Review on Image Processing Method based on AI Large Models

Hanwen Liu[1, a], Jordon Li[2, b], Qipei Tan[3, c], Jin Xiao[4, d], Hao-Hsiang Hsu[5, e], Andy Zhu[6, f], Anze Zhuge[7, g], Xingjian Zhang[8, h]

[1] Guangdong Country Garden School, No. 1, Country Garden Ave., Beijiao Town, Shunde District, Foshan, Guangdong, PRC

[2] MPW London, 90-92 Queen's Gate, South Kensington, London, UK

[3] Beijing SMIC Private School, No. 9, Liangshuihe 2nd Street, Daxing District, Beijing, China

[4] China Jiliang University, No. 258, Xueyuan Street, Xiasha Development Zone, Jianggan District, Hangzhou, Zhejiang, China

[5] University of British Columbia, 2329 West Mall Vancouver, BC, Canada

[6] Rensselaer Polytechnic Institute, 110 8th St, Troy, NY, USA

[7] Hangzhou High School, No. 238, Fengqi Road, Xiacheng District, Hangzhou, Zhejiang, China

[8] New York University, 26 Broadway 3rd Floor New York, NY, USA

[a]hanwenliu69@gmail.com, [b]hurricane07550@gmail.com, [c]170032@bjsmicschool.com, [d]3155679969@qq.com, [e]xuhaoxiang_app@163.com, [f]zhua6@rpi.edu, [g]3508445692@qq.com, [h]xz3954@nyu.edu

Abstract

The application of AI large models in image processing technology is continuously expanding and deepening. They automatically extract feature information from raw image data through deep learning technology and perform efficient analysis and processing. This article provides a review of the current state of image processing technology, focusing on the analysis of image processing techniques based on machine learning and AI large models. It is found that the introduction of AI large models has led to more rapid and intelligent development of image processing technology.

Keywords

AI, AIGC image processing, machine learning.

1. Introduction

Currently, with the popularity of AI-generated art and ChatGPT, AI-generated content (AIGC) has become a hot topic in the industry, attracting widespread attention. AIGC refers to a production method that uses AI technology to automatically or assistively generate content. By inputting instructions, humans can have AI complete complex tasks such as coding, drawing, modeling, and writing, thereby creating content. The development of AIGC is fundamentally supported by the underlying technologies of large AI models.

Large AI models, also known as pre-trained artificial intelligence models, involve feeding massive amounts of data into models with billions of parameters. The machine learns the features and structures within the data by completing tasks similar to "fill-in-the-blank," eventually becoming an AI with logical reasoning and analytical abilities. In recent years, large AI models have become the "brightest stars" in the AI field, with tech giants at home and abroad making significant efforts to develop them.

Image classification is one of the popular research directions in the field of computer vision and serves as an important foundation for applications such as object detection [1-2], face recognition [3-4], and pose estimation [5-6]. Thus, image classification technology holds significant value in both academic research and technological applications. Image classification involves determining the category of a given input image using some classification algorithm. The primary process of image classification includes image preprocessing, feature extraction, and classifier design.

Image preprocessing involves operations like image filtering, such as median filtering [7], mean filtering [8], Gaussian filtering [9], and image normalization. The main purpose is to filter out irrelevant information from the image, while maximizing the retention of useful information under data simplification, thereby enhancing the reliability of feature extraction. Feature extraction is the most critical part of the image classification task. It transforms the input image into a feature representation according to certain rules, which has specific properties. The new features typically have advantages like lower dimensionality, reduced redundancy, reduced noise, and better structure, thereby lowering the complexity requirements for the classifier and improving model performance. Finally, the classifier is trained to classify the extracted features, thus achieving image classification.

