Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO25722

Research on Image Recognition Technology Based on Deep Learning

Yiyu Pan

School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, China
s1340552@live.hkmu.edu.hk

Abstract. With the rapid development of artificial intelligence technology, image recognition technology is playing an increasingly important role in fields such as healthcare, security, and autonomous driving. Traditional image recognition methods have limited recognition accuracy and generalization ability in complex scenarios, while the introduction of deep learning technology has greatly promoted progress in this field. This paper mainly studies image recognition technology based on deep learning, including a review of traditional methods, the evolution of new technologies, the development process of deep learning in image recognition, and the optimization directions of existing technologies. This paper also explores how deep learning can enhance the performance of image recognition, analyzes its technical advantages, and explores possible directions for improvement. The study employed the methods of literature review and comparative analysis. By collating academic papers, industry reports, and technical cases from the past, it compared the performance differences between traditional methods and deep learning methods, and analyzed the applications of typical deep learning models in image recognition. This paper finds that deep learning significantly improves the accuracy and robustness of image recognition, but still faces challenges such as strong data dependence and high computational cost. Future research can focus on lightweight models, small sample learning, and multimodal fusion.

Keywords: Image Recognition, Deep Learning, Convolutional Neural Networks, Feature Extraction

1. Introduction

Image recognition refers to the technology where computers automatically identify objects or features in images through algorithms. Deep learning is a machine learning method based on neural networks, which can automatically extract high-level features and is widely applied in computer vision tasks. Traditional image recognition techniques rely on manual feature extraction and have limited generalization ability. In recent years, deep learning has performed exceptionally well in competitions such as ImageNet, driving changes in this field. However, existing research still lacks in model efficiency, small sample learning, and interpretability. The successful application of deep learning in image recognition demonstrates its great potential. However, the evolution and optimization direction of related technologies, as well as their implementation in the industry, still require systematic summarization. This study aims to fill this gap and provide a reference for

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO25722

subsequent research. This paper specifically examines: a. the limitations of traditional image recognition methods; b. the development history of deep learning technology; c. the performance comparison of traditional methods and deep learning models; d. possible optimization directions. Through literature review and case analysis, the advantages and disadvantages of different technologies are compared, and the analysis is conducted based on the experimental results of typical datasets. This research helps the academic and industrial communities to have a clearer understanding of the advantages and challenges of deep learning in image recognition, and provides directions for future studies, such as lightweight models and cross-modal learning.

2. Traditional image recognition technology

Traditional image recognition technology mainly relied on manual feature extraction and machine learning classifiers. Before the emergence of deep learning (before 2012), it was widely applied in the field of computer vision. These methods usually involve these key steps. Preprocessing including noise removal through filters such as median or Gaussian filters to reduce noise and enhance image quality, grayscale conversion where color images are converted to grayscale to reduce data volume and simplify subsequent processing, and normalization where pixel values are normalized to a specific range (e.g., 0 to 1) to improve algorithm stability and convergence speed. Feature extraction involves edge-based features where edge detection algorithms (e.g., Sobel, Canny) are used to extract edges that often represent object contours, texture features where methods such as Gray Level Co-occurrence Matrix (GLCM) are employed to capture texture information which helps distinguish different types of surfaces, shape features where geometric features like contour, area, and aspect ratio are calculated to describe object shapes, and color features where color histograms or other color-based descriptors are used to characterize the color distribution in images. Classification and recognition utilizes Support Vector Machines (SVM) which finds the optimal hyperplane to separate different classes of data suitable for binary classification and extendable to multi-class problems, K-Nearest Neighbors (KNN) which classifies a sample based on the majority vote of its k nearest neighbors determined by distance metrics, Decision Trees which classify data through a series of decision rules that are easy to interpret, and Bayesian Classifiers which use Bayes' theorem to determine the class of a sample based on posterior probabilities suitable for probabilistic models.

2.1. Common traditional image recognition methods

2.1.1. Template matching

Template Matching’s principle computes the similarity between a target image and predefined templates using metrics such as Euclidean distance or correlation coefficients to identify the image category. Advantages include simple implementation effective for simple tasks like character recognition. Disadvantages are sensitivity to scale, rotation, and noise variations in images, and computational intensity when dealing with a large number of templates.

