Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062313

EDITORIAL

Multimodal Learning in Image Processing

Zhixin Chen[1,2], **Gautam Srivastava**[3,4,5,*] **and Shuai Liu**[1,2,*]

[1]School of Educational Science, Hunan Normal University, Changsha, 410081, China
[2]Institute of Interdisciplinary Studies, Hunan Normal University, Changsha, 410081, China
[3]Department of Math and Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada
[4]Research Centre for Interneural Computing, China Medical University, Taichung, 406040, Taiwan
[5]Institute of Engineering and Technology, Chitkara University, Chandigarh, 140401, India

*Corresponding Authors: Gautam Srivastava. Email: srivastavag@brandonu.ca; Shuai Liu. Email: liushuai@hunnu.edu.cn

Received: 16 December 2024; Accepted: 31 December 2024; Published: 17 February 2025

1 Introduction on Multimodal Learning in Image Processing

IP (Image processing), as a classical research domain in computer application technology, has been researched for decades. It is one of the most important research directions in computer vision, which is the basis for many current hotspots such as intelligent transportation/education/industry, etc. Because image processing is the strongest link for AI (artificial intelligence) applying to real world application, it has been a challenging research field with the development of AI, from DNN (deep convolutional network), Attention/LSTM (long-short term memory), to Transformer/Diffusion/Mamba based GAI (generated AI) models, e.g., GPT and Sora [1]. Today, the description ability of single-model feature limits the performance of image processing. More comprehensive description of the image is required to match the computational performance of current large scale models.

It is known that multimodal information will show more detailed features of object than single ones [2]. In this way, the multimodal information fusion is introduced, providing more description ability of image to various AI models by fusing multiple features from different source or angle, which has better performance than classical methods. The fused features of complex image are learnt as a whole, which has ability to defend major classical challenges. It attracts much attention from academic community today, and many AI models are being generated to provide the fusion of various kinds of information, or fusion of various features with different granularity. These models aim to describe the object in detail, which is called multimodal learning (ML) [3].

There are two kinds of ML methods in IP. Here, which are called "Model-based ML" and "Data-based ML", which differ in the origin of the multimodal input images. "Model-based ML" means that the original input image is single mode (one from visible, infrared, radar, etc.), and the AI model processes the input image to multiple modes; "Data-based ML" means that the original input information is multiple images of one target/scene from multiple sources.

Therefore, in "Model-based ML in IP", the main concern is how to generate different modes from one input image, which need to represent different kinds of image features, so as the ML can be more robust than using single feature. Researchers need to think what kinds of features should be extracted to describe the image to face the challenges [4]. In "Data-based ML", the main concern is how to combine the image data from different source to keep their individual descriptions of the image, which can separate or recognize objects better in the image than single mode [5]. Also, model synchronism is another important

issue in the combination of images with various modes to keep all modes represent the one image under the same condition.

Since the challenges from two directions are holding back the development of ML in IP, it is the time to research on both model-based features fusion, as well as data fusion from various sources. In this way, the special issue "Multimodal Learning in Image Processing" is provided to focus on the recent progress of multimodal learning in image processing. This issue received 34 submissions and accepted 12 out of them with at least 2 rounds of strict reviews, with acceptance ratio 35.29%.

This issue focuses on two sections: "Model-based ML in IP" and "Data-based ML in IP". The first section "Model-based ML in IP" is correspond to feature generation and fusion for entire functional steps in ML model. The other section "Data-based ML in IP" focuses on the pre-processing and combination of various features extracted from image with various sources.

2 Model-Based Multimodal Learning in Image Processing

Recent progress of mode-based ML mainly focuses on the development of feature generation and fusion. Classical model-based ML simply generates various features from one image by output the original intermediate result from the encoding processing of an AI model. However, the increase in performance with the fused features is minimal, and may be reduced in some cases. Recently, the entire encoding processing of model is studied clearly, the intermediate results may be output and fused everywhere when extracting corresponding representation of the image.

For example, features with different level can be generated and fused by such Al models like U-Net and DNN to keep some kinds of features like edge [6]; generated by knowledge distillation encoding and fused with a student-teacher imitation model to strengthen the entire foreground features from complex background [7]; generated by different channels to get features with different views in attention based model and fused with cascaded attention to complete a lightweight model [8]; generated by parallel residual operators in GAN (Generative Adversarial Networks) and fused with spatial attention to enhance the truthfulness of the generated images [9].

