International Core Journal of Engineering Volume 11 Issue 12, 2025
ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

Progress in Vehicle Recognition Methods based on Machine Vision

Huiru Dai[a], Xinyu Wu[b], Xiaoxiao Niu[c], and Jingyuan He[d, *]

School of Mathematics and Computer Science, Yan'an University, Yan'an 716000, China

[a]3203859110@qq.com, [b]3446405413@qq.com, [c]3184819430@qq.com,

[d, *]18992118537@163.com

Abstract

With the continuous growth of car drivers, the problems and needs on the road increase, and vision-based intelligent transportation technology becomes more and more important. Vehicle recognition algorithms based on machine vision face many challenges in the field of intelligent transportation, such as complex visual scenes during driving, conventional deep learning models are greatly affected by environmental factors, and limited memory and computing power of on-board embedded devices. In order to deeply analyze the application and research status of deep learning networks in vehicle recognition, the application of various current object detection methods in vehicle recognition is first introduced, and then the practical application status of real-time detection algorithms is described in detail. Finally, the advantages and disadvantages of these algorithms are compared and analyzed.

Keywords

Machine Vision; Vehicle Recognition; Feature Analysis; Convolutional Neural Networks; Multimodal Fusion.

1. Introduction

With the rapid growth of global automobile ownership, road traffic conditions have become complex, posing challenges to traffic safety and traffic management. Intelligent Transportation Systems (ITS) have emerged as a key means to improve traffic management and enhance road safety, and machine vision-based vehicle recognition technology is an important component of ITS. It provides real-time traffic information and decision support through automated image analysis. Deep learning is a critical technology for machine vision and vehicle recognition. Common models include Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), which can automatically learn image features and perform classification and recognition. Machine vision algorithms are also important, such as feature extraction and object detection algorithms, which can extract feature information to achieve vehicle detection and tracking. However, vehicle recognition faces difficulties in practical applications. The complex driving environment-including lighting changes, weather conditions, occlusions, and vehicle appearance variations-increases recognition complexity. Additionally, the computational resource constraints of on-board embedded devices (such as limited memory and processing capacity) must be considered. Furthermore, the development of autonomous driving and Internet of Vehicles (IoV) has set higher standards for its speed and accuracy. Machine vision-based vehicle recognition plays a significant role in fields such as traffic management, security monitoring, and autonomous driving. Examples of its applications include traffic flow monitoring, traffic violation detection, vehicle tracking and recognition in surveillance videos, and perception and decision-making for autonomous vehicles.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

2. Research Background of Vehicle Detection

2.1 Difficulties in Vehicle Detection

Vehicle recognition faces numerous challenges. Its development is influenced by various factors (such as environment, vehicle itself, and technical limitations) and requires real-time performance, as detailed below:

2.1.1 Environmental Factors

Lighting changes: Lighting conditions vary significantly across different times, weather conditions, and environments. Strong light can overexpose vehicle images, causing loss of some details, while weak light leads to blurred images. Both scenarios increase the difficulty of accurately identifying vehicle features.

Weather impacts: Severe weather like rain, snow, and fog degrades the quality of vehicle images. Raindrops can block parts of the vehicle, and dense fog reduces image contrast and clarity, interfering with the recognition system's extraction of vehicle contours and features.

2.1.2 Vehicle-Specific Factors

Similar appearance: Many vehicles are highly similar in shape, color, and other aspects. For example, sedans from different brands may have similar body shapes and contours, making it difficult for recognition systems to accurately distinguish them through simple shape features.

Modified vehicles: Vehicle modifications alter their original appearance features, such as modifying front and rear bumpers, adding roof racks, or changing body colors. This prevents recognition systems from matching the original vehicle templates, increasing the complexity and uncertainty of recognition.

2.1.3 Technical Limitations

Image resolution: Low image resolution fails to clearly present detailed vehicle features. Key information like license plate numbers and vehicle logos becomes blurred, affecting recognition accuracy.

Algorithm limitations: Existing object detection and recognition algorithms may lack robustness in complex scenarios, making it hard to accurately extract and match vehicle features. Additionally, their detection performance for small-target vehicles or partially occluded vehicles may be unsatisfactory.

