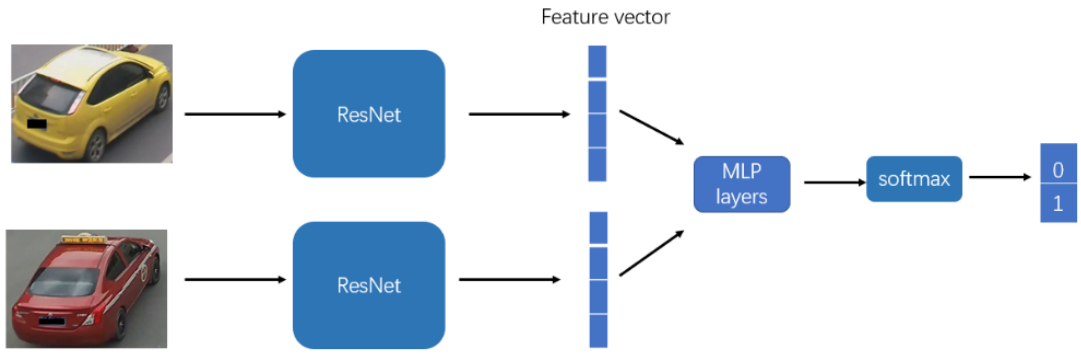

Figure 4 baseline model with Siamese neural network

The network consists of two parallel ResNet-50 models. After passing through a pair of queries image, the SNN produces a two-dimensional vector that represents the confidence score in whether the queries are of the same class or not respectively. Through applying the Cross-Entropy Loss to the training process, we can obtain a vehicle re-identification system based on their visual information.

2.3 Spatiotemporal Posterior Probability

Apart from the appearance information of vehicles, the time and geo-location information

(defined as spatiotemporal information) for each vehicle image also contains useful attributes to vehicle ReID. To effectively incorporate the spatiotemporal information into the baseline model, we propose the following two ways to compute the posterior class probability given information of time and location of queries. The weighted sum to combine Spatiotemporal Posterior Probability and output of SNN is $\lambda_1 = 0.3$ and $\lambda_2 = 0.7$.

Linear Approximation

The first approach is to compute their posterior class probability with respect to their difference in time and space:

$$P(y_i \neq y_j | i,j) = \frac{|T_i - T_j|}{T_{max}} \times \frac{|D_i - D_j|}{D_{max}} \quad (2)$$

where T_i and T_j are the time stamp of query images, $|D_i - D_j|$ represents their distances when captured, and T_{max} and D_{max} represents their largest distance in time and space. The model characterizes the probability of the queries being the same vehicles in a feasible way.

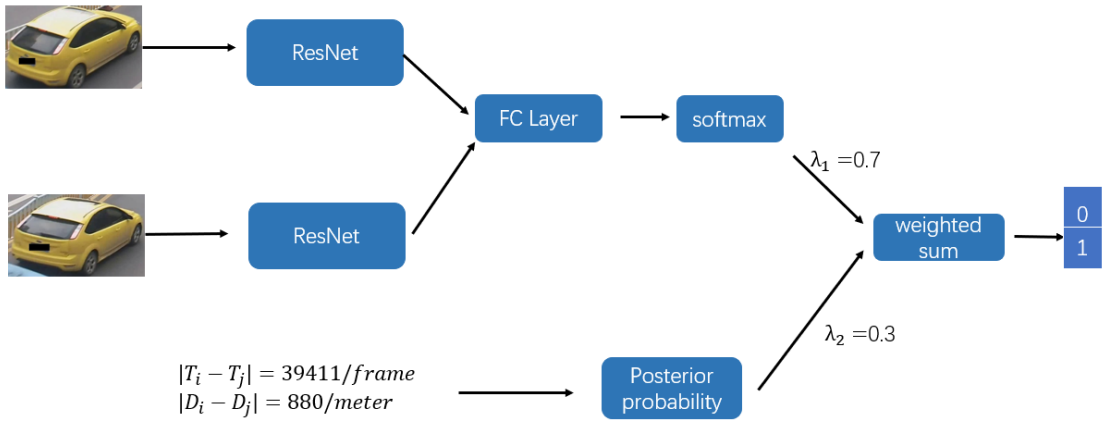

Figure 5 baseline model with posterior probability

2.4 Integrated Siamese Neural Network

In addition to solely training a deep convolutional neural network or constructing a posterior probability function, we propose an end-to-end training approach: construct a visual-spatiotemporal deep neural network (called integrated SNN) with the baseline SNN and a fully connect neural network.

