Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.23817

Computer Vision: Technologies and Applications

Yuxiang Shen

Maple Glory United School, Xiamen, China
422081782@qq.com

Abstract: Computer vision, as a crucial branch of artificial intelligence, is profoundly transforming various aspects of human society. This paper provides a systematic exploration of the key technologies, application domains, challenges, and future development trends in computer vision. We begin with a detailed analysis of core technologies including convolutional neural networks, Transformer architectures, and edge computing. Subsequently, we conduct an in-depth investigation of innovative applications in healthcare, autonomous driving, smart agriculture, and security surveillance. Furthermore, we examine critical challenges such as data scarcity, ethical privacy concerns, and computational energy consumption. Finally, we present future research directions including neuromorphic vision systems and quantum machine learning. By synthesizing insights from 15 authoritative references, this study aims to provide comprehensive technical references and application guidance for both academia and industry.

Keywords: Computer vision, Deep learning, Convolutional neural networks, Vision Transformer, Edge computing

1. Introduction

The development of computer vision can be traced back to the 1960s when research primarily focused on fundamental image processing algorithms. With the advancement of computing power and the advent of the big data era, computer vision has experienced explosive growth in the 21st century. The breakthrough performance of AlexNet in the 2012 ImageNet competition marked the successful application of deep learning in computer vision [1]. Since then, continuous innovations in computer vision technologies have facilitated the transition from laboratory research to industrial implementation.

The core objective of computer vision is to enable machines to understand and interpret visual information. By simulating the human visual system, this technology achieves automated analysis, comprehension, and decision-making regarding image and video content. Currently, computer vision has become one of the most promising technologies in artificial intelligence, with its value primarily manifested in three aspects: First, in terms of efficiency improvement, computer vision can perform numerous repetitive visual inspection tasks; second, regarding accuracy enhancement, certain visual recognition tasks have surpassed human-level performance; and third, in innovative applications, computer vision has spawned many unprecedented application scenarios.

DOI: 10.54254/2755-2721/2025.23817

This paper comprehensively analyzes the current state of computer vision development from four dimensions: technological foundations, practical applications, existing challenges, and future trends. Through systematic review of relevant research, we aim to provide technical references for researchers, guide application directions for practitioners, and discuss critical issues that require attention during the development of computer vision.

2. Core technologies

2.1. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) represent one of the most influential technologies in computer vision. CNNs employ mechanisms such as local connectivity, weight sharing, and spatial downsampling to effectively reduce network complexity while improving feature extraction capabilities. A typical CNN architecture generally consists of three layers, i.e., convolutional layers extracting local features through filters, pooling layers performing dimensionality reduction and fully connected layers accomplishing the final classification [2].

In recent years, CNN architectures have kept making improvements innovatively. As the vanishing gradient issue existed in deep network training, it can be addressed by Residual Networks (ResNet) through skip connections, which will enable network depth to reach hundreds of layers [3]. Parameter efficiency can be enhanced by DenseNet through feature reuse. Some Lightweight networks, such as MobileNet and ShuffleNet, have been optimized for mobile applications, which can significantly reduce computational complexity without accuracy compromise [4].

CNNs have performed excellently in image classification, object detection and semantic segmentation. For instance, in medical image analysis, it has been successfully applied to nodule detection in lung CT scans. The accuracy rates can be over 95% [5]. In industrial quality inspection, micron-level defects can be detected by CNN-based vision systems automatically, which subsequently can improve production efficiency and product quality.

2.2. Vision Transformers

It was a success for the application of Transformer architecture in natural language processing initially. However, after computer vision was introduced, particularly with the rise of Vision Transformer (ViT) models, which possess a larger receptive field compared to CNNs. As ViT can divide input images into fixed-size patches and establish global dependencies through self-attention mechanisms, it can model long-range dependencies with a better result [6].

