Road Information Detection Method Based on Deep Learning

Shangsheng Zheng, Jiangzhou Zhang, Xiaobo Che, Yanqiang Li

Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong, 250014, China

*Corresponding author’s e-mail: liyq@sdas.org

Abstract: Autonomous driving technology has developed rapidly in recent years, and computer vision has shown significant role in automated vehicles. Researchers use the combination of computer vision and deep learning to improve the speed and accuracy of road element detection technology. This article focuses on the semantic segmentation and road detection elements, especially in lane line segmentation and signal light detection. It summarizes the visual inspection technology and a series of visual perception algorithms based on deep learning. Although semantic segmentation and target detection technologies in general scenarios are already very mature, they are not ideal when applied to autonomous driving environments, especially where segmentation and detection are performed simultaneously. Therefore, this paper builds a fusion network of lane line segmentation and signal light detection suitable for autonomous driving scenarios.

1. Introduction

Autonomous driving technology is an important development project of today's cutting-edge science and technology. It has a significant influence on social and economic development, national defence construction, and scientific and technological development. Driverless technology involves interdisciplinary content such as cognitive science, sensor technology, computer technology, artificial intelligence, and vehicle engineering. It not only includes the research of basic theoretical methods and breakthroughs in key technologies, but also a large number of engineering design and implementation issues. With the development of the technological revolution, autonomous driving technology has become one of the important development directions of the automobile industry and artificial intelligence industry.

Nowadays, although autonomous driving technology is developing rapidly, autonomous vehicles cannot completely avoid traffic accidents. A key technology to improve safety is environmental perception, which includes drivable road surface detection, lane line detection, curb detection, motor vehicle detection, road sign detection, traffic sign detection, traffic light detection. Such a complicated road condition detection puts high demands on the perception algorithm. Computer vision based on deep learning can obtain perception capabilities closer to humans. Therefore, road element detection based on deep learning is an important topic worthy of study, which has important strategic significance and application value. In addition, few people in the field of autonomous driving combine semantic segmentation and target detection in a single network. This paper proposes a parallel network for lane line segmentation and signal detection applied to autonomous driving scenarios.
2. Environment Perception

Autonomous driving technology is not a topic that has only recently emerged, and related technologies have been studied and proposed in the 1970s and 1980s. The concept of unmanned driving proposed at that time was mainly applied to specific scenarios, such as autonomous driving engineering vehicles that use electromagnetic induction for navigation and program control in the industry. In agricultural production, with the help of high-precision global positioning system (GPS), agricultural tractors are navigated to complete agricultural activities such as automatic farming, fertilization, and harvesting. After that, visual sensors were used to assist vehicle positioning, further improve stability, and added safety protection, obstacle avoidance and fault diagnosis systems. On the other hand, in military applications, the United States has invested in the development of battlefield robots. Many companies and universities have participated in this research work, mainly developing computer-controlled wheeled or tracked armorer vehicles. However, the above studies are mainly limited to a single use scenario. Since the 1980s, autonomous driving technologies applied to ordinary people's daily lives have been proposed one after another. Some documents mainly use machine vision technology to identify lane lines, traffic lights and other structured information, and use detection and tracking technology to complete the acquisition of unstructured information such as vehicles and pedestrians. Finally, use this information to analyse and judge the state of the driverless car itself, that is, to perceive the environment. In the 21st century, under the current technical conditions, a complete unmanned driving environment perception system was proposed [1], and even tried to solve the problem of target detection under night conditions.

With the development of technology, safety has begun to become the primary issue of autonomous driving. By detecting the traffic signal lights on the road, the driving of the vehicle can be guided, so that the vehicle can comply with the traffic rules. The segmentation of lane lines can make the vehicle not deviate from the original track during driving. Based on this, reliable road information detection technology is particularly important, which in turn requires image target detection and image segmentation technology as support.

2.1 Image Target Detection Method

Traffic signal detection is a target detection problem. The main problem of target detection is to classify the target object in the image and can accurately display the position of the target object in the image, which is divided into two classification and positioning.

In the field of traditional computer vision, detection technology is usually used to select some candidate regions on a given image and use a sliding window to traverse the entire photo. After that, people use SIFT [2] manual features, HOG [3] features, etc. to extract the features of the candidate region, and finally use traditional machine learning methods to determine the final category of the target. Although traditional methods achieved the best results at the time, they were mostly based on sliding windows, hand-designed features, and shallow classifier frameworks, resulting in high algorithm time complexity and cumbersome process of designing features. Moreover, the hand-designed features have poor detection robustness and accuracy for objects with sharp scale changes, rotation, stretching, and large color changes, and the generalization of the model cannot be guaranteed. However, with the development of deep learning, traditional computer vision methods are gradually being replaced. Picture features no longer need to be manually designed but are obtained through neural network training. Since then, deep learning methods have shined in the field of detection and completely surpassed the methods in the traditional field in accuracy. Therefore, in the accurate detection of these road elements, deep learning technology has played a decisive role, which greatly improves the recognition rate and detection rate of the model algorithm.

