

Artificial Intelligence for Multimedia Signal Processing

Byung-Gyu Kim ^{1,*}  and Dong-San Jun ²¹ Department of IT Engineering, Sookmyung Women's University, Seoul 04310, Korea² Department of Computer Engineering, Dong-A University, Busan 49315, Korea; dsjun@dau.ac.kr

* Correspondence: bg.kim@sookmyung.ac.kr

1. Introduction

At the ImageNet Large Scale Visual Re-Conversion Challenge (ILSVRC), a 2012 global image recognition contest, the University of Toronto Supervision team led by Prof. Geoffrey Hinton took first and second place by a landslide, sparking an explosion of interest in deep learning. Since then, global experts and companies such as Google, Microsoft, nVidia, and Intel have been competing to lead artificial intelligence technologies, such as deep learning. Now, they are developing deep-learning-based technologies that can be applied to all industries to solve many classification and recognition problems.

These artificial intelligence technologies are also actively applied to broadcasting and multimedia processing technologies based on recognition and classification [1–3]. A vast amount of research has been conducted in a wide variety of fields, such as content creation, transmission, and security, and attempts have been made in the past two to three years to improve image, video, speech, and other data compression efficiency in areas related to MPEG media processing technology [4–6]. Additionally, technologies such as media creation, processing, editing, and creating scenarios are very important areas of research in multimedia processing and engineering. In this issue, we present excellent papers related to advanced computational intelligence algorithms and technologies for emerging multimedia processing.

2. Emerging Multimedia Signal Processing

Thirteen papers related to artificial intelligence for multimedia signal processing have been published in this Special Issue. They deal with a broad range of topics concerning advanced computational intelligence algorithms and technologies for emerging multimedia signal processing.

We present the following works in relation to the computer vision field. Lee et al. propose a densely cascading image restoration network (DCRN) consisting of an input layer, a densely cascading feature extractor, a channel attention block, and an output layer [7]. The densely cascading feature extractor has three densely cascading (DC) blocks, and each DC block contains two convolutional layers. From this design, they achieved better quality measures for the compressed joint photographic experts group (JPEG) images compared with the existing methods. In [8], an image de-raining approach is developed using the generative capabilities of recently introduced conditional generative adversarial networks (cGANs). This method could be very useful to recover visual quality when degraded due to diverse weather conditions, recording conditions, or motion blur.

Additionally, Wu et al. suggest a framework to leverage the sentimental interaction characteristic based on a graph convolutional network (GCN) [9]. They first utilize an off-the-shelf tool to recognize the objects and build a graph over them. Visual features are represented as nodes, and the emotional distances between the objects act as edges. Then, they employ GCNs to obtain the interaction features among the objects, which are fused with the CNN output of the whole image to predict the result. This approach is very useful to analyze human sentiment analysis. In [10], two lightweight neural networks



Citation: Kim, B.-G.; Jun, D.-S. Artificial Intelligence for Multimedia Signal Processing. *Appl. Sci.* **2022**, *12*, 7358. <https://doi.org/10.3390/app12157358>

Received: 15 July 2022

Accepted: 21 July 2022

Published: 22 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

with a hybrid residual and dense connection structure are suggested by Kim et al. to improve super-resolution performance. They show that the proposed methods could significantly reduce both the inference speed and the memory required to store parameters and intermediate feature maps, while maintaining similar image quality compared to the previous methods.

Kim et al. propose an efficient scene classification algorithm for three different classes by detecting objects in the scene [11]. The authors utilize a pre-trained semantic segmentation model to extract objects from an image. After that, they construct a weighting matrix to better determine the scene class. Finally, this classifies an image into one of three scene classes (i.e., indoor, nature, city) using the designed weighting matrix. This technique can be utilized for semantic searches in multimedia databases.

Lastly, an estimation method for human height is proposed by Lee et al. using color and depth information [12]. They use color images for deep learning by mask R-CNN to detect a human body and a human head separately. If color images are not available for extracting the human body region due to a low light environment, then the human body region is extracted by comparison with the current frame in the depth video.

