Technology and Application of Computer Vision in Smart Driving

Yiheng Zhou

Maynooth International Engineering College, Fuzhou University, Fuzhou, Fujian, China
YIHENG.ZHOU.2025@mumail.ie

Abstract. Intelligent driving is a systematic project that involves the realization of multiple links in a collaborative manner. Environmental perception is an important link in an intelligent driving system. Machine vision can acquire rich and accurate data through cameras, and through the processing of this information, it can provide rich environmental information for vehicles. As the core technology of intelligent driving environment perception, machine vision is promoting the evolution of automatic driving from auxiliary function to full-scene autonomous decision-making through multimodal data fusion and deep learning algorithms. This paper systematically sort out the technical system of machine vision in intelligent smart driving cars, and deeply analyze the core methods of the whole process of image acquisition, processing, and analysis, in addition to which, this paper also focuses on the analysis of deep learning-based environment perception methods, multi-sensor fusion architectures, typical application scenarios, and future development trends. This review aims to elucidate the development status of machine vision technology and application in the field of intelligent driving, and to provide a systematic reference for promoting intelligent driving technology from theoretical research to large-scale landing.

Keywords: Machine Vision; Intelligent Driving; Environment Perception

1 Introduction

With the rapid development of science and technology, intelligent driving technology is becoming more and more mature and progressive, which has just become a research favorite in the field of science and technology at present. Intelligent driving, as a complex system engineering, involves a number of complex links. The core links include environment perception, localization and mapping, path planning, decision making and control. These complex links work together to realize intelligent driving of vehicles [1]. In order to realize safe and efficient intelligent driving, accurate and comprehensive environment perception is the cornerstone.

With its unique advantages, machine vision stands out among a number of perception technologies and becomes an indispensable key component of intelligent driving systems. Machine vision uses cameras and other equipment to obtain rich information about the surrounding environment. After obtaining enough information, machine vision can process and analyze this information through advanced computer algorithms

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_6

to realize the recognition and understanding of target entities such as roads, vehicles, pedestrians, traffic signs and signals [2]. Compared to other environment-aware sensors, such as LIDAR and millimeter-wave radar, cameras have a lower cost and higher resolution, and provide rich texture and color information, which helps intelligent driving systems make more accurate decisions.

However, machine vision is not without defects, complex and changing traffic scenes will bring a lot of impacts to machine vision, such as drastic changes in lighting conditions, interference from bad weather (rain, snow, fog, etc.), partial or complete occlusion of the target object, etc., and these complex external factors of change will have an impact on the perceptual accuracy and reliability of machine vision. Therefore, how to improve the performance of machine vision in complex environments has become one of the core issues in current intelligent driving research [3]. In recent years, with the rapid development of deep learning technology, its powerful feature extraction and pattern recognition capabilities have injected new vitality into the application of machine vision in intelligent driving environment perception, and promoted the rapid progress of research in this field.

This review firstly introduces the technical foundation of machine vision in the field of intelligent driving. Then it combines and analyzes several core tasks of environment perception realized by machine vision based on its technical foundation. Finally, this paper synthesizes several typical intelligent driving scenarios to introduce the real-world application integration scheme of machine vision technology. Through the synthesis logic from theory to application, this paper aims to help readers have a more systematic and comprehensive knowledge of this technology field.

2 Machine Vision in Intelligent Driving

2.1 Image Acquisition

Image acquisition is the "data entrance" of machine vision, and its hardware performance directly affects the subsequent processing accuracy. Intelligent driving in-vehicle camera system consists of optical lenses, image sensors, ISP (Image Signal Processor) and calibration modules, of which the image sensor is the core component. Currently, the mainstream CMOS sensor has replaced CCD as the first choice for mass production due to its high integration (single-chip integration of ADC and ISP), low power consumption (<1W) and high frame rate (60-120FPS) [4]. In order to cover the whole scene perception, the vehicle camera needs to be configured according to the differentiation of functions: the main front view camera (FOV 50°-70°) is responsible for long-distance target detection (such as traffic signals); the surround view camera (FOV 120°-150°) to achieve 360 ° environmental coverage; the side view camera (FOV 90°) focuses on lateral obstacle recognition [5].