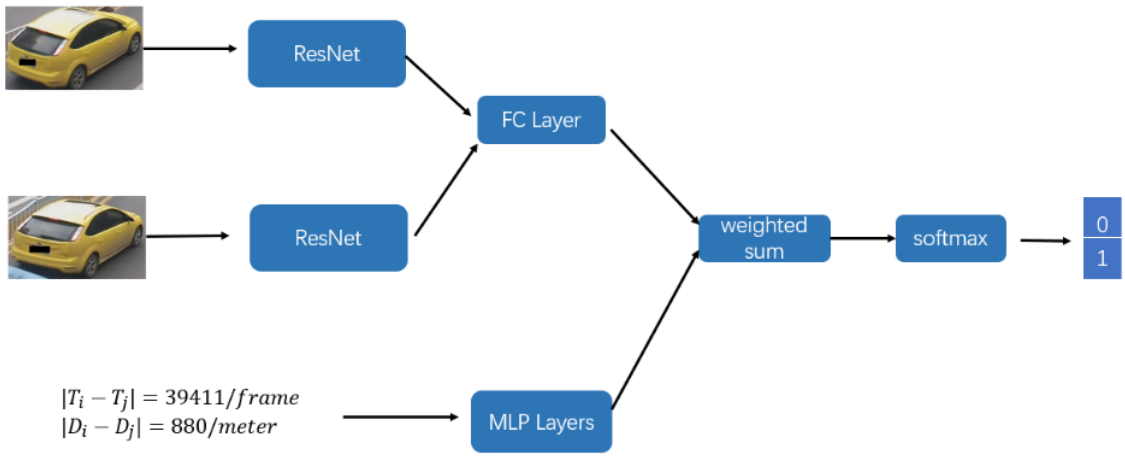

Figure 6 integrated Siamese neural network

The integrated SNN consists of two branches: the visual branch is designed as a SNN with two trained ResNet-50; the other branch consists of two fully connected neural network computing the spatiotemporal compatibility. The first branch computed the weighted sum of two pooling features, producing a two-dimensional vector. The second branch utilizes ReLu activation function after each intermediate layer and generates a two-dimensional vector. The output of the two branches are concatenated with a fully connected layer and finally input into a sigmoid function to obtain the ultimate confidence score.

3. Orientation-based Vehicle ReID

Since images of different orientations of an identical vehicle is constantly captured by surveillance camera, the phenomenon causes a huge challenge for the vehicle to be reidentified by a deep neural network. Therefore, we intend to design an orientation-based vehicle reidentification network while retaining the global feature of vehicles to improve the performance of the reidentification model.

3.1 Vehicle's Key Point detection

3.1.1 Stacked Hourglass Models

The human pose estimation has intrigued the vision community to make constantly effort to tackle the problem. One state-of-the-art technique to solve the problem is to utilize Stacked Hourglass Models [7]. The network consists of three critical structures:

i. Residual layer acts as basic building blocks. Instead of simply convoluting the input with a single kernel, the residual layer convolutes the input with several kernels of

varied sizes with the aid of activation functions.

