Journal of Real-Time Image Processing (2021) 18:1167–1171
https://doi.org/10.1007/s11554-021-01156-1

GUEST EDITORIAL

Special issue on deep learning for emerging embedded real-time image and video processing systems

Gwanggil Jeon[1] · **Abdellah Chehri**[2]

Published online: 26 July 2021

1 Introduction

One of the main aims of the multimedia as related to image and video processing is to enable real-time image super resolution or a visually pleasing high-resolution image based on low-resolution image sequences. High resolution images are composed of higher pixel density with fine and more precise details as compared with low-resolution images or video. Many related applications, such as video surveillance, ultra-high definition TV, low-resolution face recognition, and remote image sensing are based on super-resolution techniques. These techniques have attracted high interest from both academia and industry, and currently is an active area of research in image and video processing.

Previously, conventional machine learning techniques, such as supervised and unsupervised learning, reinforcement learning, Bayes classifier, K-means clustering, random forests, and decision trees, etc., have been utilized. Recently, the rapid advancements in deep learning or deep neural networks have shown a promising performance for high resolution scenarios. There remain many research issues regarding the high-resolution aspect. The objective of this special issue is to the application of deep learning for real-time super resolution image and video processing, including new objective functions, new architectures, large scale images, depth images, data acquisition, feature representation, knowledge understanding, and semantic modeling, types of corruption, and new applications. There still exists a gap between extracting representations (or knowledge) from high resolution image and video data and their practical demands.

✉ Gwanggil Jeon
gjeon@inu.ac.kr

Abdellah Chehri
achehri@uqac.ca

[1] Incheon National University, Incheon, South Korea

[2] Université du Québec à Chicoutimi, Chicoutimi, Canada

2 Themes of this special issue

This special issue on "Deep Learning for Emerging Embedded Real-Time Image and Video Processing Systems" is intended to provide representative papers in the current state-of-the-art in the field of real-time image and video processing systems. The ultimate objective is to bring together well-focused, top quality research contributions, providing to the general real time image processing community. We welcomed both theoretical contributions as well as papers describing interesting applications.

Each submitted manuscript found appropriate for this special issue was reviewed on the basis of theoretical originality, technical quality, relevance, originality, significance, and clarity, we finally selected 13 articles. These articles present novel research in Deep Learning for Emerging Embedded Real-Time Image and Video Processing Systems.

2.1 Models

Many scholars are committed to use deep learning methods to study facial expression recognition (FER). In recent year, FER has gradually been confined to psychology research from the early days to now involving knowledge of many disciplines such as physiology, psychology, cognition and medicine. With the extreme achievement of computer vision techniques, various Convolutional Neural Network structures were developed for real-time and accurate facial expression recognition (FER). There are two main problems in the existing convolutional neural network for handling FER problems: insufficient training data caused over-fitting and expression-unrelated intra-class differences. In the contribution by Liu et al. "Two-pathway attention network for real-time facial expression recognition," authors propose a Two-Pathway Attention Network (TPAN) to improve their solution [1]. The authors suppress the intra-class differences efficiently by extracting facial regions based on facial muscle movements driven by facial expressions. They prevent

deep networks from insufficient training data by extensively extracting global structures and local facial regions as the training dataset to feed a two-pathway ensemble model. Furthermore, the authors weight the whole feature maps from the global image and local regions by introducing an attention mechanism module to reweigh each part according to its contribution to FER. The authors adopt real-time facial region extraction and multi-layer feature data compression to ensure the real-time performance of the algorithm and reduce the amount of parameters used in this ensemble model. Experiments on public datasets suggest that this method certifies its effectiveness, reaches human-level performance, and outperforms current state-of-the-art methods with 92.8% on the extended Cohn-Kanade (CK+) and 87.0% on FERPLUS.

