Research on Vehicle Tracking Algorithm Based on Deep Learning

Yan Zhou, Jun Zhou and Fangli Liao
Department of Vehicle Engineering, Faculty of Mechanical and Electrical Engineering, Wuhan Business University, Wuhan, Hubei, China
Email: yan974248850@163.com

Abstract. Comparing with the advantages and disadvantages of the existing target tracking algorithms based on deep learning, a vehicle tracking algorithm based on Yolov2 and GOTURN algorithm is proposed, which is called YOLOv2-tracker vehicle tracking algorithm. The Algorithm is trained and tested by using the collected training set and test set. The results show that the YOLOv2-tracker vehicle tracking algorithm can achieve higher tracking accuracy and faster tracking speed, and can effectively overcome environmental interference. Further analysis of the test results, the algorithm found that there is “error” phenomenon, the paper discusses and analyzes the causes of this phenomenon, and put forward a reasonable solution. In addition, a “dynamic save” method is proposed to solve the “lost track” problem.

1. Introduction
In recent years, key technologies such as Internet of things, deep learning, cloud computing and artificial intelligence based on big data and the Internet has developed rapidly, driving a new round of technological revolution [1]. In view of these challenges, the development of intelligent automobile promotes the development of China’s automobile industry from two technical levels: intellectualization and networking, thus ensuring that China's automobile industry in the new round of technological revolution in the rapid completion of industrial upgrading and transformation [2]. Tracking target vehicle is the precondition of predicting the track of target vehicle, so fast and accurate tracking algorithm is very important in intelligent vehicle technology. By analyzing the performance of deep learning in classification and detection tasks, we can see that deep learning has high potential in tracking tasks, for example, deep learning's GOTURN tracking algorithm, which can achieve tracking speed of 100fps, but the tracking accuracy of this algorithm is not high [3]. According to the shortcomings of the GOTURN tracking algorithm, the tracking algorithm with superior performance should have the following three characteristics: (1) has a strong feature extraction ability; (2) a reasonable method for generating candidate regions; (3) more data sets. Therefore, this paper proposes an improved algorithm based on the GOTURN tracking algorithm and YOLOv2 detection algorithm, which can quickly and accurately complete the tracking task of the target vehicle, which has certain practical application value.

2. YOLOv2-Tracker Tracking Algorithm
2.1. Algorithm Definition
In order to achieve a tracking algorithm with superior performance, this paper proposes three improvements based on the GOTURN tracking algorithm and YOLOv2 detection algorithm: (1) A
deeper network structure is adopted to enhance the ability of the tracking algorithm to extract features. 

(2) Use YOLOv2 detection algorithm to divide the grid candidate area method to replace the previously introduced candidate area generation method. Similar to the detection algorithm, this method can improve the quality of candidate region generation and improve the accuracy of the tracking algorithm. At the same time, the two-stage tracking algorithm can also be improved into a single-stage tracking algorithm to increase the speed of the tracking algorithm. (3) Because this article mainly focuses on the tracking of target vehicles, many traffic road videos (330 videos in total) are collected and the target vehicles are marked, thereby obtaining enough training set for training the network [4]. Since the first and second points of the above improvements are based on the YOLOv2 detection algorithm, this algorithm is called the YOLOv2-tracker tracking algorithm in this article.

2.2. Algorithmic Processes

(1) Mark the tracking target vehicle in the first frame picture, and “blacken” the surrounding environment of the target vehicle, as shown in figure 1. The “blackening” process is similar to the GOTURN tracking algorithm's target object cropping method, and its main function is to mark the tracking target. But compared with the target object cropping method of GOTURN tracking algorithm, the “blackening” process has two main advantages: First, it retains the position information of the target object on the entire picture. Second, during the training and testing process, the size of the target vehicle in the picture will change, but the size of the entire picture can remain unchanged, so it is easy to input the network for feature extraction [5].

(2) Input the second frame picture and the “blackened” first frame picture into two parallel convolutional neural networks to extract the picture features.

(3) The extracted features are input into the same YOLOv2 output layer together, and the network output layer predicts the location of the target vehicle in the second frame of the picture.

(4) According to the prediction result of the network, “blacken” the second frame picture, and input the two parallel networks with the third frame picture respectively. And so on, to complete the tracking of the target vehicle. The algorithm flow chart is shown in figure 2.

3. YOLOv2-Tracker Tracking Algorithm Testing Process

3.1. Tracking Data Set

In this paper, 330 sets of urban traffic scene videos were collected by the driving recorder, of which 280 sets were used as the training set and the remaining data were used as the test set. Each set of data is a traffic scene video composed of several frames of pictures. Select a target vehicle from each video, and manually mark the position of the target vehicle in each frame.