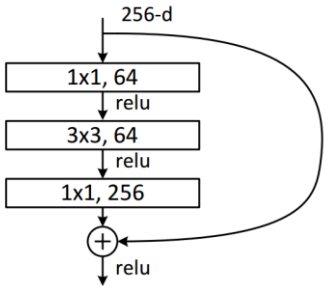

Figure 7 Configuration of a Residual Layer

ii. Bottom-up and top-down processing is applied to the whole hourglass network. By employing intermedia loss upon each scale of the layer, the network can capture global features as well as local features.

iii. By stacking several hourglass end-to-end, feeding the output of one as the input of another, the hourglass network can demonstrate better ability to detect key points. Similarly, intermediate supervision is also utilized to prevent previous network from not being functioning.

3.1.2. Vehicle's Key Point detection

3.1.2.1 Ground-truth Generation

We apply a Gaussian-like mask to the exact location of the key point to generate the ground truth. The ground-truth heatmap is actually a 2D gaussian function (with a standard deviation of 2 pixels) centered on the location of key points. These scenarios are classified as positive samples. If there are no key points appearing, the ground-truth heatmap will become zero everywhere, which is referred to negative samples.

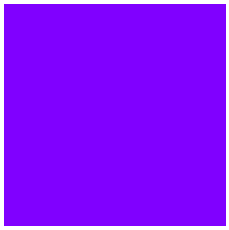

Figure 8 Negative Sample

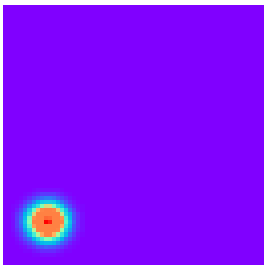

Figure 9 Positive Sample

3.1.2.2 Precision Evaluation for Key Points Detection

Since the conventional criterion for human pose estimation utilized the size of head of a person as well as normalized by size of the whole person, we make compromise and propose the evaluation criterion as follows:

if max(pred_map) > threshold && $|x - x_0| < r$ && $|y - y_0| < r$:
    return ***Correct***
else if max(pred_map) <= threshold:
    return ***Correct***
else:
    return ***False***

where pred_map is the heatmap generated by the hourglass. (x, y) is the location of the predicted heatmap. (x_0, y_0) is the location of ground truth.

Figure 10 Precision Evaluation

If the position of the maximum of predicted heatmap is close enough to the location of key points of ground-truth, then we think network correctly predicted the key point. If the maximum is smaller than a specific threshold and no such a key point appears in the ground truth, then we also regard it as a correctly predicted scenario.

3.2 Integrated Hourglass Network

Upon obtaining the Stacked Network which generates predicted heatmaps of vehicle's key points, we apply it to construct a Integrated Hourglass Network for vehicles ReID. A pretrained ResNet model generates global feature of a vehicle, which is later pointwise multiplied with the heatmap generated by the hourglass model.

