Human-Computer Interaction approach with Empathic Conversational Agent and Computer Vision

Rafael Pereira[1][0000−0001−8313−7253], Carla Mendes[1][0000−0001−7138−7124], José Ribeiro[1][0000−0003−3019−1330], Nuno Rodrigues[1][0000−0001−9536−1017], and António Pereira[1,2][0000−0001−5062−1241]

[1] Computer Science and Communications Research Centre, School of Technology and Management, Polytechnic of Leiria, 2411-901 Leiria, Portugal
{rafael.m.pereira, carla.c.mendes, jose.ribeiro, nunorod, apereira}@ipleiria.pt
[2] INOV INESC Inovação, Institute of New Technologies, Leiria Office, 2411-901 Leiria, Portugal

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

Empathy, which entails comprehending and sharing others' emotions to forge emotional connections, is vital for human relationships. Similarly, in Human-Computer Interaction (HCI), empathy is crucial in ensuring more realistic, improved, convenient and meaningful interactions. However, the typical HCI, aimed at tailoring computer systems to meet the specific needs and preferences of individuals, still lacks the users' emotional state, therefore losing crucial information during these interactions [8]. Recent Artificial Intelligence (AI) techniques, such as Emotion Recognition (ER) and empathic conversational agents, when integrated with HCI allow for continuous understanding of the user's emotions throughout interactions and empathically providing responses, greatly contribute to an increase in the quality and deepness of interactions between humans and computers, improving the user's overall experience [19].

Artificial Intelligence (AI) encompasses various techniques and methodologies aimed at enabling machines to perform tasks that typically require human intelligence, whereas Deep Learning (DL) stands out as a specialized approach relying on Artificial Neural Networks (ANNs) to process unstructured data (including images, voice, videos, and text, among others). ER, being a recent application of AI combined with DL, involves detecting human emotions through various modalities ranging from facial features, gestures, and poses, to speech and text captured through continuous interactions with the user [2]. conversational agents consist of computer programs designed to simulate human-like conversation and engage in interactions with users through natural language using

various techniques, including natural language processing (NLP), ML, and DL, to understand user input, interpret context, and generate appropriate responses.

Due to the immense potential of ER and conversational agent, individually, and the numerous benefits provided when integrated with HCI, this study offers a comprehensive guide covering ER modalities and key design and functionality aspects of a conversational agent, furthermore reviewing widely adopted datasets and methodologies. Lastly, proposing an innovative HCI approach to ensure more realistic and meaningful interactions by leveraging HCI in conjunction with ER techniques and an empathic conversational agent.

The primary findings of this study can be summarized as follows:

- Detailed guide on how DL impacts HCI nowadays;
- Performed a literature review regarding ER and conversational agents;
- Explored the main methods and datasets used for ER and conversational agents;
- Proposal of a taxonomy encompassing HCI, DL, conversational agent, and ER;
- Proposal of an architecture of an HCI approach aided with ER, through Computer Vision (CV) and Sentiment Analysis (SA), and emphatic conversational agent.

This research is organized into six sections. Section 2 presents the main concepts behind DL, ER, and conversational agent. Section **??** details the most used datasets to train, validate, and evaluate ER and conversational agent algorithms. Section **??** introduces and discusses the core AI algorithms used nowadays to build ER systems and conversational agents, while Section 3 presents the architecture, features, and characteristics of our proposed solution for HCI aided with ER and an empathic conversational agent. Lastly, Section 4 introduces the challenges and directions for future work and the conclusions.

2 Background

The evolution of technology in HCI has been significantly influenced by deep learning, a subset of machine learning. DL has enhanced performance in domains such as speech recognition and object detection, with a focus on supervised learning. Supervised Learning involves training a system with a labeled dataset to map inputs to outputs. The quality of training data is crucial for model efficacy, and techniques like data augmentation help in improving data diversity for better training outcomes [3, 10, 12, 14, 16, 20].

Convolutional Neural Networks (CNNs) are widely used for image-based tasks in HCI. These networks, with their structure of convolutional and pooling layers, process multi-array data like images efficiently, as depicted in Figure 1. This architecture is instrumental in emotion detection through computer vision, impacting HCI. CNNs not only enhance computational efficiency but also improve the robustness of the features extracted. Transfer learning plays a significant role in this context, facilitating knowledge transfer from large datasets

to specific tasks and reducing the need for extensive data and computational resources [11, 12].

