ISSN 2959-6157

Research and Analysis of Deep Learning in Environmental Perception for Autonomous Driving

Xuetong Zhao

Sichuan Agricultural University, Ya'an City, Sichuan, 625014, China
E-mail: 17318215867@163.com

Abstract:

Autonomous driving technology, as a revolutionary development in the transportation field, relies on efficient and reliable environmental perception systems. Deep learning plays a key role in the environmental perception of autonomous driving, with its capabilities widely applied in areas such as signal and sign recognition, semantic segmentation, instance segmentation, and SLAM (Simultaneous Localization and Mapping). This study conducts research and analysis on the aforementioned key technologies. It concludes that while deep learning technology has shown excellent performance in the perceptual systems of autonomous driving, it still faces challenges that require further research and resolution, such as generalization performance, real-time requirements, and safety issues.

Keywords: Autonomous Driving, Environmental Perception, Deep Learning, Computer Vision

1. Introduction

The development of autonomous vehicles began with preliminary research in the 1970s and achieved key technological breakthroughs with the impetus of the DARPA Grand Challenge in 2004 [1]. Subsequently, through the joint efforts of academia and industry, continuous technological innovation and commercial trials have gradually developed a key technological system including perception, planning, and control. Autonomous driving technology has driven technological innovation in the automotive industry, giving rise to new business models and services, such as shared mobility and intelligent transportation services. It has also posed new challenges and opportunities in areas such as the job market [2], laws and regulations [3], data security [4], and privacy protection [5].

Perception technology [6] is crucial for autonomous vehicles, as it is the foundation for the vehicle's autonomous driving. It collects and interprets information about the vehicle's surrounding environment with high precision, including identifying road signs, traffic signals, pedestrians, vehicles, etc., providing key data for the decision-making and control of the autonomous driving system. The effectiveness of perception technology is directly related to the safety and reliability of autonomous vehicles, ensuring that the vehicle can make accurate responses and decisions in various complex traffic environments and conditions. With the application of advanced technologies such as deep learning, perception technology plays a central role in improving the environmental adaptability and response speed of autonomous vehicles.

The development of perception technology in the field of autonomous vehicles has evolved from basic sensor applications to deep learning algorithms. This includes the distance measurement of LiDAR [7], visual information processing of cameras, all-weather working capability of millimeter wave radar [8], and the subsequent rise of advanced technologies based on deep learning such as object detection, semantic segmentation, object recognition, multi-sensor data fusion, and SLAM. The addition of these technologies has significantly enhanced the environmental perception capabilities of autonomous vehicles, enabling them to more accurately understand their surroundings and make safe decisions.

This study mainly explores the progress of perception technology in the field of unmanned vehicles using deep learning, including the latest developments in signal and sign recognition, SLAM, and semantic and instance segmentation.

2. Core Technology

2.1 Signal and Sign Recognition

Signal and sign recognition primarily involves detecting and identifying traffic lights, traffic signs, and other traffic control devices, as well as recognizing road types, markers, and road surface markings to assist in driving decisions.

Initial traffic light recognition employed traditional image processing techniques, converting images from the RGB color space to HSV [9] or other color spaces, and using color information for preliminary screening. Thresholds were set in the color space based on the color characteristics of traffic lights for target segmentation. Morphological operations such as dilation and erosion were used to improve segmentation results and remove noise. However, this method is sensitive to lighting and background color changes, has poor robustness, and struggles with occlusions and complex backgrounds. With the widespread application and rapid development of machine learning, image features (such as shape, texture, color) were manually extracted, and traditional machine learning algorithms like Support Vector Machines (SVM) [10] and Random Forests [11] were used for classification, improving recognition accuracy by introducing feature selection and classifiers. Using CNNs to automatically learn the features of traffic lights [12] eliminated the need for manual feature design, and recognition accuracy significantly improved with extensive data training. Research [13] compared six color spaces with three deep learning network models and concluded that the combination of RGB color space and the Faster R-CNN model was the optimal solution. Jensen M. B. et al. [14] deeply discussed the evaluation methods for TLR systems, proposed a universal evaluation framework, and released a public dataset based on stereo camera shots of the American road, including annotated video sequences under various lighting and weather conditions, with high diversity and continuity. The YOLO algorithm [15] treats the object detection task as a single regression problem, directly mapping from image pixels to target category probabilities and bounding box coordinates. With the development from YOLOv5 to YOLOv8, performance has continuously improved, introducing more advanced feature fusion mechanisms, attention mechanisms, and more refined bounding box prediction methods. Although deep learning-based object detection models have significantly improved the accuracy of vehicle detection, existing technologies still face challenges in low-light night conditions and adverse weather conditions. Huang, J. et al. [16] used the Coordinate Attention mechanism CA to enhance the model's ability to capture global and local features, and improvements to the ELAN structure made the model more adaptable to noise interference in night environments, enhancing feature extraction accuracy.

