ASMMC21: the 6th International Workshop on Affective Social Multimedia Computing

Dongyan Huang, Björn Schuller, Jianhua Tao, Lei Xie, Jie Yang

Angaben zur Veröffentlichung / Publication details:

Huang, Dongyan, Björn Schuller, Jianhua Tao, Lei Xie, and Jie Yang. 2021. "ASMMC21: the 6th International Workshop on Affective Social Multimedia Computing." In *ICMI '21: Proceedings of the 2021 International Conference on Multimodal Interaction, Montréal, QC, Canada, October 18 - 22, 2021*, edited by Zakia Hammal, Carlos Busso, Catherine Pelachaud, Sharon Oviatt, Albert Ali Salah, and Guoying Zhao, 864–67. New York, NY: ACM. https://doi.org/10.1145/3462244.3480980.

ASMMC21: The 6th International Workshop on Affective Social Multimedia Computing

Dong-Yan Huang
UBTECH Robotics, USA
dongyan.huang@ubtrobot.com

Björn Schuller
IEEE, USA
schuller@ieee.org

Jianhua Tao
National Laboratory of Pattern Recognition, China
jhtao@nlpr.ia.ac.cn

Lei Xie
Northwestern Polytechnical University, China
jyang@nsf.gov

Jie Yang
IIS, NSF, USA
jyang@nsf.gov

ABSTRACT

Affective social multimedia computing is an emergent research topic for both affective computing and multimedia research communities. Social multimedia is fundamentally changing how we communicate, interact, and collaborate with other people in our daily lives. Comparing with well-organized broadcast news and professionally made videos such as commercials, TV shows, and movies, social multimedia media computing imposes great challenges to research communities. Social multimedia contains much affective information. Effective extraction of affective information from social multimedia can greatly help social multimedia computing (e.g., processing, index, retrieval, and understanding). Although much progress have been made in traditional multimedia research on multimedia content analysis, indexing, and retrieval based on subjective concepts such as emotion, aesthetics, and preference, affective social multimedia computing is a new research area. The affective social multimedia computing aims to proceed affective information from social multi-media. For massive and heterogeneous social media data, the research requires multidisciplinary understanding of content and perceptual cues from social multimedia. This workshop served as a successful step towards this goal and attracted contributions from different research disciplines on the analysis of affective signals in interaction (multimodal analyses enabling artificial agents in Human-Machine Interaction, social Interaction with artificial agents) and social multimedia (e.g., twitter, wechat, weibo, youtube, facebook, etc). This paper provides a summary of the activities of the workshop and the accepted papers and abstracts.

ACM Reference Format:
Dong-Yan Huang, Björn Schuller, Jianhua Tao, Lei Xie, and Jie Yang. 2021. ASMMC21: The 6th International Workshop on Affective Social Multimedia Computing. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3462244.3480980

ACM ISBN 978-1-4503-8481-0/21/10.
https://doi.org/10.1145/3462244.3480980

1 HERE IS TECHNICAL PROGRAM

Date: 18 OCT. 2021

- 08:45-09:00 Opening - Welcome
- 09:00-10:00 Kenote Speaker I
- 10:00-10:30 Coffee Break
- 10:30 - 12:30 Multimodal Emotion Recognition
- 12:30 - 14:30 Lunch Time
- 14:30 - 15:30 Keynote Speaker II
- 15:30 - 16:00 Coffee Break
- 16:00-17:40 Multimodal Emotion Synthesis

2 ACCEPTED PAPERS AND ABSTRACTS

BERT Based Cross-Task Sentiment Analysis with Adversarial Learning

Zhiwei He, Xiangmin Xu, Xiaofen Xing and Yirong Chen

Abstract: Sentiment Analysis (SA) is an essential task in natural language processing. Generally, previous sentiment analysis models focus on a single subtask. However, a generalized SA agent is expected with the ability to learn knowledge from one task and use it in other relevant tasks. Consequently, we formulate this challenge as an unsupervised task adaption problem and propose TAL-IS, a simple and efficient approach to finetune cross-task SA model. In this approach, we use Task Adversarial Learning (TAL) with a BERT-specific Input Standardization (IS) scheme to obtain both emotion-discriminative and task-invariant contextual features. To the best of our knowledge, our work is the first attempt to propose a cross-task model for SA subtasks with unsupervised task adaption. Experiments show that our proposed model outperforms the general finetuning method and can learn knowledge effectively cross SA subtasks.