Real-time requirements: In practical applications such as intelligent transportation systems, real-time recognition of moving vehicles is required, which imposes high demands on computing speed and processing capacity. Completing a series of operations (including image acquisition, processing, feature extraction, and recognition) in a short time faces the challenge of balancing limited computing resources and algorithm complexity.

Since the concept of deep learning was proposed in 2006, it has promoted the development of object detection and recognition. The field of vehicle recognition has advanced rapidly, with numerous research achievements emerging both domestically and internationally, as shown in Table 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

Table 1. Machine Vision-Based Vehicle Recognition Algorithms

Time	Reference	Characteristics
2006	[1]	The ability of CNNs to learn image features has been thoroughly explored, prompting numerous scholars to conduct continuous research in this field. Deep learning-based algorithms have emerged one after another, achieving remarkable results in target detection and recognition.
2015	[2]	The combination of RPN and anchor strategy obtains candidate regions with different scales and aspect ratios. Finally, candidate region generation, feature extraction, and target detection are all integrated into a single deep learning framework, which not only significantly improves the speed of the entire model compared to previous generations but also greatly enhances detection accuracy.
2016	[3]	YOLO9000 trains the network using a new joint training algorithm and classifies objects from a hierarchical perspective, ultimately enabling the recognition of up to 9000 categories. YOLOv3 introduces multi-scale prediction through a feature pyramid network structure and replaces the backbone network with DarkNet53, achieving stronger feature extraction capabilities.
2021	[4]	Excellent data augmentation techniques such as Mosaic and Mixup are adopted. At the prediction end, prediction is performed using a decoupled prediction branch approach, resulting in considerable performance improvements. It changes from predicting three sets of anchors from one feature map to only one set, reducing the number of parameters while achieving better performance.
2011	[5]	Extended Haar-like features are used to describe vehicle characteristics, and these features are extracted from integral images. An improved AdaBoost classifier is then employed for training, which outperforms traditional methods in terms of recognition accuracy and speed.
2016	[6]	A five-layer network is used for feature extraction, and an SVM classifier trains the extracted features to recognize and classify three different types of vehicles, achieving excellent results.
2018	[7]	Image difference methods are used to extract and train vehicle salient features, and a cascaded classifier identifies targets. Block projection matching and image difference technology detect moving targets. The fusion of vehicle salient features and motion features improves the accuracy of target region selection and recognition.
2018	[8]	ResNet-18, ResNet-34, and ResNet-50 are trained and tested on comprehensive automotive datasets respectively. Without any pre-training and only using spatial weighted pooling, the final accuracy is 3.7 percentage points higher than that of traditional CNN methods.
2020	[9]	Algorithms such as YOLOv5, EfficientNet, and ResNet-50 are utilized, offering high accuracy, adaptability to complex scenarios, and support for end-to-end training. However, they require high computing resources and are sensitive to occlusion and lighting conditions.
2019	[10]	Algorithms like Vision Transformer (ViT) and Swin Transformer are applied, boasting strong global feature capture capabilities and excellent long-range dependency modeling. Nevertheless, they demand large amounts of training data and have slow inference speeds.
2022	[11]	Methods such as lidar + camera fusion and infrared + visible light fusion are adopted, enhancing robustness and adaptability to various weather and lighting conditions. However, they involve high hardware costs and complex data synchronization and alignment.
2019	[12]	Algorithms including MobileNetV3 and ShuffleNetV2 are used to achieve low computational overhead, making them suitable for edge devices and embedded systems. Yet, they suffer from accuracy loss and weak small target recognition capabilities.
2020	[13]	Algorithms like CycleGAN and StyleGAN3 are employed, featuring outstanding data augmentation capabilities and generating diverse training samples. However, their training is unstable, and the generated images may contain artifacts.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

2.2 Traditional Vehicle Detection Methods

In the field of machine learning, shallow learning mainly refers to a pattern recognition method based on traditional machine learning algorithms (such as support vector machines, decision trees, K-nearest neighbors, etc.), which extracts features through manual design or statistical methods. Among them, feature recognition based on feature analysis is its core link. It constructs discriminative feature representations by systematically decomposing and describing the features of target data. The following will systematically elaborate on its feature extraction principles and application scenarios from three dimensions: global features, local features, and 3D contour features.