Autonomous vehicles perceive their surroundings and understand road conditions and traffic situations by recognizing traffic signs. Traffic signs provide essential environmental information and instructions for autonomous driving systems, ensuring vehicles can travel safely and compliantly, making it crucial for autonomous systems to correctly recognize traffic signs. Although there are various types of traffic signs, upon careful observation and analysis, in China, traffic signs can be categorized into three major categories: directional signs, warning signs, and prohibitory signs [17]. Warning signs are mainly used to alert drivers to upcoming road conditions or hazards so they can take necessary actions. These signs typically have a yellow background with black patterns, designed to attract the driver's attention and vigilance, such as sharp curves, slopes, animal crossings, etc. Prohibitory signs are used to regulate certain behaviors under specific conditions to maintain traffic order and safety. These signs usually have a red circular background with white patterns, such as no stopping, no left turn, no overtaking, etc. Directional signs are used to instruct drivers on the actions they should take or to indicate the direction and destination of the road. These signs typically have a blue background with white patterns, such as road names, directional indicators, service areas, exits, etc. Huang Z. et al. [18], based on the above characteristics, simulated the different stimuli our eyes receive from the three types of visible light and proposed a segmentation algorithm that can effectively cope with the impact of lighting and weather conditions on the algorithm. Barnes, N. et al. [19] analyzed the edges of the image and efficiently detected the positions of regular polygons through a method of dimensionality reduction followed by dimensionality increase. Jonathan M. et al. [20] used CAE to construct a convolutional neural network to effectively recognize traffic signs, and subsequently Dan [21], Sermanet P. [22], Zeng, Y. et al. [23] and others successively optimized the model structure, continuously improving the network's generalization ability and robustness.

2.2 Semantic and Instance Segmentation

Semantic and instance segmentation are two important tasks in computer vision, both involving the classification of every pixel in an image. Semantic segmentation [24] assigns each pixel in an image to a specific category label, such as person, car, building, etc., focusing on distinguishing instances of all categories in the image, but without differentiating between different instances of the same object class. Instance segmentation [25], as a combination of object detection and semantic segmentation, not only distinguishes instances of all categories in the image but also differentiates between different instances of the same object class, assigning a unique label to each object instance in the image and performing pixel-level classification.

Traditional semantic segmentation methods are usually based on low-level image features (such as color, texture, etc.) and classic computer vision techniques (such as image segmentation algorithms and feature extraction).

These methods work well with small datasets and simple scenes but perform poorly in complex scenes and large-scale datasets. With the rise of deep learning, especially convolutional neural networks (CNNs), significant progress has been made in semantic segmentation. CNNs can end-to-end learn image features and semantic information, thereby improving the accuracy and efficiency of segmentation. For example, FCN (Fully Convolutional Network) has greatly advanced the field by converting fully connected layers to convolutional layers, enabling pixel-level semantic segmentation.

Traditional instance segmentation methods usually combine object detection and segmentation algorithms, first detecting the bounding box of an object, and then segmenting the object within each bounding box. These methods face challenges in accuracy and efficiency, especially in cases of object occlusion, overlap, and complex backgrounds. In recent years, with the advancement of deep learning technology, especially deep learning models such as Mask R-CNN [26] that combine regional proposal networks (RPN) and semantic segmentation, instance segmentation has made significant breakthroughs. These models can not only detect the location of objects but also accurately segment the pixels of each object, effectively dealing with complex scenes and situations of multiple object overlaps.

2.3 SLAM (Simultaneous Localization and Mapping)

SLAM, which stands for Simultaneous Localization and Mapping, is a key technology in fields such as robotics, autonomous driving, and unmanned aerial vehicles. SLAM is designed to enable autonomous navigation and map construction for robots or autonomous vehicles in unknown environments. As the name implies, SLAM encompasses localization and map construction, requiring the determination of the vehicle's exact location in the environment and the creation of a detailed map of the environment, including the ground, obstacles, and other features.

