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 [7]. Recent Artificial Intelligence (AI) techniques, such as Emotion Recognition (ER) and empathic Conversational Agents (CAs), 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 [18].

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]. CAs 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 CA, 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 CA, 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 CA.

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 CAs;
- Explored the main methods and datasets used for ER and CAs;
- Proposal of a taxonomy encompassing HCI, DL, CA, and ER;
- Proposal of an architecture of an HCI approach aided with ER, through Computer Vision (CV) and Sentiment Analysis (SA), and emphatic CA.

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

2 Background

The evolution of technology has led to an increased demand for advanced HCI. These include voice recognition, face recognition, and gesture recognition, which are essential for facilitating more natural and intuitive interactions between humans and computers [3]. Furthermore, the ability of machines to perceive and interact with humans, whether in physical or virtual environments, needs an understanding of human motion [13].

DL, a subset of Machine Learning (ML), has been crucial in advancing the capabilities of HCI. This has significantly enhanced performance in domains like speech recognition, visual object recognition, object detection, and even fields like drug discovery and genomics. The focus of this paper will be on supervised learning, a predominant form of DL. Supervised Learning (SL) involves training a system with a labeled dataset, where the system learns to map inputs to outputs based on example input-output pairs. During training, the system iteratively adjusts its parameters to minimize the difference between its outputs and the desired outputs [11] [15].

The DL processes rely on gathering and processing data for image-based models. The quality of data is crucial for a model's learning efficacy. To mitigate

issues such as overfitting, which refers to the phenomenon where a network learns a function with very high variance to perfectly model the training data [19], data augmentation can be used. This technique generates new training samples by applying transformations like rotation and scaling. It's essential for diversifying the training dataset, particularly in cases of low amounts of data [9].

The shift from hand-engineered features to trainable multilayer networks, namely Convolutional Neural Networks (CNNs), is a concept realized through the implementation of backpropagation. This method allows the calculation of gradients in multilayer structures by working backward from the output. CNN architecture, shown in Figure 1, is structured with convolutional and pooling layers. These networks process multi-array data, like images, using local connections and shared weights. The convolutional layers consist of units organized in feature maps connected to local patches in previous layers, highlighting the principle of location invariance in data like images that fit when detecting emotions through computer vision [11].

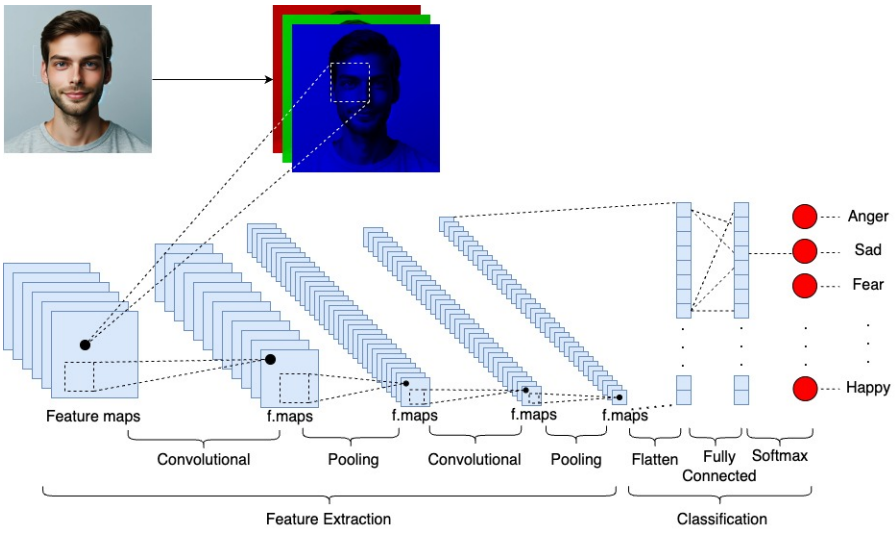

Fig. 1. An illustrative diagram of a CNN for emotion detection. The process begins with an input image of a smiling person and progresses through successive convolution and pooling layers for feature extraction. After flattening the feature maps, a fully connected network follows, leading to a final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

In CNN architecture, convolutional layers extract key features through their feature maps, while pooling layers further refine them by reducing their spatial size. This process improves the computational efficiency, and also increases the robustness of the features, a relevant aspect in the context of CV. Following this, the flattened layer transforms the two-dimensional feature maps into a

one-dimensional vector, preparing them for the classification stage. This stage primarily involves fully connected layers, which interpret the extracted features and make decisions. The last layer typically employs a softmax function, shown in equation 1, converting the output into probabilities for each class. In this context, TL emerges as a significant strategy, allowing the transfer of knowledge from one model to another, particularly beneficial when training data is limited. It involves using pre-trained models on large datasets and adapting them to specific tasks, significantly reducing the need for extensive training data and computational resources [10].

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } i = 1, 2, ..., K \tag{1}$$

Emotion detection encompasses the analysis of various human expressions gathered through different modalities. In CV, it includes the detection of macro-expressions, micro-expressions, and body expressions from images and videos. These expressions provide insights into emotional states through facial movements and body language. In the non-visual domain, emotion is detected from voice, tone, and speech spectrograms, where vocal nuances reflect emotional states. Text-based emotion detection, also known as SA, on the other hand, interprets written language to identify emotions, analyzing word choice and sentence structure. Each modality presents unique challenges and requires specialized models for effective emotion recognition [4] [20] [8].