(1) Global Features

Global features are macro representations of overall attributes. They are statistical or structural properties extracted from a global perspective of data and are often used to describe the macro characteristics of targets. These features do not focus on local details but achieve classification or recognition through the statistical laws of global distribution. Global feature analysis describes the overall information of images by extracting vehicle feature information to obtain feature vectors representing the images. Combined with shallow learning methods, it performs vehicle category judgment and prediction. It is usually suitable for tasks such as vehicle color recognition and type recognition with significant differences. Common global features include color, texture, and shape features, among which color is one of the main clues for recognition. Color histograms are widely used. Their characteristic is to take the probability of various colors in an image as features, and this probability is not sensitive to image rotation, translation, and size changes. Advantages of color histograms: In the HSB color space composed of hue, saturation, and brightness, using the H and S components to form a two-dimensional feature vector solves the problem of color feature expression. Considering that the feature information in different color channels has different importance for recognition, in the HSI color space composed of hue, saturation, and intensity, it is necessary to set statistical intervals for each color channel. To reduce the interference of non-vehicle color regions, color histograms can be extracted from different regions to construct feature vectors, achieving better vehicle color recognition performance by directly extracting vehicle color features for recognition. This method can also better adapt to changes in illumination.

The core methods of global features are as follows:

a. Color histogram: Reflects the overall hue and contrast of the target by counting the frequency of color distribution in the image. For example, in image retrieval, a histogram dominated by red can be used to identify flames or flowers.

b. Texture features: Extract attributes such as roughness and directionality of textures based on gray-level co-occurrence matrix (GLCM) or Gabor filters, suitable for material classification (such as distinguishing wood from metal).

c. Shape geometric features: Describe the basic contour shape of objects by calculating parameters such as area, perimeter, and aspect ratio of the target. For example, in industrial inspection, ellipticity is used to determine whether a part is qualified.

Global features have high computational efficiency and strong noise resistance but lack sensitivity to local changes. Typical applications include image classification (such as natural scene classification) and simple object recognition (such as traffic sign detection). For example, in early face recognition systems, principal component analysis (PCA) was used to project face images into a low-dimensional space and extract global features for matching.

(2) Local Features

Local features are refined descriptions of local structures. They focus on key local regions in the data and improve recognition accuracy by capturing highly discriminative microstructures. These features have strong robustness to occlusion, rotation, and scale changes. The method of global feature analysis is insufficient in obtaining detailed information when recognizing vehicles of the same type with small differences, leading to poor recognition accuracy and stability. Therefore, local feature

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

analysis methods need to be adopted for vehicle recognition. The Object Bank-like method proposed by Li et al., with the help of the Deformable Parts Model idea, mines discriminative occlusion patterns from training data, which well reduces the impact of occlusion on vehicle recognition. Local feature analysis first extracts and describes local features, then uses feature transformation algorithms to encode some features, integrates the features to obtain more accurate feature expressions, and thus obtains a suitable feature vector. Finally, an appropriate learning structure is selected to design a classifier. Among them, how to effectively and compactly express local features is the key to improving recognition accuracy.

The core methods of local features are as follows:

a. Edge and corner detection: Uses Canny operator or Harris corner detector to extract key points of target boundaries, often used for image registration and motion tracking.

b. SIFT (Scale-Invariant Feature Transform): Constructs rotation and scale-invariant feature descriptors through multi-scale space extremum detection and direction histogram, widely used in panoramic image stitching.

c. HOG (Histogram of Oriented Gradients): Counts the distribution of gradient directions in local regions, is sensitive to human posture and contour, and is one of the mainstream methods for pedestrian detection.

Local features can effectively handle complex backgrounds and local deformations but have high computational complexity. They perform well in fields such as target tracking (such as SIFT matching UAV aerial images) and medical image analysis (such as pathological section cell localization). For example, in autonomous driving, HOG features combined with SVM classifiers can real-time detect pedestrians on the road.