Aspect-based Sentiment Analysis with Weighted Relational Graph Attention Network

Yingtao Huo, Dongmei Jiang and HichemSahli

Abstract: The aim of aspect-based sentiment analysis (ABSA) is to determine the sentiment polarity of a specific aspect in a sentence. Most recent works resort to exploiting syntactic information by utilizing Graph Attention Network (GAT) over dependency trees, and have achieved great progress. However, the models based on traditional GAT do not fully exploit syntactic information such as

the diversified types of dependency relations. The variant of GAT called relational graph attention network (R-GAT) takes different types of dependency relations into consideration, but ignores the information hidden in the word-pairs. In this paper, we propose a novel model called weighted relational graph attention network (WRGAT). It can exploit more accurate syntactic information by employing a weighted relational head, in which the contextual information from word-pairs is introduced into the computation of the attention weights of dependency relations. Furthermore, we employ BERT instead of Bi-directional Long Short-term Memory (Bi-LSTM) to generate contextual representations and aspect representations respectively as inputs to the WRGAT, and adopt an index selection method to keep the word-level dependencies consistent with the word-piece unit of BERT. With the proposed BERT-WRGAT architecture, we improve the state-of-the-art performances on four ABSA datasets, yielding gains in accuracy of 0.58% to 3.24%.

Semantic and Acoustic-Prosodic Entrainment of Dialogues in Service Scenarios

Liu Yuning, JianwuDang, AijunLi and Di Zhou

Abstract: According to the Communication Accommodation Theory, speakers dynamically adjust their communication behaviors, converging to or diverging from their interlocutors in order to diminish or increase social distance, which is called entrainment. Most of the studies investigated the entrainment of the interlocutors in terms of linguistic and paralinguistic features respectively, but paid less attention to the (dis)entrainment relation between paralinguistic and linguistic ones. In this study, we employed BERT to extract the semantic similarities of turns within dialogues in service scenarios, and found the semantic entrainment. We also found that (dis)entrainments policies were adopted between acoustic-prosodic (paralinguistic) and linguistic (semantic)features. These findings will contribute to fully understanding the mechanism of entrainment in dialogue.

Improving Model Stability and Training Efficiency in Fast Speed High Quality Expressive Voice Conversion System

ZHIYUAN ZHAO, JINGJUN LIANG, ZEHONG ZHENG, LINHUANG YAN, ZHIYONG YANG, WAN DING, and DONGYAN HUANG

Abstract : Recently, voice conversion systems (VC) have been made great progress thanks to advanced deep learning methods. Current researches not only pursues high-quality and fast speed of audio synthesis, but also requires richer expressiveness. The most popular VC system was constructed on the concatenation of an automatic speech recognition module with a text-to-speech module (ASR-TTS). However, it suffers from errors of recognition and pronunciation and the requirements of a large amount of data for a pre-trained ASR model. To remedy the above drawbacks, we propose an approach to enhance the model stability and training efficiency of a VC system. Firstly, a data redundancy reduction method is used to balance the distribution of vocabulary for avoiding uncommon words being ignored during the training process, by adding CTC loss, the WER of our system declines to 3.02, which is 5.63% lower than that of the ASR-TTS system (8.65%), and the inference speed (e.g., real-time rate 19.32) of our VC system are much faster than those of the baseline (real-time rate 2.24). Finally, the emotional embedding is added to the pre-trained VC system to generate expressive speech conversion. The results show that after fine-tuning on the multi-emotional dataset, the system can achieve high quality and expressive speech synthesis.