Wireless Capsule Endoscopes (WCE) are revolutionary devices for noninvasive inspection of gastrointestinal tract diseases. However, it is tedious and error-prone for physicians to inspect the huge number of captured images. Artificial Intelligence (AI) supports computer-aided diagnostic tools to tackle this challenge. Unlike previous research focusing on the application of large Deep Neural Network (DNN) models for processing images that have been saved on the computer, the authors in the contribution by Wang et al. "A locally-processed light-weight deep neural network for detecting colorectal polyps in wireless capsule endoscopes" propose a light-weight DNN model that has the potential of running locally in the WCE [2]. Thus, only images indicating potential diseases are transmitted, saving energy on data transmission. Several aspects of the design are presented in detail, including the DNN's architecture, the loss function, and data augmentation. The authors explore design parameters of the DNN architecture in several experiments. These experiments use a training dataset of 1222 images and a test dataset with 153 images. The results of their study indicate that their designed DNN has an Average Precision of 91.7% on their test dataset while the parameter storage size is only 29.1 KB, which is small enough to run locally on a WCE. In addition, the real-time performance of the designed DNN model is tested on an FPGA, completing one image classification in less than 6.28 ms, which is much less than the 167 ms needed to achieve real-time operation on the WCE. The authors conclude that their DNN model possesses significant advantages over previous models for WCEs, in terms of model size and real-time performance.

Background subtraction is a substantially important video processing task that aims at separating the foreground from a video in order to make the post-processing tasks efficient. Until now, several different techniques have been proposed for this task but most of them cannot perform well for the videos having variations in both the foreground and the background. In the contribution by Imran et al. "Background subtraction in videos using LRMF and CWM algorithm," a novel background subtraction technique is proposed that aims at progressively fitting a particular subspace for the background that is obtained from L_1-low rank matrix regularization using the cyclic weighted median algorithm and a certain distribution of a mixture of Gaussian noise for the foreground [3]. The expectation maximization algorithm is applied to optimize the Gaussian mixture model. Furthermore, to eliminate the camera jitter effect, the affine transformation operator is involved to align the successive frames. The performance of the proposed method is compared with other state-of-the-art methods and it was concluded that the proposed method performs well in terms of F-measure and computational complexity.

Classification of brain tumors based on the brain magnetic resonance imaging (MRI) results of patients has become an important problem in medical image processing. A computer program that can efficiently analyze brain MRI images of patients in real-time and generate accurate classification results of the tumors in these images can significantly reduce the amount of time needed for diagnosis, which may increase the chances for patients to survive. The contribution by Li et al. "Real-time classification of brain tumors in MRI images with a convolutional operator-based hidden Markov model" proposes a new statistical method that can accurately classify three types of brain tumors based on MRI images, the three types of tumors considered include pituitary tumor, glioma, and meningioma [4]. The features for a pixel in an MRI image are obtained by applying a set of convolutional operators to the neighborhood area of the pixel. For training, a hidden Markov model (HMM) is constructed and trained with a training dataset by computing a statistical profile for the feature vectors for pixels in the tumor regions of each type of brain tumors. In addition, a statistical profile is also obtained for pixels that are in the background of a tumor. For classification, the trained HMM is used to assign labels to pixels in an MRI image with a dynamic programming approach and the classification result of the image is obtained from the labels assigned to the tumor region. Both the training and classification processes can be efficiently performed and does not require the availability of a large amount of computational resources. Experimental results on a large dataset of MRI images show that the proposed method can provide classification results with high accuracy for all three types of brain tumors. A comparison with state-of-the-art methods for brain tumor classification suggests that the proposed method can achieve improved classification accuracy. In addition, real-time analysis also reveals that the proposed approach can probably be used for real-time classification of brain tumors.

2.2 Performance improvements

Recently, numerous methods based on convolutional neural networks (CNNs) have been proposed to attain satisfactory performance in single image super-resolution (SISR). Meanwhile, diverse lightweight CNN-based networks have been proposed to achieve applicability in real-time scenarios. However, the receptive fields in lightweight networks are limited because they do not make good use of multi-scale information. In the contribution by Liu et al. "A lightweight multi-scale feature integration network for real-time single image super-resolution," authors propose a lightweight multi-scale feature integration network (MFIN) to address the above issue [5]. Specifically, to expand the receptive fields for global features, MFIN is constructed by cascading the multi-scale feature integration blocks (MFIBs) in a serial manner. Each MFIB contains a multi-scale feature extraction module (MFEM) and a feature integration unit (FIU). To enlarge the receptive fields at a granular level, the features in MFEM are cascaded in a parallel manner. To capture the full-image dependencies, FIU incorporates the dense and pixel-wise correlations from the outputs of MFEM efficiently. The conducted experiments demonstrate that this method outperforms state-of-the-art methods according quantitative and qualitative evaluation. Notably, the experimental results on run time measurement confirm that the method can achieve real-time performance.

