BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FINE-GRAINED RECOGNITION AND RE-IDENTIFICATION OF VEHICLES USING ADVANCED FEATURE EXTRACTION

ROZPOZNÁNÍ TYPŮ A RE-IDENTIFIKACE VOZIDEL S POUŽITÍM POKROČILÝCH METOD EXTRAKCE PŘÍZNAKŮ

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR **ONDŘEJ DOSEDĚL**
AUTOR PRÁCE

SUPERVISOR **Ing. JAKUB ŠPAŇHEL**
VEDOUCÍ PRÁCE

BRNO 2019

Brno University of Technology
Faculty of Information Technology

Department of Computer Graphics and Multimedia (DCGM) Academic year 2018/2019

Bachelor's Thesis Specification

22078

Student:	**Doseděl Ondřej**
Programme:	Information Technology
Title:	**Fine-Grained Recognition and Re-Identification of Vehicles Using Advanced Feature Extraction**
Category:	Image Processing

Assignment:

1. Study the available materials on the fine-grained classification of types and re-identification of vehicles using convolutional neural networks.
2. Analyze available solutions for advanced features extraction.
3. Design modifications to the architecture of selected neural networks for fine-grained vehicle classification / re-identification using key points detected on the vehicle.
4. Experiment with your implementation and improve iteratively.
5. Compare the results and discuss future developments.
6. Create a brief poster and a video presenting your bachelor thesis, its goals and results.

Recommended literature:

- According to the supervisor's instructions

Requirements for the first semester:

- The first three points of the assignment

Detailed formal requirements can be found at http://www.fit.vutbr.cz/info/szz/

Supervisor:	**Špaňhel Jakub, Ing.**
Head of Department:	Černocký Jan, doc. Dr. Ing.
Beginning of work:	November 1, 2018
Submission deadline:	May 15, 2019
Approval date:	April 3, 2019

Abstract

The aim of this theses was to analyze and improve methods used for fine-grained vehicle recognition and vehicle re-identification. The proposed method can be used both for recognition and re-identification. It was based on 3D bounding boxes, which were used to detect the vehicle on the image and then the vehicle was normalized by unpacking into 2D. Improvement of this method was done by determining direction of the vehicle and distinguishing between front and rear while unpacking the vehicle. This proposed method improved the existing method based on 3D bounding boxes for recognition, reducing error up to 13 % in single sample accuracy and up to 17 % track accuracy. However, no improvement was gained for vehicle re-identification using LFTD aggregation.

Abstrakt

Práce se zabývala analýzou a následným vylepšením metod užívaných k rozpoznávání typů vozidel a jejich re-identifikace. Navržená metoda může být využita jak pro rozeznání, tak pro re-identifikaci. Byla založena na používání tzv. 3D bounding boxes. Pomocí těchto boxů docházelo k detekci vozidla na obraze. Vozidlo bylo následně normováno rozbalením do dvojrozměrné interpretace. Tato metoda byla vylepšena určením směru vozidla a rozlišováním mezi čelní a zadní stranou vozidla během rozbalení třírozměrného modelu. Představená metoda vylepšuje stávající metodu pro rozpoznávání a snižuje její chybovost až o 13 % pro jeden vzorek a o 17% pro presnost trati. Pro re-identifikaci nedošlo k zvylepšení při použití LFTD agregovani.

Keywords

convolutional neural network, vehicle re-identification, Fine-grained vehicle recognition,3D bounding box

Klíčová slova

konvoluční neuronové sítě, rozpoznávání typů vozidel, re-identifikace vozidel, 3D bounding box

Reference

DOSEDĚL, Ondřej. *Fine-Grained Recognition and Re-Identification of Vehicles Using Advanced Feature Extraction.* Brno, 2019. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Jakub Špaňhel

Rozšířený abstrakt

V posledních letech je snaha o vytvoření co nejinteligentnějších systémů, které by šly použít v dopravě. V tomto ohledu jsou navrhnuté metody a pokusy o sestavení samořídících aut. Jiné výzkumné skupiny se zabývají zlepšením dopravních sledovacích systémů pomocí automatizace.

Pro vytvoření těchto dopravních systémů je potřeba překonat obtížné problémy, jako je rozpoznání typů vozidel a jejich následná re-identifikace. Tato práce je rozdělena na dvě části (tj. rozpoznávání typů vozidel a re-identifikace vozidel), které jsou adresovány pomocí jediné metody. Vozidlo je rozpoznané do co nejpodrobnějších detailů, tj. výrobce, model, verze a rok výroby.

Navrhovaná metoda je založena na používání takzvaných *3D bounding boxů*. Tyto boxy lze automaticky získat z videozáznamů, které jsou pořízené pomocí sledovacích kamer a které je možné použít k detekci jednotlivých částí vozidla, tj. přední/zadní, boční a vrchní část. Bohužel nelze rozlišit mezi přední a zadní částí vozidla. Tento problém je řešen v této práci, pomocí zjištění směru jedoucího vozidla. Následně dochází k rozbalení těchto boxů do jejich 2D interpretace pomocí zjištěné homografie mezi jednotlivými body. Ty jsou následně perspektivně rozbaleny.

Tato práce se zabývá rozlišováním mezi přední a zadní části pomocí detekce směru vozidla. Směr jde vypočítat jako průnik přímky, která je získána ze středu bounding boxu a úběžníku fotografie s polorovinou. Přímku lze vytvořit použitím bodů, pomocí kterých lze určit přední a zadní část. Jestliže dojde k průniku mezi zmíněnou přímkou a polorovinou, tak je vozidlo nasměrováno ke kameře (lze vidět přední část). Pokud k tomuto průniku nedojde, je vidět zadní část. V navrhovaném postupu nedochází k tomuto výpočtu, jelikož oba dva datasety, které jsou používány pro ohodnocení navržené metody, obsahují pomocné informace. V Boxcars116k datasetu, který je použit pro rozpoznávání typů, je uveden směr vozidel. Pro re-identifikaci, kde je použit CarsreId74k, není taková informace k dispozici. Avšak lze určit pomocí kamery, jestli jsou vozidla zobrazené ze přední části nebo ne.

Jestliže dojde ke zjištění, že vozidlo nesměřuje ke kameře, dochází k vertikálnímu zrcadlení jednotlivých částí. To má za následek, že vozidla jsou otočená stejným směrem, jako kdyby by byla čelně ke kameře, a tím pádem jsou stejné části vozidla na stejném místě. Takhle zpracované vozidlo, které má normalizované části, je použito jako vstup do neuronových sítí. Během trénování dochází ke dvěma modifikacím vstupních obrázků pro vytvoření větší rozdílnosti mezi trénovacími daty. V první modifikaci dochází ke změně barvy fotografií, jelikož výstup z kamer závisí na světelných podmínkách a konkrétním nastavení každé kamery. V druhé modifikaci dochází k odstranění nějaké části vozidla, jelikož ve zpracovaném záznamu nemusí být vidět vždy celé vozidlo.

Experimenty pro rozpoznávání typů vozidla byly provedeny pomocí BoxCars116k datasetu. Tento dataset je pro testování rozdělený na 2 části, tj. těžký a střední. V těžké části je nutné vozidlo identifikovat až do roku vytvoření. V předchozí práci, na kterou tato metoda navazuje, docházelo k minimálním rozdílům mezi středí a těžkou částí. Z tohoto důvodu je použita těžká část. Během trénování byly použity obě zmíněné modifikace vozidel. Pro testování byly použity 4 neuronové sítě (ResNet50, VGG16, VGG19, InceptionV2). Naše metoda překonala původní metodu a došlo k redukci chyby až o 17 % pro jeden vzorek a o 13 % pro trat. Druhý pokus byl proveden s používáním pouze jednotlivých částí vozidla a jejich kombinace. Tato část byla testovaná pouze na ResNet50. Nejhorší výsledek byl pro použití pouze přední/zadní části. To může být způsobené 2 důvody. Za prvé, vozidla v datasetu mají podobné přední masky z důvodu místa vytvoření datasetu. Za druhé, během trénování může mít určité vozidlo pouze přední část a ne zadní. Toto

podporuje i fakt, že kombinace přední části s bokem měla větší přesnost než horní část s bokem. Toto testování bylo provedeno pro možné vylepšení této metody pomocí přidání pozornostní vrstvy jako vstup pro neuronové sítě.

Experimenty pro re-identifikaci vozidel byly použité na CarsReId74k datasetu. Použité byly 2 neuronové sítě (ResNet50 a Inception-ResNet-v2). Ohodnocení bylo provedeno pomocí dvou agregačních funkcí, tj. LFTD a average. Navrhovaná metoda, při použití LFTD, nezlepšila stávající metodu, avšak pro average s použitím modifikací obrázku došlo k mírnému zlepšení. Toto může být způsobené kratším trénováním extractoru rysů. Výsledek mohlo také ovlivnit zvětšení prázdného prostoru ve vstupních obrázcích v navrhované metodě. Pro lepší porovnání by bylo vhodné najít lepší parametry metody a delší trénování.

Další práce bz mohla spočívat v přidání pozornostní vrstvy nebo zlepšené rozbalování pomocí detekovaných klíčových bodů.

Fine-Grained Recognition and Re-Identification of Vehicles Using Advanced Feature Extraction

Declaration

Hereby I declare that this bachelor's thesis was prepared as an original author's work under the supervision of Ing. Jakub Špaňhel. All the relevant information sources, which were used during preparation of this thesis, are properly cited and included in the list of references. acknowledgment

....
Ondřej Doseděl
May 14, 2019

Acknowledgements

I would like to thank my supervisor Ing. Jakub Špaňehl for providing all the necessary information and my sister for text reviewing. I would also like to thank all of my friends that helped me with completing this theses by giving usefull feedback on parts, which should be improved.

Contents

Chapter 1

Introduction

People have been trying to create a more intelligent transportation system for the last few years. There have been attempts in the creating of self-driving electrical cars. Others have been trying to improve traffic surveillance systems using automation.

Automatic traffic surveillance systems deal with many non-trivial tasks. For example, they have to be able to detect a vehicle from images and to specify its type. Fine-grained vehicle recognition deals with this problem. The goal of recognition is identification of the vehicle and its make, model, submodel and model year. The biggest problems for recognition are that images from surveillance cameras have usually lower resolution and vehicle models are more and more similar. Surveillance systems have to be capable of re-identifying vehicle, which can be useful for tracking specific vehicle. The biggest problem is color inaccuracy. Each camera has different color accuracy and a lightning condition of a given scene can change during the day. Many methods for re-identification are based on the licence plate, but it might be missing from the image.

In this thesis, we present improvements for surveillance systems by increasing accuracy of vehicle recognition and re-identification. The first part (chapter 2) informs about existing work both on recognition and re-identification. Fine-grained object recognition and specific vehicle recognition are described more in detail. Moreover, it is focused on datasets used for evaluating methods for vehicle recognition. The second part of this chapter informs about general object re-identification followed by person re-identification. Person re-identification is extremely important for surveillance systems. Last two parts deal with vehicle re-identification and datasets. The second part of this thesis (chapter 3) presents the suggested method, that can be used for vehicle recognition and re-identification. This method is based on 3D bounding boxes unpacking. Unpacking is modified by determining vehicle direction.

