APSIPA Transactions on Signal and Information Processing, 2024, 13, e406

Editorial

Editorial for Special Issue on Invited Papers from APSIPA ASC 2023

Jia-Ching Wang[1], Hsin-Min Wang[2], Wen-Hsiao Peng[3] and Chia-Hung Yeh[4]

[1] *National Central University, Taiwan*
[2] *Academia Sinica, Taiwan*
[3] *National Yang Ming Chiao Tung University, Taiwan*
[4] *National Taiwan Normal University, Taiwan*

We are delighted to present this special issue of APSIPA Transactions on Signal and Information Processing, which features a selection of outstanding papers presented at the APSIPA ASC 2023, held from October 31 to November 3 in Taipei, Taiwan. This special issue aims to highlight the significant advancements and emerging trends discussed at the conference, showcasing cutting-edge research and innovative solutions in the field of signal and information processing.

The APSIPA ASC 2023 served as a vibrant platform for the exchange of ideas, knowledge, and breakthroughs across a range of topics. The conference attracted a diverse group of experts, whose presentations and discussions led to a wealth of new insights and advancements. Following the conference, a rigorous selection process was undertaken to identify papers that represent the highest standards of research contributions in APSIPA ASC 2023. This special issue has collected five excellent articles reviewed and highly recommended by the editors and reviewers.

The first paper is titled “A Lightweight Enhancement Approach for Real-Time Semantic Segmentation by Distilling Rich Knowledge from Pre-Trained Vision-Language Model”, authored by Chia-Yi Lin, Jun-Cheng Chen and Ja-Ling Wu. This paper addresses the challenge of improving real-time semantic segmentation by leveraging the rich textual knowledge embedded in vision-language models like CLIP. Real-time semantic segmentation is crucial in applications such as autonomous driving and augmented reality, where both speed and accuracy are key. The paper proposes a lightweight framework that distills the knowledge from CLIP’s text encoder into a segmentation model,

ISSN 2048-7703; DOI: 10.1561/116.20240103

aligning visual and textual embeddings for enhanced semantic understanding. The approach introduces class-specific, learnable prompts that optimize textual guidance for each class in the segmentation model. This process improves the model's performance without significantly increasing latency, making it practical for real-time scenarios. The two-stage training procedure- first aligning the segmentation backbone with CLIP's embeddings and then optimizing the class-specific learnable prompts- demonstrates improved accuracy across multiple benchmark datasets while maintaining high processing speed.

The second paper is titled "Multi-Modal Pedestrian Crossing Intention Prediction with Transformer-Based Modell", authored by Ting-Wei Wang and Shang-Hong Lai. This paper addresses the critical problem of predicting pedestrian crossing intentions, which is essential for the safety of autonomous driving systems and advanced driver assistance systems (ADAS). It presents a cutting-edge multi-modal framework using transformer-based models, integrating various sources of data like pedestrian posture, traffic lights, crosswalks, and road signs to enhance prediction accuracy. This approach represents a significant advancement over earlier methods, as it incorporates lifted 3D human pose data and 3D head orientation to provide a more comprehensive understanding of pedestrian behavior. The paper's experimental results demonstrate that the model achieves state-of-the-art performance on benchmark datasets, making a notable contribution to improving road safety in autonomous driving systems.

The third paper is titled "Meta Soft Prompting and Learningl", authored by Jen-Tzung Chien Ming-Yen Chen, Ching-Hsien Lee and Jing-Hao Xue. This paper tackles the challenge of improving language models' generalization across unseen domains in natural language understanding (NLU). It addresses the problem of domain shift where models trained on one domain perform poorly on others, which is critical for improving few-shot learning and domain adaptation tasks. The authors introduce a novel method called Meta Soft Prompting, which enhances the adaptability of pretrained language models (PLMs) without requiring extensive retraining or domain-specific data. By utilizing a parameter efficient learning framework, the proposed approach integrates meta learning with soft prompt optimization, allowing the PLM to adjust its predictions across multiple unseen domains. This innovative approach enables robust few-shot unsupervised domain adaptation, where minimal data from new domains is needed to achieve high performance. Notably, the method improves performance in low-resource settings and demonstrates significant advancements in multi-domain sentiment classification, showcasing cutting-edge techniques for cross-domain language modeling and signal processing.

The fourth paper is titled "Estimating 3D Hand Poses and Shapes from Silhouettesl", authored by Li-Jen Chang, Yu-Cheng Liao, Chia-Hui Lin, Shih-Fang Yang-Mao and Hwann-Tzong Chen. This paper presents a novel method, Mask2Hand, for predicting 3D hand poses and shapes using only 2D binary silhouettes. This method addresses a significant challenge in hand pose estimation, traditionally requiring RGB or depth data, by using minimal input, making it more accessible and versatile. The proposed approach leverages differentiable rendering and a tailored loss function to project 3D estimations onto 2D silhouettes for end-to-end optimization. Mask2Hand's performance is shown to be comparable to state-of-the-art methods while requiring less complex input data, offering a self-supervised learning mechanism that eliminates the need for manual annotations. This makes it a promising solution for applications involving low-resolution sensors or environments with limited visual data.

The fifth paper is titled "End-to-End Singing Transcription Based on CTC and HSMM Decoding with a Refined Score Representationl", authored by Tengyu Deng, Eita Nakamura, Ryo Nishikimi and Kazuyoshi Yoshii. This paper addresses the problem of automatic singing transcription (AST), which is a crucial task in music information retrieval. The challenge lies in accurately transcribing vocal performances into symbolic musical scores, with significant potential applications in areas like music search and interpretable emotion recognition. The authors propose a novel approach that combines Connectionist Temporal Classification (CTC) with a Hidden Semi-Markov Model (HSMM) to improve transcription accuracy. Their key innovation is a metrical-position-based (MP-based) score representation that captures the onset times of notes relative to barlines, thus preserving the musical structure even when minor errors occur. This method outperforms traditional note-value-based (NV-based) approaches by ensuring a more stable estimation of score times and metrical structures. Experimental results demonstrate that this method achieves state-of-the-art performance in transcription accuracy and robustness.

Guest Editors

Jia-Ching Wang
Hsin-Min Wang
Wen-Hsiao Peng
Chia-Hung Yeh