In the past decade, single image super-resolution (SISR) based on convolutional neural networks (CNNs) represented remarkable performance. Powerful characterization of CNN is important for recent methods to learn an intricate nonlinear mapping between high-resolution and corresponding low-resolution images. However, a deeper and wider network structure brings superior performance while increasing the amount of network parameters and calculations so that it is difficult to handle the information in real-time. Hence, it can be embedded in mobile devices only with difficulty. Inspired by the above motivation, the authors' contribution by Yang et al. "Efficient local cascading residual network for real-time single image super-resolution" proposed a lightweight network for the real-time SISR by stacking efficient cascading residual blocks (ECRB), which consist of several concatenated effective modules with wide activation (WCEM) [6]. In order to further improve the network performance, with the increase of a slight number of parameters, the proposed network cooperate with a lightweight residual efficient channel attention (RECA) module to capture feature interaction between channels. Extensive experiments provide significant demonstrations that the proposed network obtains a superior trade-off between performance and parameters compared with other contemporary methods. The lightweight trait of this method allows it to be implemented for real-time image processing and can be embedded in mobile devices.

Melanoma is considered the skin cancer with highest mortality rate that in its advanced state hard to treat. Diagnoses are visually performed by dermatologists. The contribution by Francese et al. "A mobile augmented reality application for supporting real-time skin lesion analysis based on deep learning" proposes an augmented reality smartphone application for supporting the dermatologist in the real-time analysis of a skin lesion [7]. The app augments the camera view with information related to the lesion features generally measured by the dermatologist for formulating the diagnosis. The lesion is also classified by a deep learning approach for identifying melanoma. The real-time process adopted for generating the augmented content is described. The real-time performance is also evaluated and a user study is also conducted. Results revealed that the real-time process may be entirely executed on the Smartphone and that the support provided is well accepted by the dermatologists.

Vehicle detection in videos is a valuable but challenging technology in traffic monitoring. Due to the advantage of real-time detection, Single Shot MultiBox Detector (SSD) is often used to detect vehicles in images. However, the accuracy degradation caused by SSD is one of the significant problems in video vehicle detection. To address this problem in real-time, the contribution by Yang et al. "A fast and effective video vehicle detection method leveraging feature fusion and proposal temporal link" enhances the detection performance by improving the SSD and employing the relationship of inter-frame detections [8]. The authors propose a feature-fused SSD detector and a Tracking-guided Detections Optimizing (TDO) strategy for fast and effective video vehicle detection. The authors introduce a lightweight feature fusion sub-network to the standard SSD network, which aggregates the deeper layer features into the shallower layer features to enhance the semantic information of the shallower layer features. At the post-processing stage of the feature-fused SSD, the non-maximum suppression (NMS) is replaced by the TDO strategy, which links vehicles of inter-frames by fast tracking algorithm. Thus, the missed detections can be compensated by the propagated results, and the confidence of the final results can be optimized according processing time. Their approach significantly improves the temporal consistency of the detection results with computations of lower complexity. The authors evaluate the proposed method on two datasets. The experiments of their labeled highway dataset show that the mean average precision (mAP) of their method is 8.2% higher than that of the base detector. The runtime of their feature-fused SSD is 27.1 frames per second (fps), which is suitable for real-time detection in videos. The experiments on the ImageNet VID dataset prove that the proposed method is comparable with the state-of-the-art detectors as well.

In recent years, convolutional neural networks (CNNs) based methods have achieved remarkable performance for

the single-image super-resolution (SISR) task. However, huge computational complexity and memory consumption of these methods limit their applications on resource-constrained devices. In the contribution by Yang et al. "Lightweight network with one-shot aggregation for image super-resolution," authors propose a lightweight network named OAN to address this problem for image super-resolution [9]. Specifically, to take advantage of diversified features with multiple receptive fields and overcome the inefficiency of dense aggregation, which aggregates all previous feature maps to the subsequent layer, the authors propose a one-shot aggregation block (OAB) as the cascaded block to adopt one-shot aggregation strategy by aggregating the intermediate features with multiple receptive fields only once in the last feature map. Experimental results on benchmark datasets demonstrate that their proposed OAN outperforms the state-of-the-art SR methods in terms of the reconstruction quality, the number of parameters and multiply-accumulate (MAC) operations.

