1. Fajer FADHIL[1], 2. Mohammed ABDULGHANI[2], 3. Anmar SALIH[3], 4. Mohammed GHAZAL[4]

Mosul University (1,2), Northern Technical University (3,4)
ORCID: 1.0000-0002-4191-8024; 2. 0000-0003-0126-6778; 3. 0000-0002-9139-5954; 4. 0000-0003-3159-4983

doi:10.15199/48.2023.02.22

Traffic Surveillance: Vehicle Detection and Pose Estimation Based on Deep Learning

***Abstract**. Video-based traffic surveillance analysis is an important area of research with numerous applications in intelligent transportation systems. Due to camera positioning, background crowd, and vehicle orientation fluctuations, urban situations are more complex than highways. This paper provides a state-of-the-art technique for vehicle detection and orientation estimation based on the convolutional neural network CNN for detecting and determining the orientation of a vehicle from a given image to reduce traffic accidents. Different CNN model architectures have been examined to reach this approach's goal, which results in a small and fast model that is compatible with limited-resources hardware. A large-scale dataset of vehicles has been used to train the model. The dataset includes different types and views of cars; the taken images are high quality with diverse backgrounds and light conditions. To train the model, the dataset has been divided into five classes according to view: Front, Rear, Side, Front-side, and Rear-side, to fit the requirement of this work. The system achieves a high accuracy result.*

Streszczenie.** Analiza monitoringu ruchu oparta na wideo jest ważnym obszarem badań z licznymi zastosowaniami w inteligentnych systemach transportowych. Ze względu na ustawienie kamery, tłum w tle i wahania orientacji pojazdu sytuacje w mieście są bardziej złożone niż na autostradach. W artykule przedstawiono najnowocześniejszą technikę wykrywania pojazdów i szacowania orientacji w oparciu o konwolucyjną sieć neuronową CNN do wykrywania i określania orientacji pojazdu na podstawie danego obrazu w celu zmniejszenia liczby wypadków drogowych. Zbadano różne architektury modeli CNN, aby osiągnąć cel tego podejścia, co skutkuje małym i szybkim modelem, który jest kompatybilny ze sprzętem o ograniczonych zasobach. Do trenowania modelu wykorzystano wielkoskalowy zbiór danych pojazdów. Zbiór danych zawiera różne typy i widoki samochodów; wykonane zdjęcia są wysokiej jakości z różnym tłem i warunkami oświetleniowymi. Aby wytrenować model, zestaw danych został podzielony na pięć klas według widoku: przód, tył, bok, przód i tył, aby spełnić wymagania tej pracy. System osiąga wysoką dokładność wyniku.* (Nadzór drogowy: wykrywanie pojazdów i szacowanie pozycji w oparciu o głębokie uczenie się***)

Keywords: Vehicle, Convolutional Neural Network, Orientation Estimation, Surveillance system.
Słowa kluczowe: nadzór drogowy, sieć neuronowa, deep learning.

Introduction

Since surveillance cameras can be considered the biggest source of acquiring traffic flow information, cameras have recently been widely used for traffic monitoring systems [1]. Furthermore, the rapid development of computer vision, artificial intelligence, and camera technologies and progress in automatic video analysis and processing have increased the attention to video-based traffic surveillance applications [2]. The use of computer vision techniques made the transportation system more intelligent [3]. These techniques depend on the appearance of the vehicle in detection, which helps in incident detection. Although many researchers in this field have worked to improve traffic surveillance systems, there are many challenges practically facing the development of these systems. Road scale, traffic crowding, and lighting conditions represent the most challenges in designing surveillance systems. Furthermore, vehicles vary in type, pose, and size, limiting vehicle recognition [4]. One important issue to overcome some of these challenges is camera placement and how to save the lens from occlusion.

