Vehicle Classification and Tracking Based on Deep Learning

Hyochang Ahn and Yong-Hwan Lee*

Wonkwang University, Jeonbuk, Iksan City 54538, Republic of Korea
E-mail: youcu92@empas.com; hwany1458@empal.com
**Corresponding Author*

Received 27 April 2021; Accepted 24 February 2022;
Publication 18 April 2022

Abstract

Traffic volume is gradually increasing due to the development of technology and the concentration of people in cities. As the results, traffic congestion and traffic accidents are becoming social problems. Detecting and tracking a vehicle based on computer vision is a great helpful in providing important information such as identifying road traffic conditions and crime situations. However, vehicle detection and tracking using a camera is affected by environmental factors in which the camera is installed. In this paper, we thus propose a deep learning based on vehicle classification and tracking scheme to classify and track vehicles in a complex and diverse environment. Using YOLO model as deep learning model, it is possible to quickly and accurately perform robust vehicle tracking in various environments, compared to the traditional method.

Keywords: Vehicle classification, Moving object tracking, Deep learning, YOLO.

Journal of Web Engineering, Vol. 21_4, 1283–1294.
doi: 10.13052/jwe1540-9589.21412
© 2022 River Publishers

1 Introduction

With the recent diversification of data collection methods such as smartphones, cameras, and drones, the amount of multimedia data acquisition such as images and videos are increasing [1, 2]. Accordingly, interest in multimedia data has increased, and the need for image processing technology to utilize image data is increasing. Image processing technology is a method of recognizing image data through image data analysis, and this is used for data observation and abnormal situation detection [3, 4]. In addition, researches are being conducted on an automatic monitoring method that automatically derives meaningful information by applying image processing technology to an image obtained by observing an object using a camera [4].

Due to the development of technology and the concentration of population into the city, traffic volume is gradually increasing. Accordingly, traffic information is collected for continuous research and improvement, as social losses due to traffic congestion represent an astronomical amount [4, 5]. The collected traffic information is increasingly used and important enough to be used as basic data in a variety of ways, from navigation used in everyday life to various road traffic policies and recently emerged autonomous driving.

Vehicle detection and tracking using images are very helpful in providing important information, such as grasping road traffic conditions and crime situations [6]. Other autonomous vehicles and driver assistance systems also use technology to detect and track vehicles in images. Because it is used for accident prevention purposes, it is important to detect and track vehicles quickly and accurately. In addition, vehicle detection and tracking using a camera has a problem that is greatly affected by environmental factors in which the camera is installed. Therefore, in this paper, we propose a deep learning-based vehicle classification and tracking method to classify and track vehicles in a complex and diverse environment. Using the YOLO model as a deep learning model, it is possible to quickly and accurately perform robust vehicle classification and tracking in various environments.

The composition of this paper is discussed in Chapter 2 about the methods related to classification and tracking of vehicles. In Chapter 3, we will learn how to classify and track cars using deep learning. In Chapter 4, the results of the experiments that classified and tracked cars by deep learning were analysed, and finally, conclusion was made in Chapter 5.

2 Related Works

2.1 SIFT

SIFT stands for Scale-Invariant Feature Transform, and is an algorithm used to extract feature points of objects [7, 8]. SIFT is an algorithm that finds the point where the corners are maximized on the scale axis as well as the image using the Difference of Gaussian (DoG) to solve the problem of being vulnerable to changes in the image scale of the existing Harris corner algorithm [8].

SIFT generates reduced images in the form of an image pyramid through stepwise reduction of the size of the input image [9]. For each image of each scale, the corner component is inspected to find the corner point that is the largest. When each of the scale images detects a corner point, and when a corner point is detected across the scale of an adjacent image, the largest point of the corner property is searched along the scale axis to find a feature point that is invariant to the scale. By using the feature points that are invariant to the scale, the feature points can be found even when the input image scale changes.

2.2 Mean-SIFT

One of the common techniques used in many fields, including computer vision, such as object tracking and image segmentation, is an algorithm that tracks the region of interest (RoI) based on the density distribution of a data set, that is, feature points or colors [10, 11].

The operation method receives an image and designates a region of interest corresponding to an object to be tracked in the image [11]. After obtaining the color histogram from the image, the region of interest is moved using the method to find the location with the highest density in the region of interest and reset the location of the region of interest based on the location. Repeat these tasks to track movement.

Because of this, it is an algorithm that uses the Hill Climb search method as a disadvantage, so it is easy to fall into the local minimum depending on the setting of the region of interest. Also, because the region of interest is predefined, it is difficult to detect when the size of the object to be tracked changes. For this reason, it is not used alone in object tracking, but is often used in combination with other tracking methods, use in limited environments, and detectors.