Chapter 2

Related work

In this chapter different approaches used in object recognition and re-identification are introduced. Recognition is further divided into general fine-grained object recognition, vehicle recognition and vehicle recognition datasets. Re-identification is divided into general object re-identification, person re-identification, vehicle re-identification and vehicle re-identification datasets. Most approaches described in this chapter are based on neural networks. However, there are examples of non cnn-based approaches used for vehicle recognition in subsection 2.1.2.

2.1 Fine-grained Object Recognition

Fine-grained object recognition aims to distinguish various categories with very small visual differences. Example of small vehicle differences are showed in figure 2.1 Earlier methods were based on specific algorithms. Usage of Convolutional neural networks(CNN) has gained more popularity recently not just in object recognition, visual data mapping and many more. These approaches have usually better results then non CNN-based methods.

2.1.1 General Fine-grained Object recognition

There are many different approaches for object recognition. Lin *et al.* [23] created a new architecture for fine-grained image recognition called Bilinear Convolutional Neural networks (B-CNNs).The images were represented as a pooled outer product of features which were derived from two CNNs. Localized feature interactions were captured in a translationally invariant manner. Their model was tested on birds, aircraft and cars. They also showed that bilinear features are redundat and can be reduced in size while maintaining accuracy.

Wu *et al.* [55] presented a deep attention-based spatially recursive model which can be used both for fine-grained image recognition and re-identification. This model was tested on birds, aircraft and cars recognition and on person re-identification while maintaining performance of state-of-the-art methods. They achieved this results by using bilinear pooling and also spatial long-short term memory units (LSTMs). The attention model was leveraged between LSTMs and bilinear outcomes for a dynamic selection.

Lai *et al.* [17] used a different method that combined semi supervised learning with fine-grained learning. Images were cut into multi-scaled parts and then they were fed into the network for fine-grained features. These cuts were assigned with dynamic weights, so the negative impact of background information could be reduced. They also presented a

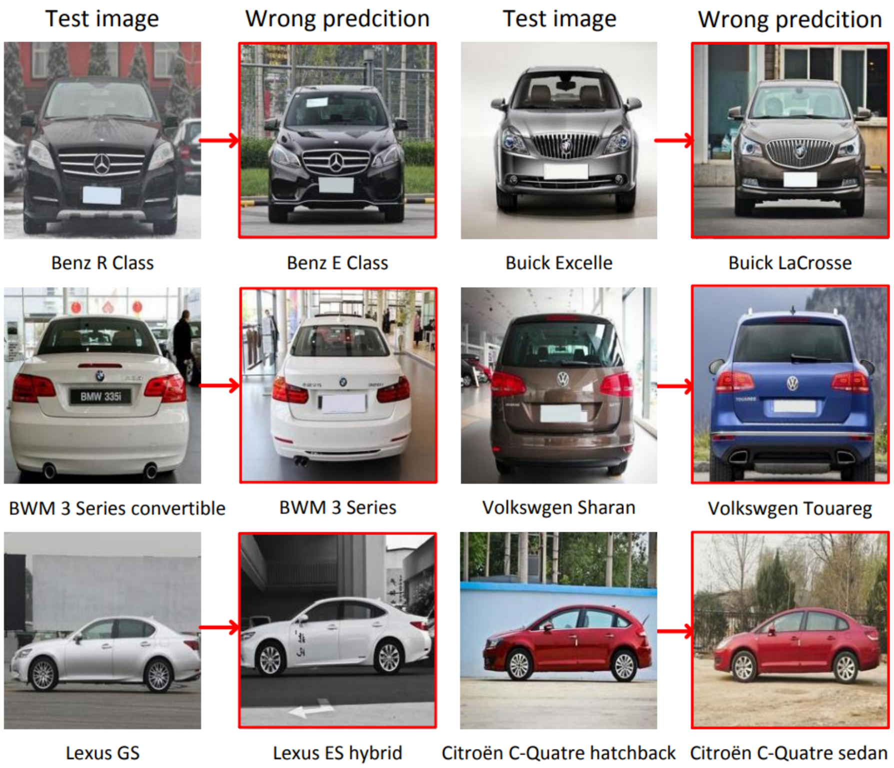

Figure 2.1: This figure shows examples of cars that are very similar. These similarities can be difficult to detect for people on high resolution photos. Cameras from surveillance systems has usually lower resolution. Image taken from [43].

voted pseudo label (VPL) which is an efficient method of semi-supervised learning. VPLS function was to pick up non-confused labels which were verified by prediction of different classification models. Using this combination of VPL and multi-scale recognition in deep CNN architecture, state-of-the-art performance in dogs and birds recognition was achieved.

Sumbul *et al.* [49] presented a novel methodology, i.e. the unified framework to simultaneously learn the alignment of data from multiple sources and the classification model. This method involved a multisource region attention network that can compute per-source feature representations. The combination of attention scores from all sources with feature representations was used for object classification.

Shi *et al.* [43] used general deep CNN and improved it in two aspects. Firstly, they replaced top fc layer of a model with fully connected layers and then trained them with the cascaded softmax loss. This was made to improve a hierarchical label structure of image classes in given dataset. Secondly, they proposed a novel loss function to make model explicitly explore that label structure and similarity regularities of image classes. They named this function as generalize large-margin loss. This can be applied to any deep CNN and was tested on a Stanford car, a fine-grained visual classification-aircraft and CUB-200-2011.

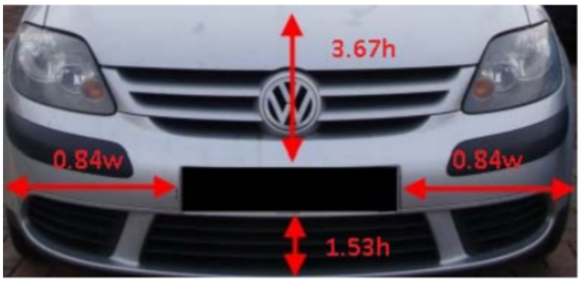
[redacted]

Figure 2.2: Example of a image used in vehicle recognition using the licence plate. Example shows extracted region of interest, where the licence plate is located. Image taken from [36].

2.1.2 Fine-grained Vehicle Recognition

The main goal of fine-grained vehicle recognition is to distinguish and recognize the exact vehicle type that involves the car manufacturer, model, sub-model and the model year. The recognition systems that are fully aimed at vehicles benefit from the fact that vehicles are rigid, have unique parts (e.g. license plates) and precise graphical models and are often published (e.g. 3D car models,CAD models).

Non CNN-Based Methods

Methods described in this section are based on different approaches than neural networks. Mostly these methods are quite old and are not used nowadays. Moreover they are generally not as accurate as CNN-based or they are harder to implement.

The most dominant part of the vehicle is the license plate. Because of that there are some methods (e.g. [18, 28, 36]) that are limited to only frontal/rear images. The general approach is to detect the license plate and extract features that are in the near area of the license plate. Example of region of interests where features are extracted can be sen in figure 2.2.

Other non cnn-methods are based on usage of 3D car model for vehicle recognition. Prokaj and Medioni [37] presented a method capable of recognizing vehicles make and model in a video clip from any viewpoint. Vehicles were classified by vehicle pose estimation in each frame. Pose was estimated by calculating 3D motion in every frame. After that the models from database were modified to the same pose as the vehicle in the video and projected to the image. Features were matched and score was computed. Exemplar from the database with the best result was identified as model of the vehicle.They evaluated their method on 20 clips and 36 different models.

Lin *et al.* [24] proposed to jointly optimize 3D active shape model fitting and fine-grained classification by providing more information (i.e., viewpoint and precise part locations). They evaluated their method on their own dataset FG3DCar achieving better results then other existing methods. They also created analyses about dependence between 3D models and fine-grained classification performance.

Stark *et al.* [48] used the deformable part model which utilized part detectors to encode category-specific information. They also showed usage of fine-grained category prediction to get information about depth from a single image. They also created a new dataset called car-types on which they evaluated their method. This is a small dataset containing 2k images of 14 different categories.

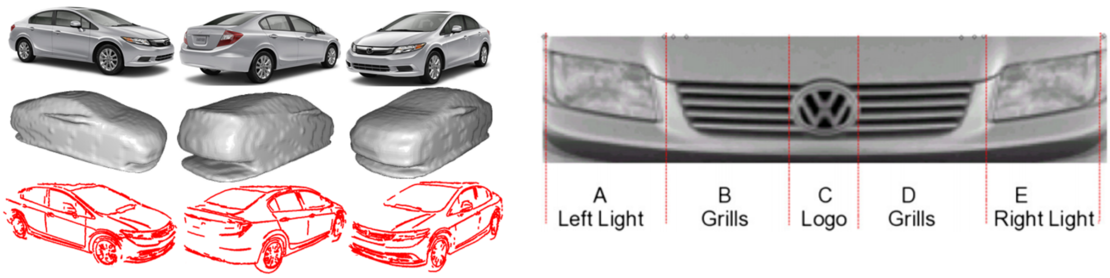

Figure 2.3: **LEFT:** Steps in the construction of a 3D car model. The vehicle in the figure is 2011 Honda Civic Sedan. The upper part consists of images used to generate the visual hull. In the middle part is shown the visual hull and on the bottom part are shown 3D space curves projected onto the visual hull. Image taken from [40].
RIGHT: Vehicle mask segmentation based on phase congruency map gradient treshold. Image taken from [38].

Hsiao *et al.* [40] created a new approach for recognizing a car from a single image from an arbitrary view. Their models consisted of 3D space curves which were obtained by back-projecting image curves onto visual hulls and then used three-view curve matching as shown in figure 2.3. Using chamfer matching of these curves on visual hull of the vehicle that is automatically generated. Images of vehicles have to have clean background and shots must be taken in regular intervals. They also presented two methods to obtain pose of a car which is necessary to the initial 3D curve matching.

Kraus *et al.* [16] lifted two state-of-the-art 2D object representation into 3D on the level of both a local feature appearance and a location. They used 3D CAD models for training geometry classifiers. They also used 3D patch sampling and rectification for improving results of methods based on 2D representation. They tested it on their BMW-10 dataset and also on existing datasets. In both cases it gained significant improvements over existing methods.

CNN-Based Methods

These methods are based on usage of neural networks to determine vehicle car make model etc. Psyllos *et al.* [38] used a relatively simple image processing technique for licence plate recognition, a vehicle mask and logo image detection. Vehicle mask was segmented based on congurency map gradient threshold. This can be seen in Figure 2.3 They used a probabilistic neural network as a classifier as it is suitable for a real-time application caused by speed of this method. They also used a hierarchical database of images containing smaller databases for each manufacturer. They tested this method on custom data and maintained accuracy around 85 %.

Yu *et al.* [59] presented a different scheme called Feature Fusion based Car Model Classification Net (FF-CMNET) based on the principle that frontal images of the car can be partitioned into uppers and lowers. These parts exhibit distinct feature distributions, but they are still correlated to allow a feature fusion. These sub-networks were called UpNet and DownNet. The fusion of features at the output of these nets was two-step in Fusion-Net.