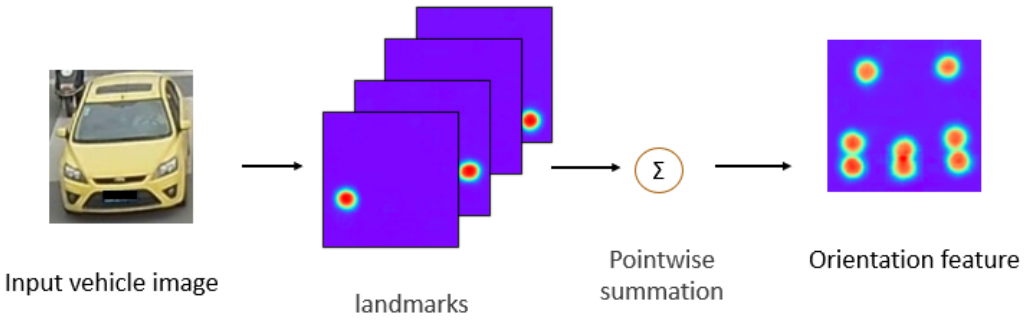

Figure 11 Orientation feature aggregation

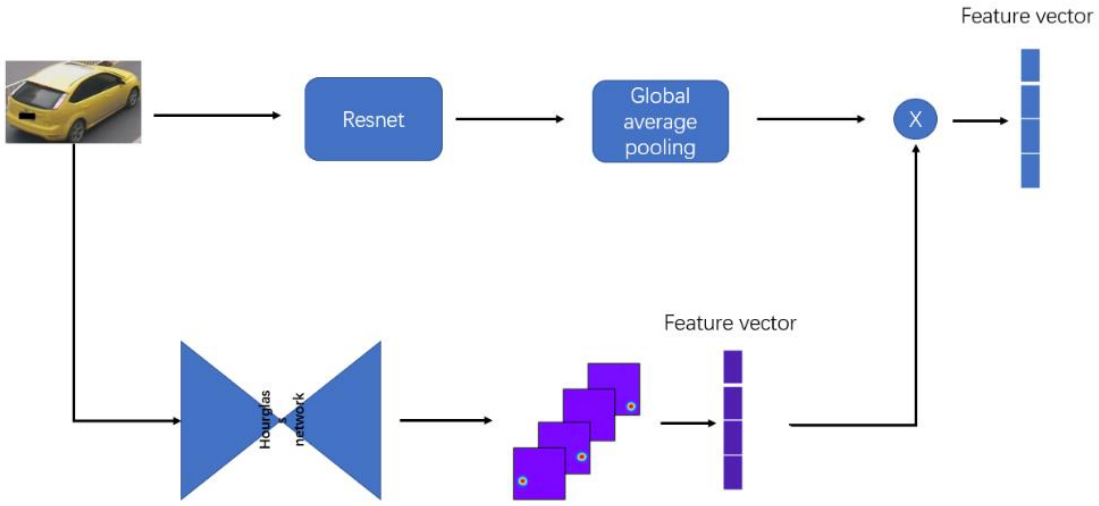

Figure 12 Integrated Hourglass Network

4. Experiments

In order to evaluate the performance of our vehicle ReID model, we conduct experiments on the VeRi-776 dataset [5], which consists of 50,000 images of 776 vehicles. Each vehicle is capture by 2 to 18 cameras. Besides the timestamps and geo-locations of cameras with respect to each image are provided as well. The entire dataset is divided into a training dataset comprised of 37,781 images of 576 vehicles, a validation dataset of 11,579 images of 200 vehicles, and a query and gallery dataset of 1678 and 6741 images for future testing.

4.1 Evaluation Criterion

The mean Average Precision (mAP) is utilized to evaluate the performance of the proposed vehicle ReID framework. The average precision for each query q is calculated by:

$$AP = \frac{\sum_{class=k} P(k) \times gt(k)}{N_{gt}} \tag{5}$$

where $gt(k)$ is an indicator function of the ground truth, which equals to 1 if the item at rank k is a matched vehicle image and 0 otherwise. Besides, $\mathrm{P(k)}$ is the retrieved accuracy at each retrieve k and N_{gt} is the number of ground truth. Therefore, the mAP for all query images is the computed using:

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{6}$$

where Q is the number of query images. For each query image, only images of the same vehicles from other cameras would be taken into account for computing the mAP. The mAP is to evaluate the overall performance on the entire queries.

4.2 Training Scheme

An alternative training strategy is utilized to train the whole Neural Network. As for the Integrated Siamese Network, (i) we first train a Resnet by giving the classification labels as supervision. (ii) With the Resnet fixed in the first 20 epochs, we train on a vehicle ReID classification task the spatiotemporal branches and fully connected layers at the output of the Siamese network. (iii) Finally, we jointly updated the weight of the whole model.

As for the Integrated Hourglass Networks, the training scheme consists of four steps: (i) We first trained a key point detector of vehicles. In this training process, data augmentation techniques such as random horizontal flipping are applied to training dataset, and mean square error is adopted as the loss function. (ii) Train a Resnet by giving the classification labels as supervision, which acts as a global branch of vehicle's feature. (iii) Jointly train the Integrated Hourglass Network in a vehicle ReID task. (iv) In the first 10 epochs, only Resnet is updated. Otherwise, since the loss is quite large in the first few epochs, we will run the risk of damaging the pretrained hourglass branch.