This paper includes two stages: the first is vehicle detection from a video stream, while the second stage is vehicle pose estimation. Video analysis is the first step of vehicle detection in several traffic monitoring systems. A conventional technique based on vehicle appearance has been used to extract the vehicle from the surrounding background scene. These techniques use visual information such as color, shape, and texture in images or videos; appearance-based algorithms detect stationary objects. Similar features use coded descriptions to describe the vehicle's visual appearance. Vehicle detection has made use of a variety of features, including local symmetry edge operators [5], Histogram of Oriented Gradient (HOG) [6], Scale Invariant Feature Transformation (SIFT) [7], and Haar-like features [8]. In computer vision, the detection of foreground regions could be matched to different natural scene objects. For example, the scene could contain various types of vehicles, persons, and other moving objects such as animals, bikes, etc. As a result, the object of interest must be isolated, distinguished, and recognized (i.e., vehicle). Recently, the Deep Learning DL technique caused conventional computer vision techniques to become obsolete. DL development is responsible for significantly increasing the ability to recognize objects [9]. This study introduces a pre-trained model for vehicle pose classification from their viewpoints using multiple car images. The purpose of this paper is to detect the vehicle on its viewpoint or orientation to determine whether the moving vehicle is reversing or accelerating toward you. Although the new cars come with advanced systems with multiple sensors or radars that have proven their reliability, This paper aims to demonstrate how important information may be taken from an image and combined with other equipment to provide considerably robustness and more reliable results. Additionally, when constructing an advanced CNN model, researchers focus on increasing the accuracy over other factors like speed and power consumption [10]. Our goal is to deliver an efficient CNN model in terms of the number of parameters and operations, speed, and accuracy.

The rest sections of this research are as follows: Section 2 discusses the related works, while Section 3 clarifies the implementation of the proposed vehicle detection and pose estimation system. The experimental work and results are given in Section 4. A comparison with other works on the CompCars dataset is in Section 5. Finally, the conclusions and future research direction are stated in Section 6

Related Works

In the context of this paper, we have focused on two important issues: object detection (e.g., Vehicle), and the orientation of this object. This section briefly outlines the papers available in the literature related to these two issues. There are many researches on vehicle detection techniques based on vehicle parts models [11,12,13] or orientations [14,15]. The multi-lane scene technique for vehicle detection has been proposed in [11]. To accommodate multi-view and partial observation, researchers employed a framework of probabilistic inference depending on models of car parts. The viewpoint maps have been constructed to estimate the car viewpoint depending on the layout of the road and the pattern of driving. The results demonstrated that part-based inference was effective in spotting obstructed cars. In [12], a learning And-Or model is used to estimate viewpoint and detect cars. The occlusion patterns and car-to-car context were categorized using the And-Or model into three levels: car pieces, single cars, and spatially aligned cars. The model parameters were jointly trained using Weak-Label Structural SVM. This technique outperformed other deformable part-based model methods significantly. The researchers in [13], proposed a recognition framework for detecting cars based on a single frame of the image captured by a surveillance camera. The framework refined the vehicle projective distortion using headlamps, license plates, and a part-based car detector as anchor points. They classified the car features using a neural network. The results demonstrated efficiency in vehicle detection and model identification. [14] suggests using a 3D vehicle model patch to localize a vehicle on a surveillance camera. The model has been applied as a kernel, and 3D vehicle geometry was used to track the kernels. During tracking, a kernel density estimator was employed to fit the 3D model. The test results indicate that vehicle localization and monitoring are effective. [15] proposed a vehicle detection system for intelligent traffic surveillance. The Haar-like features and AdaBoost have been utilized to build the classifier. While Gabor wavelet and Local Binary Pattern LBP were utilized to extract multi-scale and multi-orientation features. The test results showed that vehicle detection was highly accurate.

Hw proposwed approach of the system

System Overview

Our method's pipeline consists of two stages: vehicle key-point detection and poses estimation. First, it is required to find the Region of Interest (ROI) to select the vehicle from the 2D image and surround it with a bounding box. YOLO [16] is one of the best single-stage detectors used to find ROI and predict the coordinates of a bounding box. Single-stage detectors generate a lot of region proposals, which they must categorize as either containing objects or not. The VGG-19 [17] pre-trained CNN model is trained with multi-class to predict the visible key-point locations. This step could be used to predict the locations of visible key-points. However, the heatmaps may contain erroneous activations that correspond to invisible key-points. Thus, a second step refines the results using the sub-sampled version of the input image and the coarse key-point estimates. The hourglass network [18], which is commonly used to enhance heatmaps and eliminate artifacts resulting from hidden key-points, provides a framework for the refinement network. [19,20]. The coarse heatmaps predicted in the first step are improved using a two-stack hourglass network with skip connections. Moreover, refining the predicted key-points, the vehicle's pose is estimated using a parallel section consisting of two fully connected layers that categorize the poses into (5) classes.

