Automation and Machine Learning (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2025.060206
ISSN 2516-5003 Vol. 6 Num. 1

Review on Machine Vision-Based Detection of Pedestrians and Non-Motorized Vehicles in Autonomous Driving

Xueju Hao

School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, China
1450946940@qq.com

Keywords: Machine Vision; Autonomous Driving; Vulnerable Road Users; Pedestrian Detection; Non-Motorized Vehicle Detection; Sensor Fusion; KITTI Dataset

Abstract: Vulnerable Road Users (VRUs), including pedestrians, bicycles, and electric bikes, are the primary targets for collision risk prevention in autonomous driving systems due to their random motion and weak protection. Machine vision, as a core environmental perception technology, enables real-time detection and early warning of VRUs, which is crucial for the safety of autonomous driving. This paper focuses on the application of machine vision in VRU detection, systematically elaborates on the technical logic of detection and early warning, and emphasizes the characteristics and application values of mainstream datasets such as KITTI and Waymo Open Dataset. It deeply analyzes the detection bottlenecks in complex scenarios like nighttime driving and fast-moving pedestrians, and proposes corresponding technical optimization paths. Additionally, the integration ideas of machine vision and radar sensors are briefly discussed to improve the robustness of the detection system. The research shows that deep learning models (e.g., YOLOv8, DETR) and multi-sensor fusion technologies effectively enhance the accuracy and reliability of VRU detection. This review provides a comprehensive technical reference for the performance improvement of environmental perception systems in autonomous driving.

1. Introduction

As a core component of the intelligent transportation system, the safety performance of autonomous driving technology directly determines the feasibility and popularization of its practical application. In complex road environments, vulnerable road users (VRUs) such as pedestrians, bicycles, and electric bikes have become the key targets for collision risk prevention in autonomous driving systems due to their strong randomness in movement trajectories, variable morphological features, and weak protection capabilities. According to statistics from the World Health Organization (WHO), over 70% of casualties in global road traffic accidents are related to VRUs, among which accidents in scenarios such as nighttime driving and sudden road crossing account for more than 45%. Endowed with advantages including low cost, rich information acquisition

dimensions, and comprehensive environmental perception, machine vision technology realizes real-time detection, tracking, and risk early warning of VRUs by collecting road images via cameras and integrating image processing, deep learning, and other algorithms, thus serving as a core supporting technology for the environmental perception module of autonomous driving systems.

Focusing on three typical types of VRUs—pedestrians, bicycles, and electric bikes—this paper systematically sorts out the core technical logic of machine vision in detection and early warning links, highlights the characteristics and application values of mainstream datasets such as KITTI and Waymo Open Dataset, deeply analyzes the detection bottlenecks and technical optimization paths in complex scenarios like nighttime driving and fast-moving pedestrians, and briefly discusses the integration application ideas of machine vision and radar sensors. It aims to provide technical references for improving the performance of environmental perception systems in autonomous driving.

2. Core Technical System of Machine Vision Detection and Early Warning

2.1 Evolution of Detection Technologies

The essence of machine vision-based detection of VRUs lies in realizing target localization and category determination through image feature extraction and pattern matching, and its technical development has experienced a leap from traditional methods to deep learning-based approaches. Traditional detection methods rely on the combination of manually designed features and shallow classifiers, among which the hybrid model of Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) is the most typical. It constructs feature vectors by calculating the gradient direction distribution in local image regions and completes classification using SVM. This method can achieve an accuracy of 75% in pedestrian detection on structured roads, but it has poor adaptability to targets with large morphological differences such as bicycles and electric bikes, and is vulnerable to environmental factors like illumination and occlusion.