In both training processes, cross-entropy classification loss is adopted.

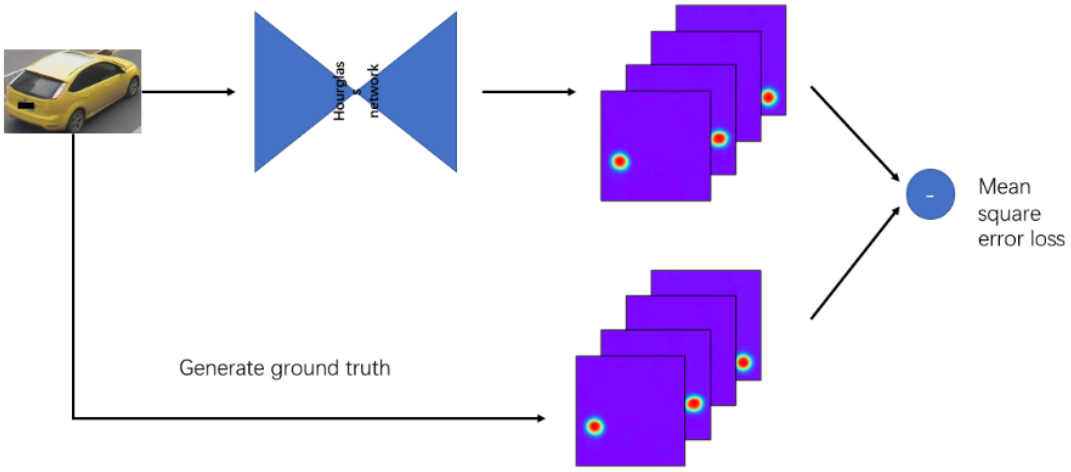

Figure 13 Training scheme for hourglass network

4.3 Experiment on Vehicle's Key Points Detection

To evaluate the performance of our Hourglass Network, we adopted the proposed precision criterion. These are examples of an input image, corresponding predicted heatmaps generated by Stacked Hourglass Network and ground truth heatmaps.

Figure 14 Vehicle Image after transformation (after rescaling and normalization)

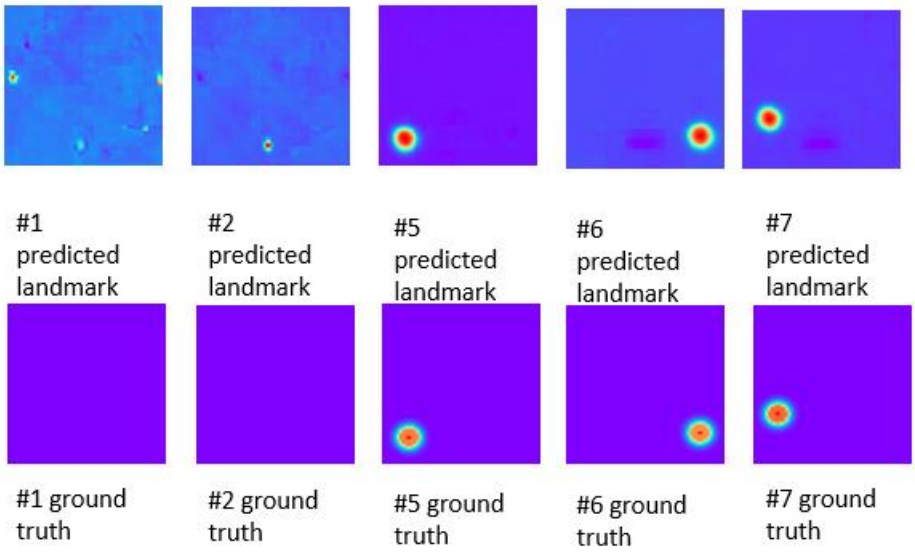

Figure 15 example prediction and ground truth

The overall precision of the predicted heatmap of the test dataset is as follows. The best performance is attained in the 75th epoch. The model can predict vehicle's key points with an accuracy of 89.6%.