2.3 Applications

Deep neural networks are widely used in computer vision, pattern recognition, and speech recognition and achieve high accuracy at the cost of remarkable computation. High computational complexity and memory accesses of such networks create a big challenge for using them in resource-limited and low power embedded systems. Several binary neural networks have been proposed that exploit only 1-bit values for both weights and activations. Binary neural networks substitute complex multiply-accumulation operations with bitwise logic operations to reduce computations and memory usage. However, these quantized neural networks suffer from accuracy loss, especially in big datasets. In the contribution by Nazari et al. "E2BNet: MAC-Free yet accurate 2-level binarized neural network accelerator for embedded systems," authors introduce a quantized neural network with 2-bit weights and activations that are more accurate compared to the state-of-the-art quantized neural networks, and also the accuracy is close to the full precision neural network [10]. Moreover, the authors propose E2BNet, an efficient MAC-free hardware architecture that increases power efficiency and throughput/W about 3.6× and 1.5×, respectively, compared to the state-of-the-art quantized neural networks. E2BNet processes more than 500 images/sec on ImageNet dataset that not only meet real-time requirements of images/video processing but also can be deployed on high frame rate video applications.

With the development of SAR imaging, the efficiency and rapidity of imaging algorithms have always been a research direction, and real-time imaging is a reflection of the speed of imaging algorithms. For missile-borne SAR, real-time imaging is particularly important because it provides radar with a sense of the battlefield environment. In the case with highly squinted angle, the azimuth space variance and the coupling between range and azimuth dimension will become seriously in imaging, thus an azimuth frequency nonlinear chirp scaling (AFNCS) algorithm is proposed to solve this problem. Based on linear range walk correction, in the contribution by Zhang and Qu "Focusing highly squinted missile-borne SAR data using azimuth frequency nonlinear chirp scaling algorithm," a novel chirp scaling algorithm is adopted to correct the range migration, and then a high-order phase filtering factor is introduced into the azimuth dimension frequency domain to decrease the azimuth space variance [11]. In addition, combined with the SPECtral Analysis method, the image is focused in the Doppler domain. The AFNCS algorithm does not need complex mathematical calculations such as interpolation and can meet the real-time requirements of missile-borne SAR imaging. Simulation results illustrate the effectiveness of the proposed algorithm. The integration of the research in this paper and the deep learning will further pave the way of real-time SAR imaging applications in disaster monitoring, security and surveillance.

The convolution neural network makes deeper and wider for better accuracy, but requires more computations. When the neural network goes deeper, some information will be lost. To improve this drawback, the residual structure was developed to connect the information of the previous layers. This is a good solution to prevent the loss of information, but it requires a huge amount of parameters for deeper layer operations. In the contribution by Hsia et al. "Convolution neural network with low operation FLOPS and high accuracy for image recognition," a fast computational algorithm is proposed to reduce the parameters and to save the operations with the modification of DenseNet deep layer block [12]. With channel merging procedures, this solution can reduce the dilemma of multiple growth of the parameter quantity for deeper layer. This approach is not only to reduce the number of parameters and FLOPs, but also to keep high accuracy. Comparisons with the original DenseNet and ResNet-110, the parameters can be efficiently reduced about 30–70%, while the accuracy degrades lightly. The lightweight network can be implemented on a low-cost embedded system for real-time application.

