Pattern Analysis and Applications (2022) 25:489–492
https://doi.org/10.1007/s10044-022-01075-2

EDITORIAL

Guest Editorial: Special issue on computer vision and machine learning for healthcare applications

Cristina Palmero[1] · Maria Inés Torres[2] · Anna Esposito[3] · Sergio Escalera[1]

Published online: 7 May 2022

Recent advances in technology have boosted the development and release of active assistive living devices, based on wearable and/or non-obtrusive visual and multi-modal signals for e-health and welfare support. These solutions are seamlessly integrated in the environment, such as sensor-based systems installed in elderly people's homes for ambient monitoring and intelligent visual warning. In addition, research on ubiquitous computing has favored the implementation of more user-centered applications such as virtual tutoring, coaching agents, physical rehabilitation, and psychological therapy systems. To allow these systems to provide features that satisfy the user's requirements, expectations, and acceptance, the tendency is now shifting toward the conception of empathic solutions, tailored to personalized user needs. The new assistive systems must be able to understand the user's behaviors, mood, and intentions, and react to them accordingly in real time, as well as detect timely changes in behaviors and health states. Furthermore, such systems are expected to infer the user's traits, attitudes, and psychological profile to deliver a more personalized user-machine interaction. These solutions require advanced computer vision and machine learning techniques, such as facial expression analysis, gaze and pose estimation, and gesture recognition, in addition to behavioral and psychological theories for modeling individual's profiles. While these tasks are currently obtaining outstanding performances in controlled and prototypical environments (e.g., detection of facial expressions of emotion on static faces), the challenge falls on their integration and application in naturalistic scenarios, where extensive sources of variability (pose, age, behaviors, moods, illumination conditions, dynamic speaking emotional faces, among others) affect the processing of the detected signals.

This special issue aimed to collect the latest approaches and findings related to machine learning and computer vision-based healthcare applications, with a special emphasis on e-health and welfare via single and multi-modal face, gesture, and pose analysis. The special issue was preceded by the First Workshop on Faces and Gestures for e-Health and Welfare (FaGEW), held in conjunction with the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG, November 2020 in Buenos Aires, Argentina). The call for papers of the special issue was open to all researchers working on the topic. All submitted articles have undergone rigorous peer-review according to the journal's high standards, with a final acceptance rate of 30%, considering only articles within the scope of the special issue. The special issue consists of 13 papers, which can be classified into face, gesture, and body pose analysis based on the visual modality only, and multi-modal processing from video, audio, text, and wearable data. Most papers propose novel deep learning-based approaches to tackle different health aspects (stress, pain, bipolar disorder, or depression detection, mood analysis, and cognitive and physical rehabilitation) or other human-oriented applications (sign language recognition, gesture analysis for task understanding, thermal comfort assessment). We briefly discuss all papers below.

✉ Cristina Palmero
crpalmec7@alumnes.ub.edu

Maria Inés Torres
manes.torres@ehu.es

Anna Esposito
iiass.annaesp@tin.it

Sergio Escalera
sergio@maia.ub.es

[1] Universitat de Barcelona and Computer Vision Center, Barcelona, Spain

[2] Universidad del País Vasco UPV/EHU, Bilbao, Spain

[3] Università della Campania 'Luigi Vanvitelli', Caserta, and International Institute for Advanced Scientific Studies, Salerno, Italy

1 Face analysis

In *Automatic Stress Analysis from Facial Videos Based on Deep Facial Action Units Recognition*, the authors propose a deep learning pipeline to perform visual recognition of stress from sequences of RGB face images. The deep learning pipeline learns from geometry and appearance features from aligned faces to regress the intensity of a set of facial Action Units (AUs) based on the Facial Action Coding System (FACS). These intensity values of face muscle activations are used to learn to classify stress bouts. Promising results are achieved for both AUs and stress category recognition. The paper also explores the relation between AUs and stress labels.

Geometric- and appearance-based AU recognition is also performed in *Facial Action Unit Detection Methodology with Application in Brazilian Sign Language Recognition*, this time to aid in the recognition of Brazilian Sign Language. More concretely, the authors identified a large set of AUs that appear in faces showing expressions when signing. The high recognition performance of a large number of facial expressions shows the utility of the method to be applied to future sign language recognition by complementing hand and body signing information.