(3) 3D Contour Features

3D contour features are three-dimensional modeling of spatial structures. By describing the geometric shape and topological structure of targets in 3D space, they break through the perspective limitations of 2D images and are suitable for tasks sensitive to depth information. Currently, limited by acquisition equipment, using 2D images for recognition will encounter many problems, such as changes in geometric shape and spatial position, making it difficult to achieve good recognition results. To improve recognition accuracy, people perform 3D modeling on vehicles in 2D images to obtain 3D data, and then conduct vehicle recognition by training fixed 3D models. Buch et al. used 3D models to extract motion contours and compare them with projected model contours to identify the geographical location and vehicle type, eliminating the impact of vehicle shadows. To a certain extent, 3D modeling methods can solve the viewpoint problem, but fixed 3D vehicle models are generally difficult to distinguish target objects of different shapes. In addition, the links of feature information extraction and model matching will become more complex with the increase in the number of models, making it relatively difficult to implement for fine-grained vehicle recognition tasks with many categories. Therefore, algorithms need to be optimized to improve the discriminative ability and adaptability of the model.

The core methods of 3D contour features are as follows:

a. Point cloud feature extraction: Obtains point cloud data based on 3D lidar or depth cameras, and describes surface concavity and convexity characteristics through normal vector estimation (such as PCA normal vector) or curvature calculation.

b. Depth map analysis: Projects 3D information into a 2D depth map, and extracts contour features using traditional image processing methods (such as edge detection).

c. Geometric descriptors: Uses descriptors such as Spin Image or 3D HOG to encode local 3D shapes, suitable for object recognition (such as mechanical part classification).

3D features can more truly reflect the spatial structure of objects but have high requirements for sensor accuracy and computing resources. Typical applications include industrial robot grasping (locating workpieces based on point cloud matching) and virtual reality (3D model retrieval). For

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

example, in Kinect somatosensory interaction, human body movements are recognized through the 3D coordinates of skeleton joints.

The comparison of the three feature extraction methods for traditional vehicle detection mentioned above is shown in Table 2, which conducts a comparative analysis from core advantages, typical algorithms, and applicable scenarios.

Table 2. Comparison of Feature Extraction Methods for Traditional Vehicle Detection

Feature Type	Core Advantages	Typical Algorithms	Application Scenarios
Global Features	High computational efficiency, global robustness	PCA, Color Histogram	Simple classification, fast retrieval
Local Features	Local invariance, high discriminability	SIFT, HOG	Target tracking, complex scene detection
3D Contour Features	Complete spatial information, stereoscopic perception	Point Cloud Normal Vector, Spin Images	Robot vision, 3D modeling

Feature-based methods once dominated shallow learning, with advantages of interpretable features and low computational cost. However, the limitations of manually designed features (such as reliance on prior knowledge and weak generalization ability) have driven deep learning to gradually become the mainstream.

Accurate vehicle recognition is crucial in intelligent transportation systems, and different road conditions have a significant impact on vehicle recognition results. This study focuses on video image processing technology and analyzes the recognition of different types of vehicles under specific road conditions. The involved road conditions include snowy days, rainy days, nights, intersections, and sunny days, while vehicle types cover trucks, buses, and cars. Through the collection and collation of vehicle recognition accuracy data for different vehicles under various road conditions, it is found that the recognition accuracy varies with different road conditions and vehicle types (as shown in Table 3). Analyzing these data helps to gain an in-depth understanding of the performance of video image processing technology in different scenarios and provides a basis for optimizing vehicle recognition algorithms.

Table 3. Vehicle Recognition Comparison Under Specific Road Conditions Based on Video Image Processing

Road Condition	Vehicle Type		
	Truck	Bus	Car
Snowy Day	87.19%	89.49%	87.49%
Rainy Day	89.76%	86.76%	89.19%
Night	90.49%	87.19%	90.04%
Intersection	89.89%	87.86%	89.19%
Sunny Day	93.49%	94.49%	95.49%
Average	90.16%	89.15%	90.28%