The technology of SLAM was first proposed in the 1980s, primarily using sensors such as LiDAR for environmental measurement and positioning [27]. LiDAR SLAM obtains accurate distance and direction information of the environment through LiDAR [28], while visual SLAM uses visual information captured by cameras for environmental understanding and positioning. Visual SLAM is mainly divided into two major categories: feature-based methods and direct methods. Feature-based methods rely on feature point matching in image sequences, while direct methods directly use image pixel information for pose estimation [29]. In visual SLAM systems, the presence of dynamic objects can significantly affect system performance, potentially causing inaccurate positioning and a decline in map construction quality. To reduce the impact of dynamic objects on visual SLAM systems, researchers have proposed various deep learning-based methods: using object detection networks to identify and locate dynamic objects in the image, and identifying the category of each object by drawing candidate frames for the objects. For example, the Detect-SLAM system uses the SSD network for object detection to improve the robustness of SLAM in dynamic environments. By using deep learning networks for pixel-level semantic segmentation, dynamic areas can be more accurately identified and removed, thereby improving the system's accuracy. For instance, DS-SLAM combines the semantic segmentation network SegNet and mobile consistency detection methods to reduce the impact of dynamic objects and generate a dense semantic octree map. Instance segmentation can also achieve pixel-level classification and locate different instances in the image, suitable for precise segmentation in complex dynamic environments. For example, DynaSLAM [30] utilizes Mask R-CNN for instance segmentation to enhance the robustness of visual SLAM in dynamic environments.

3. Conclusion

Autonomous vehicles face challenges in handling complex scenarios and enhancing real-time performance and accuracy, especially in high-speed moving traffic environments where these technologies require faster response times and higher precision. Future research and development will need to focus on integrated solutions that combine multiple perception technologies to enhance the robustness and reliability of the overall perception system. Deep learning models will place greater emphasis on the diversity of datasets and the generalization capabilities of models to improve the reliability and safety of the system in the real world. Perception is only the first step in the decision-making of autonomous driving systems; future development will pay more attention to transforming perceptual information into safe and efficient driving decisions.

References

[1] Zheng, C., Guo, M., Zhou, C. (2024). Review of SLAM Algorithms Based on LiDAR. Digital Manufacturing Science, (02), 100-105.

[2] Guo, C. (2024, July 22). Behind the Popularity of "Luobo Kuai Pao" [News Article]. China Automotive News, p. 033. doi:10.28116/n.cnki.ncqcb.2024.000115.

[3] Zhao, L. (2024). Study on the Responsibility of Autonomous Vehicle in Traffic Accidents. China-Arab States Science and

Technology Forum (Bilingual Edition), 7, 152-156.

[4] Zhang, X., Chen, H., Li, Y. (2024). Application Practices of Automotive Data Processing Security Requirements National Standards in the Automotive Industry. Information Technology and Standardization (Supplement 1), 130-135.

[5] Zhu, X., Wang, B., & Lin, Y. (2023). Risks and countermeasures of artificial intelligence data security and privacy protection. Cybersecurity (03), 30-34.

[6] Wang, S., Dai, X., Xu, N. (2017). Overview on Environment Perception Technology for Unmanned Ground Vehicle. Journal of Changchun University of Science and Technology (Natural Science Edition), 1, 1-6.

[7] Wang, H., Luo, T., & Lu, P. (2018). Development of the lidar applications in unmanned vehicles and its key technology analysis. Laser & Infrared, 12, 1458-1467.

[8] Han, B., & Wang, Z. (2019). Overview of Development of Vehicle-mounted Millimeter-wave Radar at Home and Abroad. Digital Communication World, (09), 15-16.

[9] Liu, K., Dong, M., Wang, P. (2022). A traffic light recognition method based on image enhancement. Electronic Measurement Technology, (07), 137-145. doi:10.19651/j.cnki.emt.2108559.

[10] Li, X., Guo, X., & Guo, J. (2013). HOG-Feature and SVM Based Method for Forward Vehicle Recognition. Computer Science, S2, 329-332.