2.1.2. Feature matching

Feature Matching extracts key features (e.g., SIFT, SURF, ORB) from images and uses the similarity between these features to recognize images. It is robust to scale and rotation changes but has complex feature extraction and matching processes, is computationally demanding, and sensitive to occlusions and background noise. The computational burden primarily stems from high-dimensional

DOI: 10.54254/2755-2721/2025.PO25722

feature descriptors and exhaustive search algorithms required for reliable matching. Recent binary descriptors like ORB attempt to mitigate this through efficient computation while preserving rotation invariance. Nevertheless, performance degradation still occurs when matching features across images with significant perspective distortion or homogeneous textures where discriminative keypoints are scarce.

2.1.3. Statistical methods

Statistical methods represent images as feature vectors and apply statistical learning methods (e.g., PCA, LDA) to reduce dimensionality and classify the data. They are suitable for high-dimensional image data and capable of capturing the intrinsic structure of the data but require a large amount of labeled data for training and are sensitive to data quality and distribution. These methods fundamentally assume linear relationships within data distributions, limiting their effectiveness for complex nonlinear patterns common in real-world images. While kernel extensions exist to address nonlinearity, they substantially increase computational complexity and still struggle with highly variable intra-class features like facial expressions under different lighting conditions.

2.1.4. Shallow neural networks

Shallow neural networks, such as Multilayer Perceptrons (MLP) with few hidden layers, are used to classify images, these networks can automatically learn features and handle non-linear data to some extent. While providing automatic feature learning and classification ability for non-linear data, they have limited feature extraction capability compared to deep learning models, making them less effective for complex image tasks.

2.2. Specific traditional methods

2.2.1. Scale-Invariant Feature Transform(SIFT)

The SIFT algorithm detects and describes local keypoints using Difference of Gaussian (DoG) extremum detection and gradient orientation histograms, providing scale, rotation, and illumination invariance. SIFT is widely used in panorama stitching (e.g., Google Street View) and object recognition where viewpoint variations are significant. Its effectiveness derives from multi-scale extrema detection that identifies stable keypoints across octave layers, coupled with orientation assignment using local gradient directions.

2.2.2. Histogram of Oriented Gradients (HOG)

The Histogram of Oriented Gradients (HOG) method works by computing localized gradient orientation histograms to capture object shapes. HOG is particularly effective for rigid object detection, such as pedestrians in surveillance systems. Its detection framework divides images into cells, computes directional gradients per cell, and normalizes histograms within overlapping blocks to enhance illumination invariance. The method's success with rigid objects stems from encoding edge structures that define geometric silhouettes. However, performance declines substantially for deformable objects like animals or cloth where boundary shapes vary dynamically, as the descriptor lacks flexibility to model non-rigid deformations.

DOI: 10.54254/2755-2721/2025.PO25722

2.2.3. Local Binary Patterns (LBP)

The Local Binary Patterns (LBP) technique encodes texture by comparing each pixel's intensity with its neighbors. For a 3×3 pixel block, thresholding neighboring pixels against the center pixel (1 if higher, 0 otherwise) and converting the binary result to a decimal value (0-255) to form a texture descriptor. LBP is robust for facial recognition (e.g., early OpenCV implementations) and texture classification in medical imaging.

3. Deep learning in image recognition

3.1. Convolutional Neural Networks (CNN)

The core principle of CNNs involves specialized processing of grid-structured data (e.g., images) through hierarchical feature learning. Convolutional Layers apply sliding filters (kernels) to extract spatial features. Pooling Layers (e.g., max-pooling) reduce dimensionality while preserving translational invariance. Fully Connected Layers perform classification or regression. Key innovations include local receptive fields and weight sharing to reduce parameters, while activation functions introduce non-linearity. Applications include image classification (ResNet, VGG) and medical image segmentation (U-Net). Limitations encompass computational intensity for high-resolution images and requirement for large, labeled datasets.

3.2. Recurrent Neural Networks (RNN)

The core principle of RNNs processes sequential data by maintaining hidden states that capture temporal dependencies. Vanilla RNNs suffer from vanishing or exploding gradients, while LSTM/GRU addresses long-term dependency issues via gating mechanisms. Architectural Variants include Bidirectional RNNs that capture past and future context, and Attention Mechanisms that dynamically weight relevant timesteps. Applications range from machine translation (early Seq2Seq models) to time-series forecasting (stock prices). Limitations include sequential computation limiting parallelism and struggles with very long sequences.