Facial Micro-Expression Recognition Based on Multi-Scale Temporal and Spatial Features

Hao Zhang, Bin Liu, Jianhua Tao and Zhao Lv

Abstract: Micro-expression is a kind of facial activity with weak changes and short duration that can reflect people's true feelings. For micro-expression recognition, it is not only necessary to extract the spatial feature information of the face movement changes in the image, but also to consider the time series information of the continuous image sequence. Thus, we propose a multiple aggregation networks to verify the impact of local facial regions and temporal features on micro-expression recognition in detail. It can learn the temporal and spatial feature from the whole micro expression video frame and combined the local region where the micro-expression mainly occurs with the global region. The spatial features of micro-expressions frames are extracted by 3D CNN, and the extracted video sequences features are input into LSTM processing temporal features. Experiments from two public datasets, CASME-II and SAMM, show that our method obtains higher performance than several existing studies.

FER by Modeling the Conditional Independence between the Spatial Cues and the Spatial Attention Distributions

Wan Ding, Dongyan Huang, Jingjun Liang, Jinlong Jiao and Zhiping Zhao

Abstract: This paper presents a novel approach for FER. The spatial cues, for example the locations of face components such as eyes and mouth, play an important role to guide the spatial attentions for FER. Traditional approaches define the relations between the spatial cues and the spatial attention distributions based on linear models. However there also exists non-linear relations between them, in which case the spatial cues and the spatial attention distributions can be conditional independent. In this paper we model the conditional independence based on the state-of-the arts framework of the attention models for FER. We design the spatial cues as the hyper-parameters to affect the metric for spatial attention calculation.We exploit the Global-Attention (no spatial cues), Local-Attention (spatial cues affect the attention distributions) and Self-Attention (spatial cues as the hyper-parameters to affect the attention metric) as three different configurations. The experimental results show that the Self-attention achieves the best performances (68.5% on FER2013 Dataset and 49.8% on EmotiW2017 Dataset) which improves the accuracies by 2.8 % (on FER2013) and 1% (on EmotiW2017) compared with the Global-attention. The experimental results support the idea that non-linear modeling the relations between the spatial cues and the spatial attention distributions can improve the performances for FER.

Temporal Attentive Adversarial Domain Adaption for Cross Cultural Affect Recognition

Haifeng Chen, Yifan Deng and Dongmei Jiang

Abstract : Continuous affect recognition is becoming an increasingly attractive research topic, recent works mainly focus on modeling the temporal dependency and multi-modal fusion to boost the performance. Despite recent improvement, the cross-cultural affect recognition in videos is still not well-explored. In this paper, we propose the temporal attentive adversarial domain adaption for cross cultural affect recognition. The LSTM is firstly used to encode the contextual representation for each frame. Then, a DNN based regressor is used to estimate the affective dimension arousal or valence, and optimized to promote the encoded representation is emotion discriminative. In addition, another DNN based sequence level culture classifier, which takes the fused representation of each frame as the input, is used to recognize the culture of the input sequence, and optimized to encourage the encoded representation is culture invariant. Since different frames over a video may contribute not equally in recognizing the culture, we propose to add another frame level culture classifier, which could adaptively and attentively assign more weighting scores for the important frames for recognizing the culture. The proposed method is evaluated on the benchmark dataset AVEC2019 CES. Our experimental results show that the proposed method improves the performance compared to state-of-the-art methods, with the concordance correlation coefficient (CCC) reaching 0.576 for arousal and 0.472 for valence, on the cross cultural test set.

A Multimodal Dynamic Neural Network for Call for Help Recognition in Elevators

Ran Ju, Huangrui Chu, Yechen Wang, Qi Deng, Ming Cheng and Ming Li

Abstract : As elevator accidents do great damage to people's lives and property, taking immediate responses to emergent calls for help is necessary. In most emergency cases, passengers must use the "SOS" button to contact the remote safety guard. However, this method is unreliable when passengers lose the ability of body movement. To address this problem, we define a novel task of identifying real and fake calls for help in elevator scenes. Given that the limited call for help dataset collected in elevators contains multimodal data of real and fake categories, we collected and constructed an audiovisual dataset dedicated to the proposed task. Moreover, we present a novel instance-modality-wise dynamic framework to efficiently use the information from each modality and make inferences. Experimental results show that our multimodal network improves the performance on the call for help multimodal dataset by 2.66% (accuracy) and 1.25% (F1 Score) with respect to the pure audio model. Besides, our method outperforms other methods on our dataset.