2.3 HOG-SVM

HOG-SVM algorithm is an object detection algorithm that combines Histogram of Oriented Gradient (HOG) and Support Vector Machine (SVM) [12–14]. HOG is an algorithm that extracts features for input to the SVM, a learning-based classifier, and uses a local gradient of an image as a feature of a corresponding image. The HOG is a feature that divides the object area into cells of a certain size, calculates the amount and direction of the contours in the cell, generates it as a histogram for each cell, and then concatenates them into a one-dimensional vector [14].

SVM is one of the fields of machine learning, a supervised learning algorithm for pattern recognition, and a binary classifier that classifies two labels. SVM is a non-stochastic binary linear classification learning model that determines which label new data belongs to base on a given data set. The SVM determines the linear hyperplane with the largest margin and classifies the given data according to the linear hyperplane.

In HOG-SVM, when the HOG extracts the HOG features for the objects in the image and the HOG features for the non-objects and inputs them to the SVM, the SVM learns to classify the presence or absence of the objects in the image based on the two features.

3 Proposed Method

Object recognition algorithm based on R-CNN is composed of object area proposal and object recognition. YOLO is the first one-shot architecture proposed to further speed the object recognition algorithm [15–17]. Since the process of the one-shot architecture processes the image only once with one CNN, it is not necessary to propose a region and an object region, so it can recognize an object with a simpler and much faster speed than the existing R-CNN based algorithm. YOLO divides the input image into S $\times$ S grids, and each grid cell predicts B bounding boxes and confidence scores for those boxes.

YOLO's loss function is designed to calculate the class of the object to be detected, the size of the position and bounding box, and the existence of an object per grid. Equation (1) calculates the loss of coordinates (x, y) for the bounding box j predicted for the grid cell i where the object is present. Then, the size (*horizontal*, *vertical*) loss is calculated for the bounding box j

predicted for the grid cell *i* where the object is present

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}I_{ij}^{obj}[(x_i-\hat{x}_l)^2+(y_i-\hat{y}_l)^2]$$

$$+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}I_{ij}^{obj}\left[(\sqrt{\omega_i}-\sqrt{\hat{\omega}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2\right] \quad (1)$$

Equation (2) calculates the loss of the reliability score for the predicted bounding box *j* of the grid cell *i* where the object is present, where $C_i = 1$. Equation (3) calculates the loss of the confidence score for the bounding box *j* of the grid cell *i* where the object does not exist, where $C_i = 0$ in this equation. The equation calculates the loss of class probability for the grid cell *i* where the object is present.

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}I_{ij}^{noobj}(C_i-\hat{C}_l)^2 \quad (2)$$

$$\sum_{i=0}^{S^2}I_{ij}^{obj}\sum_{c\in classes}(p_i(c)-\widehat{p_i(c)})^2 \quad (3)$$

The size of the input image for training was changed from 448 × 448 to 416 × 416, because the final output through the convolutional neural network was determined as 13 × 13 divided by 32. The size of the final output was determined to be 13 × 13 so that the middle grid cell was centred. This is because when the object is in the center, the prediction method with one grid cell is more efficient than the prediction with 4 grid cells based on the center.

Direct location prediction could solve the problem that the model becomes unstable in the early stage of learning by using the anchor box. The main cause of learning instability is that the (*x*, *y*) position of the box is predicted too randomly at the beginning of learning. The Region proposal network predicts (t_x, t_y). That is, (t_x, t_y) is the output value of the convolution operation. The center of the bounding box (*x*, *y*) is calculated by Equation (4).

$$x=(t_x\times w_a)-x_a,\quad y=(t_y\times h_a)-y_a \quad (4)$$

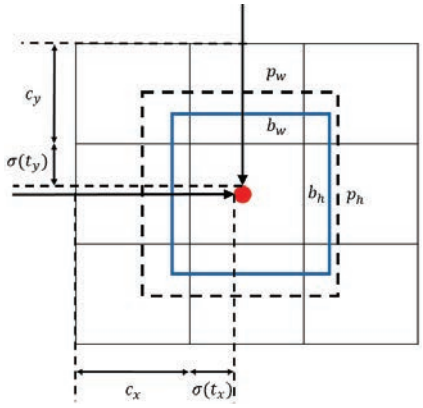

Figure 1 Prediction of bounding box coordinates.