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

3. Vehicle Recognition based on Deep Learning

3.1 Research Status of Deep Learning in Target Detection and Recognition

In 2006, Hinton et al. [1] proposed the concept of deep learning, advancing research on CNN-based image feature learning. Deep learning-based algorithms have emerged continuously, achieving remarkable results in object detection and recognition. In 2012, Alex et al. presented the AlexNet model, which won the ILSVRC2012 image classification competition with outstanding performance. This demonstrated the potential of CNNs and marked the onset of a boom in deep learning. Subsequently, a series of models such as VGG, ResNet, and DenseNet were developed, along with lightweight networks like MobileNet for mobile devices and RetinaNet based on loss functions.In 2015, He Kaiming et al. iterated the Faster R-CNN model [2] on the basis of R-CNN and Fast R-CNN. By integrating a Region Proposal Network and anchor box strategy, it consolidated multi-step processes into a deep learning framework, significantly improving algorithm speed and detection accuracy. In 2016, Redmon et al. proposed the end-to-end YOLO (You Only Look Once) model [3] based on a regression approach, achieving a substantial leap in detection speed. Over the following two years, Redmon et al. continuously optimized the YOLO network, introducing YOLO9000 in 2017 and YOLOv3 in 2018. YOLO9000 adopted a new joint training algorithm and a hierarchical perspective for object classification [14], enabling recognition of up to 9000 object categories. YOLOv3 incorporated multi-scale prediction through a Feature Pyramid Network structure [4] and replaced the backbone network with DarkNet53, enhancing feature extraction capability. In 2021, Megvii Technology proposed the YOLOX model [15] based on previous YOLO algorithms and advanced technologies. Through improvements such as data augmentation and prediction decoupling, it reduced the number of parameters while achieving superior performance.

3.2 Application of Convolutional Neural Networks in Vehicle Recognition

A convolutional neural network (CNN) is a deep learning algorithm for image classification. It processes input data through neuron nodes with different connection methods to extract features from images. The image label is compared with the network output, and the weights in the network are adjusted using algorithms based on the comparison results. The main characteristics of CNNs are local receptive fields and weight sharing. When classifying images, these two features can mitigate the interference of light, noise, and other factors on images in complex environments, greatly improving the robustness of classification. Convolutional neural networks mainly consist of convolutional layers, pooling layers, and fully connected layers [16].

Features are extracted through iterative loops of convolutional layers and pooling layers, and the final features are output through the fully connected layer. The vehicle recognition model based on convolutional neural networks is shown in Fig. 1.

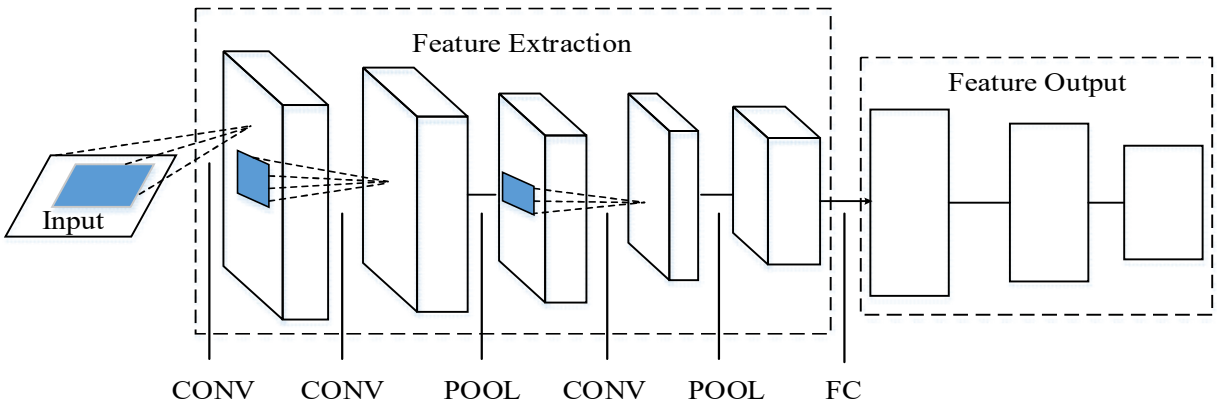

Fig. 1 Vehicle Recognition Model Based on Convolutional Neural Network

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

Based on the characteristics of convolutional neural networks, the vehicle recognition model is designed with different network layers. The number of model layers and hyperparameters are set to affect the overall model training process and final performance. Founded on meeting practical applicability (as shown in Fig. 2), the model adopts a structure of five convolutional layer modules followed by several fully connected layers. Different modules are integrated into each convolutional layer to enhance the internal network depth of the convolutional layers, enabling the overall fully connected layers to extract more global information [17].