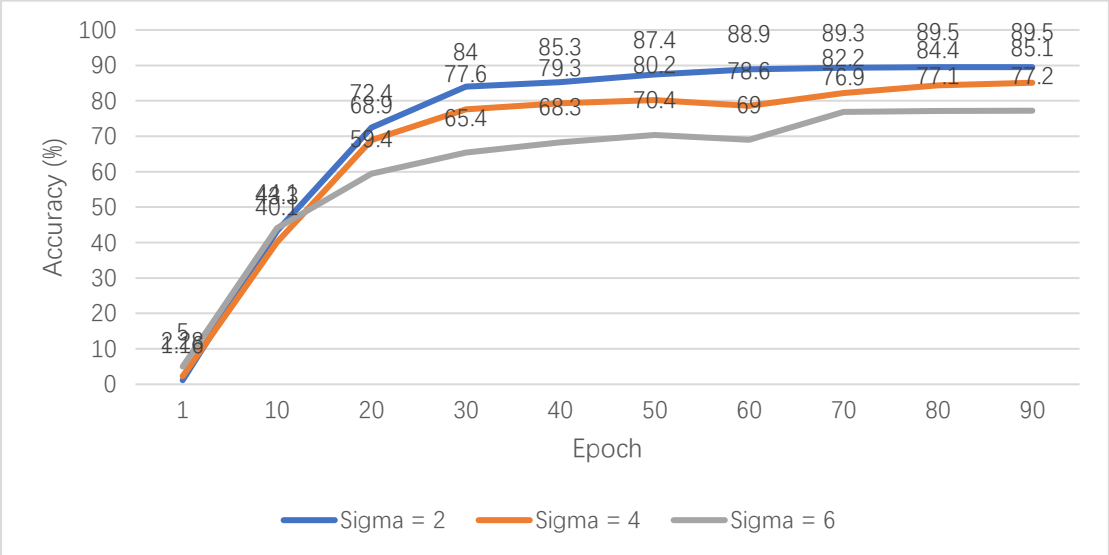

Figure 16 Average precision of the predicted key points.

Key points detection	Best average accuracy (%)
Sigma = 2	89.6
Sigma = 4	85.4
Sigma = 6	77.6

Figure 17 Prediction accuracy w.r.t. different sigma

4.4 Experiment on Vehicle ReID

In this session, we conduct our own proposed methods to vehicle ReID, testing them on the mentioned VeRi-776 dataset. By means of calculating their mAP respectively, we can evaluate their performance of ReID accuracy:

- *Single CNN* denotes our approach which makes use of ResNet-18 to generate features and evaluate their similarity in Euclidean Space.
- *Single CNN with pairwise training* is an improved version of single CNN. The network is learnt on the training dataset that is augmented with random cropping, rotation and random flapping, etc. Furthermore, it is also finetuned on pairwise data for better performance. The pairwise data consists of an anchor image and an image either sampled from the same vehicle or a different vehicle.
- *SNN* is constructed with two parallel identical ResNet-18. In the training stage, if a pair of vehicle has the same identity, it is labeled as positive sample, while the pair with a different identity will be treated as a negative sample. In the following experiment, the positive-to-negative sampling ratio is set to 1:3.
- *SNN + spatiotemporal posterior probability* combines SNN with the linear approximate model at the test stage. The ultimate score is a weighted sum of posterior probability and the similarity score in appearance.
- *Integrated SNN* denotes that the network consists of two branches, which calculates the similarity scores regarding spatiotemporal relation as well as appearances respectively.
- Integrated Hourglass Network consists of two branches: one of the branch is a ResNet that is pretrained on classification supervision. The orientation feature generated by the hourglass network is expected to amplify useful component of the global feature produced by the pretrained ResNet.