The performance and efficiency of ViT are enhanced by Swin Transformer through hierarchical design and sliding window mechanisms [7]. While maintaining global modeling capability, this architecture can significantly reduce computational complexity. It will enable the processing of high-resolution images. To combine the advantages of CNNs and Transformers, hybrid architectures like Convolutional Vision Transformer (CvT) made attempts to achieve a balance between local feature extraction and global relationship modeling [8].

Furthermore, Vision Transformers have demonstrated strong potential in image classification, object detection and image generation. Transformer architectures often lead to better results than traditional CNNs, particularly in tasks requiring global understanding, such as scene comprehension and visual question answering. However, due to larger training datasets and longer training times, ViT is facing limits in its application scope.

DOI: 10.54254/2755-2721/2025.23817

2.3. Edge computing

With the proliferation of IoT devices, the role of edge computing has become more prominent in computer vision. By offloading certain computational tasks from the cloud to terminal devices, edge computing can reduce data transmission latency and improve system response speed, resulting in the enhancement of data privacy protection [9]. In the applications of computer vision, edge computing can enable real-time processing. It is particularly critical for latency-sensitive scenarios like autonomous driving and industrial inspection.

To achieve efficient edge vision computing, various optimization techniques have been developed. Network quantization converts floating-point weights into low-bit representations. This technique can significantly reduce model size and computational requirements [10]. Moreover, network pruning streamlines network architecture by removing redundant connections or channels. Knowledge distillation can utilize large models to guide the training of smaller models. Thus, their performance can be enhanced. Additionally, specialized neural network accelerators like TPUs and NPUs, along with other hardware innovations, provide robust support for edge vision computing [11].

In industrial applications, remarkable success has been achieved in edge computing-based vision systems. For example, in intelligent security, vision algorithms deployed on camera terminals can perform real-time face and behavior recognition. This technique can protect privacy while improving response speed. Moreover, edge vision systems in agricultural drones can immediately analyze crop growth conditions to guide precision farming operations.

2.4. Deep learning and the evolution of computer vision technologies

The rise of deep learning, particularly the development of Convolutional Neural Networks (CNNs), marked a key turning point in the field of computer vision. In 2012, the success of AlexNet in the ImageNet competition showcased the power of deep learning for image classification, signaling a new era in computer vision [12]. CNNs mimic the hierarchical structure of the human visual system, enabling computers to automatically extract features and classify raw images. With the increase in training data and computational power, CNNs have surpassed traditional image processing algorithms in several tasks such as image classification, object detection, and semantic segmentation [13].

However, Convolutional Neural Networks are not the only solution. In recent years, Vision Transformers (ViTs), a new architecture, have emerged. Compared to CNNs, ViTs divide images into fixed-size patches and use self-attention mechanisms to establish long-range dependencies within images. ViTs are better suited for capturing broader contextual information within an image, and they have outperformed traditional CNNs, particularly in tasks that require global understanding, such as image generation and visual question answering [14]. While ViTs require large training datasets and longer training times, they often outperform CNNs on large-scale image datasets, especially in complex visual understanding tasks.

Besides CNNs and ViTs, hybrid architectures such as the Convolutional Vision Transformer (CvT) are gaining attention. These architectures aim to combine the advantages of CNNs (local feature extraction) and Transformers (global relationship modeling), achieving better performance [15]. This progress shows that computer vision technology is no longer limited to purely deep learning networks, but also involves more innovative network design ideas.

2.5. Multimodal fusionin computer vision

DOI: 10.54254/2755-2721/2025.23817

As computer vision technology has advanced, relying solely on visual input is increasingly inadequate for complex applications. Multimodal fusion, especially in combining visual, auditory, and textual information, has become an important trend. Multimodal learning can gather information from multiple sources, combining the strengths of different modalities to enhance the model's generalization and task completion ability.

