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 [12]. 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 [23].

Artificial Intelligence (AI) encompasses various techniques and methodologies aimed at enabling machines to perform tasks that typically require human intelligence, whereas deep learning 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 deep learning, 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), machine learning, and deep learning, 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 deep learning 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, deep learning, 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 three sections. Section 2 presents the main concepts behind deep learning, ER, and conversational agent. 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. deep learning 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, 14, 16, 18, 20, 24].

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

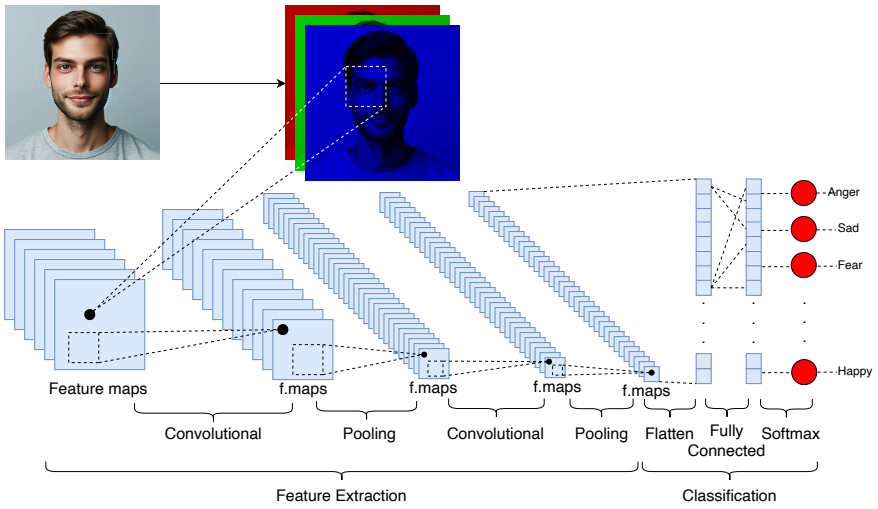

Fig. 1. An illustrative diagram of a CNN for emotion detection. The process begins with an input image of a smiling person and progresses through successive convolution and pooling layers for feature extraction. After flattening the feature maps, a fully connected network follows, leading to a final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

to specific tasks and reducing the need for extensive data and computational resources [15, 16].

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 [5, 13, 26].

For Speech Emotion Recognition (SER), deep convolutional recurrent networks have proven effective, gathering the spectrogram of the speech and feeding CNNs for classification [4]. Text-based emotion detection often employs Long Short-Term Memory (LSTM)-based models to interpret emotions from written language [13, 26] or even 1D CNNs and SVM [11].

Fusion methods integrate information from various modalities by creating a vector that combines features from these modalities. These methods are categorized into early fusion, late fusion, and cross-modality fusion based on when the fusion process takes place [25, 27]. Early fusion occurs before the application of primary learning models, consolidating all multimodal data (e.g., by concatenating incoming modality data) into a single representation. Early fusion is

preferable when there exists a strong association between each data source. Late fusion, or decision-level fusion, involves applying the fusion model after the learning models, merging the outputs of each learning model (e.g., learned features or class probabilities) to produce a final classification. Late fusion is more effective in scenarios where each modality needs to be trained with a specific algorithm and when modalities are subject to volatility. Cross-modality fusion facilitates the exchange of modality-specific data either before or during the primary learning stage. This approach enables each modality to leverage the context provided by others, enhancing the predictive capabilities of the overall model [25].

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 approaches. Each approach has its unique implementation complexity and response generation capability [1, 8, 19, 21].

As a type of generative-based conversational agent, Large Language Models (LLMs) are sophisticated algorithms designed to produce human-like language in response to prompts. They leverage vast amounts of text data and employ techniques like unsupervised learning to understand the statistical patterns of language. Through the state-of-the-art pipeline stages: of pretraining, supervised fine-tuning, reward modeling, and reinforcement learning, LLMs excel in capturing intricate linguistic patterns, nuances, and semantic relationships with the aid of transformer networks. LLMs utilize complex algorithms such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), or autoregressive models to create new textual content mirroring the characteristics of the training data. Among the popular LLMs are ChatGPT, LLaMA, Bard, Falcon, and others [9]. LLMs, although powerful, usually are computationally intensive due to the architecture and the size of the pre-trained parameters (often surpassing billions of parameters), bringing the need for a finetuning approach that reduces memory usage, number of parameters, and improves performance such as QLoRA which reduces the average memory requirements of finetuning LLaMA 65B parameter model from >780GB of GPU memory to <48GB without degrading the runtime or predictive performance [6]. Similarly, LoRa reduces the number of trainable parameters by 10,000 times and GPU memory requirement by 3 times (compared to GPT-3 175B) [10].

3 Proposed solution

Empathy is a fundamental trait in ensuring realistic, natural, and meaningful interactions, therefore in the context of HCI, the goal is to provide authentic and engaging interactions through the use of empathic conversational agents. The architecture proposed in this paper, as depicted in Figure 2, is an integrated system designed to enhance these interactions. Previous approaches in emotion recognition have often been unimodal, isolating text or facial features, despite

the inherently complex and multimodal nature of human emotional expression, which encompasses verbal and non-verbal cues such as voice, behavior, and physiological signals [7,27]. The proposed architecture addresses this complexity with a dual-component system: a conversational agent module that maintains an ongoing interaction with the user, and a multimodal emotion recognition module. This latter module employs deep learning algorithms to analyze multiple data streams (textual, auditory, and visual), in unison, improving the accuracy and robustness of emotion detection. The fusion model within this system synergizes the different modalities, enhancing the strengths and balancing the limitations of each. Therefore, an implementation of this architecture tries to achieve a more nuanced, empathic conversational experience that aligns closely with human communication, which inherently weaves together speech, text, and visual information.