In recent years, convolutional neural networks based on deep learning have been studied. Due to the release of ImageNet [4], a million-scale classification data set, convolutional neural networks have made great progress in image classification tasks. Professor Hinton and his team members developed AlexNet [5] based on convolutional neural network audit. This neural network far surpassed other teams using traditional image processing methods on the ImageNet data set and won the championship.
of the image classification competition that year. It applied ReLU activation function, DropOut, and LRN techniques in the convolutional neural network for the first time and used GPU for parallel acceleration operations. Deep learning methods have become one of the most effective and popular tools in the field of computer vision. Since then, machine vision has entered a new era, and then methods based on convolutional neural networks have gradually replaced traditional methods in various application fields. Later, VGGNet [6] was proposed in 2014, and successfully constructed a 16-19 layer deep convolutional neural network by repeatedly stacking small-size convolution kernels. In the same year, GoogLeNet [7] was proposed, which spliced convolution kernels of multiple sizes to fuse different receptive fields and used a large number of convolution kernels of 1x1 size for model parameter compression. The two networks achieved very good results on the image classification task in 2014 and have been used today. The residual network ResNet [8] helps the model converge by using the residual module, speeds up the training speed of the neural network, and improves the accuracy of the model. On top of this, a 152-layer neural network was successfully trained and won the championship in the ILVRC 2015 competition. At the same time, the number of parameters is lower than that of VGGNet, and the effect is very outstanding.

The target detection models that apply these deep learning methods have achieved very good results in the data sets and competitions of various tasks. R-CNN [9] proposed by Ross Girshick is the first algorithm to successfully apply convolutional neural networks to object detection. It follows the traditional object detection concept and uses four steps for detection: feature extraction frame, feature extraction for each frame, category prediction, and non-maximum suppression (NMS). In the feature extraction step, the traditional manual features (such as HOG, SIFT features) are replaced with features extracted by deep convolutional networks. After that, FastR-CNN [10] proposed ROI pooling, pooling feature extraction frames of different scales to a uniform scale and using multi-task loss function (multi-taskloss) to return the frames to the CNN network for training. Faster R-CNN [11] greatly improves the detection speed, using the candidate region RPN (Region Proposal Network) instead of the original selective search [12] method to generate detection frames, and share some calculation parameters, which greatly saves calculation time. Subsequently, R-FCN [13] proposed a location-sensitive score map to solve the location-sensitive problem of target detection, and used a two-stage object detection framework with a full convolutional network, which is 2-2.5 times faster than Faster R-CNN. Yuting Zhang et al. proposed a target detection model combining Bayes optimization with structured prediction. This type of algorithm can improve the positioning accuracy of the identified target through a structured loss function. Spyros Gidaris et al. proposed the MR-CNN/S-CNN/LOC model to improve the accuracy of detection through candidate target regions and depth feature maps. This method first divides the candidate region into multiple different sub-regions, and then uses MR-CNN to extract the features corresponding to these regions. The author extracts the foreground features on the feature map through the convolutional neural network S-CNN of semantic analysis, and finally combines the features obtained with the features obtained by MR-CNN to perform target recognition.

There are also many one-stage detection frameworks that pay attention to speed. YOLO [14] uses a simple convolutional neural network to directly predict and regress the category. The main idea of the YOLO algorithm is to evenly divide the picture into multiple modules, then determine whether each module contains a target, and infer the target category and boundary. This method does not need to generate candidate regions, which greatly saves computing time and improves the possibility of using the algorithm in high-speed unmanned driving scenarios.