The rise of deep learning technology has brought a qualitative leap in detection performance, and end-to-end detection models based on Convolutional Neural Networks (CNNs) have become mainstream. Two core technical paths each have their own advantages: One is the two-stage detection model represented by Faster R-CNN, which generates candidate target regions through a Region Proposal Network (RPN), then extracts features via CNN to complete classification and regression[2]. It exhibits high detection accuracy for small targets such as child pedestrians and folding bicycles, achieving an Average Precision (AP) of 89.2% in pedestrian detection on the Waymo Open Dataset. The other is the one-stage detection model represented by YOLO and SSD, which integrates target localization and classification tasks into a single regression problem, significantly improving detection speed[1]. The YOLOv8 model can achieve a pedestrian detection AP of 88.5 while maintaining an inference speed of 300 FPS, meeting the real-time requirements of autonomous driving. In addition, aiming at the diverse morphologies of non-motorized vehicles, Transformer-based models such as DETR effectively improve the discrimination accuracy between bicycles and electric bikes by means of global feature modeling, solving the problem that traditional models tend to confuse these two types of targets[3].

2.2 Construction and Implementation of Early Warning Logic

The machine vision-based early warning system constructs hierarchical early warnings based on detection results, combined with target motion state analysis and risk assessment models. Firstly, it acquires dynamic information such as the target's speed, acceleration, and motion trajectory through target tracking across multi-frame image sequences (e.g., Kalman filtering, Hungarian algorithm).

Then, based on road structures (e.g., lane lines, zebra crossings) and traffic rules, it establishes a risk assessment index system, including the distance between the target and the host vehicle, relative speed, Time-to-Collision (TTC), and conflict probability at intersections. When the indicators exceed the preset thresholds, the system triggers hierarchical warnings: Level 1 warning (low risk) prompts the driver to pay attention to the target via the in-vehicle display; Level 2 warning (medium risk) activates audio-visual alarms and slightly adjusts the vehicle speed; Level 3 warning (high risk) triggers emergency braking or evasive steering. The entire early warning response time needs to be controlled within 100 ms to reserve sufficient time for safety decision-making.

In typical application scenarios, when the system detects a pedestrian's intention to cross the road at a zebra crossing, it judges the time when the pedestrian will reach the lane through a trajectory prediction model, and triggers a Level 2 warning if the TTC is less than 3 seconds. If an electric bike suddenly moves from the non-motorized lane to the motorized lane with a relative speed exceeding 20 km/h, the system immediately initiates a Level 3 warning and links with the braking system. This closed-loop logic of "detection-tracking-prediction-warning" realizes seamless connection from environmental perception to safety decision-making.

3. Core Datasets and Benchmark Test Analysis

3.1 Characteristics and Applications of the KITTI Dataset

As a classic benchmark dataset in the field of autonomous driving environmental perception, the KITTI dataset, released by the Karlsruhe Institute of Technology in Germany, plays an irreplaceable role in the detection tasks of pedestrians and non-motorized vehicles[4]. Collected based on real road scenarios, this dataset includes 389 pairs of stereo image sequences and 155 LiDAR point cloud sequences, covering various scenarios such as urban roads, rural roads, and highways. Among them, there are 12,496 annotated pedestrian samples, 3,381 bicycle samples, and 1,927 electric bike samples. Equipped with rich annotation information including target bounding boxes, occlusion levels (0-3), and motion states (stationary/moving), it provides comprehensive support for the robustness testing of algorithms.

In the evaluation of detection algorithms, the KITTI benchmark test is divided into three levels—easy, moderate, and hard—corresponding to scenarios with different occlusion levels and target sizes. For example, in the moderate-difficulty test, the AP of the traditional HOG+SVM model for pedestrian detection is only 52%, while that of the YOLOv7 model increases to 78%, demonstrating the advantages of deep learning algorithms. However, this dataset has limitations: the proportion of nighttime scenario samples is less than 10%, and the number of electric bike samples is relatively small, which cannot meet the algorithm training needs for complex nighttime environments.