For example, in autonomous driving, relying solely on visual information from cameras is not sufficient to handle the complexities of the driving environment. By fusing data from LiDAR (Light Detection and Ranging), radar, cameras, and other sensors, autonomous systems can gain a more comprehensive perception of the environment. This data fusion not only improves the accuracy of surrounding environment perception but also enhances system robustness in adverse weather conditions. Moreover, multimodal learning can improve the interaction between speech and vision, enabling autonomous driving systems to react quickly when certain obstacles are detected [16].

In healthcare, multimodal data fusion also demonstrates immense potential. For example, by combining CT images with pathology reports and genomic data, doctors can gain more comprehensive disease analysis, improving early diagnosis accuracy [17]. This direction is continually being researched, with the goal of enhancing efficiency and precision in the medical diagnosis process.

3. Application domains

3.1. Healthcare

Computer vision demonstrates tremendous value in healthcare. In medical image analysis, deep learning algorithms assist physicians in disease diagnosis. For instance, in diabetic retinopathy screening, deep learning-based systems can match the diagnostic level of professional ophthalmologists, greatly improving screening efficiency [18]. In pathological slide analysis, computer vision systems can rapidly and accurately identify cancer cells, alleviating the workload of pathologists.

Surgical navigation represents another important application direction. By integrating computer vision with augmented reality technology, surgeons can obtain real-time anatomical guidance during procedures. In minimally invasive surgery, vision systems can precisely track surgical instrument positions to avoid damage to critical tissues. Additionally, computer vision plays an irreplaceable role in medical image 3D reconstruction and radiotherapy target delineation.

During the pandemic, computer vision technologies served crucial functions in temperature screening, mask detection, and social distance monitoring. These applications not only improved epidemic prevention efficiency but also reduced cross-infection risks. Looking ahead, with technological advancements, computer vision holds promise for creating greater value in personalized medicine and telemedicine.

3.2. Autonomous driving

Autonomous driving represents one of the most challenging application domains for computer vision. Modern autonomous systems typically incorporate multiple cameras that employ computer vision for environmental perception. Object detection algorithms identify key elements like vehicles, pedestrians, and traffic signs, while semantic segmentation algorithms understand detailed road scene compositions [19]. Multi-camera data fusion technology further enhances perception accuracy and robustness.

DOI: 10.54254/2755-2721/2025.23817

Behavior prediction constitutes a critical capability of autonomous systems. By analyzing information such as pedestrian poses and vehicle trajectories, systems can predict future behaviors of traffic participants, providing basis for decision-making and planning. Meanwhile, visual positioning technology utilizes environmental features captured by cameras to achieve centimeter-level vehicle positioning, compensating for GPS signal deficiencies.

Despite significant progress, autonomous driving still faces numerous challenges. Complex weather conditions, rare traffic scenarios, and sensor failures all impose higher demands on vision systems. Addressing these challenges requires algorithmic innovations and extensive real-world scenario data. With continuous technological refinement, autonomous driving has the potential to fundamentally transform future transportation.

3.3. Smart agriculture

Computer vision is driving intelligent transformation in agricultural production. For crop monitoring, drones equipped with multispectral cameras capture high-resolution field images, with vision algorithms analyzing crop growth and identifying pests and diseases [20]. This information helps farmers implement precise interventions, reducing pesticide and fertilizer use while improving crop yield and quality.

In precision spraying, vision-based intelligent spraying systems can identify weed locations for targeted spraying, dramatically decreasing pesticide usage. Similar principles apply to precision fertilization and irrigation. During harvesting, intelligent picking robots use vision systems to locate ripe produce for automated harvesting, addressing agricultural labor shortages.

Livestock farming also benefits from computer vision technology. By analyzing animal posture, behavior, and physiological characteristics, systems can detect early signs of disease for timely intervention. Intelligent feeding systems provide customized feed based on individual characteristics, improving breeding efficiency. These applications not only enhance agricultural productivity but also promote sustainable agricultural development.