2.2 Lane Line Segmentation Method

For lane line detection, its essence is semantic segmentation of the image. What semantic segmentation does is to assign each pixel in the image to a specific category of objects. Related neural network models need to have pixel-level dense prediction and classification capabilities. The current datasets used for semantic segmentation mainly include VOC2012 and MSCOCO.
The earliest image segmentation method is the image block classification method, which uses the pixel blocks around each pixel to divide the pixels into different categories. The reason why pixel blocks must be used is that the deep neural network at that time contains a fully connected layer, and the input image must be a fixed size. In addition, in traditional convolutional neural networks, in addition to the fully connected layer, the pooling layer also restricts the solution of the segmentation problem. The pooling layer causes the image to lose part of the position information, while semantic segmentation needs to adjust the segmentation map to the precise position, which needs to retain the position information discarded by the pooling layer. In 2014, Long and others of the University of California, Berkeley invented a fully convolutional neural network [15] (Fully Convolutional Networks), which replaced the fully connected layer in the neural network with a convolutional layer. This breaks the limitation that the input image must be a fixed size. In addition, for the first time, FCN uses skip connection structure and backwards convolution technology to fuse shallow detail information with deep semantic information, which enables the network to have the ability of pixel-level dense prediction and comprehensive global information. The processing speed of the network structure is faster than the pixel block method, and the segmentation map can generate images of any size.

Although the deconvolution technique can make the output image become the size of the input, it cannot completely recover the information lost by the pooling layer. Researchers at home and abroad have proposed two different forms of network structures to address this issue. The first network structure is the encoder-decoder structure [16]. Among them, the encoder gradually reduces the spatial dimension of the input image under the action of the pooling layer, and the decoder uses a deconvolution network to gradually restore the detailed information and spatial dimension of the image. The U-Net network constructs a direct information transfer between the encoder and the decoder, thereby reducing information loss in the pooling layer. The second type is the hollow convolution structure. The network structure removes the pooling layer. When the hole convolution ratio value is 1, the network degenerates to a normal convolution network structure.

In recent years, algorithms in the field of semantic segmentation include: FCN network, SegNet network, DeepLab, DeepLab V3+[17], etc. Among them, FCN promotes the end-to-end convolutional network to solve the semantic segmentation problem. It uses the backwards convolution layer for upsampling and uses jump connections to adjust the roughness of upsampling. SegNet moves the maximum pooling index to the decoder to improve the segmentation resolution. Both FCN and SegNet networks are encoder-decoder structures that appeared earlier, and SegNet neural network benchmark scores still cannot meet the requirements of actual use. The use of Dilated Convolution in the DeepLab V3+ network (Dilated Convolution is also known as Atrous Convolution) increases the corresponding receptive field index without reducing the spatial dimension, and uses the spatial pyramid pooling technology SPP[18]. To fuse multi-scale information, this network also introduces the commonly used encoder-decoder form of semantic segmentation. In the encoder-decoder architecture, the encoder can arbitrarily control the resolution of the features extracted by the encoder, and the accuracy and time-consuming are balanced through the hole convolution. The specific network structure is shown in the figure below.
Figure 1. Network structure of DeepLab V3+

The DeepLab V3+ network uses dilated convolution and encoder-decoder forms to enlarge the receptive field without losing information, so that each convolution output contains a larger range of information. In the Encode structure, the main DCNN deep convolutional neural network uses serial Atrous Convolution. After the picture passes through the main DCNN deep convolutional neural network, the result is divided into two parts, one is directly passed into the Decoder branch, and the other part is passed through the parallel Atrous Convolution layer, which uses different rates for feature extraction, and then performs 1x1 volume after stitching Product to compress features. Decode structure input has two parts, one is the output of DCNN, and the other is the result of parallel dilation and convolution of DCNN output. After a certain amount of processing, the two results are together, and then upsampled by bilinear interpolation.

The current lane segmentation method tends to transform the dense pixel-by-pixel prediction classification into the interval pixel classification under the grid map. Zequn Qin of Zhejiang University and others proposed an ultra-fast lane line detection method [19]. The specific method is to divide the input image into equal-area grids, then take a row of effective pixels at intervals of several rows of pixels, and finally mark the position of the lane line in the extracted pixels. This method greatly reduces the number of pixels that need to be classified and greatly improves the calculation speed. In addition, for curve detection has always been a difficult problem for lane line detection, the CurveLane-NAS [20] method proposed by Huawei Noah’s Ark Laboratory uses an adaptive search method to build a neural network, focusing on training for curves, and has achieved great results.

3. Parallel Network Structure

The detection and segmentation schemes in general scenarios based on deep learning have become mature. However, in some scenarios, the technologies of general scenarios may not be used perfectly and effectively, especially in complex unmanned driving scenarios where safety is important. So far, although the general method has realized the network of detection and segmentation at the same time, in terms of automatic driving, the network of detection and segmentation is mostly performed separately. Even if the method of detection and segmentation is performed at the same time, it simply integrates the segmentation result and the detection directly and cannot achieve the purpose of practical application.