2.2 Image Processing

Image pre-processing technology through the ROI delineation, grayscaling, filtering and edge detection, etc., to reduce noise interference, extract effective information, and lay a foundation of accurate data for subsequent image analysis.

2.2.1 Region of Interest (ROI) Delineation. Region of interest (region of interest) delineation is the first and most basic step of image preprocessing technology. Its role is to delineate the localization of the image associated with the target task (such as lanes, traffic lines), eliminate irrelevant areas (such as the sky, distant buildings), to achieve the purpose of reducing the subsequent computational load, reducing the interference of irrelevant noise, and improving the efficiency and stability of subsequent image processing.

The division of region of interest is based on the a priori knowledge of the scene and the target features. Commonly used methods can be roughly divided into two categories in principle. One class is to set the division range based on the target's typical positional feature setting in the image. For example, in highway scenes, lane lines are mostly located in the lower half of the image, so according to this experience, usually delimit the ROI to the bottom 60%-80% region of the image [6]. The other category is based on target features. Combining the color, width and other attributes of the target, the target region is separated by image segmentation techniques (e.g., threshold segmentation, region growing). For example, using the difference in gray value between the lane line and the road surface, the region where the lane line is located is extracted by threshold segmentation to realize the accurate positioning of the ROI.

2.2.2 Image Graying. Image grayscaling refers to the conversion of color images into grayscale images through specific calculations and achieves the purpose of simplifying image information and reducing the number of operations. The weighted average method is one of the most commonly used image grayscaling algorithms, which is based on the sensitivity of the human eye to different colors, giving different weights to the R, G and B channels, and calculating the weighted average to get the grayscale value. The commonly used weight combination is R:G: B=0.299:0.587:0.114, and other grayscaling methods such as the maximum value method, the minimum value method and the average value method can be used in the image grayscaling process.

2.2.3 Image Filtering. In realistic driving scenarios, images acquired by cameras often have a lot of noise due to various complex environmental factors. Image filtering can suppress the noise of the image, while retaining the target's edges, texture and other key features. Through the filtering operation, the image quality can be improved to support subsequent analysis (e.g., target detection, segmentation). Common image filtering methods include mean filtering, median filtering and Gaussian filtering. Mean filtering is the simplest filtering method, which directly takes the average value of the pixels in the image domain as the value of the current pixel; median filtering is a nonlinear filtering method, which takes the median value of the pixels in the neighborhood as the value of the current pixel and has better effect on pretzel noise processing [7]; Gaussian filtering is a linear filtering method, which uses the Gaussian function to obtain the value of the current pixel by taking the weighted average of the pixels in the neighborhood. Value of the current pixel.

2.2.4 Edge Detection. Edge detection is an image processing technique to extract the gray scale or color drastically changed areas in the image (such as object contours, texture boundaries), which provides basic structural information for target recognition, image segmentation, etc. Commonly used methods include the Sobel algorithm. Commonly used methods include the Sobel algorithm, Canny algorithm, etc. [8]. The Sobel algorithm is based on the first-order derivatives and calculates the horizontal and vertical gradient magnitude and direction through two 3x3 convolution kernels, according to which the edge points are determined. The Canny algorithm is based on the second-order derivatives, and smoothes the image by Gaussian filtering before calculating the gradient; it eliminates the spurious response through non-maximum value suppression (retaining the local maximum value). Combined with dual-threshold detection (high and low threshold distinction) to extract the real edge points.

3 Intelligent Driving Core Tasks

Machine vision mainly plays the role of environment perception in an intelligent driving system, and its core task is to transform a 2D visual signal into structured semantic information of the 3D environment, which mainly includes four categories: target detection and classification, semantic segmentation, depth estimation and 3D reconstruction, and target tracking. The technical evolution of machine vision in the core tasks directly determines the perception accuracy and robustness of the automatic driving system.