3.2 Breakthroughs and Values of the Waymo Open Dataset

Released by Waymo, a subsidiary of Google, the Waymo Open Dataset is one of the largest and most scenario-diverse autonomous driving datasets currently available, effectively making up for the deficiencies of the KITTI dataset[5]. It contains 12 million frames of images and 1.2 TB of LiDAR data, covering road scenarios in 25 cities, with a total of over 1 million VRU samples, including various target types such as pedestrians (adults, children, the elderly), bicycles, electric bikes, and electric tricycles. Compared with KITTI, its outstanding advantages are reflected in three aspects: First, scenario diversity. It includes samples of harsh environments such as heavy rain, dense fog, and nighttime, with nighttime scenarios accounting for 35%, providing sufficient data for the training of nighttime detection algorithms. Second, high annotation accuracy. It adopts a multi-

sensor fusion annotation method, with the error of target bounding boxes less than 1 pixel, and adds "target behavioral intention" annotations (e.g., whether a pedestrian intends to cross the road). Third, complete time series. Each scenario contains continuous 20-second image sequences, facilitating research on target tracking and trajectory prediction.

The benchmark test system of the Waymo Open Dataset is more comprehensive. In addition to the traditional AP indicator, it also introduces evaluation indicators such as "detection accuracy in extreme scenarios" and "long-distance tracking stability". For example, in the nighttime scenario without street lights, the AP of the detection model based on multi-spectral image fusion reaches 72%, while that of the model relying solely on visible light images is less than 50%. This dataset provides a quantitative evaluation basis for the performance optimization of algorithms in extreme scenarios. Currently, the Waymo Open Dataset has become a core training and testing platform for major autonomous driving enterprises and research institutions worldwide, promoting the adaptation of detection algorithms to complex practical scenarios.

4. Core Detection Challenges in Complex Scenarios

4.1 Detection Bottlenecks and Optimization in Nighttime Driving

Nighttime driving is a typical challenging scenario for machine vision detection, and its core problems stem from the degradation of image quality caused by insufficient illumination. First, the signal-to-noise ratio of visible light images is low, the contrast between targets and the background decreases, and features such as pedestrian clothing and non-motorized vehicle bodies become blurred, leading to easy missed detections by traditional detection algorithms based on texture features. Second, direct or reflected vehicle lights cause glare interference, and image pixels in high-glare areas are saturated, resulting in the loss of target details. For instance, the reflection of electric bike headlights may obscure the body contour. Third, the behaviors of pedestrians and non-motorized vehicles at night are more random, such as pedestrians not wearing reflective marks and electric bikes illegally using high beams, which further increases the difficulty of detection. Data shows that the missed detection rate of pedestrians in nighttime scenarios is 3-5 times higher than that in the daytime, which is the main cause of autonomous driving accidents at night.

To address the above issues, the technical optimization paths mainly focus on three directions: First, the upgrading of image preprocessing technology. The Contrast-Limited Adaptive Histogram Equalization (CLAHE) is used to enhance the local contrast of images, and the dark channel defogging algorithm is combined to eliminate the impact of light scattering, improving the clarity of target features in nighttime images by 40%. Second, multi-spectral image fusion. The data from visible light cameras and infrared cameras are fused. Infrared images can capture the temperature features of targets without being restricted by illumination conditions. The AP of the fused model for pedestrian detection in nighttime scenarios reaches 81%, which is 25% higher than that of the single visible light model. Third, the design of dedicated network structures. For example, the NightNet model realizes fast target localization by introducing a nighttime feature attention module that automatically focuses on highlight areas in images (e.g., pedestrian reflective strips, non-motorized vehicle taillights), reducing the missed detection rate to below 8% on the KITTI nighttime test set.

4.2 Detection Difficulties and Solutions for Fast-Moving Pedestrians

The detection challenges in scenarios involving fast-moving pedestrians (e.g., running, crossing the road suddenly) are mainly reflected in two aspects: motion blur and abrupt trajectory changes. When the moving speed of pedestrians exceeds 5 m/s, the images collected by the camera will have

obvious motion blur, and the target bounding boxes will be deformed. Traditional CNN models find it difficult to extract effective features, leading to an increase in classification error rates. At the same time, the trajectories of fast-moving pedestrians are highly random, and traditional tracking algorithms based on linear prediction (e.g., Kalman filtering) tend to have trajectory drift and cannot accurately predict collision risks. In intersection scenarios, the missed detection rate of fast-running pedestrians can reach 15%, which is a major safety hazard for autonomous driving systems.