In recent years, large AI models have made significant progress in the field of image processing and demonstrated powerful capabilities across various application scenarios. Pre-trained models based on deep learning, such as VGG, ResNet, and Vision Transformer (ViT), have been widely applied to tasks like image classification, object detection, image segmentation, and style transfer. These models, trained on massive datasets, are capable of automatically extracting high-level features from images, significantly improving accuracy and efficiency in image processing. For example, in medical imaging analysis, large models can assist doctors in automatically diagnosing diseases. In autonomous driving, object detection models can identify pedestrians, vehicles, and obstacles in real-time road conditions. Additionally, large AI models are extensively used in generative adversarial networks (GANs)[10] and image restoration, enabling the generation of high-quality images, super-resolution reconstruction, and denoising, thus enhancing visual quality and detail restoration. Overall, the application of large AI models in image processing is driving innovation and development across multiple fields, from industrial manufacturing to everyday life, by enhancing automation and improving the quality of results.

2. Image Processing Technology

The development of image processing technology can be divided into three stages: traditional image processing methods, machine learning methods, and deep learning technologies based on large models. Each of these stages has had a profound impact on the field of image processing.

2.1. Traditional Image Processing Methods

In the early stages of image processing technology, traditional methods were primarily based on mathematical techniques and handcrafted designs. These methods typically included filtering, edge detection, image enhancement, and image transformations. Fourier transform was widely used for frequency domain analysis to extract useful information from signals [11]. The Sobel operator is a classic edge detection method that detects object edges by calculating the gradient of image intensity [12]. Additionally, histogram equalization is a common image enhancement technique often used to improve image contrast, making the information more easily recognizable [13]. These methods effectively solved some simple early image processing tasks, but they relied heavily on manually set parameters and features, which often lacked robustness and generalizability.

2.2. Machine Learning Methods

With the increase in computational power and data volume, feature-based machine learning methods gradually became mainstream in image processing. Machine learning techniques built on traditional methods, using statistical and machine learning algorithms for image classification and recognition. In this stage, feature extraction became crucial. Handcrafted feature extraction methods, such as SIFT (Scale-Invariant Feature Transform) [15] and HOG (Histograms of Oriented Gradients) [16], were widely used for object recognition and facial recognition. These features had good descriptive capabilities, allowing machine learning classifiers like support vector machines (SVM) [14] and k-nearest neighbors (KNN) to perform image classification effectively. However, the limitations of handcrafted features include poor adaptability to different scenarios, requiring expert knowledge to design, and needing to be redesigned for different tasks.

2.3. The Rise of Deep Learning and Large Models

The emergence of deep learning brought image processing into a new stage. Convolutional Neural Networks (CNNs) automatically learn hierarchical features from images through multiple layers of convolution operations, overcoming the limitations of handcrafted feature extraction [17]. In 2012, the introduction of AlexNet marked the success of deep learning in computer vision, achieving a significant performance boost in the ImageNet Large Scale Visual Recognition Challenge . Subsequently, networks such as VGG [18] and ResNet [19] further advanced the application of deep learning in image processing. The residual learning framework introduced by ResNet greatly deepened network depth, making it possible to train deeper networks, which significantly improved the accuracy of image classification and recognition.

2.4. The Application of Large AI Models

In recent years, the concept of large AI models has emerged, and their applications in image processing have become increasingly widespread. Vision Transformer (ViT)-based models broke through the limitations of traditional CNNs by using self-attention mechanisms similar to those in natural language processing to handle image data, thus improving performance on large-scale datasets [20]. Furthermore, Generative Adversarial Networks (GANs) have brought revolutionary progress in image generation and enhancement by achieving high-quality image generation and restoration through the interplay between the generator and the discriminator [21].

In medical image analysis, large AI models have been used to assist doctors in disease diagnosis by automatically extracting and learning features, thereby achieving high-precision lesion identification and segmentation [22]. In the field of autonomous driving, large models based on deep learning are capable of detecting pedestrians, vehicles, and obstacles in real time, providing safety assurance for autonomous driving technologies [23].

In summary, the evolution of image processing technology from traditional methods to machine learning, deep learning, and large AI models has significantly improved the performance and adaptability of image processing, driving a leap from theoretical research to widespread applications in computer vision.

3. Image Processing Techniques based on AI Large Models

3.1. Image Generation and Editing

Image generation and editing technology is one of the important research directions in the field of computer vision. In recent years, with the rapid development of deep learning and large models, significant progress has been made in image generation and editing. The core objective

of image generation is to create realistic images from scratch through algorithms, while image editing involves modifying existing images to achieve specific goals.