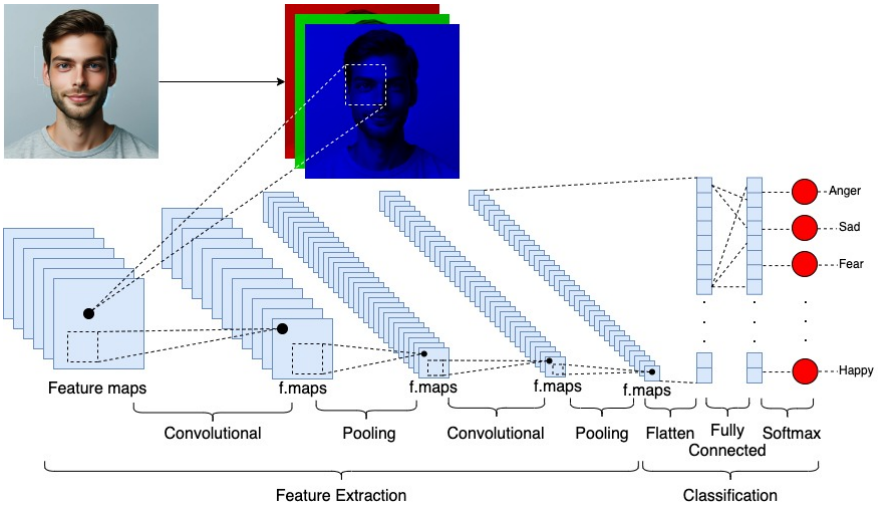

Fig. 1. An illustrative diagram of a CNN for emotion detection. The process begins with an input image of a smiling person and progresses through successive convolution and pooling layers for feature extraction. After flattening the feature maps, a fully connected network follows, leading to a final classification layer that categorizes the detected emotions into anger, sadness, fear, and happiness.

Emotion detection, related to analyzing human expressions and classifying them into emotions, is often studied individually and encounters challenges in real-life scenarios. This paper addresses these challenges through a multi-modal approach to emotion detection, aiming to mitigate problems associated with individual analysis methods. In computer vision, it includes facial movements and body language analysis from images and videos. Non-visual domains use vocal nuances and text analysis to detect emotions. Different modalities present unique challenges and require specialized models for effective recognition [4,9,21].

For Speech Emotion Recognition (SER), deep convolutional recurrent networks have proven effective, combining CNNs' feature extraction with Long Short-Term Memory (LSTM) networks' sequential data processing. Text-based emotion detection also employs LSTM-based models to interpret emotions from written language [9, 21].

In the context of HCI, conversational agents are really important by interacting with users through simulated voice and/or text messages. These agents can either be domain-specific or versatile in handling various interaction types. The effectiveness of conversational agents largely depends on their response delivery mechanisms, which include rule-based, retrieval-based, and generative-based ap-

proaches. Each approach has its unique implementation complexity and response generation capability [1, 6, 15, 17].

3 Proposed solution

Empathy is a fundamental trait in ensuring realistic, natural, and meaningful interactions in the field of HCI, therefore this section will delve into the proposed HCI approach architecture consisting of six modules as detailed in Figure 2. Early emotion recognition mainly focused on text or facial features alone, however, humans express emotions in a highly complex way, depending on both verbal and non-verbal cues like facial expression, behavior, voice, text, and physiological signals [5] [22]. Therefore, this article proposes an architecture, for mobile devices, to diminish this issue, composed of a conversational agent module, to interact continuously with the user, and a multimodal emotion recognition module, where multiple modalities (text, audio, and video) are analyzed simultaneously with DL algorithms to better recognize human emotions and gathered together to a final emotion using a fusion model which capitalizes on the strengths of each modality and compensates for potential limitations in individual modalities. Therefore, an implementation of our proposed architecture can lead to a more engaging, empathic, and natural conversational experience with a more robust and accurate emotion recognition while being suitable for applications where natural communication involves text, speech, or visual cues simultaneously.

Upon the capture of interaction with the user, the textual input obtained from a chat section (although if obtained as speech from the microphone it must be first converted to text using automatic speech recognition (ASR) algorithm aided with the speech spectrogram) is passed on to the NLP stage, where the input goes to some pre-processing stages such as segmentation, tokenization, lemmatization, and PoS tagging. Then, the processed data is passed through the Natural Language Understanding (NLU) stage, to extract the intents and entities while turning the pre-processed data into a structured representation. Afterward, the data is passed on to the Dialogue Management stage, which aims to maintain and incorporate the context of the current and past conversations [18]. Finally, by analyzing the contextual data, the structured representation of the user's input, and the emotion obtained from the fusion model, the Large Language Model (LLM) provides an accurate response which is converted into a user-readable format and presented back to the user.

As previously detailed, the system could gather the speech or textual data resulting from ongoing, real-time, interactions with the conversational agent, complementing this, visual data could be obtained from the camera embedded in the mobile device, either in real-time or in video, capturing facial and body pose information. Hence, the multimodal ER module analyzes in parallel all the present modalities, starting with textual data being analyzed by an SA model to obtain the sentiment conveyed in the textual input, following preprocessing, feature extraction, and classification steps. Nowadays, SA is shifting from the

early key-word comparisons approach to the adoption of ML (Support Vector Machines (SVM), Naive Bayes, Maximum Entropy), DL algorithms (CNN, ensemble of CNNs) and NNs (LSTM, DNNs), and even hybrid approaches [7].