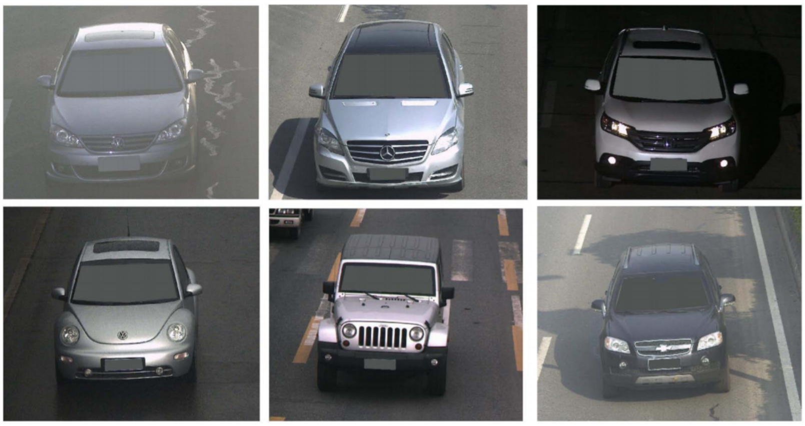

Figure 2.4: Examples of surveillance-nature data used in Compcars dataset which have large appearance variations caused by different lightning, weather and traffic conditions. Image taken from [58].

This method was tested on the CompCars dataset [58] and it outperformed state-of-the-art methods.Examples of images from CompCars dataset,which is mostlz used for evaluating methods can be seen in figure 2.4.

Biglari *et al.* [2] proposed the cascaded part-based system that used a latent support vector for automatically finding the discriminative parts. At the same time, it also learnt a part-based model for each category. A new training procedure, a novel greedy parts localization and a multi-class data mining algorithms were also employed. Experiments were made on their own data and also on CompCars where it achieved average accuracy of 95.55%.

Rachmadi *et al.* [39] used a different method to modify CNN to get a better performance. They removed the last layer of existing CNN classifier and attached the hierarchical spatial pyramid pooling. This pooling consisted of 2 independent pyramids that pooled features from two different layers of CNN. In their experiment they modified Alex-Net and ResNet50 and used Boxcars dataset. This method showed overall better accuracy.

Tian *et al.* [53] used an iterative discrimination CNN based on selective multi-convolutional region feature extraction. Features, which can be divided into global or local features, were extracted from the original image with a higher activation response value. Both global and local features were used to localize pivotal features. Pivotal features were concatenated into a fully-connected fusion layer to predict the vehicle category. This method gained 96.2% accuracy on CompCars and a little less on Stanford Cars-196 datasets.

Zhang *et al.* [62] used a lightweight CNN with combined learning strategy. Firstly, the lightweight CNN for only fine-grained vehicle recognition was designed. Secondly, a combined learning strategy including pre-training, fine-tuning training and transfer training was created. The main purpose of this was to optimize parameters of the lightweight CNN. The main advantage of their approach was to maintain accuracy of existing methods while reducing complexity of the network.

Ma *et al.* [30] proposed the channel max pooling scheme with a new layer inserted between fully connected layers and the convolutional layers. This scheme divided the feature maps into several sub-groups to reduce the number of parameters via reducing the number of channels in the CNN.

Experiments were created using Stanford Cars-196 and the Compcars dataset confirming that this approach gained a slight improvement to existing methods while significantly reducing the number of model parameters.

Fang *et al.* [9] dealt with recognizing problems by locating discriminative parts. Parts were selected by localizing the most significant appearance variation according to the large-scale training set. They also proposed a method based on feature maps taken from the CNN. In their method both global and local features were extracted from the original image. This method was evaluated on CompCars and the best accuracy was 98.29%.

Summary of Methods

The methods mentioned above usually have significant limitations. Images captured by traffic cameras are often small or they can have different viewpoints, so methods based on Front/rear parts of the vehicle cannot be used in many cases. Other methods based on 3D information of vehicles can be hard to collect for a large-scale surveillance system, especially for new models that will be released in the future. These new cars might not have a public 3D model or a number of models would be too enormous to be used. Our proposed method does not have such limitations, because it works on the 3D bounding boxes that can be created from video surveillance data [8, 7, 46].

2.1.3 Datasets for Fine-Grained Vehicle Recognition

There are some existing datasets (e.g. [41, 31]) that are used for vehicle detection and other tasks. Unfortunately, these datasets cannot be used for fine-grained vehicle recognition, because of a lack of annotation for the precise vehicle's make and model.

For CNN training and for future deployment of real-world traffic surveillance applications, it is best to find as many samples or classes as possible. That is why some datasets (e.g., [16, 20, 25]) that are used for the fine-grained recognition are not very practical for training.

Yang *et al.* [58] presented the large CompCars dataset containing images from web-nature and surveillance-nature. This dataset contains over 136k images capturing entire cars and over 27k images just car parts. These images are from 163 different car makes with 1716 car models. These images were taken from different viewpoints. The surveillance-nature data contained 50k images captured from the front view.

Sochor *et al.* [46] created another large dataset Boxcars116k containing images from various viewpoints all taken by numerous surveillance cameras placed on various locations around Brno, Czech Republic. See figure 2.5. In this dataset there were 116k images. Each vehicle was captured in multiple images to have more visual information.

Summary of Datasets

Most of datasets mentioned in this chapter has at least one limitation. They are not either as precisely annotated as it is needed to be or they do not have many different classes to be fully tested. Very popular is CompCars dataset, however, it does not contain information about 3D bounding boxes. BoxCars116k is the most suitable dataset for our proposed method, because it has information about bounding boxes, it is not limited to front/rear images and it is big enough to be used for testing surveillance systems.

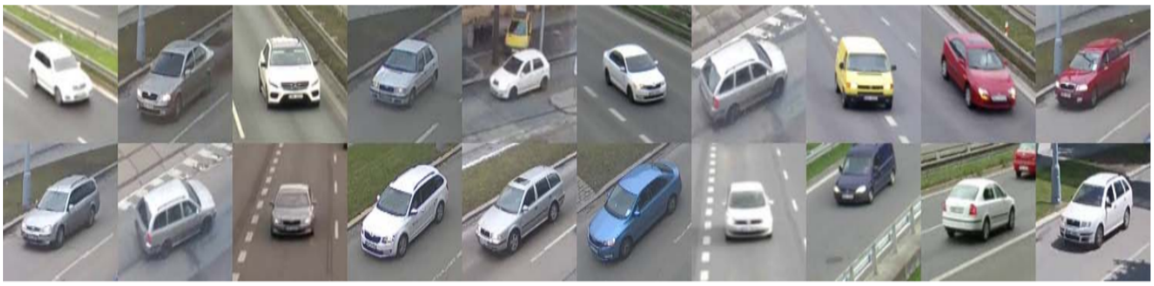

Figure 2.5: Examples of images used in boxcars116k dataset. Image taken from [46].

2.2 Object Re-identification

Object re-identification aims to identify a specific object across multiple camera views in different angles and places. This is a fundamental task for modern surveillance systems. Mostly it is used to determine two different groups. Many modern approaches deal with re-identification of people (chapter 2.2.2) and other approaches deal with vehicle re-identification (chapter 2.2.3). Prior mentioned two groups are usually necessary to track in surveillance systems. It is almost impossible for human to track the specific person or the car in huge surveillance systems.

2.2.1 General Object Re-identification

There are several approaches to object re-identification, which can be used for different objects. Ayedi *et al.* [1] created the multi-scale covariance image descriptor and also the quadtree based scheme. It was used to describe any object of interest such as a person, a group of people, vehicles etc. They evaluated their method only on person-reidentification using the VIPer dataset showing that multi-scale approach outperforms the already existing mono-scale.

Tahir *et al.* [52] proposed the approach that generates camera-invariant object matching scores. They were based on score variations in multiple camera pairs. This pair was presented with two parametric distribution models. They tested this method on person re-identification and gain improvement on two publicly available datasets.

Noor *et al.* [35] presented a novel aggregation scheme where subspaces were captured on the position of key points. Grassmannian distance metric was used to discriminate the aggregated information. They demonstrated that their scheme captured information complementary to the Fisher Vector aggregation. A significant improvement to the matching performance was shown in comparison to existing methods. Examples of different objects and their differences that were used to evalueate their method can be found in figure 2.6.

Špaňhel *et al.* [47] proposed a method based on aggregation of features in the temporal domain because there are usually multiple observations of the same object. Weighting different elements by different weights of the feature vectors was used for the aggregation. It was trained in an end-to-end manner using the Siamense network. It was tested on vehicle and person re-identification and it outperformed already existing methods for a feature extraction.

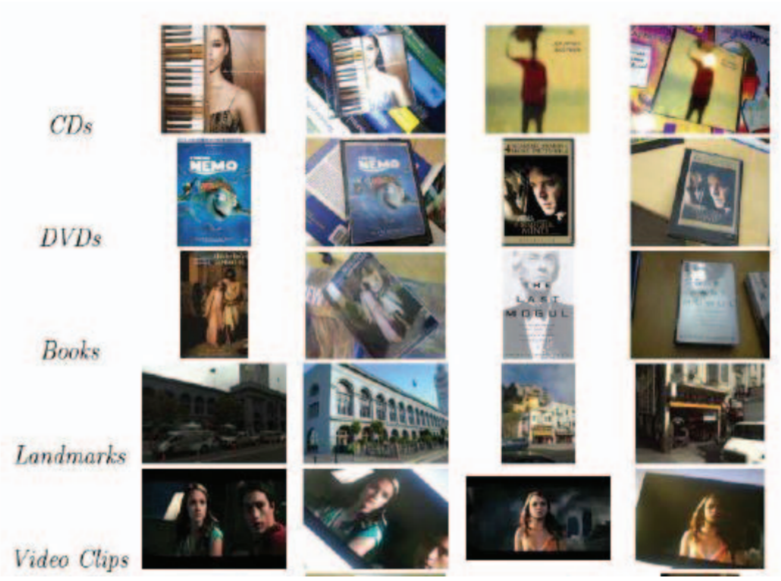

Figure 2.6: Examples of different shots of same object with various differences. These images are taken from various datasets which are used for object re-identification. Image taken from [35]

2.2.2 Person Re-identification

In recent years attention to person re-identification (Re-ID) has increased due to its usefulness in surveillance applications. This problem is not trivial, because of person pose changes, occlusion or illumination variations and low resolution of images produced by surveillance systems. Various Datasets are used to evaluate these methods. Some examples of datasets and images taken from them can be seen in figure 2.7 Approaches used for person Re-ID can be mostly divided into two groups. First group deals with this problem by extracting features. Zhao *et al.* [64] created a method based on unsupervised salience learning. They extracted distinctive features without requiring identity labels while training. They firstly applied adjacency constrained patch matching in order to build dense correspondence between image pairs. This effectively handles misalignment which is caused by pose and viewpoint variations. Secondly, they learnt human salience and incorporated it in patch matching. They tested their approach on the VIPer and the ETHZ dataset.

Li *et al.* [19] introduced the algorithm to hierarchically cluster image sequences. Representative data samples were used to learn a feature subspace so Fisher criterion was maximized. Both clustering and subspace learning process were iteratively applied to obtain discriminative features. Then a learning step was applied to bridge appearance difference between cameras. They evaluated this method on three datasets and outperformed existing methods. Lin *et al.* [21] proposed the consistent-aware deep learning approach. The aim of their work was to obtain maximal correct matches for a camera network. They exploited inter-camera and intra-camera cosistent-aware information in both training and testing phases. They also created algorithm based on a gradient descent to obtain globally optimal matching. The method was evaluated on multiple datasets showing significant improvements.