Early image generation methods were mainly based on traditional probabilistic models and handcrafted features, such as Markov Random Fields (MRF) and sparse representations. While these methods could generate reasonable image content to some extent, the quality of the generated images was often limited by feature design and algorithmic complexity. With the rise of deep learning, Generative Adversarial Networks (GANs) [21] brought a revolutionary change to image generation. GANs, through adversarial training between the generator and the discriminator, significantly improved the quality of generated images, especially for generating natural scenes and human faces with rich details. In recent years, conditional GANs (Conditional GAN) [24] have also been used to generate corresponding images based on specific conditions (e.g., text descriptions or labels), providing users with more flexible ways to generate images.

In terms of image editing, techniques such as image inpainting [25], style transfer [26], and super-resolution reconstruction [27] are important research topics. Image inpainting aims to repair missing or damaged parts of an image to restore completeness and naturalness. Deep learning-based image inpainting methods, such as those based on Convolutional Neural Networks (CNNs) and GANs, can effectively restore image content in complex scenes. Style transfer technology learns the features of a specific artistic style and applies them to other images, allowing users to transform ordinary photos into works of art. Super-resolution reconstruction technology reconstructs the details of low-resolution images to generate high-resolution images, which is crucial for applications such as video enhancement and medical imaging.

With the application of large models and self-attention mechanisms (e.g., Vision Transformer, ViT) [28], image generation and editing technologies have made further advancements in quality and flexibility. For example, Diffusion Models [29] have been widely used for high-quality image generation, enabling the generation of clearer and more diverse image content. Additionally, systems like DALL-E [30] and Stable Diffusion [31], based on large models, can generate images from text descriptions, providing users with an intuitive way to create content. The development of these technologies has not only promoted artistic creation and entertainment applications but has also had a profound impact on fields such as healthcare and industrial design.

In summary, the development of image generation and editing technologies, from traditional methods to deep learning, and then to large models and self-attention mechanisms, has significantly improved the quality of generated images and the flexibility of editing, making the application scenarios of image generation and editing more extensive and in-depth.

3.2. Image Recognition and Understanding

Image recognition and understanding are critical research directions in the field of computer vision, aiming to enable computers to automatically identify objects in images and understand their semantic content. The goal of image recognition is to determine the categories of objects within an image, while image understanding extends further to interpret the relationships between objects and provide semantic information about the scene. With the rise of deep learning and large models, technologies for image recognition and understanding have made substantial advancements in accuracy and scope of applications.

Initially, image recognition mainly relied on manually crafted features and traditional machine learning classifiers. Techniques such as color histograms, edge detection, and texture descriptors (e.g., Gabor filters) were commonly used for feature extraction, and these features were then fed into classifiers like Random Forests and k-means clustering to achieve object categorization [32]. Although these methods performed well in controlled environments, their

reliance on handcrafted features led to limited robustness and generalizability when dealing with complex scenes and noisy environments.

With the advent of deep learning, the application of Convolutional Neural Networks (CNNs) in image recognition and understanding achieved breakthrough progress. CNNs automatically extract image features from low-level to high-level through multiple convolutional layers, significantly improving the performance of image recognition. Architectures like GoogLeNet [33] and DenseNet [34] achieved remarkable success in the ImageNet Large Scale Visual Recognition Challenge. These deep networks made automated feature extraction possible, greatly enhancing the accuracy and efficiency of image classification. Additionally, object detection algorithms such as R-FCN [35] and Mask R-CNN [36] combined object detection with semantic segmentation, making object recognition and localization more accurate, especially in handling complex backgrounds and multiple objects.

In image understanding, semantic segmentation and scene parsing technologies are particularly important. Semantic segmentation aims to assign a category label to every pixel in an image, enabling a deeper analysis of the scene. Deep learning-based networks such as U-Net [37] and PSPNet [38] use fully convolutional structures and spatial pyramid pooling mechanisms, significantly improving the precision and applicability of semantic segmentation. Moreover, Generative Adversarial Networks (GANs) [39] have demonstrated strong capabilities in applications like image inpainting and synthesis, showcasing the ability to complete and enhance missing or blurred information. These methods can generate realistic details within images, enabling better scene understanding and inference.