[11] Gao, X., Liao, Y., & Liu, L. (2024). Road condition detection based on adaptive random forest migration learning. Journal of Ningde Normal University (Natural Science Edition), (02), 143-151. doi:10.15911/j.cnki.35-1311/n.2024.02.003.

[12] Qin, Y., Cai, Y., Huang, P. (2023). Detection Method of Railway Signal Lights and Parking Carriages Based on Improved Faster R-CNN. Journal of Xihua University (Natural Science Edition), (02), 62-69.

[13] Kim, H.-K., Park, J. H., & Jung, H.-Y. (2018). An Efficient Color Space for Deep-Learning Based Traffic Light Recognition [Article]. Journal of Advanced Transportation, Article 2365414. https://doi.org/10.1155/2018/2365414

[14] Jensen, M. B., Philipsen, M. P., Mogelmose, A., Moeslund, T. B., & Trivedi, M. M. (2016). Vision for Looking at Traffic Lights: Issues, Survey, and Perspectives [Article]. Ieee Transactions on Intelligent Transportation Systems, 17(7), 1800-1815. https://doi.org/10.1109/tits.2015.2509509

[15] Zhang, F., Yang, F., & Li, C. (2019). Fast Vehicle Detection Method Based on Improved YOLOv3. Computer Engineering and Applications, (02), 12-20.

[16] Huang, J., Liu, P., & Tang, W. (Year not specified). MMD-YOLOv7: Vehicle detection method under dark conditions. Computer Engineering, 1-12.

[17] Liang, X. (2023). Research on Traffic Sign Recognition for Autonomous Driving Scenarios [Master's Thesis, Guangzhou University]. https://link.cnki.net/doi/10.27040/d.cnki.ggzdu.2023.001153

[18] Huang, Z., Sun, G., & Li, F. (2004). Traffic Sign Segment Based on RGB Vision Model. Microelectronics & Computer, 10, 147-148+152. doi:10.19304/j.cnki.issn1000-7180.2004.10.040.

[19] Barnes, N., Loy, G., & Shaw, D. (2010). The regular polygon detector [Article]. Pattern Recognition, 43(3), 592-602. https://doi.org/10.1016/j.patcog.2009.09.008

[20] Cireşan, D., Meier, U., Masci, J., & Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. Neural Networks, 32, 333-338. https://doi.org/https://doi.org/10.1016/j.neunet.2012.02.023

[21] Masci, J., Meier, U., Cireşan, D., & Schmidhuber, J. (2011). Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In T. Honkela, W. Duch, M. Girolami, & S. Kaski, Artificial Neural Networks and Machine Learning – ICANN 2011 Berlin, Heidelberg.

[22] Sermanet, P., & LeCun, Y. (2011, 31 July-5 Aug. 2011). Traffic sign recognition with multi-scale Convolutional Networks. The 2011 International Joint Conference on Neural Networks,

[23] Zeng, Y., Xu, X., Fang, Y., & Zhao, K. (2015). Traffic Sign Recognition Using Extreme Learning Classifier with Deep Convolutional Features.

[24] Tian, X., Wang, L., & Ding, Q. (2019). Review of Image Semantic Segmentation Based on Deep Learning. Journal of Software, (02), 440-468. doi:10.13328/j.cnki.jos.005659.

[25] Su, L., Sun, Y., & Yuan, S. (2022). A survey of instance segmentation research based on deep learning. Journal of Intelligent Systems, (01), 16-31.

[26] Yao, M., Deng, H., Fu, W. (2021). An Improved Mask R-CNN Image Instance Segmentation Algorithm. Software, (09), 78-82.

[27] Yuan, C., Lai, J., Zhang, J., & Lyu, P. (2018, 9-13 Jan. 2018). Research on an autonomously tightly integrated positioning method for UAV in sparse-feature indoor environment. 2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST),

[28] LIU Peng, REN Gongchang, HE Zhou. Method for extracting corner feature from 2D laser SLAM[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2021, 53(3): 366-372.

[29] Engel, J., Schöps, T., & Cremers, D. (2014). LSD-SLAM: Large-Scale Direct Monocular SLAM. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars, Computer Vision – ECCV 2014 Cham.

[30] Bescos, B., Facil, J. M., Civera, J., & Neira, J. (2018). DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes [Article]. Ieee Robotics and Automation Letters, 3(4), 4076-4083. https://doi.org/10.1109/lra.2018.2860039