The refinement network can use this multi-task learning to produce reliable predictions of visible key-points while minimizing the response of invisible key-points [21]. Figure 1 depicts the procedures of the suggested system stages. Most vehicles have similar landmarks with some small differences in details, such as colors, edges, lights, and mirrors. Therefore, the ROI could be determined using landmarks and thus locate the car pose. The orientation of a vehicle is classified into five categories: front, rear, side, front-side, and rear-side. Indeed, no utter boundary exists between two adjacent orientations. To accomplish this, the common points between adjacent orientations have been grouped, the probability of any orientation group is determined during inference, and the group with the greatest possibility is chosen.

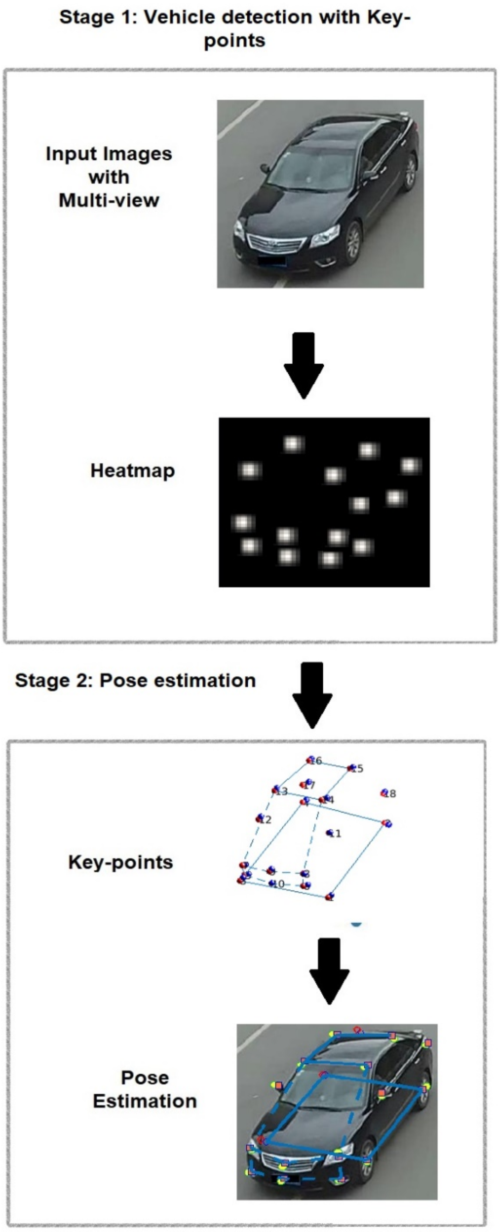
Stage 1: Vehicle detection with Key-points
Input Images with Multi-view
Heatmap
Stage 2: Pose estimation
Key-points
1
2
3
4
5
6
7
8
10
Pose Estimation

Fig.1. The stages of the proposed vehicle pose estimation model.

This is performed to demonstrate the significance of the surrounding regions around the key-points, which could include discriminatory information. Our CNN model has trained on the Comprehensive Cars (CompCars) dataset [22], which has about 208,826 images of vehicles. The images were obtained from two different cases: web-nature and surveillance-nature. CompCars dataset contains various vehicle perspectives, internal and external parts, and extensive information. Although the dataset contains

208,826 images, our model is trained with only 12000 images captured for the full car, divided into five classes, each class with 2400 images. All the images are high-resolution and were taken using a variety of backgrounds and lighting conditions.

The distance between the label and CNN's output is used to define the loss function during the training stage and expressed according to the following equation:

$$(1) \quad loss = \sum_{i=1}^{i} \parallel P_{label} - P_{output} \parallel$$

Where P_{label} and P_{output} denote the 2D coordinates of 12 key-points in the label and model output, respectively, and (i) represents the index of key-points. Notably, the CNN model learns each key-points of the vehicle in the same way. So that all the model outputs have a 2D Gaussian Distribution (GD). The probability of the key-point position is indicated by the value of GD. In general, hidden regions in output heatmaps are usually inaccurate, and varied viewpoints influence CNN model output. Therefore, in some situations, a single image could cause an incorrect pose estimate with some key-points that differ significantly from true values. In light of this phenomenon, multi-view images are combined to prevent depending on a single erroneous image.