3.1 Target Detection

Target detection, as one of the core tasks of intelligent driving, aims to realize real-time recognition of the categories and locations of traffic participants (vehicles, pedestrians, bicycles) and traffic facilities (signals, signs, guardrails), which is the basis for realizing intelligent driving. Traditional detection methods have relied on manual features, such as HOG+SVM. However, this method has only 72% accuracy on the MIT pedestrian dataset and is sensitive to occlusion [9]. With the rapid development of deep learning, target detection has realized a breakthrough compared to traditional methods: YOLOv8 adopts CSPDarknet53 backbone network with PAN-FPN feature fusion, which achieves 53.9% mAP on the COCO dataset with a frame rate of 140FPS to meet real-time demand [9]; DETR realizes end-to-end target detection with Transformer encoder-decoder structure. DETR realizes end-to-end detection through the transformer encoder-decoder structure, and the recognition rate of small targets (e.g., distant speed limit signs) is 18% higher than that of YOLOv5 [10]. For the low-light environment, YOLO-LLTS optimizes feature extraction through a priori guided enhancement (PGFE) module, and the mAP50 reaches 78.3% on the TT100K-night dataset, which is 1.9% higher than that of the traditional method, and effectively solves the problem of detecting traffic signs at night [11].

3.2 Semantic Segmentation

The value of semantic segmentation is in parsing out the key elements of road structure, such as lane lines, drivable areas, curbs, etc., through pixel-level classification, which

provides a key basis for path planning. This is exactly what is needed for car navigation. Specifically on the application of the model, DeepLabv3 + is cleverly designed. It adopts the method of empty space pyramid pooling (ASPP) to fuse multi-scale features, which makes his performance on the Cityscapes dataset shine, and the segmentation of 19 types of road elements can be accurately identified, with an mIoU of 82.1% [12]; SegNeXt optimizes the efficiency of the interaction between features through dynamic convolution, and from the actual situation, under the same amount of computation SegNeXt optimizes the interaction efficiency between features through dynamic convolution [12]; in practical terms, with the same amount of computation, SegNeXt improves the mIoU by 2.8% compared to SegFormer, which means that it is more suitable for in-vehicle low-computing-power platforms [13]. In lane line detection, the Canny algorithm improves the edge localization accuracy by 15%-20% compared with the Sobel algorithm through Gaussian smoothing (σ=0.5), gradient computation, and non-maximal value suppression, which is widely used in the lane keeping system of mass-produced models [14].

3.3 Depth Estimation and 3D Reconstruction

Depth estimation and three-dimensional reconstruction are the core technologies to make machines "understand" the three-dimensional world from two-dimensional images. These two technologies are like putting "spatial perception eyes" on the intelligent body, which can transform flat images into three-dimensional spatial information. Let's look at depth estimation first. The monocular depth estimation model, BTS, is cleverly designed to extrapolate depth through an encoder-decoder structure, achieving a relative error (δ<1.25) of 87.6% on the KITTI dataset, but it needs to be combined with the ground plane assumption to calibrate the absolute scale [15], or else the calculated distances may be inaccurate. In contrast, the binocular stereo matching algorithm SGM takes another route. It optimizes the parallax calculation through dynamic planning, and the parallax error can be controlled within 1 pixel on the Middlebury dataset, which makes it very suitable for 0-50m close range obstacle detection [16]. Recently, the popular BEV (Bird's Eye View) perception technology models the surrounding environment as a 360-degree stereo model through multi-camera image projection. For example, the LSS (Lift-Splat-Shoot) algorithm achieves 45.3% AP for 3D detection on the nuScenes dataset [17]. BEVFormer introduces temporal attention to merge historical frame information, which further improves the prediction accuracy of dynamic targets by 18% [18].