The ground truth of (x, y) is always between 0 and 1 because YOLO is restricted to the position of the (x, y) coordinates only inside the grid cell. When predicting, the range of values was [0–1] by using σ (sigmoid function). The network's output feature map contains five pieces of information from the bounding box. (C_x, C_y) is the offset of the top left corner of each grid cell.

$$b_x = \sigma(t_x) + c_x, \quad b_y = \sigma(t_y) + c_y, \quad b_w = p_w e^{t_n}, \quad b_h = p_h e^{t_h}$$
$$P_r(object) \times IoU(b, object) = \sigma(t_0) \tag{5}$$

The 13 × 13 feature map is large enough to detect large objects, but may be insufficient for small objects. A simple skip-layer was used to solve this problem. Skip the 26 × 26 middle feature map and paste it onto the 13 × 13 layer. The 26 × 26 feature map contains high-resolution features compared to 13 × 13. In each grid cell, 5 boundary box candidates and class probability of each boundary box are independently proposed.

4 Experimental Results

The environment used in this paper was implemented using Windows 10 Education, 64-bit environment running Python. For experimental evaluation, we trained with the proposed model and set the vehicle in the image and track the classified vehicle.

A confusion matrix was used to evaluate the proposed method. Confusion ordinances can be expressed as shown in Table 1.

Table 1 Confusion matrix

		True Condition	
		Condition Positive	Condition Negative
Predicted Condition	Condition Positive	True Positive	False Positive
	Condition Negative	False Negative	True Negative

Table 2 Vehicle classification and comparison of detection performance

	Average Precision
K-NN	82.12
Faster R-CNN	88.10
YOLO	90.32

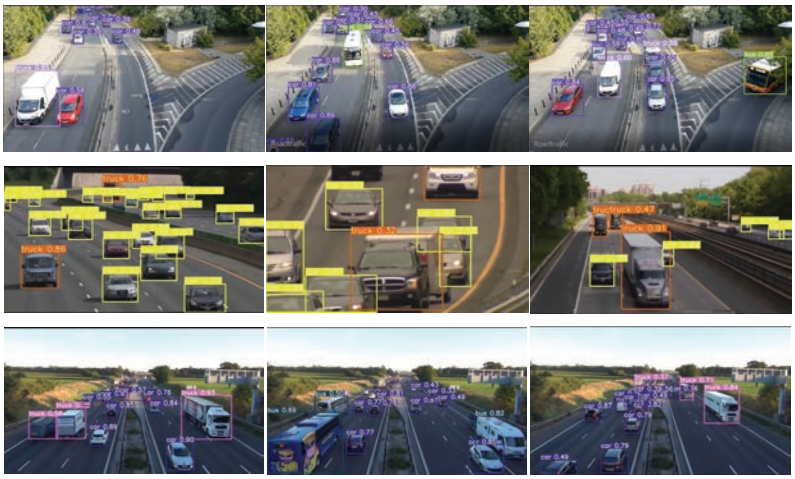

Figure 2 Results of Vehicle classification and detection.

In this paper, Average Precision (*AP*) is used to compensate for problems related to Precision (*P*) and Recall (*R*) values. Equation (6) represents the average precision.

$$\textit{Average Precision} = \sum_{n} (R_n - R_{n-1}) \times P_n \tag{6}$$

As a result of the experiment in this paper, the average precision of vehicle classification and detection showed 90.32%, and the performance of the deep learning-based vehicle classification and tracking technique using YOLO improved compared to other existing methods. Table 2 shows the experimental results compared to the conventional method. Figure 2 also shows the results of vehicle classification and tracking using YOLO.

5 Conclusion

Recently, interest in multimedia data has increased and the need for image processing technology to utilize image data is increasing. Researches are being conducted on an automatic monitoring method that automatically derives meaningful information by applying image processing technology to an image obtained by observing an object using a camera.

Vehicle detection and tracking using images are very helpful in providing important information, such as grasping road traffic conditions and crime situations. Because it is used for accident prevention purposes, it is important to detect and track vehicles quickly and accurately. In addition, vehicle detection and tracking using a camera has a problem that is greatly affected by environmental factors in which the camera is installed. Therefore, in this paper, we proposed a deep learning based vehicle classification and tracking method to classify and track vehicles in a complex and diverse environment. As a result of the experiment, the average precision of automobile classification showed 90.32%, so the proposed deep learning-based vehicle classification and tracking showed improved results compared to other existing methods.

Future tasks require research on traffic control systems, such as detecting vehicles and recognizing license plates to extract text.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2021R1A2C1012947).

References

[1] Tian, B., Morris, B. T., Tang M. et al., “Hierarchical and Networked Vehicle Surveillance in ITS: A Survey”, IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 557–580, 2014.

