Analysis of Visual Navigation Systems in Autonomous Driving

Yizhan Zhang

School of Automotive and Transportation Engineering, Hefei University of Technology, Hefei, 230009, China
2023213033@mail.hfut.edu.cn

Abstract. As a key technology for environmental perception and precise positioning in autonomous driving, the performance of visual navigation systems directly impacts vehicle safety and efficiency in complex environments. This system primarily relies on visual cameras to capture rich road scene information, utilizing image processing and deep learning algorithms to identify lane markings, traffic signs, pedestrians, and obstacles. It offers advantages such as low cost and broad information dimensions. However, its performance is susceptible to interference from lighting variations, adverse weather, and dynamic scenes, leading to reduced recognition reliability and insufficient positioning accuracy. To enhance system robustness, current research focuses on multi-sensor fusion techniques. By integrating the ranging stability of millimeter-wave radar or the 3D spatial perception capabilities of lidar, these approaches compensate for the limitations of single-visual sensors. This paper systematically outlines the core principles and technical architecture of visual navigation systems, analyzes two major challenges—adaptability and real-time performance—in complex environments, explores solutions through algorithm optimization and hardware collaboration, and examines high-precision semantic mapping and end-to-end learning for autonomous driving. It aims to provide theoretical support for the research and engineering implementation of visual navigation systems in autonomous driving.

Keywords: Visual navigation; Autonomous driving; Environmental perception; Multi-sensor fusion.

1 Introduction

Autonomous driving technology, exemplifying the deep integration of artificial intelligence and the automotive industry, is reshaping traditional transportation. Through innovative perception systems, intelligent decision-making algorithms, and precise actuation mechanisms, autonomous driving systems achieve environmental perception, path planning, and autonomous control, offering new technical pathways to enhance driving safety and mobility efficiency.

In recent years, breakthroughs in deep learning technology have increasingly highlighted the central role of visual navigation systems in autonomous driving. As a

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_21

critical component of environmental perception, camera-based visual systems capture rich road semantic information. Through advanced neural network algorithms, high-precision recognition of lane lines, traffic signs, pedestrians and obstacles is achieved. Compared with other perception methods, the visual system has a significant cost advantage and a more comprehensive ability to obtain information.

However, visual systems still face numerous challenges in practical applications. In complex urban road scenarios, dense traffic flow, dynamic obstacles and diverse traffic environments pose extremely high requirements for the robustness of the system. Under conditions of drastic changes in light or adverse weather conditions such as rain, snow and fog, the performance of the system often fluctuates significantly. In addition, technical bottlenecks such as long-distance positioning accuracy and real-time data processing capabilities also urgently need to be broken through.

In response to these challenges, current research focuses on two key directions. On the one hand, through continuous optimization at the algorithm level, including improving the neural network architecture, enhancing feature extraction capabilities, and boosting model generalization performance; On the other hand, through the collaborative innovation of hardware systems, including the research and development of high-performance image sensors, the improvement of computing platform efficiency, and the adaptive improvement of optical systems. The breakthroughs in these technological paths will effectively enhance the perception reliability of the visual system in complex environments, laying a solid foundation for the commercial application of autonomous driving technology.

This paper explores key technologies in visual navigation systems, focusing on their performance in complex environments. It further investigates pathways to enhance system performance through coordinated algorithmic and hardware optimization. The research findings provide theoretical references and practical guidance for designing and implementing autonomous driving perception systems.

2 Core Principles and Technical Architecture of Visual Navigation Systems

2.1 Image Acquisition in Pure Vision-Based Perception Systems

Pure vision perception systems can be categorized into monocular visual object detection, stereo vision detection techniques, and multi-view fusion detection. Monocular visual object detection captures scene textures and geometric features through a single camera, utilizing deep learning networks for target recognition. T. Wang et al. proposed FCOS3D, a fully convolutional one-stage monocular 3D object detection framework that simplifies the 2D-3D correspondence in traditional methods by performing 3D detection directly in the image domain [1]. L. Peng introduced the OccupancyM3D method, aiming to improve monocular 3D detection by learning occupancy information. This approach combines sparse LiDAR point clouds with voxelized occupancy labels to extract more discriminative 3D features during training, thereby enhancing monocular 3D detection accuracy. Experimental results demonstrate