Another thought line is to generate new features by training an assisted model, then complete a two-stage fusion. For example, depth information of image can be extracted from 2D image by using AI model like LSTM, and the fused 2D and 3D semantic descriptor can better estimate gait factors than classical 2D or 3D models [10]. Also, a 3D image can be encode to extract 2D features to reduce the computation. Then it can add newer module like attention mechanism, and the fused features achieves superior performance compared to previous models [11].

3 Data-Based Multimodal Learning in Image Processing

Data based ML focuses on the fusion mechanism and representation ability of different kinds of images from various sources. Common modes include text, audio, visible, infrared, radar image, etc., which has different ability to describe the object in different view [12]. Therefore, how to enhance the positive ability from each mode and reduce the negative ability in fusion is the main concern in this domain. Additionally, the mode synchronism from different sources is also a research hotspot because the time or view change of the image extraction may cause uncertain impact to the final performance. In this way, open dataset of multimodal image becomes one of the most important resource too.

For example, infrared light captures significant thermal radiation data, whereas visible light excels in presenting detailed texture. The fusion of the two modes allows for merge their respective strengths to result in high-quality images with enhanced contrast and rich texture details [13]. Also, fusion of visual and LiDAR

point-cloud projection to reduce the missing rate of single visible feature [14], and the fusion of voice and facial (image) features can enhance the generation quality by GAN [15]. It is worth mentioning that one multimodal remote image dataset for fine-grained recognition based on ships with visible and near-infrared image is proposed [16].

Another research domain in this direction is collaborative perception and recognition with multi-agents, that is, using many parallel perceptrons to describe an entire system. For example, GCN (graph convolutional network) is an effective model to describe an entire status by using a group of information as a graph [17]. Also, the pre-segmentation for image containing multiple objects can divide a classic IP task to a ML task, and it has higher probability of better performance than classical algorithms [18].

4 Conclusion and Future Directions

Multimodal feature can describe more information of one image than single feature. With the processing and fusion of multiple modes, it can more comprehensively understand the information in complex scene and provide more accurate analysis results by using cross-modal interactive learning. It has shown great potential and value in many research fields.

However, there are technical challenges of ML in IP that are waiting for solution. First is the data fusion, that is, how to reasonably process multimodal information to reduce the information redundancy and enhance the information complementarity. Another is AI model improvement, that is, how to generate and improve the ML model to get cross-modal information, as well as to reduce the computational scale (parameters scale) to contain more valuable multimodal information. The last is the application of the ML model in real industry, that is, how to effectively integrate and use the multimodal information to achieve accurate and robust IP for tasks in actual scene.

Therefore, there are three research directions for this research in future development trend. The first is generation and translation within multimodal image. Currently, there are many research on relations between text-image, text-audio, image-audio. NLP (Natural Language Processing) is widely used in ML models since text features show overwhelming performance in actual tasks. So, using text mode as intermediate mode to connect multimodal information is a meaningful attempt.

Secondly, ML model is required to improve for cross-modal image understanding. Researchers are exploring the architecture of ML model to achieve better semantic understanding between different modes. This also includes GAI (generative AI) model and model compression. GAI is used to enhance the cross-modal image understanding by using prompt engineering to learning the relation between modes. Model compression can fuse multiple large scale models by compressing each of them.

Finally, basic mechanisms of image analysis and retrieval are required to study more deeply. Deep convolutional AI models often lack interpretability, and the features extracted by DNN are difficult to understand, which limits the sustainability and reliability of research in related domains. The clearer meaning of the extracted/generated feature will make the feature selection and fusion more targeted, improve the performance of image understanding, and achieve more accurate image classification, object detection and other tasks.

Acknowledgement: The guest editors are thankful to the authors for their trust to submit the manuscripts, and reviewers for their effort in reviewing the manuscripts. We also thank the Edit-in-Chiefs, Prof. Ankit Agrawal, Prof. Timon Rabczuk, and Prof. Guoren Wang for their supportive guidance during the entire process. This work was supported by 2023 Key Supported Project of the 14th Five Year Plan for Education and Science in Hunan Province with No. XJK23AXX001, and 2021 Supported Project of the Educational Science Plan in Hunan Province with No. XJK21BXX010.