The technical breakthrough directions focus on motion modeling and network optimization: First, the introduction of optical flow estimation technology. By calculating the pixel motion vectors of adjacent frame images, the motion information of targets is obtained, and image deblurring is realized by combining motion blur kernel estimation, which improves the accuracy of feature extraction for fast-moving pedestrians by 30%. Second, the adoption of dynamic target detection networks. For example, the Motion-YOLO model adds a motion feature branch to the traditional YOLO, fusing optical flow features with appearance features to realize accurate modeling of motion states, and achieves an AP of 85% on the Waymo fast-moving pedestrian test set. Third, the optimization of trajectory prediction algorithms. Temporal models based on LSTM or Transformer can capture the non-linear features of pedestrian movement and predict the trajectory within the next 2 seconds, reducing the prediction error by 50% compared with linear models and providing more sufficient response time for the early warning system.

4.3 Analysis of Other Typical Challenges

In addition to nighttime and fast-moving scenarios, occlusion and variable morphologies are also common detection difficulties. In occlusion scenarios (e.g., pedestrians blocked by vehicles, bicycles and electric bikes traveling side by side), the target features are incomplete, and traditional models tend to misclassify partially occluded targets as the background. By introducing a context-aware module and using the correlation between the road environment (e.g., sidewalks, non-motorized lanes) and targets for auxiliary judgment, the AP of detection in occlusion scenarios can be increased by 18%. For the problem of variable morphologies of non-motorized vehicles (e.g., bicycles carrying people, electric bikes equipped with sunshades), lightweight semantic segmentation networks are used to realize complete extraction of target contours, and the recognition accuracy of electric bikes in morphologically variable scenarios reaches 89% by combining part-level feature (e.g., wheels, handlebars) recognition to improve the adaptability of the model to deformed targets. Meanwhile, Deformable DETR, an improved version of the DETR model, solves the problem of insufficient detection accuracy of the original DETR for small and occluded targets by introducing a deformable attention mechanism, which can further enhance the model's ability to handle complex occlusion scenarios in VRU detection[7].

5. Integration Application of Machine Vision and Radar Sensors

5.1 Core Values and Technical Paths of Integration

The integrated application of machine vision and radar sensors (LiDAR, millimeter-wave radar) is an effective means to solve the limitations of a single sensor. Machine vision has advantages in target category recognition and semantic understanding, but it is insufficient in distance measurement accuracy and adaptability to harsh weather. Radar sensors can accurately obtain the 3D coordinates, speed, and other distance information of targets, and are not affected by illumination and weather, but they cannot realize fine category discrimination. The integration of the two can achieve a "1+1>2" effect and improve the robustness and reliability of the detection system.

The integration technical paths are divided into three levels: Data-level fusion registers and fuses the data from cameras and radar point clouds at the raw data level. It projects radar point clouds to the image pixel coordinate system through coordinate transformation to generate fused feature maps rich in texture and distance information. This method retains the most complete raw information but has a large data processing load. Feature-level fusion extracts visual features from images and distance features from radar separately, adaptively adjusts the weights of the two types of features through the attention mechanism, and then inputs them into the classification and regression network. This method balances feature integrity and processing efficiency and is currently the mainstream integration method. Decision-level fusion evaluates the confidence of the visual detection results and radar detection results respectively, and fuses the two types of results through Bayesian inference or D-S evidence theory to output the final detection conclusion. This method has strong fault tolerance and is suitable for scenarios with extremely high safety requirements.

5.2 Application Cases of Typical Integrated Systems

Tesla's Autopilot system adopts an integration scheme of "vision + millimeter-wave radar". The millimeter-wave radar is responsible for detecting the distance and speed of targets, and the camera completes target category recognition and trajectory prediction. In intersection scenarios, the accuracy of pedestrian detection after integration is 22% higher than that of the single vision system, effectively reducing the risk of misjudgment. Waymo's autonomous driving system adopts a multi-sensor integration architecture of "vision + LiDAR + millimeter-wave radar". The 3D point clouds generated by LiDAR can accurately locate the spatial positions of non-motorized vehicles, and the texture features of visual images are combined to realize fine category recognition (e.g., distinguishing between electric bikes and electric tricycles). In heavy rain scenarios, the missed detection rate of the system for VRUs is less than 3%, which is much better than the single-sensor scheme.