[2] Han, D., Cooper, D. B., and Hahn, H. S., “Bayesian Vehicle Class Recognition using 3-D Probe”, International Journal of Automotive Technology, vol. 14, no. 5, pp. 747–756, 2013.

[3] Ahn, H., and Lee, Y. H., “Performance Analysis of Object Recognition and Tracking for the Use of Surveillance System”, Journal of Ambient Intelligence and Humanized Computing, vol. 7, no. 5, pp. 673–679, 2016.

[4] Li, Q. L., and He, J. F., "Vehicles Detection based on Three-Frame-Difference Method and Cross-Entropy Threshold Method", Computer Engineering, vol. 37, no. 4, pp. 172–174, 2011.
[5] Munroe, D. T., and Madden, M. G., "Multi-Class and Single-Class Classification Approaches to Vehicle Model Recognition from Images", Proceedings of the 16th Irish Conference on Artificial Intelligence and Cognitive Science, pp. 1–11, 2005.
[6] Morris, B., and Trivedi, M., (2006, November). "Improved vehicle classification in long traffic video by cooperating tracker and classifier modules. In 2006 IEEE International Conference on Video and Signal Based Surveillance (pp. 9–11). IEEE.
[7] Prahara, A., "Car Detection based on Road Direction on Traffic Surveillance Image", International Conference on Science in Information Technology, pp. 344–349, 2016.
[8] Sakai, Y., Oda, T., Ikeda, M., and Barolli, L., "An Object Tracking System based on SIFT and SURF Feature Extraction Methods", International Conference on Network-Based Information Systems, pp. 561–565, 2015.
[9] Moranduzzo, T., and Melgani, F., "A SIFT-SVM Method for Detecting Cars in UAV Images", International Geoscience and Remote Sensing Symposium, pp. 6868–6871, 2012.
[10] Sotheeswaran, S., and Ramanan, A., "Front-View Car Detection using Vocabulary Voting and MEAN-SHIFT Search", International Conference on Advances in ICT for Emerging Regions, pp. 16–20, 2015.
[11] Lou, Z., Jiang, G., Jia, L., and Wu, C., "Monocular 3D Tracking of MEAN-SHIFT with Scale Adaptation based on Projective Geometry", International Conference on Multimedia Technology, pp. 1–4, 2010.
[12] Prahara, A., "Car Detection based on Road Direction on Traffic Surveillance Image", International Conference on Science in Information Technology, pp. 344–349, 2016.
[13] Guzman, S., Gomez, A., Diez, G., and Fernández, D. S., "Car Detection Methodology in Outdoor Environment based on Histogram of Oriented Gradient and Support Vector Machine, 2015.
[14] Bougharriou, S., Hamdaoui, F., and Mtibaa, A., "Linear SVM classifier based HOG Car Detection", International Conference on Sciences and Techniques of Automatic Control and Computer Engineering, pp. 241–245, 2017.
[15] Nie, Y., Sommella, P., O'Nils, M., Liguori, C., and Lundgren, J, "Automatic Detection of Melanoma with YOLO Deep Convolutional Neural

Networks", International Conference on e-Health and Bioengineering, 2019.
[16] Xu, Z., Shi, H., Li, N., Xiang, C., and Zhou, H., "Vehicle Detection Under UAV Based on Optimal Dense YOLO Method", International Conference on Systems and Informatics, pp. 407–411, 2018.
[17] Lou, L., Zhang, Q., Liu, C., Sheng, M., Zheng, Y., and Liu, X., "Vehicles Detection of Traffic Flow Video Using Deep Learning", International Conference on Data Driven Control and Learning Systems Conference, pp. 1012–1017, 2019.

Biographies

Hyochang Ahn received the M.S. degree and Ph.D. in Electronics and Computer Engineering from Dankook University, South Korea, in 2006 and 2012, respectively. He was a Research Professor at Dankook University, South Korea, from 2014 to 2016. Currently, he is working as research director in R&D at Innogru, Korea. His research interests include Image Processing, Computer Vision, Embedded system and Mobile Programming.

Yong-Hwan Lee received the MS degree in computer science and PhD in electronics and computer engineering from Dankook University, Korea, in 1995 and 2007, respectively. He is an active member of International Standard committees of ISO/IEC JTC1 SC29 responsible for Image Retrieval and Coding issues. Currently, he is a Professor at the Department of Digital Contents, Wonkwang University, Korea. His research areas include Image Retrieval, Image Coding, Computer Vision and Pattern Recognition, Augmented Reality, Mobile Programming and Multimedia Communication.