In recent years, large pre-trained models, such as the Swin Transformer [40], have further advanced image recognition and understanding. These models introduced the self-attention mechanism, which allows them to capture long-range dependencies in images, thereby enhancing the understanding of complex scenes. For instance, SETR [41] applied the Transformer architecture to semantic segmentation tasks, effectively improving segmentation accuracy. Beyond the Transformer architecture, multimodal models like BLIP [42] combine visual and linguistic features, enabling joint understanding and generation of images and text in cross-modal tasks, enhancing the capabilities of image recognition and semantic description.

Image recognition and understanding technologies have found widespread applications in fields such as autonomous driving, intelligent surveillance, and medical image analysis. In autonomous driving, image recognition technologies are used for real-time detection of vehicles, pedestrians, and traffic signs, ensuring safe driving [43]. In the medical field, image understanding technologies assist doctors in precisely identifying and localizing lesions, improving diagnostic accuracy [44]. Despite the significant progress made, image recognition and understanding technologies still face challenges, such as generalizability to complex scenes, robustness to occlusion and noise, and high computational requirements. These issues require further research and improvements to better address the complexities of real-world applications.

In summary, with the development of deep learning and large models, image recognition and understanding technologies have evolved from traditional handcrafted feature methods to modern methods using deep neural networks and large-scale pre-trained models, significantly enhancing the system's recognition capabilities and semantic understanding. In the future, with the advancement of self-supervised learning, multimodal learning, and other emerging technologies, image recognition and understanding are expected to be applied more extensively and deeply in more complex and diverse scenarios.

4. Conclusion

Image processing technology is entering a new era of intelligent development, largely driven by the rapid rise of deep learning and AI large models. From traditional handcrafted feature-based methods to the application of machine learning, and now to modern image processing techniques based on deep neural networks and large-scale pre-trained models, there has been a qualitative leap in image recognition, understanding, generation, and editing capabilities. AI large models, with their powerful feature learning and adaptability, have demonstrated significant advantages in image processing tasks. These models are capable of learning complex image features from vast datasets, exhibiting strong generalizability and robustness, thereby significantly improving the accuracy and efficiency of image classification, object detection, semantic segmentation, and image generation. Moreover, the introduction of large models has expanded image processing beyond single modalities, achieving joint understanding and generation of images and text through multimodal learning, which opens new possibilities for intelligent creation and human-computer interaction.

In the future, with advancements in self-supervised learning, multimodal integration, and computational hardware, the application of AI large models in image processing will deepen further. We have reason to expect that these technologies will play a key role in more complex scenarios, not only advancing applications such as autonomous driving and medical image analysis but also bringing new breakthroughs to fields like industrial manufacturing and creative arts. As AI large models continue to improve in terms of accuracy, efficiency, intelligence, and adaptability to complex environments, image processing will evolve towards greater intelligence and diversity, bringing more innovation and convenience to various sectors of society.

References

[1] OUYANG W, ZENG X, WANG X, et al. DeepID-Net: object detection with deformable part based convolutional neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(7): 1320-1334.

[2] DIBA A, SHARMA V, PAZANDEH A, et al. Weakly supervised cascaded convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: ACM Press, 2017: 5131-5139.

[3] HU G, YANG Y X, YI D, et al. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition[C]//International Conference on Computer Vision, December 11-18, 2015, Santiago, Chile. Piscataway: IEEE Press, 2015: 142-150.

[4] LAWRENCE S, GILES C L, TSOI A C, et al. Face recognition: a convolutional neural-network approach[J]. IEEE Transactions on Neural Networks, 1997, 8(1): 98-113.

[5] CAO Z, SIMON T, WEI S, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. EprintArxiv, 2017: 1302-1310.

[6] TOSHEV A, SZEGEDY C. DeepPose: human pose estimation via deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: ACM Press, 2014: 1653-1660.

[7] PERREAULT S, HEBERT P. Median filtering in constant time[J]. IEEE Transactions on Image Processing, 2007, 16(9): 2389-2394.