The other approach is by learning the distance using the projections of data items from different cameras. Ni *et al.* [34] created the relative distance metric leaning algorithm. It

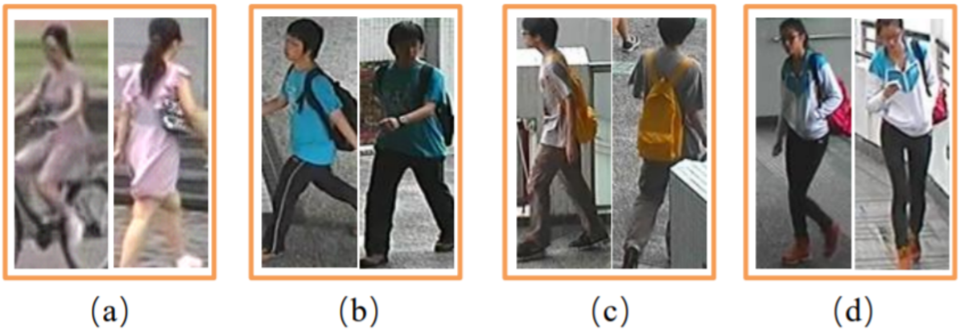

Figure 2.7: Examples of pedestrian images from different datasets used for person Re-ID (a), Market-1501 dataset (b), CUHK03 dataset (c), and CUHK01 dataset (d). Image taken from [22].

was based on projection vectors learning and clustering centralization. It improved performance and efficiency of methods while reducing a storage space and a computation complexity. A conjugate gradient method was also used in the projection vector learning. They evaluated their method on multiple datasets and shown improvement over existing methods. Ding *et al.* [5] presented the scalable distance drive feature learning framework. This framework was based on a deep neural network and was trained using a set of triplets. The also developed a triplet generation scheme making the computational load dependent on the number of original images instead of triplets. This helped with dealing with cubically growing number of triplets. Lin *et al.* [22] proposed usage of reinforcement learning based on recurrent neural network. The aim of this method was to simulate a local co-attention matching. For image pair, model selected an optimal sequence of co-attention regions. Experiments on three different datasets showed state-of-the-art performance mainly in robustness to occlusion cases.

Kansal *et al.* [14] proposed the Re-ID network called *Hdrnet*. This network consisted of an encoder and a multi-resolution decoder, which could learn embedding invariant to all problems mentioned above. They also proposed a new hybrid sampling strategy to boost effectiveness of the training loss function.

Navaneet *et al.* [33] dealt with Re-ID from a video. They proposed a change to the loss function, termed rank loss, to achieve a better performance and robustness of a system. Chen *et al.* [4] dealt with video-based Re-ID differently by using a joint attentive spatial-temporal feature aggregation network (JAFN). They introduced two different mechanisms to feature aggregation.

Wu *et al.* [54] addressed the view discrepancy by optimizing the cross-entropy view confusion by empowering the network with a similarity discriminator to promote highly discriminant features in positive and negative pairs.

2.2.3 Vehicle Re-identification

There are various ways to re-identificate a vehicle. One of the first approaches in a vehicle Re-ID was by licence plate recognition. However, this approach has huge disadvantages as the licence plate can be completely missing from the image or it can be unrecognizable due to car speed or low image quality.

Bulan *et al.* [3] presented a new automatic licence plate recognition workflow with novel methods for segmentation and annotation. They also improved plate localization and automation for a failure. They firstly localize a licence plate region in a two-stage approach. This was done by extracting set of candidates and then filtering them using a CNN classifier. Images that failed were classified to identify reason of failure.They evaluated it on custom data from US under realistic conditions.

Gou *et al.* [11] created a method based on character-specific extremal regions and hybrid discriminative restricted Boltzmann machines. The licence plate was detected by top-hat transformation and various validations. After a successful detection character-specific extremal regions were extracted and then the pattern classifier was applied to recognize characters. Their method is robust to illumination and weather changes.

Kluwak *et al.* [15] presented an approach for recognizing the licence plate from video. They focused on eliminating errors caused by poorly or covered visible licence plates. It is based on a combination of single frame licence plate recognition with object tracking methods. They tested it on five video sequences from surveillance cameras in different weather and lightning conditions and showed a significant improvement over existing methods.

Feature Representation

Other often used approach is by feature representation. Feature representation can be divided into two classes. The first class deals with hand-crafted representations. Herout *et al.* [60] presented a method based on a 3D bounding box from a video. Linear regressors were used to re-identify the vehicle, color histograms and histograms of oriented gradients. Features were used separately in order to get the best accuracy while maintaining as short computation time as possible. With this method they were able to fluently re-identify 12 vehicles per second.

Feris *et al.* [10] created the end-to-end system for vehicle retrieval based on attributes taken from video surveillance cameras. A feature pool containing many different feature descriptors was used. In their work they used time, date, color, direction and speed of the vehicle to re-identify. Their system was robust for crowded scenes and for environmental changes. The other method used a deep learning feature representation. They evaluated this method on real surveillance data.

Lie *et al.* [26] proposed the Deep Relative Distance Learning method. Two-branch deep CNN was exploited to project raw vehicle images into an Euclidean space. Distance could be there directly used to measure the similar arbitrary of two vehicles. They used CompCars and also a custom dataset called VehicleId to evaluate their method. Experiments showed that their approach outperformed sever state-of-the-art methods.

Shen *et al.* [42] created a two-stage framework that included complex spatio-temporal information. It was used for more effective regularizing results. A candidate visual-spatio-temporal path was generated from a pair of images with corresponding spatio-temporal information. The Chain Markov Random Fields model was used for generation. After that the Siamense-CNN+Path- LSTM model took that path to generate a similarity score. They used the VeRi-776 dataset and their method outperformed state-of-the-art methods.

Yan *et al.* [56] exploited multi-grain ranking constrains by two different methods. The first approach generalized the conventional pairwise ranking by considering binary relations to multiple relations. The second used permutation of a multi-grain list and it optimized the ranking using the likelihood loss function. Both of this approaches were implemented with multi-attribute classification in a multi-task deep learning framework. They created

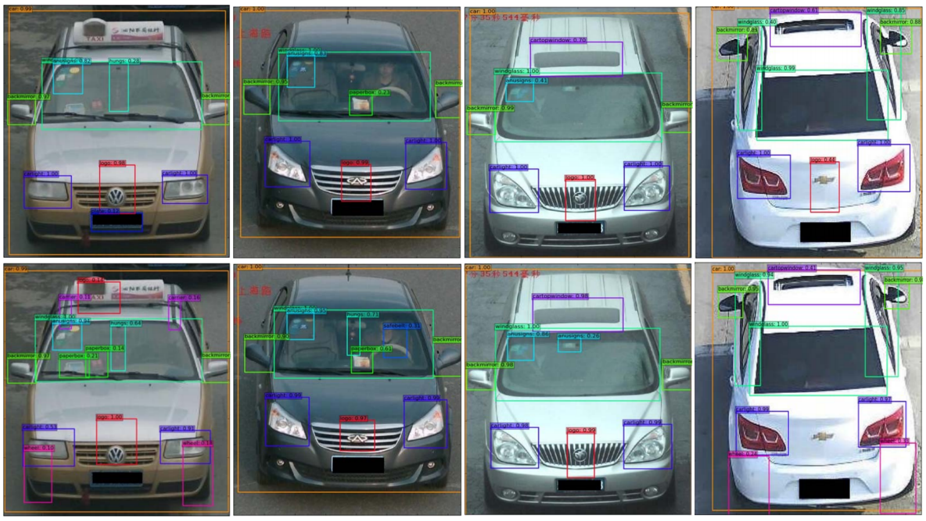

Figure 2.8: Example of Detecting features on VehicleId dataset. These are the results of single shot detection with focal loss and single shot detection with lower layer respectively. This image is taken and method proposed in [65].

datasets VD1 and VD2 to evaluate this method and for comparison with other methods the VehicleID dataset was used.

Zhang *et al.* [63] improved the triplet-wise training in two ways. They augmented a stronger constraint namely classification-oriented loss with the original triplet loss. Relative similarities between images in a tripled were ensured by the triplet stream. The classification stream is to strengthen the constraint from information of vehicle ID. They also created a new triplet sampling method. It was based on pairwise images and aimed to eliminate a misleading problem of triplet loss. It was tested on the VehicleID dataset showing improvements over existing methods.

Zhu *et al.* [68] designed quadruple directional deep learning networks to extract quadruple directional deep learning features of vehicle images. These networks had very similar overall architecture but they differed in a directional feature pooling level (e.g. horizontal average, vertical average, diagonal, etc.) to compress the basic feature maps into horizontal. These maps were then spatially normalized and concatenated together. VeRi and VehicleId datasets were used to show that these methods outperformed multiple state-of-the-art methods.

Other Methods

Zhao *et al.* [65] used the Single-shot detection as a baseline model for detecting attributes. They added more proposals from low-level to increase accuracy of a small object and also employed the focal loss improving the mean average precision. Example of all detected attributes for specific vehicle can be seen on figure 2.8.

Zhou *et al.* [66] combined CNN with LSTM to model transformations across different viewpoints of vehicles. They also presented in [67] the Viewpoint-aware Attentive Multiview Inference model. This model extracted the single-view feature for each input image aiming to transform these features into a global multiview so it could be better optimized.

Lou *et al.* [29] proposed the end-to-end embedding adversarial learning network capable of generating samples localized in the embedding space.

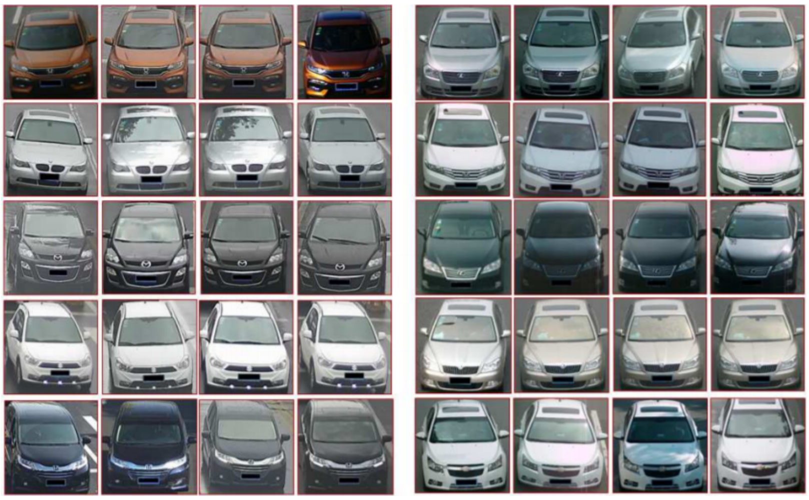

Figure 2.9: Examples of images in the PKU-VD dataset. Images on the left side are part of the VD1 dataset and on the right side are from VD2. Images of the same vehicle are shown in each row. Image taken from [56].

In their learning scheme, hard negative samples automatically generated in the specified embedding space could improve the capability of the network.

Guo *et al.* [12] created the Two-level attention network that was supervised by Multi-grain Ranking loss to learn an efficient feature embedding.

The attention network was created by a hard part-level attention and a softy pixel attention combination. This network could learn a feature space where both intra-class compactness and inter-class discrimination were well guaranteed.