Figure 1. “Blackening” processing.
3.2. Evaluation Method
In this paper, the coincidence rate is used to evaluate the accuracy of the tracking algorithm. Due to the small change in vehicle shape during the tracking process, the spatial robustness evaluation method is used [6]. In the data set collected in this paper, in order to verify the performance of the tracking algorithm, after the target vehicle in some videos leaves the video screen, the video is still not over. At this time, the tracking algorithm with superior performance should not mark the “target” in the subsequent video pictures. Therefore, in order to evaluate the accuracy of the algorithm, the method of evaluating the accuracy of the coincidence rate needs to be appropriately modified: that is, when the target leaves the video screen, if the algorithm does not predict the “target”, it means that the current frame is tracked successfully, otherwise it is not.

3.3. Network Framework
This paper trains two kinds of architectures: double network and single network. The dual network framework is shown in table 1. The single network framework is shown in table 2. It is worth noting that because of the single network frame two images are superimposed into the network. Therefore, the input channel of the single network framework is 6.

3.4. Web Training
The loss function of the network uses the loss function of the YOLOv2 detection algorithm. This article trains two architectures, dual network and single network. The difference is that the tracking algorithm in this paper only completes the tracking of the target vehicle (single category). Therefore, there is only one category neuron.

In the loss function, $\lambda_{\text{obj}} = 5$, $\lambda_{\text{noobj}} = 1$, $\lambda_{\text{coor}} = 1$ and $\lambda_{\text{class}} = 1$. The number of anchor boxes per grid is 5. The initial learning rate is 0.0001, and the learning rate is set to 0.00001 when the entire training set traverses the 100th round (Epoch). The whole training set is trained 160 times in total. The momentum is 0.9 and the weight decay value is 0.0005. Except for the output layer, all convolutional layers use batch normalization [7].

The data collected in this article contains a total of 330 sets of videos. Select 280 groups of videos as the training set for training the network. Single and double network training loss curve is shown in figure 3.
**Table 1.** Dual network framework.

| Input: 608x608x3 | Input: 608x608x3 |
|------------------|-----------------|
| Conv3-32         | Conv3-32        |
| Maxpool          | Maxpool         |
| Conv3-64         | Conv3-64        |
| Maxpool          | Maxpool         |
| Conv3-128        | Conv3-128       |
| Conv1-64         | Conv1-64        |
| Conv3-128        | Conv3-128       |
| Maxpool          | Maxpool         |
| Conv3-256        | Conv3-256       |
| Conv1-256        | Conv1-256       |
| Conv3-512        | Conv3-512       |
| Conv1-256        | Conv1-256       |
| Conv3-512        | Conv3-512       |
| Conv1-256        | Conv1-256       |
| Conv3-512        | Conv3-512       |
| Maxpool          | Maxpool         |
| Conv3-1024       | Conv3-1024      |
| Conv1-512        | Conv1-512       |
| Conv3-1024       | Conv3-1024      |
| Conv1-512        | Conv1-512       |
| Conv3-1024       | Conv3-1024      |
| Conv3-1024       | Conv3-1024      |
| Route-16         | Route-16        |
| Conv1-64         | Conv1-64        |
| Reorg-2          | Reorg-2         |
| Route-27/24      | Route-27/24     |
| Conv3-1024       | Conv3-1024      |
| Conv1-30         | Conv1-30        |

**Table 2.** Single network framework.

| Input: 608x608x6 |
|------------------|
| Conv3-32         |
| Maxpool          |
| Conv3-64         |
| Maxpool          |
| Conv3-128        |
| Conv1-64         |
| Conv3-128        |
| Maxpool          |
| Conv3-256        |
| Conv1-128        |
| Conv3-256        |
| Maxpool          |
| Conv3-512        |
| Conv1-256        |
| Conv3-512        |
| Conv1-256        |
| Conv3-512        |
| Conv1-256        |
| Conv3-1024       |
| Conv1-512        |
| Conv3-1024       |
| Conv1-512        |
| Conv3-1024       |
| Conv3-1024       |
| Conv1-512        |
| Conv3-1024       |
| Conv1-512        |
| Conv3-1024       |
| Conv1-512        |
| Conv3-1024       |
| Conv1-512        |
| Conv3-1024       |
| Conv1-512        |
| Conv3-1024       |
| Conv1-512        |
| Conv3-1024       |
| Route-16         |
| Conv1-64         |
| Reorg-2          |
| Route-27/24      |
| Conv3-1024       |
| Conv1-30         |