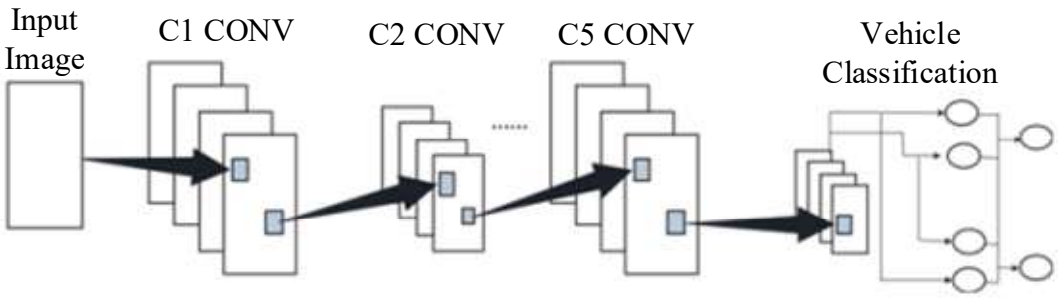

Fig. 2 Vehicle Recognition Model Based on Convolutional Neural Network

The vehicle recognition technology based on convolutional neural networks is less affected by different scene environments and can recognize different types of vehicles without being influenced by scenes. It has strong resistance to external environmental interference and the recognition results remain stable even in harsh environments. As the number of test vehicles increases, its recognition accuracy is not affected by external factors, which can effectively meet the needs of recognizing different types of vehicles in various environments.

3.3 Vehicle Target Recognition based on Attention Mechanism

3.3.1 Attention Mechanism

The attention mechanism is a crucial mechanism for the human brain to process images input by the human eye. When the human eye receives an image, this mechanism enables the visual system to focus on salient regions of interest, acquire detailed information, and ignore useless interfering information [18]. In computer vision, the core of the attention mechanism is to make the system focus on key points and overlook irrelevant information, which plays a significant role in tasks such as image classification, segmentation, and recognition.

Tyan et al. introduced a new attention unit called Squeeze-and-Excitation [19]. This unit processes input feature maps through Squeeze and Excitation operations, establishes interdependencies between feature channels, and enhances network representation capability. Inserting it into mainstream networks can improve network performance. SWoo et al. proposed a convolutional attention mechanism module [20]. It introduces spatial information encoding through large-scale convolution kernels and extracts representative attention features from two aspects: channel and spatial [21]. The channel attention module identifies and weights important channels, while the spatial attention module captures regions worthy of attention. Non-local networks focus on constructing relationships between long-distance pixels to achieve a global receptive field for pixels [22], making up for the deficiency of CNN convolution operations that only focus on local regions and ignore the contribution of global pixels.

3.3.2 Comparative Experiments Based on Attention Mechanism

In this paper, attention mechanism modules are added to both Faster R-CNN and the improved Faster R-CNN for comparative experiments. The experimental results are shown in Table 4.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

Table 4. Experimental Results of Attention Mechanism Modules

Method	mAP
Faster R-CNN	91.92%
Improved Faster R-CNN	93.45%
Faster R-CNN + Attention Mechanism	93.02%
Improved Faster R-CNN + Attention Mechanism	94.63%

As can be seen from Table 4, after adding the attention mechanism module to Faster R-CNN, the mAP value increased by approximately 1.1% compared with the original Faster R-CNN method. For the improved Faster R-CNN with the attention mechanism module added, the mAP value rose by 1.18%. It is evident that the feature maps extracted by the VGG-16 with dilated convolution already have an enhanced receptive field. These generated feature maps are weighted through the attention mechanism and then fused with multi-scale information via ASPP, enabling the feature information of vehicle positions to account for a larger weight in the final decision-making. Thus, the two components exert a certain synergistic promotion effect, further improving the performance of the proposed method.