Summary of Methods

As described there are plenty of methods to re-identify a vehicle. Approaches using vehicle's licence plates have huge limitations. The mostly used approach is by a feature representation or by modifying neural networks. In our method we propose a similar method as is used in vehicle recognition based on 3D bounding box and direction of a car. Both attributes can be detected without the necessity of different external hardware.

2.2.4 Datasets for Vehicle Re-identification

Liu *et al.* [27] published the large dataset VeRi-776 for vehicle re-identification. This dataset contained over 40k bounding boxes made from 619 vehicles that were captured by 20 cameras. Each vehicle was captured by 2 18 cameras. These cameras were in different viewpoints, illuminations and resolution. The dataset also included important vehicle information such as type, color or bounding box.

Yan *et al.* [56] created the *PKU-VD* dataset. This dataset consisted of 2 smaller datasets which were both well-annotated and high-quality named *VD1* and *VD2*. VD1 consists of images from high resolution traffic cameras while images for VD2 were obtained from surveillance videos. Examples of images are showed on figure 2.9. If both datasets are combined together there would be nearly 2 million images from over 2 thousand vehicle models. Due to privacy protection cars in this dataset did not have readable licence plates.

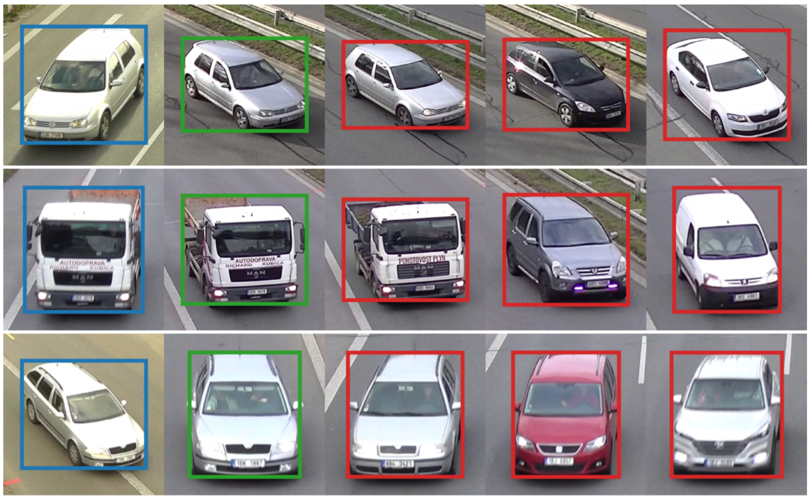

Figure 2.10: Examples of images in the CarsReId74k dataset. The left side anotated with a blue color is query, a green color is used in positive samples and a red color in negative. Negative samples are sorted from hard to easy (left to right). Image taken from [47].

Identity number, vehicle model and color and even made year were used to determine the identity of a car. To maintain consistency all images that belonged to the same vehicle ID were annotated with a same color and a vehicle model. To address color issues in different lightning situations 11 common colors were annotated.

Liu *et al.* [26] created the *VehicleID* dataset where images were taken from a frontal or rear viewpoint above the road. This dataset contained over 221k images of 26267 vehicles captured by non-overlapping surveillance cameras. That means, there were over 8 images of one vehicle in average. The dataset was split into two similar-sized parts for model training and testing.

Spaňhel *et al.* [47] created a new *CarsReId74k* dataset containing almost 74k images of a vehicle track from 66 cameras. Examples can be found on figure 2.10. These tracks were precisely annotated using licence plates and were also manually verified to eliminate errors. This dataset contained information about 3D bounding box created using method which was described by [8]. The dataset was not suitable for methods based on licence plates, because it contained images from non zoomed-in cameras where licence plates are almost unreadable.

Summary of Datasets

There are various datasets, which can be used for vehicle re-identification as described above. In VehicleID and PKU-VDs datasets images are taken from frontal/rear viewpoints. As our proposed method is not limited to viewpoints, these are not suitable. The Veri-776 dataset is too small to be used for testing surveillance systems. CarsReId74k does not have these limitations as there are enough different vehicles and does not limit to front/rear viewpoints. It also has information about bounding boxes which are utilized in our proposed method.

Chapter 3

Proposed Methodology

In this chapter, a change made to vehicle recognition and re-identification is presented. The method is very similar for both of these challenging tasks. The only difference is implementation in detection of the vehicle direction in the picture. The goal of this part is to fully inform about changes that were made in both fine-grained vehicle recognition and re-identification that increased accuracy of both methods.

3.1 3D Bounding Boxes

This method, that will be used both for classification and re-identification was based on the work proposed by Sochor et al. [46] utilizing 3D bounding boxes. These boxes can be obtained by traffic surveillance cameras. 2D bounding box, vehicle contour, directions to vanishing points and different versions of 3D bouding boxes are showed on figure 3.1. These are used for estimation of 3D bouding boxes that was metntion in work [46]. 3D bounding boxes allow us to determine individual parts of the vehicle, such as roof, side and rear (or front) viewpoint. In this method we distinguished between the front and rear part of the vehicle by determining direction of the car.

3.2 Direction of a Vehicle

Vanishing points can be used to determine the direction of a car. As mentioned in [46] the exact location of a vanishing point is not necessary while estimating 3D bounding boxes. The same rule applies to direction determination of the vehicle, because we calculate whether the line from center of the bounding box to the vanishing point intersect a plane that is created from points used in 3D bounding box for creation of the front part. If there is an intersection the car towards the camera that means we can see front of the vehicle. If there is no intersection, we can see rear side of the vehicle.

For fine-grained vehicle recognition, we did not have to calculate directions of the vehicle using this method thanks to data provided in *BoxCars116k* dataset 4.1.1. This dataset contains information about the camera and simultaneously vehicle positioning which explains whether vehicles in photos taken by this camera face towards camera or not. However, some of the cameras that were used for obtaining of this data are positioned in a big angle, therefore accuracy of this value was tested. After few testing images, this value and bounding boxes were corrected, so we used this value for determining the direction of a vehicle.

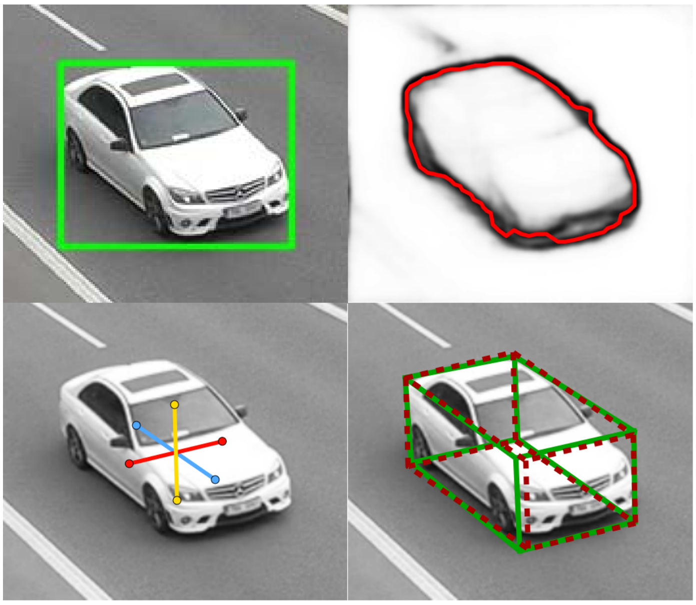

Figure 3.1: Example of automatically obtained 3D bounding box used for fine-grained vehicle classification. Top left: vehicle with 2D bounding box annotation, top right: estimated contour, bottom left: estimated directions to vanishing points, bottom right: 3D bounding box automatically obtained from surveillance video (green) and estimated 3D bounding box Image taken from [46]

Direction of vehicle for re-identification is implemented a bit differently. CarsReId74k dataset does not contain any information about direction of the cars. However, Photos from this dataset can be divided into front-facing and rear-facing cameras. Each image has information for which camera was it taken from. Then we can determine if we can see fron or rear of the car only by that value.

Image Normalization

The normalization was done using unpacking the image into a plane. The unpacking is done by retrieving homography between points and then the parts of the image are perspectivly wrapped. The plane consist of rectified versions of front (**F**), rear (**R**), top (**T**) and side (**S**). This rectified versions are then formed into matrix.This Matrix can be seen in equation 3.1.

$$U = \begin{pmatrix} 0 & T & 0 \\ F & S & R \end{pmatrix} \tag{3.1}$$

Difference is made in this matrix in determining front or rear part of the vehicle by direction. As there can be only front or rear part of the vehicle, when is missing, it is filled with 0 as are corners of this matrix.

The top corners of this submatrix are filled with zeros. The next part is determined by direction of the vehicle, as there can by only front or rear part of the vehicle. If vehicle is facing towards the camera (front mask is showed in the image), nothing is changed from previous work (Sochor *et al.* [46]).

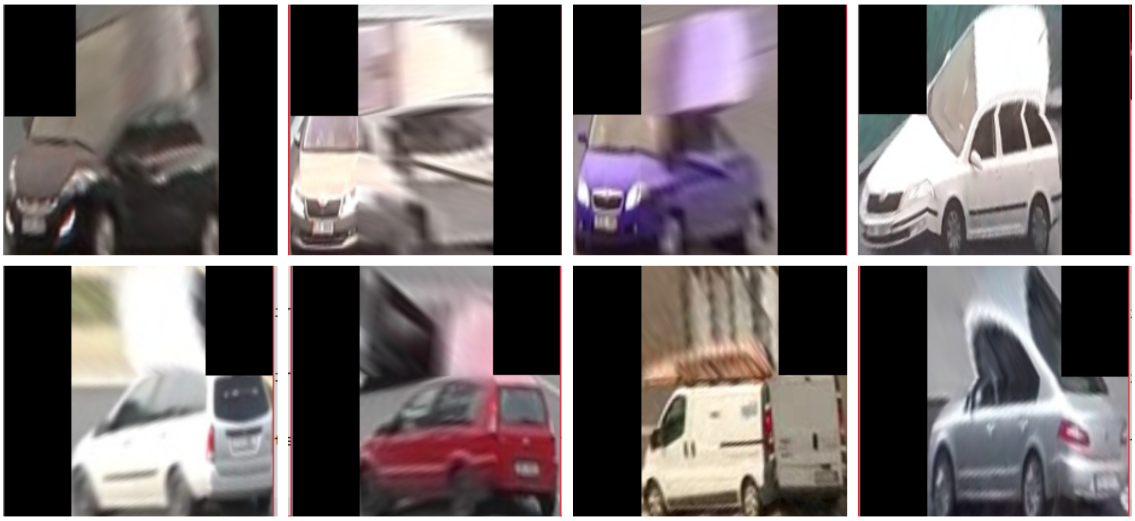

Figure 3.2: Examples of images that are used as an input to the neural network **Top** Vehicles face towards the camera **Bot** Direction of vehicles are against the camera. Vehicles on these photos are vertically flipped to be as much corresponding as possible with images facing towards the camera.