that OccupancyM3D is particularly effective in complex environments, significantly improving detection performance on the KITTI dataset [2]. In stereo vision, Y. Chen et al. proposed DSGN, an end-to-end stereo 3D detection network that jointly estimates scene depth and detects 3D objects, achieving outstanding performance on the KITTI dataset [3]. Meanwhile, C. Li et al. introduced StreamDSGN, a streaming stereo 3D detector based on the DSGN++ architecture, suitable for real-time applications [4]. Finally, multi-view fusion detection captures panoramic information through a network of surround cameras. Building upon this, Y. Zhang et al. proposed MVSDet, a geometry-aware 3D object detection method utilizing planar scanning, suitable for indoor multi-view scenarios [5] . The three detection approaches based on different routes exhibit distinct characteristics, limitations, and require different core solutions and innovations during development (as shown in Table 1).

Table 1. Characteristics of Three Visual Detection Methods in Pure Vision Perception Systems

Detection Type	Core Solutions and Innovations	Accuracy Characteristics	Inherent Limitations
Monocular Visual Detection	FCOS3D, OccupancyM3D	Excellent real-time performance, limited depth perception	Insufficient positioning accuracy due to lack of depth information
Stereo Vision Inspection	DSGN, StreamDSGN	Geometric depth enhancement, balancing real-time performance and accuracy	Occlusion failure, stringent calibration requirements, limited field of view
Multi-view fusion detection	MVSDet	Multi-view complementarity, optimal spatio-temporal fusion robustness	High hardware cost, complex spatio-temporal alignment

2.2 Image Processing Based on Image Acquisition

Following image acquisition, the processing stage is critical as it determines the ability to extract meaningful information from the captured data. With advancements in sensors and computational power, image processing methods continue to evolve, providing essential support for subsequent tasks such as object detection, classification, and localization. The goal of image processing is to enhance image quality, remove noise, strengthen features, and extract useful information to deliver accurate, real-time data for decision-making and control in autonomous driving systems.

W. Benjilali, W. Guicquero, L. Jacques, and H. P. A. Lensch proposed a hardware-friendly compressed imaging method. This approach randomizes image data through stochastic modulation and permutation, then reduces storage and transmission

requirements by combining permutations and combinations. It enables efficient image acquisition and classification, making it suitable for resource-constrained devices [6].

Other approaches optimize image processing using deep learning or image reflection models. For instance, J. Guo, Z. Zhou, and L. Wang's highlight-aware dual-stream network analyzes highlight regions within images, leveraging a dual-channel network architecture to separate and process illumination variations. This method excels in scenarios requiring highlight removal—such as material reconstruction and computational photography—reducing unnecessary illumination interference while preserving image quality [7].

They also proposed a highlight removal method based on sparse and low-rank reflection models. By modeling the reflective components of an image, this model effectively removes highlight regions, thereby enhancing image quality. This technique holds significant importance for post-capture information processing, particularly in improving the accuracy of image reconstruction and analysis, offering an effective highlight removal solution [8].

2.3 Post-Capture Image Processing

Pure vision-based navigation systems analyze captured images through algorithms and large-scale data models to generate commands for autonomous vehicles, determining their next actions while ensuring real-time processing and accuracy throughout the process.

To address traffic accident risks stemming from rising vehicle ownership, a vision-based intelligent driving system aims to enhance road safety, focusing primarily on lane departure warning and collision avoidance functions. The L-K optical flow method dynamically evaluates video frames to preliminarily locate the approximate range of lane lines. Subsequently, composite operators compute and detect road edge features, while polynomial fitting technology identifies specific lane lines. Finally, the recognition results are annotated and output [9].

Zhao Hui, Li Xu, Zhang Ying, and Liu Zhi proposed an edge flow data processing framework specifically tailored to the data processing demands of autonomous driving systems. The innovation of this framework lies in migrating computing and storage capabilities from the central processing unit of the vehicle to edge devices. can reduce the computational burden on vehicles and significantly enhance the real-time performance and reliability of data processing. By processing sensor data at edge nodes, the system can respond quickly and make decisions without increasing computing pressure. This method effectively enhances the response speed of the autonomous driving system and has demonstrated excellent application results in multiple autonomous driving scenarios [10].