3.3. Generative Adversarial Networks (GAN)

The core principle of GANs consists of two competing networks trained adversarially. The Generator learns to synthesize realistic data from noise, while the Discriminator distinguishes real vs. generated samples. Notable Variants include DCGAN with CNN-based architecture for stable image generation, and Wasserstein GAN which improves training stability via Earth Mover's distance. Applications involve photorealistic image synthesis (StyleGAN) and data augmentation for medical imaging. Challenges include mode collapse (limited diversity in outputs) and difficulty in achieving Nash equilibrium.

4. Comparative analysis

This is a systematic comparison between conventional computer vision techniques and modern deep learning methods, evaluating six critical performance dimensions through documentary analysis and theoretical analysis. The comparative analysis reveals that while traditional methods maintain advantages in computational efficiency and small-data scenarios, deep learning provides unparalleled accuracy and robustness for complex recognition tasks. Emerging hybrid approaches

DOI: 10.54254/2755-2721/2025.PO25722

(e.g., combining SIFT with CNN features) demonstrate promising results in domains requiring both precision and interpretability. These hybrids leverage efficient handcrafted features for initial region proposals while applying deep learning for fine-grained classification, balancing speed and accuracy.

4.1. Accuracy benchmarks

Traditional feature-engineered methods demonstrate constrained recognition capabilities, with SIFT achieving 68.7% accuracy on the Caltech-101 dataset and HOG-based classifiers attaining 72.3% precision in pedestrian detection [1-2]. In contrast, deep learning architectures exhibit superior performance, as evidenced by ResNet-152's 98.2% top-5 accuracy on ImageNet and Vision Transformers reaching 88.5% top-1 accuracy when pretrained on JFT-300M [3-4]. The performance gap widens significantly in complex scenarios - while LBP achieves 81.4% face recognition accuracy under controlled lighting, ArcFace networks maintain 99.8% accuracy across varying illumination conditions [5-6].

4.2. Feature representation capability

Handcrafted feature descriptors rely on manually designed algorithms. SIFT uses keypoint detection through Difference of Gaussian (DoG) extremum. HOG computes gradient orientation histograms in localized cells. LBP generates binary texture patterns in 3×3 neighborhoods. These methods require domain expertise for feature engineering and demonstrate limited adaptability. Conversely, CNN architectures automatically learn hierarchical feature representations. Initial layers detect elementary visual elements (edges, corners), intermediate layers recognize texture patterns, and deep layers assemble complex object parts. As demonstrated in Zeiler and Fergus' research through deconvolutional visualization, this learned hierarchy enables superior generalization across diverse recognition tasks [7].

4.3. Computational requirements

The computational characteristics reveal complementary strengths (see Table 1):

Table 1. Computational performance comparison

Metric	Traditional Methods	Deep Learning
Training Time	Minutes (CPU)	GPU-days
Inference Speed	150 FPS (CPU)	50 FPS (GPU)
Memory Footprint	<10MB	~100MB
Hardware Dependency	CPU-optimized	GPU/TPU-accelerated

Notably, modern optimizations like MobileNet's depthwise separable convolutions and YOLO's single-pass detection have significantly improved deep learning efficiency for edge deployment [8-9].

4.4. Data dependency analysis

Traditional methods function with minimal data (100-1,000 samples), whereas deep learning requires orders of magnitude more data. Minimum Effective Dataset Size is ~500 images for SIFT

DOI: 10.54254/2755-2721/2025.PO25722

versus ~1M images for AlexNet [1,10]. Regarding Label Efficiency, HOG requires only bounding box annotations while Faster R-CNN needs pixel-level segmentation for optimal performance. Recent advances in self-supervised learning and few-shot adaptation are reducing deep learning's data requirements [11-12].

4.5. Robustness evaluation

This research assesses robustness through three challenge scenarios. In Occlusion Tests, CNN models maintain 85% accuracy with 40% occlusion versus 62% for SIFT. Under Lighting Variations, HOG performance drops 35% under extreme illumination versus <10% for Vision Transformers. For Scale Changes, LBP fails completely beyond ±20% scaling while RetinaNet maintains 90% AP. The superior robustness of deep learning stems from its learned invariance properties during hierarchical feature extraction [13]. The comparative analysis reveals that while traditional methods maintain advantages in computational efficiency and small-data scenarios, deep learning provides unparalleled accuracy and robustness for complex recognition tasks. Emerging hybrid approaches (e.g., combining SIFT with CNN features) demonstrate promising results in domains requiring both precision and interpretability [14].