Instead of AUs, discrete emotion categories are used in *Deep Transfer Learning in Human–Robot Interaction for Cognitive and Physical Rehabilitation Purposes*. The paper proposes an assistive living framework to support Traumatic Brain-Injured (TBI) patients in their daily living activities and rehabilitation exercises. Patients' expressions were collected using RGB, depth, and thermal sensors while performing daily activities like physiotherapy, rehabilitation, and social communication activities. A deep convolutional-recurrent network is first trained on a set of existing face datasets, and then fine-tuned on the TBI dataset via transfer learning. Results show state-of-the-art performance of the facial expression recognition model, capable of recognizing emotions of people with facial paralysis. Furthermore, the model is integrated into a SoftBank Pepper robot to decode patients' emotions and help staff members and care workers to better understand patients' emotional reaction during rehabilitation, enhancing physiotherapy effectiveness and social interaction.

Facial expressions are also leveraged in *Real-Time Pain Detection from Facial Expressions Using Domain Adaptation Technique*. The authors propose a method for facial pain recognition that can adapt to new subjects with very little data. A basic deep learning pipeline extracts appearance features from face detected regions that are mapped to recognize 5 pain categories on existing datasets. This pre-trained model is then fine-tuned on a specific pain dataset using a meta-learning model, showing a fast adaptation to recognize pain to new subjects with just limited data, e.g., 1 or 5 samples.

On a different note, the work presented in *Assessing Facial Symmetry and Attractiveness using Augmented Reality* uses 10 specific facial landmarks and 7 facial features for assessing facial symmetry and attractiveness. As stated in the paper, this could serve as an aid in several key application areas ranging from plastic surgery to rehabilitative health-related tasks. The authors embed the proposed solution into an augmented reality-based smartphone application. The proposed solution outperforms other baseline methods and achieves remarkable performance across the multiple datasets considered, while retaining comparable computation time to class-leading methods.

Finally, the study performed in *Age and Gender Effects on the Human's Ability to Decode Posed and Naturalistic Emotional Faces* investigates factors affecting human abilities to decode emotional facial expressions. The goal is to understand how the decoding accuracy of facial emotional expressions affects human social functioning abilities, and to investigate users' requirements and expectations to the demand of implementing socially believable technologies devoted to assisting vulnerable people. The factors investigated are the age and gender of the interpreters, and age, gender, and type of stimuli. The results from the proposed experiments add knowledge on the ability of differently aged individuals to decode differently aged faces, and shed light on the different patterns of relations existing among emotion recognition skills, social competences, and social-emotional behaviors.

2 Gesture analysis

The paper entitled *JSE: Joint Semantic Encoder for Zero-Shot Gesture Learning* proposes to use the zero-shot learning paradigm for gesture recognition (ZSL). Contrarily to other object classification problems where ZSL has been used due to the large number of available examples, data for gestures are scarce. The lack of large datasets for gestures and the complexity associated with the fact that gestures are dynamically context-embedded makes the proposed approach particularly interesting. The authors propose to apply ZSL to three different feature extraction techniques, which are shown to be relevant for the task: a) velocity features concatenating a fixed number of frames using interpolation; b) heuristic features exploiting prior knowledge on gesture recognition tasks and their semantic descriptors; c) latent features consisting in categorizing the gesture to be recognized into a fixed number of classes.

The work presented in *Assessing Task Understanding in Remote Ultrasound Diagnosis* via *Gesture Analysis* combines two elements: 1) the Multi-Agent Gestural Instructor Comparer (MAGIC) framework, used to represent and compare gestures performed by helper-worker pairs during ultrasound training activities; with 2) the Physical Instructions Comparison (PIA) metric, used to evaluate how well gestures are being used to communicate and execute physical instructions. MAGIC considers the users' skeletal information to represent the gestures' shape and movement, in addition to extracted audio information to provide meaning and context to the gesture. The work compares the use of MAGIC and PIA with other gesture representations and metrics, in addition to different gesture matching approaches. The proposed pipeline outperforms other approaches, and PIA was found to be significantly correlated with other usual metrics as well as with participants' overall task understanding, demonstrating that gestures can be used to estimate task understanding for physically based instructions.