3.4 Target Tracking

Target tracking realizes dynamic target trajectory prediction through temporal association, and supports the decision-making of lane changing, following cars, etc. DeepSORT combines appearance feature embedding (CNN extracts 128-dimensional features) and Kalman filtering, which reduces the ID switching rate by 45% on the MOT16 dataset, and can effectively track vehicles and pedestrians in consecutive frames [19]; for the occlusion scenario For the occlusion scene, OcclusionFusion estimates the motion state of the occluded area by graph neural network, and reduces the trajectory prediction error of large vehicles occluding pedestrians by 42% in the CVPR 2022 evaluation.

4 Typical Application Scenarios of Intelligent Driving with Machine Vision

For the application of the above core tasks in intelligent driving, it is necessary to match the technical solutions with the characteristics of the scene to form a closed-loop support from perception to decision-making.

4.1 Perception of Complex Intersections on Urban Roads

The environment of urban intersections has always been more complex, not only needing to pay attention to multiple types of dynamic targets, such as pedestrians crossing, non-motorized vehicles, etc., but also paying attention to the static elements (signals, crosswalks), which is a great test for the perception ability of intelligent driving. Baidu Apollo's response is more comprehensive, it adopts the "target detection + semantic segmentation + BEV fusion" program. Specifically, the front-view camera uses YOLOv8 to detect the signal status, such as whether the light is red or green; the surround-view camera segments the crosswalk area with DeepLabv3 + segmentation; and finally BEVFormer fuses the multiview features to generate a 360° environment model, which has a power of 98.3% for unprotected left turns at intersections on the OpenLane-V2 dataset [20], which is a good performance in the scenario of complex environmental intersections.

4.2 Highway Pilot Assistance

Compared with the complex intersections of urban roads, the requirements of highway scenarios for intelligent driving are more specific, focusing on the higher requirements for long-distance target detection and lane line tracking. The strategy of Tesla FSD in this field is quite targeted, through the combination of front-view camera (FOV 50°) and BEV sensing, first using YOLOv8 to detect the front car within 150m, and at the same time, relying on Canny's algorithm to extract the edge of the lane line, so as to realize the automatic control of lane centering and following distance. In terms of actual performance, the takeover rate of this system in North American high-speed scenarios is less than 0.1 times / thousand kilometers, which is a good performance. For rainstorms, the system optimizes the target confidence through multi-frame feature accumulation, which reduces the leakage rate by 35% compared with single-frame detection [3].

4.3 Automatic Parking System

Automatic parking relies on the fine perception of the near environment. The surround view camera (FOV 120°-150°) generates a panoramic view through fisheye distortion correction and image stitching, combined with the SGM algorithm to estimate the distance to the parking space, and the semantic segmentation model to recognize the parking space lines and obstacles, and ultimately realizes the success rate of entering a non-standard parking space (e.g., diagonal parking space) up to 95.6% [21].

4.4 Multi-sensor Fusion to Enhance Robustness

Relying on vision sensors alone may not be as reliable in extreme environments, such as fogging up the lens in heavy rain or overexposing the image under bright light, and

the performance will be greatly reduced at these times. Multi-modal fusion is now recognized as an inevitable technological development trend in the industry, and the MV3D algorithm can significantly improve its performance by fusing visual features with the LiDAR point cloud. In rainy and snowy weather, the point cloud noise can be reduced by 60%, and the AP value of target detection using this scheme is improved by 25% compared to using only visual sensors [22].

5 Conclusions

Machine vision is the core technology of intelligent driving environment perception, and its optimization of the whole process technology from image acquisition to processing and analysis is the foundation to support the development of intelligent driving to a higher order, while the algorithm iteration of target detection, semantic segmentation and other core tasks provides a key path to improve the accuracy of environment perception. At present, although there are a variety of technical solutions applied to the actual scene, such as Baidu Apollo's intersection sensing system, Tesla's high-speed pilot program, but there are still obvious limitations: the robustness of extreme weather is insufficient, the performance of the visual sensors in the rain, snow and bright light environment is significantly degraded, and the adaptability of most of the systems to the complex occlusion of the scene is limited. On the basis of the existing perception technology, how to realize the depth of machine vision algorithms and hardware adaptation and full-scene adaptive, is still the core direction of future research. At present, although the visual perception based on deep learning has made breakthroughs in specific scenes, the real-time response mechanism to dynamic lighting and sudden occlusion is still imperfect, and only part of the parameters can be dynamically adjusted in the structured scene. There is still a gap between this and the demand for full-scene autonomy of intelligent driving. In the future, how to optimize the robustness of the algorithm based on the characteristics of the visual data and improving the adaptability of the hardware environment will still face many challenges, and we need to continue to explore breakthroughs.