For speech, sound, and text processing, Lin et al. improve the raw-signal-input network from other research using deeper network architectures [13]. They also propose a network architecture that can combine different kinds of network feeds with different features. In the experiment, the proposed scheme achieves an accuracy of 73.55% in the open audio dataset, "Dataset for Environmental Sound Classification 50" (ESC50). A multi-scale discriminator that discriminates between real and generated speech at various sampling rates is devised by Kim et al. to stabilize GAN training [14]. In this paper, the proposed structure is compared with conventional GAN-based speech enhancement algorithms using the VoiceBank-DEMAND dataset. They show that the proposed approach can make the training faster and more stable.

To translate the speech, a multimodal unsupervised scheme is proposed by Lee and Park [15]. They make a variational autoencoder (VAE)-based speech conversion network by decomposing the spectral features of the speech into a speaker-invariant content factor and a speaker-specific style factor to estimate diverse and robust speech styles. This approach can help second language (L2) speech education. To develop a 3D avatar-based sign language learning system, Chakladar et al. suggest a system that converts the input speech/text into corresponding sign movements for Indian Sign Language (ISL) [16]. The translation module achieves a 10.50 SER (sign error rate) score in the actual test.

Two papers concern content analysis and information mining. The first one, by Krishna Kumar Thirukokaranam Chandrasekar and Steven Verstockt, regards a context-based structure mining pipeline [17]. The proposed scheme not only attempts to enrich the content, but also simultaneously splits it into shots and logical story units (LSU). They demonstrate quantitatively that the pipeline outperforms existing state-of-the-art methods for shot boundary detection, scene detection, and re-identification tasks. The other paper outlines a framework which can learn the multimodal joint representation of pins, including text representation, image representation, and multimodal fusion [18]. In this work, the authors combine image representations and text representations in a multimodal form. It is shown that the proposed multimodal joint representation outperforms unimodal representation in different recommendation tasks.

For ECG signal processing, Tanoh and Napoletano propose a 1D convolutional neural network (CNN) that exploits a novel analysis of the correlation between the two leads of the noisy electrocardiogram (ECG) to classify heartbeats [19]. This approach is one-dimensional, enabling complex structures while maintaining reasonable computational complexity.

I hope that the technical papers published in this Special Issue can help researchers and readers to understand the emerging theories and technologies in the field of multimedia signal processing.

Funding: This research received no external funding.