Method	**mAP (%)**	**top 1 accuracy (%)**	**top 5 accuracy (%)**
Single ResNet-18	33.5	77.0	83.3
Single ResNet-50	37.1	79.5	87.4
SNN (ResNet-18)	41.1	89.4	95.2
SNN (ResNet-18) + Spatiotemporal model	42.4	89.9	96.3
SNN (ResNet-50)	45.9	93.3	98.6
SNN (ResNet-50) + Spatiotemporal model	47.6	94.7	99.5
Integrated SNN	49.7	94.8	100

(ResNet-18)			
Integrated SNN (ResNet-50)	**58.6**	**96.6**	**100**
Hourglass (σ = 2)	45.1	88.6	94.8
Hourglass (σ = 6)	46.7	89.2	95.4

4.5 Experiment Observation

Not surprisingly, after adopting data augmentation and pairwise training, we can further improve the mAP to 33.0%. Moreover, the SNN baseline model would further give us 41.4% mAP. Additionally, when the linear approximate model is combined with the scores output by SNN, the mAP is increased by 1%. It shows us that the assumption of the linear approximation is reasonable: the probability that two vehicles belongs to the same class decreases with respect to the increment of distance in time and space.

Under the condition that the MLP network fails to converge to a satisfactory point, we choose to train it with the aid of a trained SNN. By means of concatenating fully connected them with a fully connected layer and fix the parameter of the SNN, the neural network can be trained and have a much better performance than previous. In the test stage, the integrated SNN can achieve a mAP of 49.7%, which outperforms all previous methods dramatically.

The Stacked Hourglass Model can predict the key points of an image, with the precision of 89.6%. The Integrated Hourglass Network seems to give reasonable results. The Hourglass model generates orientation feature. The orientation-based model has better performance than Siamese ResNet. Orientation based model has a better performance than the model learning only appearance information generally.

4.6 Parameters Used in the Experiments

Parameter	Value
positive-to-negative ratio	1:3
weighted sum to combine spatiotemporal information and output of SNN	$\lambda_1 = 0.3$
	$\lambda_2 = 0.7$
training images	31535
validation images	6243
query images	1678
gallery images	11579
learning rate for vehicle ReID	0.015
learning rate for Stacked Hourglass Network	0.001
batch size for vehicle ReID	256

batch size for key points detection	320
loss function for vehicle ReID	cross-entropy loss
loss function for key points detection	means square error (MSE) loss
epoch for vehicle ReID	50
epoch for key points detection	90
devices	4 Tesla K80 Servers

5. Conclusion

In this project, we make use of several frameworks for vehicle ReID task, with both visual and spatiotemporal information. After obtaining the baseline models, we input the spatiotemporal information into the neural network. On the other hand, with the aid of a pretrained SNN, the whole integrated neural network can have a much better performance. In this way, the overall mAP is augmented dramatically as shown in the results experiments. Nevertheless, even though Stacked Hourglass Network extraordinarily detects the key points of vehicles, the whole Integrated Hourglass Network still doesn't give satisfactory prediction. Therefore, further investigation will be made to improve the vehicle ReID framework.

6. Reference

1. Shen Y, Xiao T, Li H, et al. Learning Deep Neural Networks for Vehicle Re-ID with Visual-spatio-temporal Path Proposals[J]. arXiv preprint arXiv:1708.03918, 2017.
2. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
3. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
4. Bromley J, Guyon I, LeCun Y, et al. Signature verification using a" siamese" time delay neural network[C]//Advances in Neural Information Processing Systems. 1994: 737-744.
5. Liu X, Liu W, Mei T, et al. A deep learning-based approach to progressive vehicle re-identification for urban surveillance[C]//European Conference on Computer Vision. Springer International Publishing, 2016: 869-884
6. D. Zapletal and A. Herout. Vehicle re-identification for automatic video traffic surveillance. //In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 25–31, 2016.
7. Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//European Conference on Computer Vision. Springer, Cham, 2016: 483-499.
8. Wang Z, Tang L, Liu X, et al. Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 379-387.