[8] SLOT K, KOWALSKI J, NAPIERALSKI A, et al. Analogue median/average image filter based on cellular neural network paradigm[J]. Electronics Letters, 1999, 35(19): 1619-1620.

[9] DIREKOGLU C, NIXON M S. Image-based multiscale shape description using Gaussian filter[C]//2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, December 16-19, 2008, Bhubaneswar, India. Piscataway: IEEE Press, 2009: 673-678.

[10] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. arXiv:1406.2661, 2014.

[11] Gonzalez, R. C., & Woods, R. E. (2008). Digital Image Processing (3rd ed.). Pearson.

[12] Sobel, I. (1970). "An Isotropic 3x3 Gradient Operator for Image Processing." Stanford University.

[13] Pizer, S. M. et al. (1987). "Adaptive histogram equalization and its variations." Computer Vision, Graphics, and Image Processing.

[14] Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer.

[15] Lowe, D. G. (2004). "Distinctive image features from scale-invariant keypoints." International Journal of Computer Vision.

[16] Dalal, N., & Triggs, B. (2005). "Histograms of Oriented Gradients for Human Detection." IEEE Conference on Computer Vision and Pattern Recognition.

[17] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks." Advances in Neural Information Processing Systems.

[18] Simonyan, K., & Zisserman, A. (2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv preprint arXiv:1409.1556.

[19] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep Residual Learning for Image Recognition." IEEE Conference on Computer Vision and Pattern Recognition.

[20] Dosovitskiy, A. et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations (ICLR).

[21] Goodfellow, I. et al. (2014). "Generative Adversarial Nets." Advances in Neural Information Processing Systems.

[22] Esteva, A. et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks." Nature.

[23] Chen, L. C. et al. (2018). "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." IEEE Transactions on Pattern Analysis and Machine Intelligence.

[24] Mirza, M., & Osindero, S. (2014). "Conditional Generative Adversarial Nets." arXiv preprint arXiv:1411.1784.

[25] Pathak, D. et al. (2016). "Context Encoders: Feature Learning by Inpainting." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[26] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). "Image Style Transfer Using Convolutional Neural Networks." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[27] Ledig, C. et al. (2017). "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[28] Dosovitskiy, A. et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations (ICLR).

[29] Ho, J., Jain, A., & Abbeel, P. (2020). "Denoising Diffusion Probabilistic Models." Advances in Neural Information Processing Systems.

[30] Ramesh, A. et al. (2021). "Zero-Shot Text-to-Image Generation." International Conference on Machine Learning (ICML).

[31] Romero, J. (2022). "Stable Diffusion: A Step Towards High-Fidelity and Versatile Image Generation." arXiv preprint arXiv:2204.06125.

[32] Belongie, S., Malik, J., & Puzicha, J. (2002). "Shape matching and object recognition using shape contexts." IEEE Transactions on Pattern Analysis and Machine Intelligence.

[33] Szegedy, C., Liu, W., Jia, Y., et al. (2015). "Going deeper with convolutions." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[34] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). "Densely Connected Convolutional Networks." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[35] Dai, J., Li, Y., He, K., & Sun, J. (2016). "R-FCN: Object Detection via Region-based Fully Convolutional Networks." Advances in Neural Information Processing Systems (NIPS).

[36] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). "Mask R-CNN." IEEE International Conference on Computer Vision (ICCV).

[37] Ronneberger, O., Fischer, P., & Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).

[38] Zhao, H., Shi, J., Qi, X., et al. (2017). "Pyramid Scene Parsing Network." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[39] Ledig, C., Theis, L., Huszár, F., et al. (2017). "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[40] Liu, Z., Lin, Y., Cao, Y., et al. (2021). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." IEEE International Conference on Computer Vision (ICCV).

[41] Zheng, S., Lu, J., Zhao, H., et al. (2021). "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[42] Li, J., Selvaraju, R. R., Gotmare, A. D., et al. (2022). "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." arXiv preprint arXiv:2201.12086.

[43] Chen, X., Ma, H., Wan, J., et al. (2017). "Multi-view 3D object detection network for autonomous driving." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[44] Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks." Nature.