4. Challenges and future perspectives

4.1. Current challenges

Data scarcity remains a primary challenge for computer vision. Acquiring high-quality labeled data is costly, particularly in specialized fields like healthcare. Few-shot learning and self-supervised learning offer potential solutions but require further research [21]. Another significant issue is inadequate model generalization, with performance often unsatisfactory beyond training data distributions.

Ethical privacy concerns are becoming increasingly prominent. Widespread applications of technologies like facial recognition have sparked intense debates about personal privacy protection. Algorithmic bias also warrants attention, as some vision systems exhibit performance disparities across demographic groups. Addressing these issues requires both technological improvements and regulatory oversight.

Computational resource consumption presents another major challenge. Training large vision models demands substantial energy, resulting in significant carbon footprints. Approaches like model compression and efficient architecture design can partially mitigate this problem, but fundamental solutions remain elusive. Applications with stringent real-time requirements also need to balance algorithmic efficiency and computational resources.

DOI: 10.54254/2755-2721/2025.23817

4.2. Future trends

Neuromorphic vision represents a noteworthy frontier. Inspired by biological visual systems, neuromorphic vision sensors can process visual information more efficiently. These sensors typically employ event camera designs that only respond to scene changes, offering advantages like low latency and high dynamic range. Compared to conventional frame-based cameras, neuromorphic vision systems demonstrate superior performance in high-speed motion and low-light conditions.

Quantum machine learning may bring breakthrough advancements. The parallel computing capability of quantum computers could accelerate certain vision algorithms. Although still in early research stages, quantum-enhanced vision algorithms show potential in optimization problem solving and pattern recognition. With progress in quantum computing hardware, this direction may yield significant impacts.

Multimodal fusion represents another important trend. Integrating visual, auditory, and textual information enables more comprehensive understanding of complex scenarios. Cross-modal learning techniques allow systems to leverage data from one modality to improve task performance in another. Such fusion will propel computer vision toward more intelligent and versatile development.

5. Conclusion

As a core technology of artificial intelligence, computer vision has profoundly altered development trajectories across multiple industries. From CNNs to Transformers, algorithmic innovations continuously push technological boundaries; from medical diagnosis to autonomous driving, application scenarios keep expanding. However, challenges like data dependency, ethical issues, and computational costs still require solutions. Looking ahead, emerging technologies like neuromorphic vision and quantum computing may bring breakthrough advancements. The development of computer vision necessitates coordinated progress in technological innovation, practical implementation, and ethical considerations. Through interdisciplinary collaboration and sustained research, computer vision holds promise for creating greater value for human society.

References

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25: 1097-1105.

[2] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[4] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[5] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis[J]. Medical image analysis, 2017, 42: 60-88.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[7] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[8] Wu H, Xiao B, Codella N, et al. Cvt: Introducing convolutions to vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 22-31.

DOI: 10.54254/2755-2721/2025.23817

[9] Shi W, Cao J, Zhang Q, et al. Edge computing: Vision and challenges[J]. IEEE internet of things journal, 2016, 3(5): 637-646.

[10] Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2704-2713.

[11] Chen Y H, Emer J, Sze V. Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks[J]. ACM SIGARCH Computer Architecture News, 2016, 44(3): 367-379.

[12] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25: 1097-1105.

[13] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[15] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[16] Shi W, Cao J, Zhang Q, et al. Edge computing: Vision and challenges[J]. IEEE internet of things journal, 2016, 3(5): 637-646.

[17] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis[J]. Medical image analysis, 2017, 42: 60-88.

[18] Bojarski M, Del Testa D, Dworakowski D, et al. End to end learning for self-driving cars[J]. arXiv preprint arXiv:1604.07316, 2016.

[19] Zhang C, Kovacs J M. The application of small unmanned aerial systems for precision agriculture: a review[J]. Precision agriculture, 2012, 13(6): 693-712.

[20] Wang Y, Yao Q, Kwok J T, et al. Generalizing from a few examples: A survey on few-shot learning[J]. ACM computing surveys (CSUR), 2020, 53(3): 1-34.

[21] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.