In the future, environment perception algorithms for autonomous driving should integrate semantic segmentation and detection functions. Inspired by Mask-rcnn, considering the real-time requirements of the autonomous driving environment, for the target detection part, a one-stage detection network
should be used as much as possible, so the detection part uses the YOLO series. For the segmentation part, to ensure the segmentation speed and accuracy, consider replacing the original FCN with DeepLab V3+. Therefore, this article attempts to provide a new network structure suitable for autonomous driving environments, which can realize lane segmentation and traffic signal detection together. The network structure is shown in the figure below.

![Diagram of parallel structure of object detection and semantic segmentation](image)

**Figure 2. Parallel structure of object detection and semantic segmentation**

For the input two-dimensional image, this network first enters the deep convolutional network for feature extraction. The extracted feature map enters the YOLO detection branch on the one hand, and the DeeplabV3+ segmentation branch on the other hand. Finally, the detected and segmented images are fused and output.

In addition, most of the existing data sets are for a single scene, such as pedestrian detection and lane line segmentation. The driverless driving environment is complex, and a single scene data set cannot cope with the complex driving environment. Therefore, for the development of future data sets, it should tend to multi-scene, and segmentation and detection should be integrated.

4. **Conclusion**

This article summarizes the common methods of image segmentation and target detection. So far, although a network of simultaneous detection and segmentation has been implemented in general scenarios, most of the networks of detection and segmentation are performed separately in terms of automatic driving. Therefore, this paper predicts the development of deep learning network structure suitable for autonomous driving scenarios and proposes an improved parallel network for segmentation and detection. In addition, this article puts forward requirements for the automatic driving data set suitable for this network. The data set should be added with multiple scene annotations to meet the requirements of simultaneous segmentation and detection.

**Acknowledgment**

The work is supported by 2019 Shandong Province Key R&D Plan (Awards: 2019JZZY010126, 2019JZZY010443), the China National Key R&D Plan (Awards: 2018YFE0197700), and the National Natural Science Foundation of China (Item Number: 52072212). It is also grateful for Specially Recruited Experts of Taishan Scholar.
References

[1] Desheng Xie, Youchun Xu and Rendong Wang.: Obstacle detection and tracking method for autonomous vehicle based on three-dimensional LiDAR. International Journal of Advanced Robotic Systems, 1–13(2019).

[2] David G. Lowe.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer VisionNovember 60(2),91-110(2004).

[3] Navneet Dalal, Bill Triggs.: Histograms of oriented gradients for human detection.CVPR 2005, vol. 1, pp. 886-893. IEEE(2005).

[4] Jia Deng, Wei Dong, Richard Socher; Kai Li; Li Fei-Fei. ImageNet: A large-scale hierarchical image database.CVPR 2009, pp. 248-255. IEEE(2009)

[5] Alex Krizhevsky, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM 60(6), 84-90 (2017)

[6] Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition.ICLR 2015.

[7] Christian Szegedy; Wei Liu; Yangqing Jia; Pierre Sermanet; Scott Reed; Dragomir Anguelov; Andrew Rabinovich.: Going deeper with convolutions.CVPR 2015, abs/1409.4842.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun.: Deep Residual Learning for Image Recognition.CVPR 2016, abs/1512.03385.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. CVPR 2014, pp. 580-587.

[10] Ross Girshick.: Fast R-CNN. ICCV, 1440-1448 (2015).

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS 2015:91-99.

[12] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders.: Selective Search for Object Recognition. International Journal of Computer Vision104(2), 154-171 (2013).

[13] Jifeng Dai, Yi Li, Kaiming He, Jian Sun.: R-FCN:Object Detection via Region-based Fully Convolutional Networks. NIPS 2015.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi .You Only Look Once: Unified, Real-Time Object Detection.: CVPR 2016:779-788.

[15] J.Long, E.Shelhamer, and T.Darrell.: Fully Convolutional Networks for Semantic Segmentation .CVPR, 2015, pp. 3431-3440.

[16] Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence 39(12), 2481 – 2495 (2017).

[17] Yujie Wang, Simon Sun, Jahow Yu, Dr. Limin Yu.: Skin Lesion Segmentation Using Atrous Convolution via DeepLab v3, arXiv:1807.08891.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence,2015, 1904 – 1916.

[19] Zequn Qin, Huanyu Wang, Xi Li: Ultra Fast Structure-aware Deep Lane Detection. https://arxiv.org/abs/2004.11757, last revised 2020/08/05

[20] Hang Xu, Shaoju Wang, Xinyue Cai, Wei Zhang, Xiaodan Liang, Zhengu Li. CurveLane-NAS: Unifying Lane-Sensitive Architecture Search and Adaptive Point Blending. ECCV,2020. abs/2007.12147.