3.4 Multimodal Fusion Technology

Deep learning-oriented multimodal fusion technology integrates information from text, images, speech, video and other domains to achieve conversion and fusion, thereby improving model performance. Its development benefits from the universality of modalities and the popularity of deep learning. Multimodal fusion generally involves information from two or more modalities. Through data associations between different modal information, it realizes mutual information conversion and can effectively ensure the accuracy of information transmission even when certain modalities are missing [23]. Multimodal information fusion-based environmental perception technology is of great significance for enhancing vehicle environmental perception capabilities. Compared with traditional single-modality methods in object detection, it can obtain more comprehensive, accurate and environmentally compatible information about the environment around vehicles. This technology plays a crucial role in autonomous cruise control, collision warning and path planning of intelligent vehicles.Accurately, in real-time and effectively detecting vehicles on the road ahead, tracking surrounding moving objects and predicting their movement directions and positions can provide reliable information for vehicle path planning and decision-making, which is the core issue of the vehicle environmental perception system. The performance of vehicle detection by the environmental perception system directly affects the driving safety of intelligent vehicles, so the research on vehicle detection algorithms is highly valuable.In specific research and practice, multimodal data such as sound and images can be collected to construct a dataset containing images and audio of special and ordinary vehicles, so as to improve the accuracy of the recognition model. At the same time, in-depth research on data fusion between computer vision and other sensors should be conducted to explore effective methods for processing synchronization and matching of multi-sensor data, thereby enhancing the accuracy and stability of vehicle recognition.

The multimodal information fusion of vehicle image recognition and audio recognition is carried out, and the accuracy after fusion is calculated through the decision tree algorithm. The comparison of accuracy before and after fusion is shown in Table 5.

As can be seen from Table 5, the fused recognition accuracy has been improved across the board. Among them, the recognition accuracy of ambulances and police cars both reached 99.9%. The main reason is that these two types of vehicles have distinct features, which can be effectively distinguished from other vehicles. The recognition rates of other vehicles such as engineering vehicles, fire trucks, and ordinary vehicles have also been improved compared with single image recognition or audio

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

recognition, with a significant increase in accuracy. The average recognition accuracy of all vehicles has reached 97.4%, which can be effectively applied to the research on special vehicle recognition.

In the field of vehicle recognition algorithms, different methods have their own advantages and disadvantages (as shown in Table 6). In-depth comparison of these algorithms is conducive to better selection and application.

Table 5. Modal Fusion Recognition Effect of Special Vehicles

Name	Number of Test Samples	Image Recognition Accuracy	Audio Recognition Accuracy	Image Fusion Accuracy
Engineering Vehicle	30	92.8%	92.3%	93.3%
Fire Truck	30	95.8%	92.8%	96.6%
Ambulance	30	95.9%	99.9%	99.9%
Ordinary Vehicle	30	96.5%	91.7%	97.2%
Police Car	30	99.9%	99.9%	99.9%
Average Performance	-	96.2%	95.4%	97.4%

Table 6. Comparative Analysis of Current Vehicle Recognition Algorithms

Method	Advantages	Disadvantages
HOG SIFT	The design of portable systems provides convenience for practical applications	Traditional image processing technology has poor adaptability to complex scenes, which may lead to reduced recognition accuracy. The hardware performance of portable systems limits their ability to process high-resolution videos.
YOLO	Small computational load, fast speed, and strong generalization ability	Low accuracy in small target detection; inconsistent image sizes between training and test sets.
YOLO3	Good performance in small target detection, high accuracy, and fast speed	Slightly inferior positioning accuracy for large objects.
CNN	Automatic feature extraction, weight sharing, strong representation ability; the model has better robustness and generalization ability	High demand for data volume, excessive consumption of computing resources, difficulty in intuitively understanding model decisions, and limited ability to process non-grid data.