![Figure 3](image-url)  
**Figure 3.** Single and double network training loss curve.
3.5. Test Results

Using 50 sets of videos as the test set, the single and dual networks after training are tested, and the performance of the tracking algorithm is evaluated according to the evaluation method. The main evaluation indicators include: (1) Accuracy of the coincidence rate of the merge ratio threshold of 0.5 (denoted as OP0.5); (2) Spatial robustness, that is, the initial frame label is randomly shifted by 10% (denoted as shift10%) and randomly scaled by 10% (Recorded as scale10%); (3) Tracking speed (fps) [8]. An example of test results is shown in figure 4. The test results are shown in table 3.

![Figure 4. Single and dual networks sample training test results.](image)

**Table 3. The tracking accuracy and robustness test results.**

|            | OP0.5 | Shift10% | Scale10% | fps |
|------------|-------|----------|----------|-----|
| Single network | 0.863 | 0.856    | 0.860    | 13  |
| Double network  | 0.867 | 0.861    | 0.862    | 6   |

It can be found from table 3 that for single and dual network frameworks, the tracking accuracy and robustness of the tracking algorithm are not much different. However, the tracking speed of a single network is faster than that of a dual network. Therefore, this article uses a single network framework as the final network framework.

By analyzing the test set of this article, it is found that 11 of the 50 videos have the following characteristics: when the target vehicle leaves the video, the video is still not over. Therefore, this article further divides the test set into two categories: “Leave” and “Not Leave” [9]. The test results are shown in table 4.

**Table 4. Tracking accuracy of the tracking algorithm test results.**

|            | OP0.5- “Leave” | OP0.5-“Not Leave” | OP0.5 |
|------------|----------------|-------------------|-------|
| Single network | 0.820          | 0.875            | 0.863 |
| Double network  | 0.825          | 0.878            | 0.867 |

It can be found from table 4 that the tracking accuracy of the tracking algorithm in the “away” video group is significantly different from that in the “not leaving” video group. In the next section, an optimization method will be proposed to improve the tracking accuracy of the algorithm.

3.6. Optimization for the Phenomenon of “Mistake”

During the test, it was found that the tracking algorithm in this paper has a “misfollowing” phenomenon [10]: when the target vehicle leaves the video, the algorithm will still predict the target at the target vehicle’s departure position.

The analysis found that the main reason for this phenomenon is that the position of the target vehicle changes continuously in the video. Due to the use of “blackening” processing, the position
information of the target object is retained, resulting in the network learning both the feature information and the position information of the target vehicle during the training process. When the network is sensitive to location information, even if the target vehicle leaves the video, the network will still predict the vehicle in the vicinity of the target vehicle leaving.

In response to the above problems, this paper uses enhanced feature information to solve the algorithm defects [11]. Strengthen the feature information, that is, strengthen the ability of the network to learn the characteristics of the target vehicle. In the training process, not only two consecutive frames of pictures are input to the network for training, but also one frame of pictures is randomly selected from the two video groups and input into the network for training together, so that the network does not “predict” the target vehicle. That is, the prediction value of all confidence neurons in the network output layer is 0 as shown in figure 5.

![Figure 5. The process of enhanced feature information training.](image)

Through this method of increasing “negative samples”, the network can be biased towards learning the characteristic information of the target vehicle, thereby reducing the “mistakes” phenomenon [12]. In table 5 shows experimental results.

|                     | OP0.5- “Leave” | OP0.5-“Not Leave” | OP0.5 |
|---------------------|----------------|-------------------|-------|
| Single network-optimized | 0.867          | 0.872             | 0.871 |
| Single network- not optimized | 0.820          | 0.875             | 0.863 |

It can be found from table 5 that this method can greatly reduce the “mis-following” phenomenon of the algorithm and improve the tracking accuracy of the algorithm.

3.7. Data Enhancement
Since only 280 sets of data are used for training pictures, it is necessary to use data augmentation to expand the data set and improve the robustness of the algorithm. This article mainly uses four methods to enhance the data: (1) random changes in lighting, contrast, and saturation, as shown in figure 6; (2) random dithering, randomly adding dithering to the “blackened” pictures to improve the robustness of the algorithm, as shown in figure 7; (3) Random sorting, in the training process, the selected two consecutive frames of pictures, randomly change their order; (4) Because the difference between the two consecutive frames of pictures in the video stream is not big, in order to improve the robustness of the algorithm, the random interval Frame (n is less than 10) select two frames for training [13].