A Web-Based Longitudinal Mental Health Monitoring System

Zhiwei Chen, Weizhao Yang, Jinrong Li, Jiale Wang, Shuai Li, Ziwen Wang and Lei Xie

Abstract: Depression disorder has become an increasingly prominent problem, which brings heavy burdens to individuals, families and society. Clinical assessments are heavily relied on the questionnaire tables where patients' daily behavior, sleeping, and mood performance in the past two weeks are important metrics. However, the information obtained through the patient's review of the past two weeks' experience is neither timely nor objective. Moreover, while patients have medicine at home, doctors lose the way of monitoring and intervening them on time. In this paper, we propose and implement a web-based longitudinal mental health monitoring system that can be used on mobile phones, pads or other platforms. It consists of the user end of patient, the server end, and the doctor end. The patients can report their daily information through ecological momentary assessment (EMA), share their emotions in speech or face video, test their depression severities through the PHQ-9 questionnaire table or face videos recorded while going through a semi-structured interview, and check their recent history of activity, sleeping, emotion log, and depression severity etc. The server end performs emotion recognition and depression severity estimation on the pre-trained deep learning models. The doctor can manage the information of all the patients under his(her) supervision, monitor their recent status, and edit their depression severity scales after clinical diagnosis.

Noise Robust Singing Voice Synthesis Using Gaussian Mixture Variational Autoencoder

Heyang Xue, Xiao Zhang, Jie Wu, Jian Luan, Yujun Wang and Lei Xie

Abstract: Generating high-quality singing voice usually depends on a sizable studio-level singing corpus which is difficult and expensive to collect. In contrast, there is plenty of singing voice data can be found on Internet. However, the found singing data may be mixed by accompaniments or contaminated by environmental noises due to recording conditions. In this paper, we propose a noise robust singing voice synthesizer which incorporates Gaussian Mixture Variational Autoencoder (GMVAE) as the noise encoder to handle different noise conditions, generating clean singing voice from lyrics for target speaker. Specifically, the proposed synthesizer learns a multi-modal latent noise representation of various noise conditions in a continuous space without the use of an auxiliary noise classifier for noise representation learning or clean reference audio during the inference stage. Experiments show that the proposed synthesizer can generate clean and high-quality singing voice for target speaker with MOS close to reconstructed singing voice from ground truth mel-spectrogram with Griffin-Lim vocoder. Experiments also show the robustness of our approach under complex noise conditions.

Unaligned multimodal for depression assessment from speech

Keru Wang and Ziping Zhao

Abstract: A growing area of mental health research is how to automatically assess depression degree according to multimodal-based objective markers. However, when combined with machine learning, this research can be challenging due to unaligned multimodal sequences and a limited amount of annotated training

data. In this paper, a novel cross-modal framework is proposed for automatic depression severity assessment. The low-level descriptions (LLDs) from multiple clues, such as text, audio and video are extracted, and multimodal fusion by using cross-modal attention mechanism is utilised to learn more accurate feature representations. For extracted feature from each modality, the cross-modal attention mechanism is utilised to continuously update the input sequence of the target mode, and finally the score of the patient's health questionnaire (PHQ) -8 can be obtained. Moreover, Self-Attention Generative Adversarial Networks (SAGAN) is employed to increase the amount of the training data for depression severity analysis. Experimental results on the depression sub-challenge dataset of the Audio/Visual Emotion Challenge (AVEC) 2017 and AVEC 2019 demonstrate the effectiveness of proposed method.

FACE MASK DETECTOR : A SINCERE EFFORT TO PUT AN END TO SPREAD OF DEADLY CORONAVIRUS

Ranjith Paul

Abstract: By the end of 2019 the world had witnessed a deadly disease named COVID-19 which is caused by deadly Coronavirus. There is no effective way to put an end to coronavirus till now. Coronavirus spreads if someone gets in contact with the person who was infected by Coronavirus. To put an end to this deadly virus there is no effective way rather than wearing a face mask and maintaining social distancing. So to determine weather a person have mask on his face or not we use face mask detection algorithm which is the extinction of convolutional neural networks. For developing face mask detector ,we use. Tensor flow, Keras, Matplotlib, OpenCv , NumPy and some other python in built modules. We train the system with data set which consists of image with mask and images without mask. Experimental values shows accuracy of the mask worn in percentages.