On the other hand, the speech input can be transcribed and analyzed with ASR algorithms, including deep learning techniques such as Hidden Markov Models (HMMs), Support Vector Machines (SVM), various types of Artificial Neural Networks (ANNs) such as Radial basis functions (RBF), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs), employed to extract para-linguistic information, such as intonation, duration, prosody, pitch, and rhythm, from the speech signal [13].

On the other hand, the video input data, obtained from the CV algorithm that would capture, also continuously and in real time, the user's macro-facial expressions and body pose are passed on to a FER network and BER to further extract the emotional data conveyed in both input types.

The final stage to obtain a congregated emotion implies the usage of fusion models e.g. a late fusion model (also known as decision-level fusion) where firstly the emotion predictions are obtained from each network, and then the results based on these predictions are integrated into the final result through different decision-making methods e.g. average, majority, weighted, or other statistical strategies, this approach is usually lightweight, flexible, and versatile to change in modalities [22].

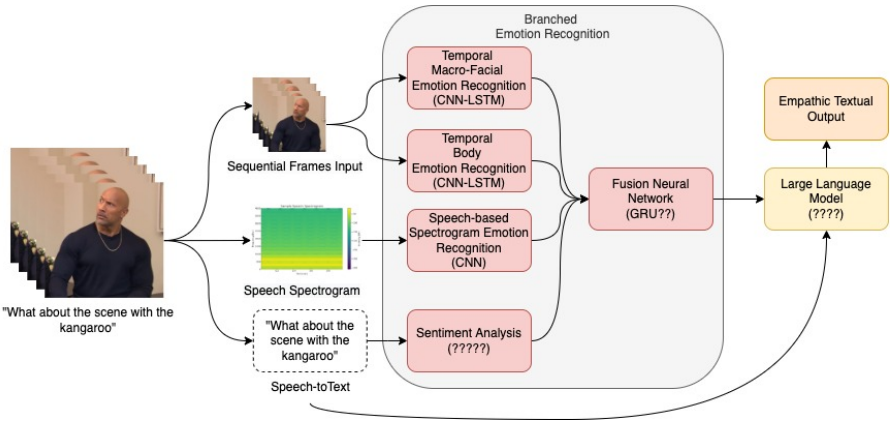

Fig. 2. Multimodal emotion recognition framework using a branching neural network and large language model for enhanced empathic response in human-computer interaction

4 Future work and conclusions

Acknowledgments. This work was supported by national funds through the Portuguese Foundation for Science and Technology (FCT), I.P., under the project UIDB/04524/2020

and was partially supported by Portuguese National funds through FITEC-Programa Interface with reference CIT "INOV-INESC Inovação-Financiamento Base"

Disclosure of Interests. The authors have no competing interests.

References

1. Aljaroodi, H.M., Adam, M.T.P., Chiong, R., Teubner, T.: Avatars and embodied agents in experimental information systems research: A systematic review and conceptual framework. Australasian Journal of Information Systems **23** (Oct 2019). https://doi.org/10.3127/ajis.v23i0.1841, https://journal.acs.org.au/index.php/ajis/article/view/1841, publisher: Australian Computer Society
2. Alrowais, F., Negm, N., Khalid, M., Almalki, N., Marzouk, R., Mohamed, A., Al Duhayyim, M., Alneil, A.A.: Modified Earthworm Optimization With Deep Learning Assisted Emotion Recognition for Human Computer Interface. IEEE Access **11**, 35089–35096 (2023). https://doi.org/10.1109/ACCESS.2023.3264260, https://ieeexplore.ieee.org/document/10091537/
3. Alrowais, F., Negm, N., Khalid, M., Almalki, N., Marzouk, R., Mohamed, A., Duhayyim, M.A., Alneil, A.A.: Modified earthworm optimization with deep learning assisted emotion recognition for human computer interface. IEEE Access **11**, 35089–35096 (2023). https://doi.org/10.1109/ACCESS.2023.3264260
4. Chul, B., Id, K.: A brief review of facial emotion recognition based on visual information. Sensors 2018, Vol. 18, Page 401 **18**, 401 (1 2018). https://doi.org/10.3390/S18020401, https://www.mdpi.com/1424-8220/18/2/401/htmhttps://www.mdpi.com/1424-8220/18/2/401
5. Ezzameli, K., Mahersia, H.: Emotion recognition from unimodal to multimodal analysis: A review. Information Fusion **99**, 101847 (Nov 2023). https://doi.org/10.1016/j.inffus.2023.101847, https://www.sciencedirect.com/science/article/pii/S156625352300163X
6. Fernandes, S., Gawas, R., Alvares, P., Femandes, M., Kale, D., Aswale, S.: Survey on Various Conversational Systems. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). pp. 1–8 (Feb 2020). https://doi.org/10.1109/ic-ETITE47903.2020.126
7. Hung, L.P., Alias, S.: Beyond Sentiment Analysis: A Review of Recent Trends in Text Based Sentiment Analysis and Emotion Detection. Journal of Advanced Computational Intelligence and Intelligent Informatics **27**(1), 84–95 (Jan 2023). https://doi.org/10.20965/jaciii.2023.p0084, https://www.fujipress.jp/jaciii/jc/jacii002700010084
8. Jaiswal, A., Krishnama Raju, A., Deb, S.: Facial Emotion Detection Using Deep Learning. In: 2020 International Conference for Emerging Technology (INCET). pp. 1–5 (Jun 2020). https://doi.org/10.1109/INCET49848.2020.9154121, https://ieeexplore.ieee.org/document/9154121
9. Karna, M., Juliet, D.S., Joy, R.C.: Deep learning based text emotion recognition for chatbot applications. Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020 pp. 988–993 (6 2020). https://doi.org/10.1109/ICOEI48184.2020.9142879
10. Khalifa, N.E., Loey, M., Mirjalili, S.: A comprehensive survey of recent trends in deep learning for digital images augmentation. Artificial Intelligence Review **55**, 2351–2377 (3 2022). https://doi.org/10.1007/S10462-021-10066-4/TABLES/5, https://link.springer.com/article/10.1007/s10462-021-10066-4

11. Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review 2020 53:8 **53**, 5455–5516 (4 2020). https://doi.org/10.1007/S10462-020-09825-6, https://link.springer.com/article/10.1007/s10462-020-09825-6
12. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 2015 521:7553 **521**, 436–444 (5 2015). https://doi.org/10.1038/nature14539, https://www.nature.com/articles/nature14539
13. Malik, M., Malik, M.K., Mehmood, K., Makhdoom, I.: Automatic speech recognition: a survey. Multimedia Tools and Applications **80**(6), 9411–9457 (Mar 2021). https://doi.org/10.1007/s11042-020-10073-7, https://doi.org/10.1007/s11042-020-10073-7
14. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 **2017-January**, 4674–4683 (11 2017). https://doi.org/10.1109/CVPR.2017.497
15. Mohamad Suhaili, S., Salim, N., Jambli, M.N.: Service chatbots: A systematic review. Expert Systems with Applications **184**, 115461 (Dec 2021). https://doi.org/10.1016/j.eswa.2021.115461, https://www.sciencedirect.com/science/article/pii/S0957417421008745
16. O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G.V., Krpalkova, L., Riordan, D., Walsh, J.: Deep learning vs. traditional computer vision. Advances in Intelligent Systems and Computing **943**, 128–144 (2020). https://doi.org/10.1007/978-3-030-17795-9_10/COVER, https://link.springer.com/chapter/10.1007/978-3-030-17795-9_10
17. Ramesh, K., Ravishankaran, S., Joshi, A., Chandrasekaran, K.: A Survey of Design Techniques for Conversational Agents. In: Kaushik, S., Gupta, D., Kharb, L., Chahal, D. (eds.) Information, Communication and Computing Technology. pp. 336–350. Communications in Computer and Information Science, Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-6544-6_31
18. Rizou, S., Paflioti, A., Theofilatos, A., Vakali, A., Sarigiannidis, G., Chatzisavvas, K.C.: Multilingual Name Entity Recognition and Intent Classification employing Deep Learning architectures. Simulation Modelling Practice and Theory **120**, 102620 (Nov 2022). https://doi.org/10.1016/j.simpat.2022.102620, https://www.sciencedirect.com/science/article/pii/S1569190X22000995
19. Santos, B.S., Júnior, M.C., Nunes, M.A.S.N.: Approaches for Generating Empathy: A Systematic Mapping. In: Latifi, S. (ed.) Information Technology - New Generations. pp. 715–722. Advances in Intelligent Systems and Computing, Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-54978-1_89
20. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**, 1–48 (12 2019). https://doi.org/10.1186/S40537-019-0197-0/FIGURES/33, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0
21. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings **2016-May**, 5200–5204 (5 2016). https://doi.org/10.1109/ICASSP.2016.7472669
22. Zhu, L., Zhu, Z., Zhang, C., Xu, Y., Kong, X.: Multimodal sentiment analysis based on fusion methods: A survey. Information Fusion **95**, 306–325 (2023).

https://doi.org/https://doi.org/10.1016/j.inffus.2023.02.028, https://www.sciencedirect.com/science/article/pii/S156625352300074X