3.8. Target Recovery
During the test, the algorithm can “follow” the target object in a frame of pictures, that is, the algorithm does not predict the target vehicle position in a frame. Because the “blackening” of the current frame requires the algorithm to predict the position of the target vehicle in the current frame, when the algorithm “follows” the target, the algorithm will fail [14]. In order to enable the algorithm to re-track the target after “following” the target, the algorithm dynamically saves the latest frame of
the “blackening” processing mark when tracking the target vehicle. If it is detected that the algorithm “follows” the target vehicle, that is, the algorithm does not predict the target vehicle position in a certain frame. Then use the saved “blackening” processing mark map to replace the “blackening” map of the current frame.

Figure 6. Random interference.

Figure 7. Random jitter.

4. Conclusion
This article discusses the advantages and disadvantages of the existing target tracking algorithm based on deep learning, and proposes a vehicle tracking algorithm based on GOTURN and YOLOv2 algorithm YOLOv2-tracker vehicle tracking algorithm. Secondly, it analyzes the advantages of YOLOv2-tracker algorithm, such as the advantage of using “blackening” processing instead of “cutting” pictures. The tracking algorithm flow of YOLOv2-tracker is introduced in detail, and single and dual network frameworks are proposed, and the performance of the two networks is verified through experiments. The experimental results show that the single network framework guarantees both faster tracking speed and higher tracking speed. Based on the experimental test results of the single network framework, it is found that the algorithm has “mis-following” phenomenon, and it is proposed to use enhanced feature information to solve. In addition, the article uses 4 data enhancement methods to improve the robustness of the algorithm. At the same time, a way to dynamically save “blackened” pictures is proposed to complete target recovery and improve the tracking accuracy of the algorithm.

Acknowledgments
This paper is a 2016 “Jingchu Excellent Talents” collaborative education program project of Hubei Ordinary University (Automobile Service Engineering Major of Wuhan Business University).

References
[1] Wang Y, Li Q and Wang J 2014 Research on maneuvering model algorithm based on vehicle tracking IEEE Workshop on Electronics pp 189-195.
[2] Zeng F, Hu S and Xiao K 2018 Research on partial fingerprint recognition algorithm based on deep learning Neural Computing & Applications 202 100-108.
[3] You F and Liang Z 2018 Research on driver fatigue monitoring algorithm based on deep learning Wuhan Ligong Daxue Xuebao 417-421 + 429.
[4] Lin Y and Lin L 2018 Research on the algorithm of LAN performance evaluation based on deep learning Neural Computing and Applications 203 178-186.

[5] Xu H and Jiang C 2019 Research on context-aware group recommendation based on deep learning Neural Computing and Applications 153 189.

[6] Pei W, Zhu Y and Zuo X 2015 A multi-scale vehicle tracking algorithm based on structured output SVM 8th International Congress on Image and Signal Processing (CISP) (IEEE) pp 138-143.

[7] Gu O, Bi-Neng Z, Bing B, et al. 2018 Recent research advances and application of object tacking algorithm based on deep neural network Journal of Chinese Computer Systems 201 125.

[8] Jiasong Z, Ke S, Sen J, et al. 2018 Bidirectional long short-term memory network for vehicle behavior recognition Remote Sensing 10 887.

[9] Zhao X, Dawson D, Sarasua W A, et al. 2017 Automated traffic surveillance system with aerial camera arrays imagery: Macroscopic data collection with vehicle tracking Journal of Computing in Civil Engineering 31 11.

[10] Yang W, Zhang J and Wang H 2018 A vehicle real-time detection algorithm based on YOLOv2 framework Real-time Image & Video Processing 129-134.

[11] Fan Z, Weng S, Jiang J, et al. 2018 Particle filter object tracking algorithm based on sparse representation and nonlinear resampling Journal of Beijing Institute of Technology 27 103.

[12] Hannan M A, Hossain Lipu M S and Hussain A 2018 Neural network approach for estimating state of charge of lithium-ion battery using backtracking search algorithm IEEE Access 201 69.

[13] Rajesh N, Selvakumar A A L 2018 Association rules and deep learning for cryptographic algorithm in privacy preserving data mining Cluster Computing 20 18.

[14] Liu F, Xu F and Yang S 2017 A flood forecasting model based on deep learning algorithm via integrating stacked autoencoders with BP neural network IEEE Third International Conference on Multimedia Big Data (BigMM) (IEEE) pp 228-231.