Multimedia Tools and Applications (2023) 82:4781–4785
https://doi.org/10.1007/s11042-022-14278-w

GUEST EDITORIAL

Multimodal interaction and IoT applications

Yogesh Kumar Meena[1] · **K. V. Arya**[2]

Accepted: 23 November 2022 /
Published online: 7 December 2022

Abstract
Multimodal interaction enables multiple modes for users to interact with the system. Whereas the multimodal interaction with IoT applications depends on multimedia systems' input/output which may limit their scalability and expressiveness. Furthermore, multimodal interaction with IoT applications could also be limited by input/output integration, security, privacy, and portability issues. This special issue focuses on recent state-of-the-art advances in multimodal interaction and IoT applications that could address these challenges. We aim at bringing together the latest industrial and academic progress, research, and development efforts within the rapidly maturing multimodal interaction and IoT applications. This article summarizes the research contribution of the accepted papers (27.94% acceptance rate) along with possible future directions emanating from these papers.

1 Introduction

Multimodal interaction exploits the synergic use of different modalities to optimize the interactive tasks accomplished by the users. This allows a user to use several input modes such as speech, touch, and visual to interact with multimedia systems to output such as text, audio, and learning. Many Internet of Things (IoT) applications have become a fundamental part of modern society. Despite the significant progress on multimodal interaction systems and IoT applications in recent years, much work remains to be done before sophisticated multimodal interaction with IoT applications become a commonplace, indispensable part of computing.

The dependency of multimodal interaction with IoT applications on multimedia systems' and their modalities limits their scalability and expressiveness. For instance, while intelligent voice assistants (e.g., Amazon's Alexa, Google Home etc.) are great in processing a lot of natural language commands, they may face challenges in interpreting and executing

✉ Yogesh Kumar Meena
y.k.meena@essex.ac.uk

K. V. Arya
kvarya@iiitm.ac.in

[1] School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK

[2] Department of Computer Science and Engineering, ABV-Indian Institute of Information Technology & Management, Gwalior, India

complex intents e.g., if a person reading a book wants to dim all the ambient light except the overhead reading light, it will be quite a challenge for the reader to give such a command using simple voice commands. Further, voice assistants may not be able to recognise and execute such commands effectively. Voice input alone, however, is limited for issuing commands with spatially distributed IoT applications. Further, such verbose commands would be challenging for the assistant to recognize and execute. Therefore, it needs to address this by designing and investigating a more convenient way of human interaction with the physical environment.

The major challenges of such multimodal interaction with IoT applications are input/output integration, security, privacy, affordability, portability, and practicality issues. For example, the integration of touch and eye gaze, speech and eye gaze, facial expression and haptics input, touch-based gesture and prosody-based affect – may not have obvious points of similarity and straightforward ways to connect. One such solution is to design an effective multimodal interaction with IoT applications that can use a synergetic combination of input modalities based on user intent, needs, comfort, and adaptability. This special issue is requesting articles that address these challenges.

2 Contributions

This special issue includes 19 papers that describe work inspired by interaction and IoT processes. In particular, we received 68 manuscript submissions and accept 19 for publication. We maintained a 27.94% acceptance rate considering that each paper should be reviewed by at least three reviewers at each stage of the review process. We briefly highlight each paper's main contributions and its field of application. We pooled the accepted manuscripts and analysed them to identify themes and categories corresponding to the above major challenges. From 19 accepted articles, 11 papers address input/output integration and recognition challenges, 06 papers address security and privacy issues, and 2 papers address energy consumption of IoT devices challenges, as follows.

2.1 Input/output integration and recognition

A novel multi-modal approach is proposed by combining task-based and mobile crowd sensing (MCS) based methods to classify the depressed and non-depressed subjects (Thati et al., 2022). This approach could provide additional advantages over the existing depression detection strategies that use visual, audio, and smartphone usage data. This could open up multiple possibilities such as exploring semantic cues in verbal, head pose and eye gaze features in visual, along with skin conductance and heartbeat in physiological modalities along with a focus on more advanced smartphone usage variables.