Funding Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Epstein Z, Hertzmann A. Investigators of human creativity. Art and the science of generative AI. Science. 2023;380(6650):1110–11. doi:10.1126/science.adh4451.
2. Liu S, Huang S, Wang S, Muhammad K, Bellavista P, Ser JD. Visual tracking in complex scenes: a location fusion mechanism based on the combination of multiple visual cognition flows. Inf Fusion. 2023;96(22):281–96. doi:10.1016/j.inffus.2023.02.005.
3. Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2018;41(2):423–43. doi:10.1109/TPAMI.2018.2798607.
4. Liu S, Luo Z, Fu W. Fcdnet: fuzzy cognition-based dynamic fusion network for multimodal sentiment analysis. IEEE Trans Fuzzy Syst. 2024;1–12. doi:10.1109/TFUZZ.2024.3407739.
5. Joshi N, Baumann M, Ehammer A, Fensholt R, Grogan K, Hostert P, et al. A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. Remote Sens. 2016;8(1):70. doi:10.3390/rs8010070.
6. Almujally NA, Chughtai BR, Mudawi NA, Alazeb A, Algarni A, Alzahrani HA, et al. UNet based on multi-object segmentation and convolution neural network for object recognition. Comput Mater Contin. 2024;80(1):1563–80. doi:10.32604/cmc.2024.049333.
7. Ramachandran A, Sendhil Kumar KS. Border sensitive knowledge distillation for rice panicle detection in UAV images. Comput Mater Contin. 2024;81(1):827–42. doi:10.32604/cmc.2024.054768.
8. Chong N, Yang F. CMMCAN: lightweight feature extraction and matching network for endoscopic images based on adaptive attention. Comput Mater Contin. 2024;80(2):2761–83. doi:10.32604/cmc.2024.052217.
9. Tian C, Gao H, Wang P, Zhang B. An enhanced GAN for image generation. Comput Mater Contin. 2024;80(1):105–18. doi:10.32604/cmc.2024.052097.
10. Luo J, Xu B, Tjahjadi T, Yi J. A novel 3D gait model for subject identification robust against carrying and dressing variations. Comput Mater Contin. 2024;80(1):235–61. doi:10.32604/cmc.2024.050018.
11. Lu Y, Chen W, Yu Z, Wang J, Yang C. Vehicle abnormal behavior detection based on dense block and soft thresholding. Comput Mater Contin. 2024;79(3):5051–66. doi:10.32604/cmc.2024.050865.
12. Liu S, Gao P, Li Y, Fu W, Ding W. Multi-modal fusion network with complementarity and importance for emotion recognition. Inf Sci. 2023;619:679–94. doi:10.1016/j.ins.2022.11.076.
13. Wang X, Zhang J, Tao Y, Yuan X, Guo Y. BDPartNet: feature decoupling and reconstruction fusion network for infrared and visible image. Comput Mater Contin. 2024;79(3):4621–39. doi:10.32604/cmc.2024.051556.
14. Li Y, Wang Y, Xie J, Xu C, Zhang K. Target detection on water surfaces using fusion of camera and lidar based information. Comput Mater Contin. 2024;80(1):467–86. doi:10.32604/cmc.2024.051426.
15. Mao J, Zhou Y, Wang Y, Li J, Liu Z, Bu F. Attention-enhanced voice portrait model using generative adversarial network. Comput Mater Contin. 2024;79(1):837–55. doi:10.32604/cmc.2024.048703.
16. Song S, Zhang R, Hu M, Huang F. Fine-grained ship recognition based on visible and near-infrared multimodal remote sensing images: dataset, methodology and evaluation. Comput Mater Contin. 2024;79(3):5243–71. doi:10.32604/cmc.2024.050879.
17. Wang Y, Xia Y, Liu S. BCCLR: a skeleton-based action recognition with graph convolutional network combining behavior dependence and context clues. Comput Mater Contin. 2024;78(3):4489–507. doi:10.32604/cmc.2024.048813.
18. Qureshi AM, Butt AH, Alazeb A, Mudawi NA, Alonazi M, Almujally NA, et al. Semantic segmentation and YOLO detector over aerial vehicle images. Comput Mater Contin. 2024;80(2):3315–32. doi:10.32604/cmc.2024.052582.