Finally, in *SL-Animals-DVS: Event-Driven Sign Language Animals Dataset*, the authors propose a sign language dataset based on an event-based Dynamic Vision Sensor (DVS). The DVS records the upper body of non-fluent signers performing a small set of isolated words derived from sign language of various animals as a continuous spike flow at very low latency. The paper benchmarks the recognition performance on the provided dataset using three state-of-the-art Spiking Neural Networks (SNN) recognition systems. SNNs are naturally compatible to make use of the temporal information that is provided by the DVS where the information is encoded in the spike times. Initial results are encouraging and show the utility of the proposed data and models.

3 Body pose and activity analysis

The paper entitled *Automatic Estimation of Clothing Insulation Rate and Metabolic Rate for Dynamic Thermal Comfort Assessment* is the only paper of the special issue that explicitly performs body pose estimation as part of its proposed approach. The authors deal with the detection of the dynamic individual thermal comfort as a key factor to be considered to reduce the energy of the heating, ventilation, and air-conditioning facilities while keeping the comfort of the room occupants. To this end, they propose the use of a thermal camera-based method to estimate the critical factors to assess personal thermal comfort, namely: the individual clothing insulation rate and metabolic rate. Specifically, a convolutional network is implemented to recognize an occupant’s clothes type and activity type simultaneously achieving more than 95% of accuracy, which leads to the segmentation of the human body into skin-bare regions and the clothing-covered regions allowing, all in all, the estimation of the individual clothing insulation rate and metabolic rate. In the experimental phase, the authors introduce a novel thermal dataset, which allows the evaluations of the proposed methodologies proving the feasibility, effectiveness, and automation of the proposed approach.

4 Multi-modal analysis

The paper *Multimodal Temporal Machine Learning for Bipolar Disorder and Depression Recognition* proposes a multi-modal approach that combines audio, video, and textual modalities for the recognition of mental disorders. The approach is based on a bidirectional long short-term memory autoencoder to model each modality features, derivative computation to consider evolution between subsequent frames, temporal serialization, and the use of fisher vectors to unify the frame-based features and the labels at whole video level. In addition, the authors also consider vectorized information extracted from the speech-to-text transcriptions at paragraph level. The authors make an interesting and detailed analysis of the contribution of each modality to the recognition of bipolar disorder and depression. The multimodal adaptive nonlinear judge classifier neural network was able to weight modalities leading to very good prediction results in terms of accuracy, unweighted average recall, and prediction, as well as F1 scores, which clearly outperform the state of the art.

The paper *Harnessing Emotions for Depression Detection* also aims to detect depression using a multi-modal approach using video, audio, and text. To this end, the authors explore unimodal datasets of emotions, the labels of which are transformed into positive and negative emotions. Then, binary classifiers are developed for each of the modalities using specific pre-processing and network architectures, which are fine-tuned with a depression dataset. The final decision is taken on the basis of a one-of-three scheme: a depressed output for at least one modality results in the final depressed output. Experimental results show that the use of unimodal datasets for transfer learning can improve the performance on the multi-modal datasets by fine-tuning.

Finally, the paper entitled *Probabilistic Elderly Person's Mood Analysis Based on its Activities of Daily Living Using Smart Facilities* is aimed at the elderly population, with the goal of estimating mood based on their everyday activities. To this end, the proposed approach extracts information about daily activities from a set of sensors in the house, from data about the use of the smartphone, and data provided by a wrist band. The combination of these data allows for the definition of an activity set. A team of psychologists visited the voluntary users to annotate their mood through a set of questionnaires. The authors propose a multi-level probabilistic

Bayesian network to model the relationships between the user's mood and their everyday activities at different levels. The experiment, carried out during five months, shows that the proposed Bayesian model is able to estimate the person's mood as well as their everyday activities. Experiments also identify interesting relationships between mood and activities.

Acknowledgements We thank all contributors for their interest in the special issue, and all reviewers for their insightful comments and support.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.