Experimental results

In this part, we discussed the experimental result of the vehicle detection and pose estimation system based on a convolutional neural network. The system is built on an ASUS TUF computer with the following specifications: The eleventh generation of Intel microprocessor Core i7-11370H, RAM of 16GB, Graphics card type GeForce RTX 3070 with 8GB RAM, and SSD Hard disk of 1 TB, as well as the Ubuntu version 22 operating system and python programming language version 3.9. In spite of the high hardware specification and the latest version of software employed in the proposed system, we must note that the training model's execution time is related to the number of images in the dataset. In the experiment, we used a dataset containing 12000 images of various vehicles divided into five classes, and the YOLO CNN model was utilized for vehicle detection, while the Vgg19 model was used for pose estimation.

The transfer learning technique with fine-turning has been applied to train the adapted VGG19 model. Experimentally, we set the model parameters to fit the requirements of the research, by training the model with 100 epochs, and 32 batch size, using an RMSprop optimizer with a learning rate set to 0.001 and 0.5 dropout probability. The system was able to recognize multi-types of vehicles accurately under normal conditions. The system performance has been evaluated based on the benchmark evaluation metrics; the Accuracy value is 92.33 %, the Precision value is 93.3 %, and the Recall/Sensitivity value is 95.78 %. An example of the detection system is shown in Fig.2. Indeed, the error rate of the system is concentrated on the difficulty of distinguishing the sides of the vehicle because the original dataset is concentrated on car model classification. To solve this, the network will be able to distinguish different perspectives if the dataset format is revised to indicate whether each image is from the right or left side of the vehicle.

Comparison witj related works on Compcars datasert

In order to validate the robustness of our proposed model, we compared the result with many related works with the same dataset (CompCars). Indeed, most of the related work concentrates on vehicle detection only. The proposed model has a good performance in vehicle detection and poses estimation as shown in Table 1. Our research's results indicate the accuracy is 92.33 %.

Table 1.

Ref.	Method	Classification	Pose Estimation	Result-acc. %
Ref [23]	InceptionV3	Yes	No	85
	ResNet50			84
Ref [24]	Mask RCNN	Yes	Yes	79
Ref [25]	Faster R-CNN + ResNet	Yes	No	91.28
Our	YOLO + VGG19	Yes	Yes	92.33

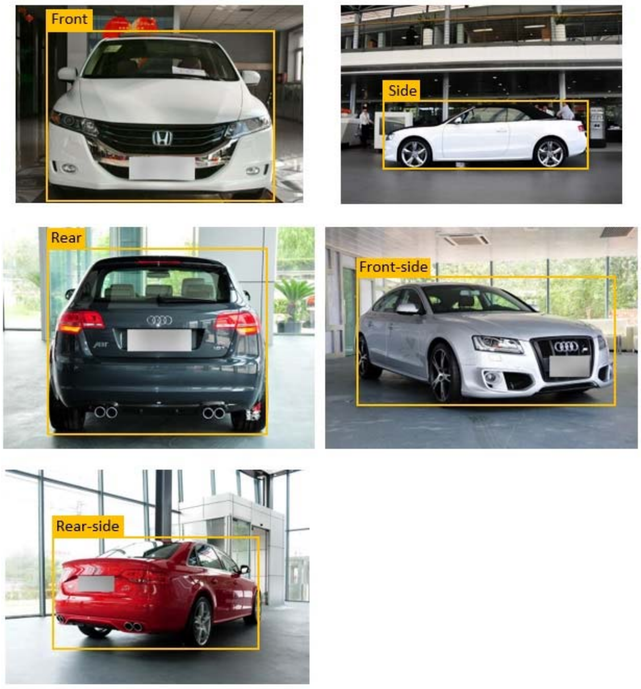

Fig.2. Experimental result of the proposed system.

Conclusion

In this research, we outlined a practical method for detecting and estimating the poses of vehicles as an application of a traffic surveillance system that contributes to reducing traffic accidents. This research methodology is based on convolutional neural network CNN pre-trained models, which enhance the performance of detection in comparison with related works. The system has been successfully tested using images of cars in our city in different places. We will enhance the system in the future by including more features like car tracking and recognition.