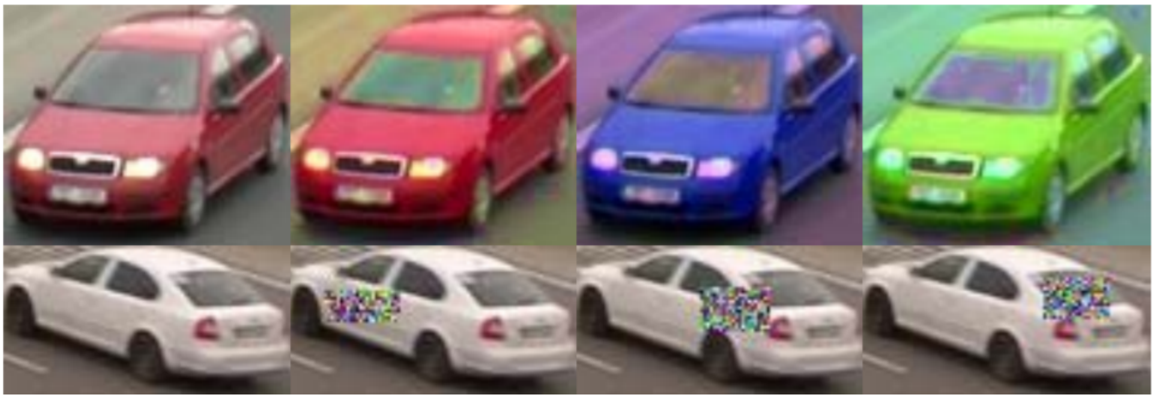

Figure 3.3: Examples of data augmentation techniques used while training data. On the left side is the original image and other images are examples of different versions of augmentation. Top of the are images modified by **Color** and on bottom **ImageDrop**. (Taken from [46])

However, if vehicle is not facing towards the camera, each part of the vehicle (i.e., top, side and rear) is vertically flipped and filled on corresponding parts on the matrix. Using this, same part of the vehicle will be on the exact place in both cases. This can be seen in figure 3.2.

This is the version of the vehicle, which is used as an input to the neural network. The biggest advantages of normalization to the 2D is that there are localized parts of the vehicle. Other advantage is that position of the vehicle is normalized and all of this can be done without the usage of other algorithms or Deformable Parts Model (DPM).

3.3 Additional Training Data Augmentation

To create bigger diversity of the training data additional data augmentation was proposed.

The first one called **ImageDrop** deals with possibility that some parts of the car would not be available all the time. This is inspired by work of Zeiler and Fergus [61]. in this work they evaluated what infualce will it cause, if a part of the input image will be covered on the probability of the ground truth class. In our modification, we take a random rectange in the image and fill it with random noise. This will effectibely drop any important information that was in this part of the image.

The second one called **Color** deals with the color inaccuracy caused be different lightning situation or color settings of each camera this causes that in fine-grained vehicle recognition color is irrelevant so to increase color variability between shots of cars in training samples, random alternation of the color of the images was proposed. The alternation is done in the HSV color space by adding random values to each pixel of the image. The value is the same for one picture and each HSV channel is proccessed separately.

Examples of both used augmentations can be found on Fig. 3.3.

Chapter 4

Experiments

4.1 Vehicle Recognition

4.1.1 Dataset

To train and validate presented method, Boxcars116k dataset was used. As mentioned in section 3, presented method is not limited to front/rear images. This method is based on 3D bounding boxes which are also included in this dataset.

Images in this dataset were taken from surveillance cameras. The aim of this dataset is to help to test methods aimed at traffic surveillance applications. Images appearing in this dataset can be from front or rear side. In some occasions it is hard to automatically detect directions of the car.

Statistics

There are over 116k different images captured by 137 different cameras with a large variation of viewpoints. In this pictures there are over 27k vehicles of 45 different makes. That means there are 693 fine-grained classes. Eeach class differ in make & model & submodel & year the car was manufactured. For each image datataset includes important information necessary for this method. Each photo has information about which camera took the photo. From that we can get values of car directions and it also has information about 3D bounding box that is necessary for unpacking vehicle.

Training & Test Splits

BoxCars116k dataset was split into a training and a testing part by randomly selected cameras for training. From these cameras all tracks are used while training and the rest are used for testing. Because of this approach the classification algorithms are tested on images of vehicles which are from previously unseen cameras. This also means that some vehicles that appeared in a training phase are tested from a different viewpoint. Testing part is divided into two splits. The first one is called **medium** and in this case it is necessary to recognize a precise type and also all vehicles of the same subtype are present in the same class. The other is called **hard** and in this case nothing is omitted so the precise type must be recognized including the model year. The hard split contains 107 fine-grained classes with over 11k tracks. That means more than 51k images are used for training. For testing there are nearly 40k images of 11k tracks. More detailed results can be found on table 4.1 where are showed detailed statistics for each split and also overall statistics of this dataset.

	hard	medium
classes	107	79
train+val cameras	81	81
test cameras	56	56
training tracks	11 653	12 084
training samples	51 691	54 653
validation tracks	637	611
validation samples	2 763	2 802
test tracks	11 125	11 456
test samples	39 159	40 842

tracks	27 496
samples	116 286
cameras	137
make	45
make model	341
make model submodel	421
make model submodel year	693

Table 4.1: Detailed statistic of the Boxcars dataset. **Left:** Detailed statistics of training split. **Right:** Overall statistics of the Boxcars dataset.

4.1.2 Experiments

We tested our proposed methodology on the Boxcars116k dataset. Proposed changes were evaluated and performance using different nets were compared. Classification accuracy was compared with existing work to determine usefulness of proposed change. This method was tested following neural networks:

- ResNet50 [13]
- VGG16, VGG19 [44]
- InceptionV3 [51]

Images taken from BoxCars116k dataset were resized to 224 x 224 to be used as an input to these nets.

NVIDIA RTX 2070 was used for training and testing all networks. We used the same batch size, the same initial learning rate and learning rate decay and the same hyperparameters for every net. Initial learning rate was set to 0.01, weight decay was $5 \cdot 10^{-4}$.

Standard data augmentation as horizontal flip and randomly moving bounding box, which were presented in [44] were also used while testing.

The accuracy improvement was evaluated for each net in the classification error reduction and also in percentage points. Sochor in their methodology [46], which was also tested on Boxcars116k dataset, showed that the results were almost identical for the medium and the hard split. That is why our method was mainly tested on the hard split. As shown in table 4.3, this method showed promising results as it has nearly 85% accuracy across 3 tested CNN. The only exception is InceptionV3.

There were experiments of usage only parts of the cars while unpacking vehicle. There were attempts on only front, side, top/rear parts. and also combination front/rear part with side and top with side. This was tested only on ResNet50 CNN and as showed on the second part of table 4.3. Using only parts did not improve accuracy of any method. Parts were re-sized to fill the whole matrix, with exception of front/rear where there was each time only half of the matrix filled. As data shows using only front or rear part of the vehicle gain worst accuracy. This is probably caused by the nature of data in given dataset and more importantly, car which was trained on front part can have rear part in testing phase. This supports the fact that combination of front with side was more accurate than usage of top with side.

Net	Error Reduction	
	Single Sample	Track Acc
ResNet50	13,25/2,1%	13,03/4,53%
VGG16	8,51/6,73%	5,41/7,07%
VGG19	8,39/11,47%	7.54/17,45%
InceptionV3	— /10,44%	— /14,04%

Table 4.2: This table shows error reduction between proposed method and results taken from [46]/[45]. Our modification outperformed an original approach in each net. Error was reduced up to 13,25 % on a single sample using ResNet50. Track accuracy error was reduced up to 17.45 %, that was gained on VGG19.

net	Accuracy[%]		time[ms]
	Single sample acc	Track acc	processing time
Resnet50	84.62	91.99	3.7
VGG16	85.17	92.65	3.4
VGG19	85.26	92.76	3.8
Inceptionv3	83.44	91.37	2.6
Front/rear	54.83	61.64	3.7
Side	80.55	89.73	3.8
Top	58.13	72.15	3.9
Front/rear+side	83.82	91.23	4.0
top+side	80.80	89.59	3.7
ResNet50 Sochor	82.27/84.29	90.79/91.61	— /5.8
VGG16 Sochor	83.79/84.10	92.23/92.09	— /5.4
VGG19 Sochor	83.91/83.35	92.17/91.23	— /5.4
InceptionV3 Sochor	— / 81.51	— /89.96	— /6.1

Table 4.3: Summary of accuracy of modifying images depending on the direction of vehicle. It was tested on hard split BoxCars116k dataset with image modification using 3D bounding boxes that are included in dataset. Top part of this table shows result of presented method where input was whole vehicle. Middle part shows result where input was only part of the vehicle. These results were gained on ResNet50. Last part informs about results taken from [46]/[45]

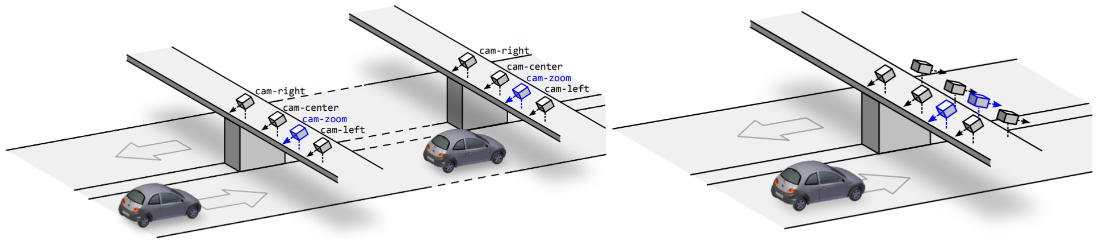

Figure 4.1: Recording setup for aquision CarsReId74k dataset.Data was recorede on two bridges using multiple cameras as is showed in the picture on left side. Some of the images are taken from one bridge as shows image on the right side. There was one camera zoomed in to detect licence plate. That was used for constructiong the ground truth labeling. Image taken from [47])

Comparison With Other Methods

The proposed method is based on the method presented by Sochor *et al.* [46]. They presented Boxcars116k dataset in this work. Results in the last part of table 4.3 are as results taken from paper [46] / results taken from GitHub [45]. These are same methods but differ in implementation. Values are taken from hard split with image modifiers on. Unfortunately, there were no results for IncpetionV3 in their paper so values can be taken only from Github. As is showed in this table, our modification of the method outperformed theirs in every net. Error calculation between this proposed method and results taken from their paper/website can be found in 4.2. It gained up to 13,25 % error reduction on single sample using ResNet50 and up to 17.45 % reduction on track accuracy on VGG19.

4.2 Vehicle Re-identification

4.2.1 Dataset

For training and testing CarsReId74k dataset was used. There are more datasets that can be used as mantioned in section 2.2.4, but others failed to succed in condition needed for this method.Firstly, different identities (different licence plates) must be presented. Secondly, there should not be limited to front/rear images and finally, it should be big enough to be tested for surveillance systems. This dataset was presented by Špaňhel *et al.* [47].

Photos were taken by 4 cameras on bridge above highway and four cameras on another bridge as showed on 4.1. One camera was zoomed in to read the licence plate of vehicles. Licence plates from zoomed video were detected using ACF detector [6] and vehicle were recognized by [69]. Vehicles were manually verified to eliminate any errors.

Statistics

Dataset is divided into 3 parts (i.e., training, testing, validation). There are over 3 million images of about 17k unique vehicles. More detailed statistics can be found in table 4.4. These were taken from 66 different cameras. The dataset contains almost 74k images of vehicle tracks. All of them have precise identity annotation which was acquired using license plates.

	training	test	validation	total
cameras	30	30	6	66
unique vehicles	7,658	9,678	1,100	17,681
tracks	32,163	36,535	5,278	73,976
images	1,469,494	1,467,680	305,539	3,242,713
positive pairs	125,086	129,774	22,376	277,236
negative pairs	1,149	1,459	881	1283

Table 4.4: Statistics of CarsReId74k dataset. The sum of unique vehicles is higher than the total number of unique vehicles. This is caused by a small number of vehicles which appear in all sets.