Multimodal and multi-user-based interaction approaches are presented for digital TV applications (Barreto et al., 2022). In particular, the Nested Context Language (NCL) is extended to support new interaction events and multiple users, allowing them to identify which user has interacted with a DTV application. This work includes multiple new event types such as Touch (touch interaction), Motion (body movement), EyeGaze (fixation), Pointer (pointing), VoiceRecognition (voice recognition), GestureRecognition (gesture recognition), and FaceRecognition (facial expression recognition). A demonstration of Ginga-NCL middleware is presented for the VoiceRecognition and EyeGaze events. A synergetic combination of event types could provide multiple opportunities to users. It

could also provide multiple opportunities to researchers e.g., usability and user experience studies with such approaches.

An interactive integration of multi-sensory and volumetric content is proposed for music education in Taiwan into applications for children (Ho et al., 2022). This study introduces a technological multi-sensory pop-up sketchbook created in collaboration with the National Taiwan Symphony Orchestra (NTSO) and Industrial Technology Research Institute (ITRI). Both organizations collaborate to integrate emerging media technologies, including augmented reality (AR) and volumetric capture for content production, and creative music teaching methods, derived from traditional pop-up sketchbooks. In future, the AR multi-sensory pop-up sketchbook with interactive music learning materials could be a promising product that bridges the physical and the digital.

A portable, light-weighted, convolutional neural network (CNN) based network is presented for precise and efficient hand gesture recognition (Bhaumik et al., 2022). In particular, these modules could capture the refined edge information of hand gestures by incorporating hybrid feature attention blocks which focus on efficient salient features from multi-receptive fields and acquire knowledge of discriminable semantic structure for hand poses.

In previous studies, emotion-based interactions have been studied for multiple purposes such as human-robot interaction, assessing entrepreneurial education, ubiquitous decision making etc. However, the accuracy of the detection of emotions using electroencephalogram (EEG) signals is always challenging and it needs further refinements at both the software and hardware levels. A hybrid method for human emotion recognition using EEG signals is proposed by combining convolution neural network (CNN) and long short-term memory (LSTM) based hybrid models (Iyer et al., 2022). This work would contribute towards designing an effective ensemble learning-based EEG emotion recognition system.

An automated attention deficit classification method is proposed to analyse and predict cognitive attention or its deficit with less computational power and adaptable in real-time (Salankar et al., 2022). EEG signals have been split into six windows of varying time duration. Robust and computationally less expensive features hurst and power have been used for the designing of feature space. The objective of this proposed work is to provide a robust methodology for the classification of the attentive and non-attentive categories of subjects for real-time screening. It could provide a significant marker to identify the beginning of the mental fatigue stage by changing the frequency bands from delta to (delta + theta) at the prefrontal region and (delta + alpha) at the occipital region.

Perceptual computing help people in making subjective decisions. However, the perception of the word suffers from imprecision and uncertainties. Typically two types of uncertainties are present in the word: 1) intra-level uncertainty (perception associated with a person), 2) inter-level uncertainty (perception associated with a group of people). A novel pipeline is presented to construct the GIT2FSUM (Gaussian IT2FSs with Uncertain Mean) following statistical (i.e. pruning) and heuristical (i.e. selection of underlying T1FS) steps on data intervals (Mishra et al., 2022). GIT2FSUM provides satisfactory performance at capturing uncertainties present in the perception of words. The GIT2FSUM method can be used in applications where human-like decision-making is needed, such as selecting an appropriate candidate from the interview, selection of a restaurant depending on the person's preferences, recommendation-based dialogue system, IoT, etc.

A Long-Short-Term Memory (LSTM) based technique is proposed for unobtrusive handedness prediction in one-handed smartphone interaction (Chen et al., 2022). The LSTM-based neural network is trained upon the motion-sensor data from 13 users' single-handed smartphone picking, holding and operating in the setting of sitting, standing, and

walking. Based on it is offline experimental results (average accuracy of 92.6%), this technique could use for automatic user interface adjustment to accommodate different sides of single-handed smartphone usage in real-time applications.

A convolutional neural network (CNN) based lightweight lane detection model is designed for autonomous driving vehicles (Singal et al., 2022). This model can detect lanes in a day, night, and rainy conditions with high accuracy and low execution time. The size of the model has been kept short to make it hardware deployable and perform in real-time. This model could directly transfer to pedestrian detection on the road while driving the car and add both models that can be used for an autonomous self-driving car. Furthermore, it could also be used alone or with other sensor-based solutions for pothole detection.