Acknowledgments: We thank all authors who submitted excellent research work to this Special Issue. We are grateful to all reviewers who contributed evaluations of scientific merits and quality of the manuscripts and provided countless valuable suggestions to improve their quality and the overall value for the scientific community. Our special thanks go to the editorial board of MDPI Applied Sciences journal for the opportunity to guest edit this Special Issue, and to the Applied Sciences Editorial Office staff for the hard and precise work required to keep to a rigorous peer-review schedule and complete timely publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kim, J.-H.; Hong, G.-S.; Kim, B.-G.; Dogra, D.P. deepGesture: Deep Learning-based Gesture Recognition Scheme using Motion Sensors. *Displays* **2018**, *55*, 38–45. [[CrossRef](#)]
2. Kim, J.-H.; Kim, B.-G.; Roy, P.P.; Jeong, D.-M. Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure. *IEEE Access* **2019**, *7*, 2907327. [[CrossRef](#)]
3. Jeong, D.; Kim, B.-G.; Dong, S.-G. Deep Joint Spatio-Temporal Network (DJSTN) for Efficient Facial Expression Recognition. *Sensors* **2020**, *20*, 1936. [[CrossRef](#)] [[PubMed](#)]
4. Lee, Y.; Jun, D.; Kim, B.-G.; Lee, H. Enhanced Single Image Super Resolution Method using a Lightweight Multi-scale Channel Dense Network for Small Object Detection. *Sensors* **2021**, *21*, 3351. [[CrossRef](#)] [[PubMed](#)]
5. Park, S.-J.; Kim, B.-G.; Chilamkurti, N. A Robust Facial Expression Recognition Algorithm Based on Multi-Rate Feature Fusion Scheme. *Sensors* **2021**, *21*, 6954. [[CrossRef](#)] [[PubMed](#)]
6. Choi, Y.-J.; Lee, Y.-W.; Kim, B.-G. Residual-based Graph Convolutional Network (RGCN) for Emotion Recognition in Conversation (ERC) for Smart IoT. *Big Data* **2021**, *9*, 279–288. [[CrossRef](#)] [[PubMed](#)]
7. Lee, Y.; Park, S.-H.; Rhee, E.; Kim, B.-G.; Jun, D. Reduction of Compression Artifacts Using a Densely Cascading Image Restoration Network. *Appl. Sci.* **2021**, *11*, 7803. [[CrossRef](#)]
8. Hettiarachchi, P.; Nawaratne, R.; Alahakoon, D.; De Silva, D.; Chilamkurti, N. Rain Streak Removal for Single Images Using Conditional Generative Adversarial Networks. *Appl. Sci.* **2021**, *11*, 2214. [[CrossRef](#)]
9. Wu, L.; Zhang, H.; Deng, S.; Shi, G.; Liu, X. Discovering Sentimental Interaction via Graph Convolutional Network for Visual Sentiment Prediction. *Appl. Sci.* **2021**, *11*, 1404. [[CrossRef](#)]
10. Kim, S.; Jun, D.; Kim, B.-G.; Lee, H.; Rhee, E. Single Image Super-Resolution Method Using CNN-Based Lightweight Neural Networks. *Appl. Sci.* **2021**, *11*, 1092. [[CrossRef](#)]
11. Yeo, W.-H.; Heo, Y.-J.; Choi, Y.-J.; Kim, B.-G. Place Classification Algorithm Based on Semantic Segmented Objects. *Appl. Sci.* **2020**, *10*, 9069. [[CrossRef](#)]
12. Lee, D.-S.; Kim, J.-S.; Jeong, S.C.; Kwon, S.-K. Human Height Estimation by Color Deep Learning and Depth 3D Conversion. *Appl. Sci.* **2020**, *10*, 5531. [[CrossRef](#)]
13. Lin, Y.-K.; Su, M.-C.; Hsieh, Y.-Z. The Application and Improvement of Deep Neural Networks in Environmental Sound Recognition. *Appl. Sci.* **2020**, *10*, 5965. [[CrossRef](#)]
14. Kim, H.Y.; Yoon, J.W.; Cheon, S.J.; Kang, W.H.; Kim, N.S. A Multi-Resolution Approach to GAN-Based Speech Enhancement. *Appl. Sci.* **2021**, *11*, 721. [[CrossRef](#)]
15. Lee, Y.K.; Park, J.G. Multimodal Unsupervised Speech Translation for Recognizing and Evaluating Second Language Speech. *Appl. Sci.* **2021**, *11*, 2642. [[CrossRef](#)]
16. Das Chakladar, D.; Kumar, P.; Mandal, S.; Roy, P.P.; Iwamura, M.; Kim, B.-G. 3D Avatar Approach for Continuous Sign Movement Using Speech/Text. *Appl. Sci.* **2021**, *11*, 3439. [[CrossRef](#)]
17. Thirukokaranam Chandrasekar, K.K.; Verstockt, S. Context-Based Structure Mining Methodology for Static Object Re-Identification in Broadcast Content. *Appl. Sci.* **2021**, *11*, 7266. [[CrossRef](#)]
18. Liu, H.; Deng, S.; Wu, L.; Jian, M.; Yang, B.; Zhang, D. Recommendations for Different Tasks Based on the Uniform Multimodal Joint Representation. *Appl. Sci.* **2020**, *10*, 6170. [[CrossRef](#)]
19. Tanoh, I.-C.; Napoletano, P. A Novel 1-D CCANet for ECG Classification. *Appl. Sci.* **2021**, *11*, 2758. [[CrossRef](#)]