For each vehicle, there were constructed 3D bounding boxes. Sochor [46] showed that they are beneficial for fine-grained recognition and in this method they are used for re-identification.Also for each vehicle, its specific line was assigned. In the end matching of vehicles from zoomed images to non-zoomed was made and all vehicles that did not match were omitted. Most of vehicles from non-zoomed-in cameras have unreadable licence plate, so this datasest is not souitable for methods based on licence-plates due to maintain anonymity of the vehicles.

4.2.2 Experiments

This proposed method was tested on the CarsReId74k dataset, as this dataset is not limited to front/rear viewpoints and there is information about 3D bounding box for each vehicle shot. This method was tested on ResNet50 [13] and Inception-ResNet-v2 [50]. Images are resized to 331 x 331. yielding feature vectors. Their length is 1536 for Inception-ResNet-v2 and 2048 for ResNet50. This is used for each input image. Same image modifiers are used as in 4.1. For evaluation proposed method on dataset, **mAP** and **hit at rank** were used as metrics. Others mostly report hit rates at rank 1, 5, 10 and 20.

The feature extractor was fine-tuned with Adam optimizer, learning rate 0.0001 batch size 16. Each network was learned for 50 epochs. Standard augmentation techniques (shifting of the bounding box and random flip) were also applied. There are several techniques that can be applied to feature aggregation in temporal domain such as standard average pooling of feature method, RNN [32], Neural Aggregation Network [57] etc. In this evaluation, Learning Features in Temporal Domain, which was presented by Špaňhel *et al.* [47] will be used as it outperformed in their work other aggregation methods.

For average pooling final features were 1536 dimentional, while fro LFTD it was 128. Training of LFTD method was done by Adam optimizer with learning rate 1e-4.4 and margin 2.

Results can be found in 4.5 and Cumulative Matching Curve is shown in figure 4.2, proposed method gained better results for Inception-ResNet-v2 than for ResNet50. Networks weere tested with LFTD_WE and avg aggregation methods. Unpacking with direction was used for all tests. Results where image modifiers were used are noted as -IM. In that table there are results for Incpetion-ResNet-v2 -original. These results were taken from [47]. However, they did not report results for ResNet50. Proposed Method gained Hit@20 93.0 % for LFTD_WE using Inception-ResNet-v2 for feature vector length 128. ResNet50 gained Hit@20 91.3 %. These results were gain with using image modifiers.

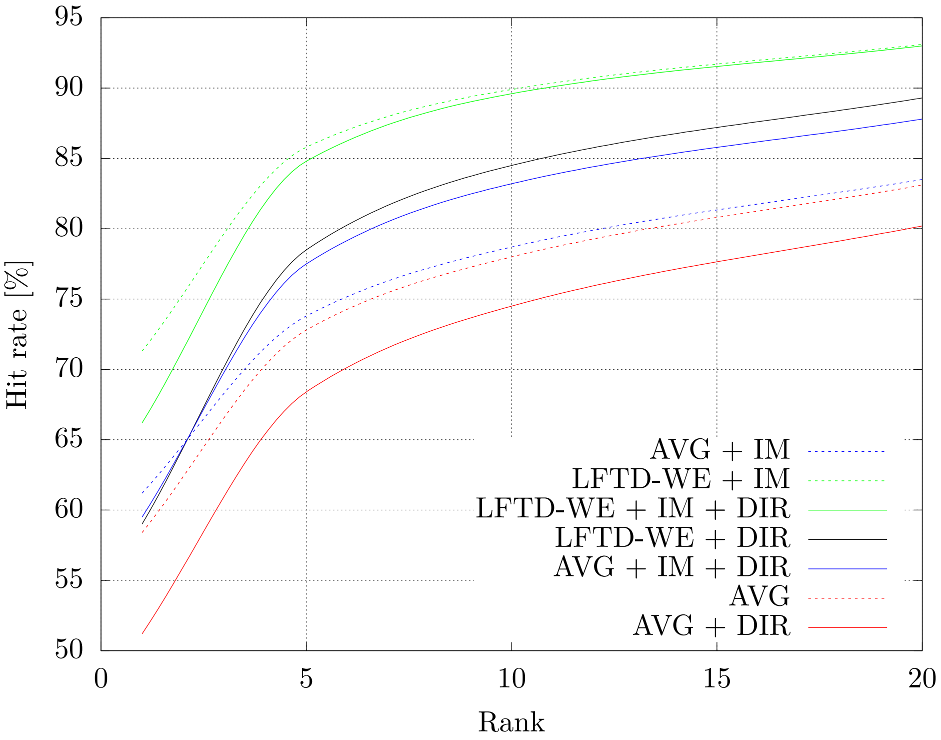

Figure 4.2: Cumulative matching curve for different methods. Results are taken using Incpetion-ResNet-v2 on CarsReId74k dataset. Proposed method is named DIR as direction. Reported results are with Image modifiers (IM) and without image modifiers. Results for comparation (dotted) were taken from [47]. For LFTD method, 128 dimentional feature vector was used.

4.2.3 Comparison with original method

The best way to compare presented method is with method created by Špaňhel *et al.* [47]. This method is based on theirs as it also use 3D bounding boxes with same feature aggregation method. The difference between these methods is in unpacking bounding box. In their approach, they do not differ between front/rear part of the vehicle. As showed in table 4.5 their method slightly outperformed ours for LFTD aggregation, however when using image modifiers and average aggregation, proposed method outperformed theirs. This might be caused mainly by two reasons. They trained feature extractor with smaller batch size (4) for higher number of epochs (300) and LFTD aggregation method choose which parts of the image will be used. Vehicles modified by proposed method has more empty spaces as there is different in positioning between front and rear part of vehicle.

			Hit@Rank			
Net	Aggregation	mAP	1	5	10	20
ResNet50 -IM	avg	0.618	53.6	71.0	77.2	82.9
Inception-ResNet-v2 -IM	avg	0.678	59.5	77.5	83.2	87.8
ResNet50	avg	0.558	47.1	65.5	72.2	78.5
Inception-ResNet-v2	avg	0.593	51.2	68.4	74.5	80.2
ResNet50 IM	LFTD_WE	0.728	64.9	82.3	87.4	91.3
Inception-ResNet-v2 IM -	LFTD_WE	0.745	66.2	84.8	89.6	93.0
ResNet50	LFTD_WE	0.694	61.0	79.2	84.8	89.4
Inception-ResNet-v2	LFTD_WE	0.679	59.0	78.5	84.5	89.3
Inception-ResNet-v2 -orginal	avg	0.652	58.4	72.8	78.0	83.1
Inception-ResNet-v2 -orginal	LFTD_WE					
Inception-ResNet-v2 -orginal -IM	avg	0.672	61.2	73.8	78.7	83.5
Inception-ResNet-v2 -orginal -IM	LFTD_WE	0.779	71.3	85.8	89.9	93.1

Table 4.5: Summary Results for vehicle re-identification. Results for net with original was taken from [47]. leanght of feature vector for LFTD_WE aggregation was 128. Results shows proposed method on two networks with Image Modifiers or without Image modifiers with avg aggregation or LFTD_WE aggregation.

Chapter 5

Conclusion

The aim of this thesis is to present a new method that can be used for both fine-grained vehicle recognition and vehicle re-identification. The proposed method is based on unpacking 3D bounding boxes into 2D pane where vehicles are normalized. The change to unpacking is presented using direction of the vehicle to determine if there is a front or rear part of the vehicle.

The proposed method outperformed the original method reducing error up to 14 % on single sample accuracy and up to 17,45 % on track accuracy. There is also showed that using only parts for vehicle recognition did not increase accuracy on given method. Using only front/rear part gained lowest accuracy. This might be caused by data nature of used datasets or that some vehicles were captured only by front/back while training as using side with front/rear gained higher accuracy than using top with side. Improvements for fine-grained vehicle recognition can be done by adding attention layer to neural networks or by appending car mask as an input to the neural networks.

For vehicle re-identification proposed method did not gain any improvements over existing method for LFTD aggregation. However, it improved for avg aggregation when image modifiers were used. This can be cause be lower number of training epochs and higher batch size. Finding optimal parameters could improve performance of this method. Another reason why it happened could be caused by smaller vehicle-sub-parts as there are more parts of the images filled with zeroes due to front/rear separation.

Bibliography

[1] Ayedi, W.; Snoussi, H.; Abid, M.: A fast multi-scale covariance descriptor for object re-identification. *Pattern Recognition Letters*. vol. 33, no. 14. October 2012: pp. 1902–1907.

[2] Biglari, M.; Soleimani, A.; Hassanpour, H.: A Cascaded Part-Based System for Fine-Grained Vehicle Classification. *Transactions on Intelligent Transportation Systems*. 2017.

[3] Bulan, O.; Kozitsky, V.; Ramesh, P.; et al.: Segmentationand annotation-free license plate recognition with deep localization and failure identification. *IEEE Transactions on Intelligent Transportation Systems*. vol. 18, no. 9. 2017: page 2351–2363.

[4] Chen, L.; Yang, H.; Gao, Z.: Joint Attentive Spatial-Temporal Feature Aggregation for Video-Based Person Re-Identification. *IEEE Access*. vol. 7. 2019: pp. 41230 – 41240.

[5] Ding, S.; Lin, L.; Wang, G.; et al.: Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*. vol. 48, no. 10. October 2015: pp. 2993–3003.

[6] Dollár, P.; Appel, R.; Belongie, S.; et al.: Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 36, no. 8. August 2014: pp. 1532 – 1545. doi:10.1109/TPAMI.2014.2300479.

[7] Dubská, M.; Herout, A.; Juránek, R.; et al.: Fully automatic roadside camera calibration for traffic surveillance. *IEEE Trans. Intell. Transp. Syst.* vol. 16, no. 3. June 2015: page 1162–1171.

[8] Dubská, M.; Sochor, J.; Herout, A.: Automatic camera calibration for traffic understanding. *BMVC*. 2014: pp. 1–12.

[9] Fang, J.; Zhou, Y.; Du, Y. Y. Y. .: Fine-grained vehicle model recognition using a coarse-to-fine convolutional neural network architecture. *Transactions on Intelligent Transportation Systems*. vol. 18, no. 7. July 2017: pp. 1782 – 1792.

[10] Feris, R. S.; Siddiquie, B.; Petterson, J.; et al.: Large-Scale Vehicle Detection, Indexing, and Search in Urban Surveillance Videos. *IEEE Trans. on Multimedia*. vol. 14(1). 2012: pp. 28–42.

[11] Gou, C.; Wang, K.; Yao, Y.; et al.: Vehicle license plate recognition based on extremal regions and restricted boltzmann machines. *IEEE Transactions on Intelligent Transportation Systems,*. vol. 17, no. 4. 2016: page 1096–1107.

[12] Guo, H.; Zhu, K.; Tang, M.; et al.: Two-Level Attention Network with Multi-Grain Ranking Loss for Vehicle Re-Identification. *IEEE Transactions on Image Processing*. 2019.