Automated medical imagining is growing rapidly for advanced clinical treatment and intervention in medical diagnosis. The segmentation of nuclei in digital histopathology is considered the most crucial aspect in diagnosis and evaluating the severity of the disease. To encounter such an issue, an automated nuclei segmentation method is introduced for histopathological images (Vijh et al., 2022). The proposed segmentation method uses a new hybrid algorithm of lion optimization and cat swarm optimization to provide an optimal threshold value for efficient multi-level image thresholding segmentation. However, there is a research gap where the proposed algorithm parameters could be further explored and the computational complexity could be improved for higher dimensional data.

In research, the sensitivity of single beam ultrasound (SBUS) is shown as a convenient, non-invasive, and radiation-free method that is relatively simple compared to other imaging modalities. Therefore, SBUS is investigated towards microstructural tissue changes during Local hyperthermia (LHT) treatment (Manaf et al., 2022).

2.2 Security and privacy

Nowadays, protecting digital data during transmission is a very big challenge. For example one of the most important medical data is the electrocardiogram (ECG) signal which detects cardiovascular diseases and any alteration in the signal may affect the diagnosis ECG signal. To address this issue, an ECG watermarking based on redundant discrete wavelet transform and singular value decomposition is developed to preserve data integrity and privacy along with source authentication (Sharma et al., 2022). The audio steganography problem is addressed by presenting a new deep neural networks-based technique to hide images within the audio using deep generative models (Paul & Mishra, 2022). This steganographic network could transfer to video steganography problems.

A transfer learning-based pipeline is proposed to recognize the face along with gender classification and facial expression recognition in Near Infrared (NIR) spectrum (Salim et al., 2022). Such approaches could directly help in under variable and low illumination conditions to protect against security breaches in low or minimum light conditions. A watermarking scheme is presented and shows its efficacy under various signal-processing operations to protect the image database used in the presentation layer of IoT applications (Verma et al., 2022). This would provide additional security to interfaces that connect IoT and the real world.

A hybrid image encryption method based on watermarking and cryptographic techniques is presented to provide a higher level of security and encrypts small amounts of confidential data by using a small key size and encryption rounds (Gupta et al., 2022). This approach is based on two-level security for the secure and error-free transmission of images between IoT-enabled devices. A discrete wavelet transform (DWT) based watermarking scheme is

used at the first level, while an efficient image encryption technique based on a hybrid of logistic chaotic map and crossover is applied at another level.

An automatic and multimodal method for real-time nuisance detection is proposed for security threats inside ATM cabins (Srivastava et al., 2022). In a CCTV stream, region-wise local motions have been captured through a motion projection matrix and local motion histograms to extract features. The nuisance is detected in real-time with the help of a motion projection matrix and localized motion information in form of local motion histograms. An ensemble learning technique, tree-bootstrap-aggregator (or tree-bagger), is used for classification. However, state-of-the-art machine learning methods could implement for effective feature classification in real-time scenarios.

2.3 Energy consumption of IoT devices

In most Internet of Things (IoT) applications, network nodes are limited in terms of energy sources. Therefore, the need for innovative methods to eliminate energy loss which shortens the life of networks is fully felt in such networks. An efficient energy approach is proposed for routing on the IoT in which the focus is on the sleep-wake schedule of nodes using chaos fuzzy grasshopper optimization algorithm (Mir et al., 2022). Linear optimization and fuzzy-based clustering technique are presented for WSNs-assisted IoTs (Maratha & Gupta, 2022). Such approaches could enhance the lifetime of IoT systems in data transmission.

3 Conclusion

Overall, the special issue attracted research from several exciting and diverse domains ranging from social sciences to medical applications, from proof of concepts to applications. The articles presented in this special issue represent some of the current challenges and implications of multimodal interaction and IoT applications. These articles show state-of-the-art results for various applications and settings: recognition (gesture, face, motion etc.), detection (lane, object, depression, attention, nuisance etc.), image processing, deep learning methods, decision-making applications, interactions, interfaces and IoT applications. All these articles somehow addressed the major challenges of multimodal interaction and IoT applications such as input/output integration, recognition, security, privacy, portability, practicality, and energy consumption of such IoT devices. With our special issue, we hope to provide an informative foundation for further research in this field.

Acknowledgements We wish to thank all authors who contributed and the set of capable reviewers who helped us select and shape the manuscripts that compose them. Yogesh Kumar Meena, K. V. Arya, guest editors.