References

1. Li, K., Dai, Y., Li, S., et al.: Development Status and Trends of Intelligent Connected Vehicle (ICV) Technology. Journal of Automotive Safety and Energy 8(1), 1 (2017)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
3. Xiao, Y., Yang, H.: A Review of Object Detection Algorithms in Traffic Scenarios. Journal of Computer Engineering & Applications 57(6) (2021)
4. Zhang, G.: Machine Vision. Science Press, Beijing (2005)
5. Redmon, J., Divvala, S., Girshick, R., et al.: You Only Look Once: Unified, Real-Time Object Detection. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788. IEEE (2016)
6. Huang, A.S., Moore, D., Antone, M., et al.: Finding Multiple Lanes in Urban Road Networks with Vision and Lidar. Autonomous Robots 26(2), 103–122 (2009)

7. Qatawneh, M., Massad, Y., Musaddaq, M., et al.: A Uniform Noise Median Filter Based on a New Type of Filtering Window. Information Journal 15(2), 699–706 (2012)
8. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Proc. of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893. IEEE (2005)
9. Swathi, Y., Challa, M.: YOLOv8: Advancements and Innovations in Object Detection. In: International Conference on Smart Computing and Communication, pp. 1–13. Springer, Singapore (2024)
10. Carion, N., Massa, F., Synnaeve, G., et al.: End-to-End Object Detection with Transformers. In: European Conference on Computer Vision, pp. 213–229. Springer, Cham (2020)
11. Sun, X., Liu, K., Chen, L., et al.: LLTH-YOLOv5: A Real-Time Traffic Sign Detection Algorithm for Low-Light Scenes. Automotive Innovation 7(1), 121–137 (2024)
12. Chen, L.C., Papandreou, G., Kokkinos, I., et al.: Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. on Pattern Analysis and Machine Intelligence 40(4), 834–848 (2017)
13. Geng, Q., Zhou, Z., Cao, X.: Survey of Recent Progress in Semantic Image Segmentation with CNNs. Science China Information Sciences 61(5), 051101 (2018)
14. Canny, J.: A Computational Approach to Edge Detection. IEEE Trans. on Pattern Analysis and Machine Intelligence 6, 679–698 (2009)
15. Eigen, D., Puhrsch, C., Fergus, R.: Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. Advances in Neural Information Processing Systems 27 (2014)
16. Hirschmuller, H.: Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Trans. on Pattern Analysis and Machine Intelligence 30(2), 328–341 (2007)
17. Philion, J., Fidler, S.: Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In: European Conference on Computer Vision, pp. 194–210. Springer, Cham (2020)
18. Li, Z., Wang, W., Li, H., et al.: Bevformer: Learning Bird's-Eye-View Representation from Lidar-Camera via Spatiotemporal Transformers. IEEE Trans. on Pattern Analysis and Machine Intelligence (2024)
19. Wojke, N., Bewley, A., Paulus, D.: Simple Online and Realtime Tracking with a Deep Association Metric. In: Proc. of 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649. IEEE (2017)
20. Huang, X., Cheng, X., Geng, Q., et al.: The Apolloscape Dataset for Autonomous Driving. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 954–960. IEEE (2018)
21. Jiang, H., Shen, Z., Ma, S., et al.: Intelligent Recognition of Parking Spaces in Automatic Parking Systems Based on Information Fusion. Journal of Mechanical Engineering 53(22), 125–133 (2017)
22. Chen, X., Ma, H., Wan, J., et al.: Multi-View 3D Object Detection Network for Autonomous Driving. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1907–1915. IEEE (2017)