Meanwhile, Yang Bo, Wang Chen, Zhou Xiao and Sun Liang discussed the application of multi-sensor information fusion technology in autonomous driving. This study integrates data from various sensors, including cameras, lidars, and millimeter-wave radars, etc., and can provide more accurate and robust environmental perception capabilities. Through the fusion of multi-sensor data, the system can better handle complex traffic scenes after image acquisition, improving the accuracy of obstacle

detection, path planning and decision-making. This information fusion technology is particularly suitable for the requirements of high-precision environmental perception in autonomous driving systems, and can significantly enhance the robustness and stability of the system [11].

3 High-Precision Semantic Maps

3.1 Advantages and Principles of High-Precision Maps

High-definition maps (HDMaps) are a core element in autonomous driving. As a scarce resource and critical data foundation, they are essential for autonomous driving. They enable vehicles to anticipate complex road surface information in advance, such as slope, curvature, and heading[12] . Current standard electronic navigation maps suffer from three critical limitations: First, they lack rich map features, providing only abstract road information without key details like lane width, traffic signals, or road obstacles. Second, their low accuracy (approximately 10 meters) fails to support lane-level navigation. Finally, dynamic information updates lag, making it difficult to reflect real-time road conditions (e.g., signal status, tidal lane changes). To meet the real-time and accuracy requirements of autonomous driving, high-precision map assistance is essential, with varying data needs corresponding to different levels of automation. Levels 0 to 2 (L0-L2) require road and traffic information with 1-10 meter accuracy, maintaining human driver control. Level 3 (L3) conditional automation demands accuracy of 0.2-0.5 meters and begins to disengage driver control. Levels 4 and 5 demand full autonomy in specific or all scenarios, requiring data accuracy further refined to 0.05–0.2 meters. Consequently, to meet the rigid demand for 0.05–0.5-meter high-precision data in Levels 3–5 autonomous driving, high-definition maps—with their centimeter-level positioning capability and rich semantic information—have become indispensable for realizing advanced autonomous driving.

The core value of high-precision maps lies in the two major stages of perception and planning. Firstly, in the perception stage, high-precision maps act as virtual digital sensors. By providing a complete static environment description such as lane boundaries, curvature, slope, height limit signs, etc., and then combining real-time data from vehicle sensors, they achieve an accurate understanding of the surrounding environment, assisting in obstacle detection and positioning verification. During the planning stage, high-precision maps can independently generate the optimal path based on semantic information such as lane-level road networks and traffic rules. Meanwhile, it can also use static data to pre-control vehicle speed and integrate dynamic information such as real-time traffic flow to adjust driving strategies, with the vehicle control system ultimately performing the operation. Overall, high-precision maps provide accurate static environmental benchmarks through the perception stage, thereby supporting lane-level decision-making and dynamic strategy optimization in the planning stage, and ultimately achieving a safe closed loop of vehicle control.

3.2 Limitations and Solutions for High-Precision Maps

High-precision maps mainly provide accurate road geometry and static object positions, supporting vehicles in basic lane-level path planning and positioning. On this basis, high-precision semantic maps deeply integrate road traffic rules and scenario-based information. For instance, in rainy or snowy weather, it can combine real-time meteorological data to automatically adjust the speed limit of specific sections prone to icing from 60km/h in dry conditions to 40km/h, and mark high-risk areas, enabling the autonomous driving system not only to plan routes It can better understand rules based on environmental changes, predict risks and dynamically adjust driving strategies, thereby achieving more anthropomorphic, safer and more reliable advanced decision-making and behavioral planning.

The core deficiency of a single high-precision map lies in its static attribute, that is, it is essentially a pre-collected and constructed road environment database, which makes it difficult for it to respond in real time to the dynamic changes in the real driving environment. Although it can accurately record the geometric features of the road, such as lane lines, slopes and the positions of fixed facilities like traffic signs, due to the lag in data updates, it is unable to perceive sudden situations, let alone predict dynamic behaviors such as pedestrians crossing the road or adjacent vehicles braking suddenly. This limitation makes autonomous driving systems that rely solely on high-precision maps lack decision-making capabilities in complex or unexpected scenarios, posing a challenge to their safety.