The key technologies of the integrated system lie in sensor calibration and time synchronization. The Zhang Zhengyou calibration method is used to realize the spatial coordinate calibration of cameras and radars, with the error controlled within 2 cm. The time stamp synchronization technology is adopted to ensure that the time deviation of multi-sensor data is less than 10 ms, avoiding detection deviations caused by unsynchronized data. With the decrease in sensor costs and the optimization of integration algorithms, multi-sensor integration has become a standard configuration for mid-to-high-end autonomous driving systems.

6. Future Development Directions and Prospects

6.1 Optimization Directions of Algorithms and Models

In the future, machine vision detection algorithms will develop in the direction of "high accuracy, high efficiency, and strong robustness". In terms of model structure, lightweight networks (e.g., MobileViT, EfficientNet) will become mainstream[6]. Through technologies such as channel pruning and quantization, the inference speed of the model is increased by more than 50% while ensuring detection accuracy, meeting the deployment requirements of embedded devices. In terms of training methods, federated learning technology can realize the joint training of multi-institutional data, solve the problems of data privacy and insufficient samples, and improve the adaptability of the model to road scenarios in different regions. In terms of target modeling, data augmentation technology based on Generative Adversarial Networks (GANs) can generate extreme scenario samples (e.g., nighttime heavy rain, sudden pedestrian falls), further enhancing the generalization ability of the model.

6.2 In-depth Application of Multi-Technology Integration

In addition to integration with radar sensors, machine vision will be deeply combined with more technologies. Vehicle-infrastructure cooperation technology realizes the detection of VRUs in road blind spots through data interaction between roadside cameras and on-board cameras, expanding the perception range. Digital twin technology can build virtual road scenarios to conduct large-scale simulation tests on detection algorithms, reducing the cost of real-vehicle tests. The application of large artificial intelligence models can realize end-to-end reasoning from target detection to behavioral intention understanding. For example, the combination of GPT-4V and detection models can judge whether a pedestrian intends to cross the road through image semantic analysis, improving the forward-looking nature of early warning.

7. Conclusion

Machine vision technology plays a core role in the detection of pedestrians and non-motorized vehicles in autonomous driving. Through the technical logic of "detection-tracking-prediction-warning", it realizes comprehensive perception and risk prevention of VRUs. Datasets such as KITTI and Waymo Open Dataset provide important support for the research, development, and testing of algorithms. The detection challenges in complex scenarios such as nighttime driving and fast-moving pedestrians have promoted the continuous breakthrough of technologies such as multi-spectral fusion and motion modeling. The integrated application of machine vision and radar sensors has further improved the robustness and reliability of the system, laying a foundation for the safe implementation of autonomous driving.

In the future, with the optimization of algorithms, the integration of multiple technologies, and the improvement of industry standards, the performance of machine vision detection systems will continue to improve. They will not only accurately respond to various complex road scenarios but also realize in-depth understanding of target behavioral intentions, providing key support for the leap of autonomous driving technology from Level 2 to Level 4. At the same time, attention should be paid to safety, privacy, and compliance in the process of technological development to promote the coordinated development of autonomous driving technology and urban transportation systems, and ultimately achieve the goal of "zero-accident" intelligent transportation.

References

[1] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arxiv preprint arxiv:1804.02767 (2018).
[2] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." IEEE transactions on pattern analysis and machine intelligence 39.6 (2016): 1137-1149.
[3] Carion, Nicolas, et al. "End-to-end object detection with transformers." European conference on computer vision. Cham: Springer International Publishing, 2020.
[4] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite." 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012.
[5] Sun, Pei, et al. "Scalability in perception for autonomous driving: Waymo open dataset." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[6] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arxiv preprint arxiv:1704.04861 (2017).
[7] Xizhou Zhu, et al. "Deformable detr: Deformable transformers for end-to-end object detection." arxiv preprint arxiv:2010.04159 (2020).