5. Enhanced image recognition techniques

Recent advances in deep learning have spurred the development of various enhancement techniques to improve image recognition performance. This section systematically examines state-of-the-art enhancement approaches, including data augmentation and feature optimization, while analyzing their technical principles and empirical effectiveness.

5.1. Data augmentation techniques

Data augmentation serves as a fundamental approach to enhance model generalization through artificial or automated expansion of training datasets. Conventional augmentation methods include Geometric Transformations such as rotation (±30°), translation (max 20%), scaling (0.8-1.2x), and flipping (horizontal/vertical). Empirical studies demonstrate that appropriate geometric transformations can improve model accuracy by 3-5% on ImageNet. Photometric Adjustments such as brightness variation (±30%), contrast modification (0.5-1.5x), and color jittering (±10% in HSV space) enhance model robustness to varying illumination conditions. Recent progress in generative augmentation has shown remarkable results. GANs including DCGAN and StyleGAN can synthesize photorealistic training samples. In medical imaging, GAN-generated data improves small-sample classification accuracy by 8-12%. Diffusion Models such as Stable Diffusion generate high-resolution synthetic images. Industrial defect detection experiments show that diffusion-based augmentation increases the F1-score from 0.82 to 0.89. Advanced Mixing Strategies like MixUp employing feature-space linear interpolation and CutMix using regional replacement typically yield 2-4% accuracy gains in classification tasks.

5.2. Feature enhancement and selection

Feature-level optimization is crucial for recognition performance. Feature Selection techniques include PCA Dimensionality Reduction which reduces facial recognition features from 2048D to 512D with only a 1.2% accuracy drop but 75% computation reduction, and Attention Mechanisms where SENet's channel attention improves ImageNet top-1 accuracy by 1.8%. Multimodal Fusion

DOI: 10.54254/2755-2721/2025.PO25722

techniques show that autonomous driving systems combining RGB and LiDAR features achieve 15% higher obstacle detection accuracy, and CLIP's joint visual-language training enables exceptional zero-shot recognition.

6. Conclusion

This paper mainly explores image recognition technology based on deep learning, analyzing its development history, technical advantages, and optimization space. Deep learning has significantly improved the performance of image recognition, especially the convolutional neural network (CNN) and Transformer architecture, which perform well in complex tasks. However, existing methods still rely on a large amount of labeled data and have high computational costs. This study does not delve into the application of unsupervised learning in image recognition, nor does it conduct empirical research in combination with specific industry cases (such as medical image analysis). Future research can focus on: (1) few-shot or self-supervised learning; (2) model lightweighting and edge computing; (3) multimodal (image + text) joint recognition; (4) the application of explainable AI in image recognition.

References

[1] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91-110. https: //doi.org/10.1023/B: VISI.0000029664.99615.94
[2] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. IEEE Computer Society Conference Vision and Pattern Recognition (CVPR), 1, 886-893. https: //doi.org/10.1109/CVPR.2005.177
[3] He, K., Zhang, X., Ren, S., & Sun, J.(2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778. https: //doi.org/10.1109/CVPR.2016.90
[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR). https: //arxiv.org/abs/2010.11929
[5] Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971-987. https: //doi.org/10.1109/TPAMI.2002.1017623
[6] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. CVPR, 4690-4699. https: //openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html
[7] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. European Conference on Computer Vision (ECCV), 818-833. https: //doi.org/10.1007/978-3-319-10590-1_53
[8] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv: 1704.04861. https: //arxiv.org/abs/1704.04861
[9] Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. ICLR 2016. https: //arxiv.org/abs/1511.06434
[10] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25. https: //papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
[11] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning (ICML), 119, 1597-1607. https: //proceedings.mlr.press/v119/chen20j.html
[12] Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys, 53(3), 1-34. https: //doi.org/10.1145/3386252
[13] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. https: //mitpress.mit.edu/9780262035613/deep-learning/

DOI: 10.54254/2755-2721/2025.PO25722

[14] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). DeCAF: A deep convolutional activation feature for generic visual recognition. Proceedings of the 31st International Conference on Machine Learning (ICML), 32(1), 647-655. http: //proceedings.mlr.press/v32/donahue14.html