To overcome the contradiction between the static limitations of high-precision maps and the complexity of dynamic environments, Shen Chen, Wang Jingping, and Ma Dongdong proposed to enhance by integrating V2X communication technology and computer vision: V2X technology establishes a real-time communication network between vehicles and everything, enabling vehicles to break through the visual range limitations of their own sensors and proactively obtain dynamic information in areas beyond visual range or obstructing, such as accident warnings ahead and vehicle positions at intersections. This fills the gap of high-precision maps in real-time acquisition of surrounding information, significantly expanding the perception range and collaborative capabilities. Provide advanced information support for decision-making[13] .

Meanwhile, computer vision uses multiple sensors to simulate human visual perception, analyze environmental details in real time, and observe the trajectories of dynamic objects and key state changes. While verifying the static data of high-precision maps, it endows the system with autonomous adaptability to sudden scenarios.

Overall, when the two work together: V2X is responsible for "hearing" real-time messages from other vehicles and roadside devices; Computer vision uses its own "eyes" to carefully observe the situation around the vehicle, such as whether the road is slippery and whether there are pedestrians beside it. Combined with high-precision maps, a real-time panoramic view of traffic conditions that is both clear and updated at any time can be pieced together.

4 End-to-End Learning for Vision-Based Navigation Systems

End-to-end learning directly maps the raw sensor input to vehicle control instructions through a deep learning architecture, eliminating the intermediate modeling link in traditional methods. This method has strong generalization ability due to simulating the intuitive driving mode of human "sensory-action", can adapt to complex road conditions such as no lane signs, and significantly reduces the development cost at the same time. However, although end-to-end learning methods have great potential in theory, they still face many challenges in practical applications, especially in terms of adaptability to dynamic scenarios, robustness of the system, and interpretability of the model. Therefore, in recent years, many studies on end-to-end visual navigation methods have emerged in the academic circle. These studies have further promoted the development of this field by introducing new algorithms, optimizing network structures, and enhancing data processing capabilities.

Firstly, Shyr-Long Jeng and Chienhsun Chiang proposed an end-to-end autonomous navigation method based on deep reinforcement learning. This study designed a system capable of performing autonomous navigation in a dynamic environment by combining the Deep Deterministic Policy Gradient (DDPG) and Double-Delay DDPG (TD3) algorithms. This algorithm demonstrates high adaptability in practical applications and is particularly suitable for scenarios that require rapid decision-making and real-time actions [14].

Eder A. Rodriguez Martinez et al. proposed A learning navigation method based on feature RGB-D pose estimation and topological maps. By integrating the lightweight multi-layer perceptron (MLP) strategy, the research team optimized the robot's navigation ability in the environment, especially in various complex environments. This method avoids the high computational costs of SLAM systems based on large-scale learning or metrics. Instead, it enables end-to-end learning methods to be more efficient and reliable in practical applications through efficient feature estimation and map construction [15].

Finally, Y Sheng et al. proposed an autonomous navigation algorithm for unmanned aerial vehicles based on deep reinforcement learning. This method can perform autonomous path planning in high-density and dynamic environments and can dynamically adjust the reward function, thereby enhancing the navigation ability of unmanned aerial vehicles in complex flight environments. Especially in high-density cities or dynamic environments, this method enables unmanned aerial vehicles (UAVs) to respond quickly to environmental changes by optimizing the representation of the state space, demonstrating high navigation accuracy and system robustness [16].

Through these studies, the application prospects of end-to-end learning methods in visual navigation systems can be seen. Despite challenges such as complex environments, high real-time requirements, and poor system robustness, end-to-end learning still demonstrates great potential. The current research directions mainly focus on the following aspects: On the one hand, deep reinforcement learning and deep neural networks are utilized to optimize the decision-making ability of the system, enabling it to adapt to more complex and dynamic environments; On the other hand, by enhancing

image processing capabilities, introducing spatio-temporal information and other means, the robustness and real-time response ability of the system are strengthened.

In conclusion, end-to-end learning methods, as an emerging visual navigation technology, are gradually transforming the architecture of traditional autonomous driving systems. Although there are still many technical bottlenecks in practical applications, with the continuous progress of deep learning and computer vision technologies, the application of end-to-end learning in autonomous driving will become more widespread in the future, providing a more solid theoretical support and technical guarantee for the commercial implementation of autonomous driving technology.