In the contribution by Lee et al. "A real-time high-speed autonomous driving based on a low-cost RTK-GPS," the feasibility of a low-cost real-time kinematic GPS (RTK-GPS) sensor for localization of an autonomous vehicle to achieve a high-speed driving is studied [13]. For achieving high-speed autonomous driving, although the image-based method combined with a GPS can be a good approach for localization, the RTK-GPS position can be utilized on low-cost. On high-speed driving, it is important to acquire an accurate localization because a less accurate position may

degrade the performance to follow the waypoints on a target path with proper driving stability. Thus, in this study, the RTK-GPS position was applied to reduce position errors in a test vehicle with a low-cost GPS and the RTKLIB. A modified adaptive look-ahead distance pure pursuit algorithm was implemented to control the test vehicle. An autonomous driving experiment using the RTK-GPS position was carried out to verify the performance of the vehicle at a high-speed of 130 kph on a track being 2477 m long with various corners and inclinations in a racing circuit. The test vehicle with the proposed real-time autonomous driving system using the RTK-GPS position achieved 111 s to complete a full lap on the racing track without departures from the track and noticeable lateral errors. This result was 32 s slower than the record accomplished by a professional human driver's 79 s.

Intelligent search techniques and an intelligent agent for smart search are useful in many application domains. In the contribution by Seo et al. "Realtime fire detection using CNN and search space navigation," the authors develop a state space navigational model for intelligent agents aimed at industrial surveillance from fire hazards [14]. They focus is on fire detection using the convolution neural network then proactively search the area which are more likely to have routes toward the target. This problem can be simulated into an optimization problem over a state space, which can be figure out effectively through a greedy algorithm. The authors also compare this approach with both uninformed and informed search algorithms. They evaluate this proposed system using various search algorithms for search and rescue agent. The analysis of the results obtained demonstrate the efficiency of the system.

3 Conclusion

The articles presented in this special issue provides insights in fields related to Deep Learning for Emerging Embedded Real-Time Image and Video Processing Systems, including models, performance evaluation and improvements, and application developments. The guest editors wish that the readers can benefit from insights of these papers, and contribute to these rapidly growing areas. We also hope that this special issue would shed light on major developments in the area of Real-Time Image Processing and attract attention by the scientific community to pursue further investigations leading to the rapid implementation of these technologies.

Acknowledgements We would like to express our appreciation to all the authors for their informative contributions and the reviewers for their support and constructive critiques in making this special issue possible. Finally, we would like to express our sincere gratitude to the Editors-in-Chief, for providing us with this unique opportunity to present these works in *Journal of Real-Time Image Processing.*

References

1. Wang, L., He, Z., Meng, B., et al.: Two-pathway attention network for real-time facial expression recognition. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01123-w
2. Wang, Y., Yoo, S., Braun, J.M., et al.: A locally-processed lightweight deep neural network for detecting colorectal polyps in wireless capsule endoscopes. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01126-7
3. Munir, W., Siddiqui, A.M., Imran, M., et al.: Background subtraction in videos using LRMF and CWM algorithm. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01120-z
4. Li, G., Sun, J., Song, Y., et al.: Real-time classification of brain tumors in MRI images with a convolutional operator-based hidden Markov model. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01072-4
5. He, Z., Liu, K., Liu, Z., et al.: A lightweight multi-scale feature integration network for real-time single image super-resolution. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01142-7
6. Yang, H., Dou, Q., Liu, K., et al.: Efficient local cascading residual network for real-time single image super-resolution. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01134-7
7. Francese, R., Frasca, M., Risi, M., et al.: A mobile augmented reality application for supporting real-time skin lesion analysis based on deep learning. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01109-8
8. Yang, Y., Song, H., Sun, S., et al.: A fast and effective video vehicle detection method leveraging feature fusion and proposal temporal link. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01121-y
9. Tang, R. et al.: Lightweight network with one-shot aggregation for image super-resolution. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01127-6
10. Mirsalari, S.A. et al.: E2BNet: MAC-free yet accurate 2-level binarized neural network accelerator for embedded systems. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01148-1
11. Zhang, Y., Qu, T.: Focusing highly squinted missile-borne SAR data using azimuth frequency nonlinear chirp scaling algorithm. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01135-6
12. Hsia, S.C., Wang, S.H., Chang, C.Y.: Convolution neural network with low operation FLOPS and high accuracy for image recognition. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01140-9
13. Park, S., Ryu, S., Lim, J., et al.: A real-time high-speed autonomous driving based on a low-cost RTK-GPS. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01084-0
14. Rahmatov, N., Paul, A., Saeed, F., Seo, H.: Realtime fire detection using CNN and search space navigation. J. Real-Time Image Proc. (2021). https://doi.org/10.1007/s11554-021-01153-4