[13] He, K.; Zhang, X.; Ren, S.; et al.: Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*. 2015.

[14] Kansal, J.; Subramanayam, A. V.: Hdrnet: Person Re-Identification Using Hybrid Sampling in Deep Reconstruction Network. *IEEE Access*. vol. 7. 2019: pp. 40856 – 40865.

[15] Kluwak, K.; Segen, J.; Kulbacki, M.; et al.: ALPR - Extension to Traditional Plate Recognition Methods. *Intelligent Information and Database Systems*. vol. 9622. 2016: page 755–764.

[16] Krause, J.; Stark, M.; Deng, J.; et al.: 3D Object Representations for Fine-Grained Categorization. *IEEE International Conference on Computer Vision Workshops*. 2013.

[17] Lai, D.; Tiam, W.; Chen, L.: Improving classification with semi-supervised and fine-grained learning. *PATTERN RECOGNITION*. 2019.

[18] Lee, S.; Gwank, J.; Jeon, M.: Vehicle Model Recognition in Video. *International Journal of Signal Processing Image Processing & Pattern Recognition*. 2013.

[19] Li, Y.; Wu, Z.; Karanam, S.; et al.: Multi-shot human reidentification using adaptive fisher discriminant analysis. *Proc. Brit. Mach. Vis. Conf.*. 2015.

[20] Liao, L.; Hu, R.; Xiao, J.; et al.: Exploiting effects of parts in fine-grained categorization of vehicles. *Proc. Int. Conf. Image Process*. September 2015: pp. 745–749.

[21] Lin, J.; Ren, L.; Lu, J.; et al.: Consistent-aware deep learning for person re-identification in a camera network. *Proc. CVPR*. June 2017: pp. 3396–3405.

[22] Lin, L.; Luo, H.; Huang, R.; et al.: Recurrent models of visual coattention for person re-identification. *IEEE Access*. vol. 7. 2019: pp. 8865–8875.

[23] Lin, T.-Y.; RoyChowdhury, A.; S., M.: Bilinear CNN Models for Fine-grained Visual Recognition. *EEE Transactions on Pattern Analysis andMachine Intelligence*. 2018.

[24] Lin, Y.-L.; Morariu, V. I.; Hsu, W.; et al.: Jointly optimizing 3D model fitting and fine-grained classification. *ECCV*. 2014: pp. 1–15.

[25] Lin, Y.-L.; Morariu, V. I.; Hsu, W.; et al.: Jointly optimizing 3D model fitting and fine-grained classification. *Proc. ECCV*. 2014: pp. 1–15.

[26] Liu, H.; Tian, Y.; Yang, Y.; et al.: Deep relative distance learning: Tell the difference between similar vehicles. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. june 2016.

[27] Liu, X.; Liu, W.; Ma, H.; et al.: Large-scale vehicle re-identification in urban surveillance videos. *IEEE Int. Conf. Multimedia Expo*. July 2016: pp. 1–6.

[28] Llorce, D. F.; Colás, D.; Daya, I. G.; et al.: Vehicle model recognition using geometry and appearance of car emblems from rear view images. *International IEEE Conference on Intelligent Transportation Systems (ITSC)*. 2014.

[29] Lou, Y.; Bai, Y.; Liu, J.; et al.: Embedding Adversarial Learning for Vehicle Re-Identification. *IEEE Transactions on Image Processing*. 2019.

[30] Ma, Z.; Chang, D.; Xie, J.; et al.: Fine-Grained Vehicle Classification with Channel Max Pooling Modified CNNs. *Transactions on Vehicular Technology*. 2019.

[31] Matzen, K.; Snavely, N.: NYC3DCars: A dataset of 3D vehicles in geographic context. *Proc. Int. Conf. Comput. Vis.* 2013: page 761–768.

[32] McLaughlin, N.; Rincon, J. M.; Miller, P.: Recurrent convolutional network for video-based person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. doi:10.1109/CVPR.2016.148.

[33] Navaneet, K. L.; Todi, V.; Babu, R. V.; et al.: All for One: Frame-wise Rank Loss for Improving Video-based Person Re-identification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[34] Ni, T.; Ding, Z.; Chen, F.; et al.: Relative distance metric leaning based on clustering centralization and projection vectors learning for person re-identification. *IEEE Access*. vol. 6. 2018: pp. 11405–11411.

[35] Noor, D. F.; Li, Z.; Nagar, A.: Robust object re-identification with Grassmann subspace feature for key points aggregation. *IEEE International Conference on Image Processing (ICIP)*. September 2016.

[36] Pearce, G.; Pears, N.: Automatic make and model recognition from frontal images of cars. *IEEE International Conference on Advanced Video and Signal Based Surveillance*. 2011.

[37] Prokaj, J.; Medioni, G.: 3-D model based vehicle recognition. *Workshop on Applications of Computer Vision (WACV)*. 2009.

[38] Psyllos, A.; Anagnostopoulos, C. N.; Kayafas, E.: Vehicle model recognition from frontal view image measurements. *Comput. Standards Interfaces*. vol. 33, no. 2. Feb 2011: pp. 142–151.
Retrieved from:
https://www.sciencedirect.com/science/article/abs/pii/S0920548910000838

[39] R., F. R.; uchimura, K.; Koutaki, G.; et al.: Hierarchical Spatial Pyramid Pooling for Fine-Grained Vehicle Classification. *International Workshop on Big Data and Information Security (IWBIS)*. May 2018.

[40] Ramnath, K.; Sinha, S. N.; Szeliski, R.; et al.: Car make and model recognition using 3D curve alignment. *IEEE Winter Conference on Applications of Computer Vision*. 2014.

[41] Russakovsky, O.; et al.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* vol. 115, no. 3. December 2015: pp. 211–252.

[42] Shen, Y.; Xiao, T.; Li, H.; et al.: Learning Deep Neural Networks for Vehicle Re-ID with Visual-spatio-Temporal Path Proposals. *The IEEE International Conference on Computer Vision (ICCV)*. October 2017.

[43] Shi, W.; Gong, Y.; Tao, X.; et al.: Fine-Grained Image Classification Using Modified DCNNs Trained by Cascaded Softmax and Generalized Large-Margin Losses. *IEEE Transactions on Neural Networks and Learning Systems*. vol. 30, no. 3. July 2019: pp. 683 – 694.

[44] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2014.
Retrieved from: https://arxiv.org/abs/1409.1556

[45] Sochor, J.: BoxCars Fine-Grained Recognition of Vehicles. May 2019 (accessed May 10, 2019).
Retrieved from: https://github.com/JakubSochor/BoxCars

[46] Sochor, J.; Špaňhel, J.; Herout, A.: BoxCars: Improving Fine-Grained Recognition of Vehicles Using 3-D Bounding Boxes in Traffic Surveillance. *IEEE Transactions on Intelligent Transportation Systems*. vol. 20. January 2018: pp. 97–108.

[47] Spanhel, J.; Sochor, J.; Juránek, R.; et al.: Learning Feature Aggregation in Temporal Domain for Re-Identification. *CoRR*. vol. abs/1903.05244. 2019. 1903.05244.
Retrieved from: http://arxiv.org/abs/1903.05244

[48] Stark, M.; Krause, J.; Pepik, B.; et al.: Fine-Grained Categorization for 3D Scene Understanding. *British Machine Vision Conference*. 2012.

[49] Sumbul, G.; Cinbis, R. G.; Aksoy, S.: Multisource Region Attention Network for Fine-Grained Object Recognition in Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*. 2019.

[50] Szegedy, C.; Ioffe, S.; Vanhoucke, V.; et al.: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv:1602.07261*. 2017.

[51] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; et al.: Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567*. 2015.
Retrieved from: https://arxiv.org/abs/1512.00567

[52] Tahir, S. F.; Cavallaro, A.; Rinner, B.: Re-identification with multiple source-cameras. *IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*. April 2015.

[53] Tian, Y.; Zhang, W.; Znang, Q.; et al.: Selective Multi-Convolutional Region Feature Extraction based Iterative Discrimination CNN for Fine-Grained Vehicle Model Recognition. *International Conference on Pattern Recognition*. 2018.

[54] Wu, L.; Hong, R.; Wang, Y.; et al.: Cross-Entropy Adversarial View Adaptation for Person Re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*. 2019.

[55] Wu, L.; Wang, Y.; Li, X.; et al.: Deep Attention-Based Spatially Recursive Networks for Fine-Grained Visual Recognition. *IEEE TRANSACTIONS ON CYBERNETICS.* 2019.

[56] Yan, K.; Tian, Y.; Wang, Y.; et al.: Exploiting Multi-grain Ranking Constraints for Precisely Searching Visually-similar Vehicles. *The IEEE International Conference on Computer Vision (ICCV)*. october 2017.

[57] Yang, J.; Ren, P.; Zhang, D.; et al.: Neural Aggregation Network for Video Face Recognition. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*. 2017.

[58] Yang, L.; Luo, P.; Loy, C. C.; et al.: A large-scale car dataset for fine-grained categorization and verification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015: pp. 1–9.

[59] Yu, Y.; Jin, Q.; Chen, C. W.: FF-CMnet: A CNN-Based Model for Fine-Grained Classification of Car Models Based on Feature Fusion. *IEEE International Conference on Multimedia and Expo*. 2018.

[60] Zapletal, D.; Herout, A.: Vehicle Re-identification for Automatic Video Traffic Surveillance. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016: pp. 25–31.

[61] Zeiler, M. D.; Fergus, R.: Visualizing and understanding convolutional networks. *Proc. Eur. Conf. Comput. Vis.* 2014: pp. 818–833.

[62] Zhang, Q.; Zhuo, L.; Zhang, S.; et al.: Fine-grained Vehicle Recognition Using Lightweight Convolutional Neural Network with Combined Learning Strategy. *IEEE Fourth International Conference on Multimedia Big Data*. 2018.

[63] Zhang, Y.; Liu, D.; Zha, Z.-J.: Improving triplet-wise training of convolutional neural network for vehicle re-identification. *Multimedia and Expo (ICME), 2017 IEEE International Conference*. 2017: page 1386–1391.

[64] Zhao, R.; Ouyang, W.; Wang, X.: Unsupervised salience learning for person re-identification. *Proc. CVPR*. Jun 2013: pp. 3586–3593.

[65] Zhao, Y.; Shen, C.; Wang, H.; et al.: Structural Analysis of Attributes for Vehicle Re-Identification and Retrieval. *Transactions on Intelligent Transportation Systems*. 2019.

[66] Zhou, Y.; Liu, L.; Shao, L.: Vehicle Re-Identification by Deep Hidden Multi-View Inference. *IEEE Transactions on Image Processing*. vol. 27, no. 7. 2018: page 3275–3287.

[67] Zhou, Y.; Shao, L.: Viewpoint-aware attentive multi-view inference for vehicle re-identification. *" Computer Vision and Pattern Recognition*. 2018.

[68] Zhu, J.; Zeng, H.; Huang, J.; et al.: Vehicle Re-Identification Using Quadruple Directional Deep Learning Features. *Transactions on Intelligent Transportation Systems*. 2019.

[69] Špaňhel, J.; Sochor, J.; Juránek, R.; et al.: Holistic recognition of low quality license plates by CNN using track annotated data. *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2017. doi:10.1109/AVSS.2017.8078501.