5 Conclusion

This study analyzes the application of visual navigation systems in autonomous driving, with a focus on discussing its core technologies, challenges and solutions. The visual navigation system collects environmental information through cameras and, in combination with deep learning algorithms, can accurately identify roads, traffic signs and obstacles. Although visual navigation systems have the advantages of low cost and extensive information acquisition, they still face challenges in terms of robustness, real-time performance and positioning accuracy in complex environments.

To address these issues, current research is focused on multiple directions, such as enhancing the generalization ability and real-time response capability of visual systems by optimizing deep learning algorithms and network architectures. Or, multi-sensor fusion technology can be adopted, combined with millimeter-wave radar, lidar and other devices, to make up for the limitations of a single visual sensor and enhance the perception ability and stability of the system. In addition, high-precision semantic mapping technology provides accurate road information and traffic rules for autonomous driving, supporting more intelligent path planning and decision-making. And through hardware collaborative innovation, the performance of image sensors and the efficiency of the computing platform are enhanced, further strengthening the real-time processing capability of the system.

Looking ahead, with the advancement of deep learning, sensor technology and computing platforms, visual navigation systems will achieve breakthroughs in enhancing system robustness, adapting to dynamic scenarios and real-time response capabilities. Especially in the fields of multi-sensor fusion and spatio-temporal information processing, continuous innovation will promote the commercialization of autonomous driving technology and provide the industry with safer and smarter solutions.

References

1. Wang, T., Zhu, X., Pang, J., et al.: FCOS3D: Fully convolutional one-stage monocular 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 913–922 (2021)

2. Peng, L., Xu, J., Cheng, H., et al.: Learning occupancy for monocular 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10281–10292 (2024)
3. Chen, Y., Liu, S., Shen, X., et al.: DSGN: Deep stereo geometry network for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12536–12545 (2020)
4. Li, C., Gu, Z., Chen, G., et al.: Real-time stereo-based 3D object detection for streaming perception. Advances in Neural Information Processing Systems 37, 115468–115490 (2024)
5. Zhang, X., Gong, W., Xu, X.: Magnetic ring multi-defect stereo detection system based on multi-camera vision technology. Sensors 20(2), 392 (2020)
6. Benjilali, W., Guicquero, W., Jacques, L., et al.: Hardware-compliant compressive image sensor architecture based on random modulations and permutations for embedded inference. IEEE Transactions on Circuits and Systems I: Regular Papers 67(4), 1218–1231 (2020)
7. Guo, J., Lai, S., Tao, C., et al.: Highlight-aware two-stream network for single-image SVBRDF acquisition. ACM Transactions on Graphics 40(4), 1–14 (2021)
8. Guo, J., Zhou, Z., Wang, L.: Single image highlight removal with a sparse and low-rank reflection model. In: Proceedings of the European Conference on Computer Vision, pp. 268–283 (2018)
9. Song, Q., Cai, D., Li, X., et al.: Research on vehicle-mounted intelligent driving systems based on visual navigation. China New Technology and New Products 2022(19), 1–3 (2022)
10. Zhao, H., Yao, L., Zeng, Z., et al.: An edge streaming data processing framework for autonomous driving. Connection Science 33(2), 173–200 (2021)
11. Yang, B., Li, J., Zeng, T.: A review of environmental perception technology based on multi-sensor information fusion in autonomous driving. World Electric Vehicle Journal 16(1), 20 (2025)
12. Goren: Research on reinforcement learning-based visual navigation for autonomous driving. Beijing University of Posts and Telecommunications (2025)
13. Shen, C., Wang, J., Ma, D., et al.: Research and application of high-precision maps in intelligent navigation. Electronics Technology and Software Engineering 2021(17), 177–178 (2021)
14. Jeng, S.L., Chiang, C.: End-to-end autonomous navigation based on deep reinforcement learning with a survival penalty function. Sensors 23(20), 8651 (2023)
15. Rodríguez-Martínez, E.A., Miranda-Vega, J.E., Achakir, F., et al.: Efficient learning-based robotic navigation using feature-based RGB-D pose estimation and topological maps. Entropy 27(6), 641 (2025)
16. Sheng, Y., Liu, H., Li, J., et al.: UAV autonomous navigation based on deep reinforcement learning in highly dynamic and high-density environments. Drones 8(9), 516 (2024)