4. Summary

The purpose of this paper is to comprehensively analyze and understand the current application of various deep learning networks in vehicle detection. Based on the analysis of traditional recognition methods and the application of various object detection methods in deep learning to vehicle detection, this paper outlines several commonly used object detection algorithms at present, and elaborates on the practical application of various single-stage detection algorithms in vehicle detection. Meanwhile, the advantages and shortcomings of these algorithms are listed. Currently, there are still several aspects of vehicle recognition technology that require in-depth research, such as adaptability to complex environments, diversity of vehicle features, multimodal information fusion, and data security and privacy protection.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

Acknowledgments

Project Funding: 2024 Undergraduate Innovation and Entrepreneurship Training Program of Yan'an University(D2024113); 2025 Shaanxi Provincial Undergraduate Innovation and Entrepreneurship Training Program(202510719045).

References

[1] G.E. Hinton, R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science. Vol. 313 (2006) No. 5786, p. 504-507.

[2] S. Ren, K. He, R. Girshick, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems. Vol. 28 (2015), p. 1137-1149.

[3] J. Redmon, S. Divvala, R. Girshick, et al. You only look once: Unified, realtime object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 27-30 Jun. 2016. Vol. (2016) , p. 779-788.

[4] J. Redmon, A. Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767. Vol. (2018).

[5] Wen, X. Z., Fang, W., Zheng, Y. H., et al. A vehicle recognition algorithm based on Haar-like features and improved AdaBoost classifier. Acta Electronica Sinica. Vol. 39 (2011) No. 5, p. 1121-1126.

[6] Deng, L., Wang, Z., et al. Research on vehicle type recognition based on deep convolutional neural network. Application Research of Computers. Vol. 33 (2016) No. 3, p. 930-932.

[7] Cheng, Q., Fan, Y., Liu, Y. C., et al. Vehicle recognition technology based on multi-feature fusion. Infrared and Laser Engineering. Vol. 47 (2018) No. 7, p. 316-321.

[8] Watkins R, Pears N, Manandhar S. Vehicle classification using ResNets, localisation and spatially-weighted pooling. Elsevier, 2018.

[9] Jocher, G., Liu C Y, Hogan A, et al. Ultralytics YOLOv5: State-of-the-Art Object Detection. GitHub Repository, 2020.

[10] Tan, M., & Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ICML 2019.

[11] Chen, X., et al. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. CVPR 2022.

[12] Howard, A., et al. Searching for MobileNetV3. ICCV 2019.

[13] Chen, T., et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020.

[14] J. Redmon, A. Farhadi. YOLO9000: Better, faster, stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, 21-26 Jul. 2017. Vol. (2017) , p. 7263-7271.

[15] Z. Ge, S. Liu F. Wang, Z. Li, et al. YOLOX: Exceeding YOLO Series in 2021. arXiv preprint arXiv:2107.08430. Vol. (2021).

[16] Gao, Y. T., Ning, H., Wang, W., et al. Research on vehicle model recognition based on convolutional neural network. Applied Science and Technology. Vol. 45 (2018) No. 6, p. 53-58+62.

[17] Guan, D. Y., Ju, M., An, L. H. Research on vehicle body color recognition technology based on convolutional neural network. Journal of Shandong Jianzhu University. Vol. 33 (2018) No. 1, p. 25-31.

[18] L. Itti, C. Koch, E. Niebur,et al. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol.20(1998) No.11, p.1254-1259.

[19] J. Hu, L. Shen, G. Sun,et al. Squeeze-and-excitation networks. 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA, 18-22 Jun. 2018, p.7132-7141.

[20] S. Woo, J. Park, J.-Y. Lee, et al. Cbam,et al. Convolutional block attention module. 2018 European Conference on Computer Vision. Munich, Germany, 8-14 Sep. 2018, p.3-19.

[21] B. Zhou, A. Khosla, A. Lapedriza, et al. Learning deep features for discriminative localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 27-30 Jun. 2016, p.2921-2929.

[22] X. Wang, R. Girshick, A. Gupta, et al. Non-local neural networks. 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA, 18-22 Jun. 2018, p.7794-7803.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0008

[23] Garing S, Brand K, Raake A, et al. Extended features using machine learning techniques for photo liking predict. Tenth International Conference on Quality of Multimedia Experience. Cagliari, Italy, 2018, p.1-6.