Authors: *Fajer Fadhil her M.Sc. degree from Computer Engineering Technology, Northern Technical University, Mosul, Iraq in 2012, E-mail: fajrfehr@uomosul.edu.iq; Mohammed Moath Abdulghani Obtained his M.Sc. degree from UKM, MALAYSIA in 2016, work in the college of agriculture and forestry at Mosul university, Email: albakri2@uomosul.edu.iq; Anmar B. Salih; obtained his P.hD. degree from Florida Institute of Technology. Florida, USA in 2020. Email: anmar.salih@ntu.edu.iq; Mohammed Talal Ghazal obtained his M.Sc. degree from Computer Engineering Technology, Northern Technical University, Mosul, Iraq in 2016, E-mail: mohammed.ghazal@ntu.edu.iq;*

REFERENCES

[1] X. Liu, W. Liu, T. Mei, and H. Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban

surveillance. IEEE Transactions on Multimedia, 20(3):645–658, 2018.

[2] S. Sivaraman, and M.M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 4, pp. 1773-1795, 2013.

[3] Szypuła, Ernest. Using deep learning to recognize the sign alphabet. Diss. Instytut Elektrotechniki Teoretycznej i Systemów Informacyjno-Pomiarowych, 2021.

[4] Kim, J. U., & Kang, H. B., "A new 3D object pose detection method using LIDAR shape set. Sensors," 18(3), 882, 2018.

[5] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 11, pp. 1475-1490, 2004.

[6] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," in proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, Jun. 2005, pp. 886-893.

[7] D.G. Lowe, "Object recognition from local scale-invariant features", in proc. IEEE 7th international conference on Computer vision, Jan. 1999, vol. 2, pp. 1150-1157.

[8] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features," in proc. of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, vol. 1, pp. I-511.

[9] N. J. T. Abdullah, M. Ghazal, and N. Waisi, "Pedestrian age estimation based on deep learning," vol. 22, no. 3, 2021.

[10] M. Ghazal, R. Albasrawi, N. Waisi, and M. Al Hammoshi,"Smart Meeting Attendance Checking Based on A multi-biometric Recognition System," PRZEGLĄD ELEKTROTECHNICZNY, vol. 2022, no. 3, pp. 93-96, 2022

[11] C. Wang, Y. Fang, H. Zhao, C. Guo, S. Mita, and H. Zha, "Probabilistic Inference for Occluded and Multiview On-road Vehicle Detection," IEEE Trans. Intell. Transp. Syst., vol. 17, no. 1, pp. 215–229, Jan. 2016.

[12] T. Wu, B. Li, and S. C. Zhu, "Learning And-Or Models to Represent Context and Occlusion for Car Detection and Viewpoint Estimation," IEEE Trans. Pattern Anal. Mach. Intell., vol. PP, no. 99, p. 1, 2015.

[13] H. He, Z. Shao, and J. Tan, "Recognition of Car Makes and Models From a Single Traffic-Camera Image," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 6, pp. 3182–3192, Dec. 2015.

[14] K. H. Lee, J. N. Hwang, and S. I. Chen, "Model-Based Vehicle Localization Based on 3-D Constrained Multiple-Kernel Tracking," IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 1, pp. 38–50, Jan. 2015.

[15] Y. Tang, C. Zhang, R. Gu, P. Li, and B. Yang, "Vehicle detection and recognition for intelligent traffic surveillance system," Multimed. Tools Appl., pp. 1–16, 2015.

[16] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[18] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In European Conference on Computer Vision, pages 483–499. Springer, 2016.

[19] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In European Conference on Computer Vision, pages 717–732. Springer, 2016.

[20] S. Tulsiani and J. Malik, "Viewpoints and Keypoints," in Proc. 2015 IEEE Conf. on Comput. Vision and Pattern Recognition (CVPR), Boston, USA, 2015, pp. 1510-1519.

[21] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle reidentification. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 379–387, 2017.

[22] Yang, L., Luo, P., Loy, C. C., and Tang, X. 2015. A largescale car dataset for fine-grained categorization and verification. In Proceedings of CVPR

[23] Kuhn, Daniel M., and Viviane P. Moreira. "BRCars: a Dataset for Fine-Grained Classification of Car Images," 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, 2021, p. 231-238.

[24] Ponimatkin, G., Labbé, Y., Russell, B., Aubry, M., & Sivic, J. "Focal Length and Object Pose Estimation via Render and Compare," 2022 In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 3825-3834.

[25] Sang, J.; Guo, P.; Xiang, Z.; Luo, H.; Chen, X. Vehicle detection based on faster-RCNN. J. Chongqing Univ. (Nat. Sci. Ed.) 2017, 40, 32–36.