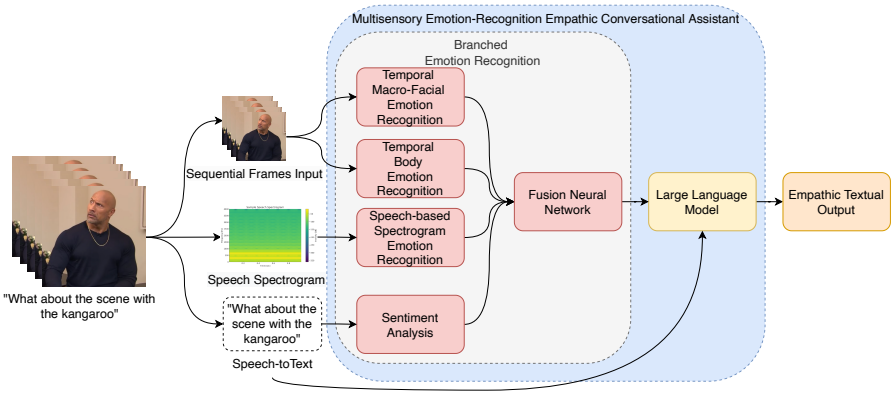

Fig. 2. Schematic representation of the proposed HCI approach architecture. Illustrates the flow from multimodal input through sequential frames, speech spectrograms, and text to the empathic conversational agent. The process integrates temporal macro-facial and body emotion recognition, speech-based spectrogram emotion recognition, and sentiment analysis, leading to an empathic textual output generated by the fusion neural network and large language model.

As previously detailed, the system could gather the visual and speech or textual data resulting from either in real-time or in video. 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 done with a 1D CNN classifier [11]. On the other hand, the speech input can be firstly transcribed with ASR algorithms, namely HMM-SVM Hybrid Model [17], and be passed on to the SA model and to the SER module composed of a CNN that analyzes the speech spectrogram (which represents the strength or loudness of

a signal over time at different frequencies in a particular waveform) to extract the emotion conveyed in the speech [4].

On the other hand, the visual data, obtained from computer vision, 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 the final emotion prediction involves employing a late fusion model (using a neural network followed by a softmax layer as the final classifier), where the outputs (an array of emotion predictions) of the previously described models are combined, leveraging the strengths of each modality while compensating for their weaknesses. Hence, in some instances prioritizing certain modality data over the others. Late fusion models are advantageous over single modality models, which often struggle to generalize across diverse scenarios and may be sensitive to modality-specific noise. In contrast, late fusion models increase robustness to noise and variability, resulting in more reliable emotion recognition across situations and improved performance [25, 27]. Therefore, the fusion model's output would consist of a single emotion resulting from the aggregation of the four context-based emotions obtained from each of the models. This approach benefits from the ability to train each modality with a specific algorithm and may make it easier to add or exchange different modalities in the future, but lacks the sharing of cross-modality data, which could hinder learning the relationships between modalities [25].

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 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 [22]. Finally, by analyzing the contextual data, the structured representation of the user's input, and the emotion obtained from the fusion model, the LLM provides an accurate response which is converted into a user-readable format and presented back to the user.

The proposed architecture aims to significantly enhance accuracy and efficiency in emotion detection compared to previous unimodal approaches, with the integration of multiple modalities: text, auditory, and visual cues, and increased robustness achieved with a late fusion model. The propagation of the final emotion prediction obtained from the late fusion modal to the LLM ensures that the latter considers the user's emotional state when generating the output response. Therefore, the proposed solution aims to enable more authentic, meaningful, and empathic interactions between a conversational agent and the user and our architecture can aid many areas such as education through personal-

ized learning experiences, and healthcare, by assisting in patient monitoring and mental health assessment among others.

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. Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W.: Speech emotion recognition from spectrograms with deep convolutional neural network. 2017 International Conference on Platform Technology and Service, PlatCon 2017 - Proceedings (3 2017). https://doi.org/10.1109/PLATCON.2017.7883728
5. 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
6. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient Finetuning of Quantized LLMs (May 2023). https://doi.org/10.48550/arXiv.2305.14314, http://arxiv.org/abs/2305.14314, arXiv:2305.14314 [cs]
7. 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
8. 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

9. Hadi, M.U., tashi, q.a., Qureshi, R., Shah, A., muneer, a., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., Mirjalili, S.: A survey on large language models: Applications, challenges, limitations, and practical usage (Jul 2023). https://doi.org/10.36227/techrxiv.23589741.v1, http://dx.doi.org/10.36227/techrxiv.23589741.v1
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models (Oct 2021), http://arxiv.org/abs/2106.09685, arXiv:2106.09685 [cs]
11. 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
12. 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
13. 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
14. 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
15. 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
16. 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
17. 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
18. 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
19. 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
20. 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

21. 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
22. 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
23. 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
24. 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
25. Sleeman, W.C., Kapoor, R., Ghosh, P.: Multimodal Classification: Current Landscape, Taxonomy and Future Directions. ACM Computing Surveys **55**(7), 150:1–150:31 (Dec 2022). https://doi.org/10.1145/3543848, https://dl.acm.org/doi/10.1145/3543848
26. 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
27. 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