For Speech Emotion Recognition (SER), deep convolutional recurrent networks have emerged as highly effective. These networks merge the feature extraction efficiency of CNNs with the sequential data processing ability of Long Short-Term Memory (LSTM) networks. The integration of CNNs and LSTMs in these networks shows a notable advancement over traditional methods reliant on hand-engineered features [20]. When detecting emotion from text, unlike other sources of expression where CNNs are used, LSTM-based models are the most used in this context. This method is effective in identifying emotional states from textual data, showcasing the ability of DL models to comprehend and process human emotions in written form [8].

Emotion recognition enables CAs to engage authentically by understanding and responding to users' emotions, fostering a more empathetic and natural interaction experience. A CA can be defined as a computer program that aims to interact with users realistically and naturally through simulated voice and/or textual messages, possessing either a virtual or physical body with appearances resembling humans, animals, objects, and abstract shapes, amidst others [1]. Nowadays, CAs are applied to numerous fields with purposes ranging from entertainment, task completion, knowledge seeking, and in some instances automating previous human tasks, where they can either initiate and hold interactions in a particular domain and switch to another anytime or be bound to a single domain [5] [16].

The capability to deliver accurate and meaningful responses to user requests is a critical factor that greatly influences the user experience with a CA.

These responses can be achieved through various approaches, namely rule-based, retrieval-based, or generative-based. In the rule-based approach, the responses are static, where the model merely uses human-made rules to match the input to a pre-defined response from a limited set, therefore being the simplest to implement where the responses often lack meaningfulness and accuracy [14].

Similarly, a retrieval-based model queries a pre-constructed conversation repository and uses NNs to choose the response that best matches the input. This approach, although better than the first by replacing human-made rules with AI techniques, still suffers from the issue of a static conversation repository where the responses are predefined [16]. A generative-based approach uses NLP algorithms to extract the interaction's intent, entities, and context from the user's inputs while employing advanced AI techniques for text generation therefore providing consecutive unique responses, with the only downside of implementation complexity due to requiring a considerable portion of training and testing data [16].

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 [**?**] [21]. Therefore, this article proposes an architecture, for mobile devices, to diminish this issue, composed of a CA 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

[17]. 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 CA, complementing this, visual data could be obtained from the camera embedded in the mobile device, either in real-time or in video, capturing only and facial and body pose information. Hence, the multimodal ER module could analyze 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. Nowadays, emotion detection regarding textual data is shifting from the early key-word comparisons to ML (Support Vector Machines (SVM), Naive Bayes, Maximum Entropy) and DL algorithms (CNN, ensemble of CNNs) and NNs (LSTM, DNNs), and even hybrid approaches [6], which any could be chosen as the SA model. 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 [12].

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 [21].

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.

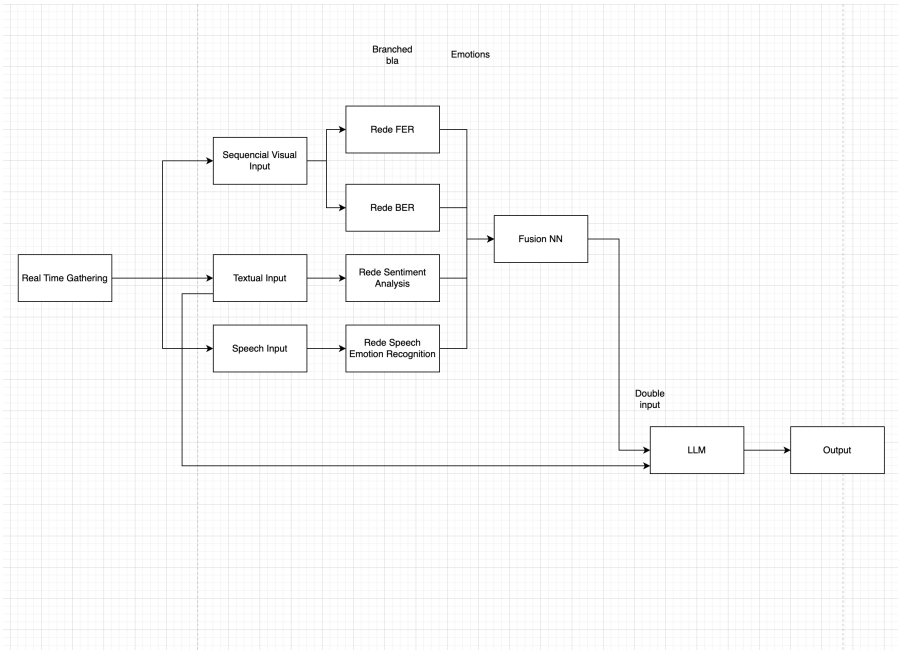

Fig. 2. HCI approach modules.

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/htm https://www.mdpi.com/1424-8220/18/2/401
5. 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

6. 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
7. 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
8. 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
9. 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
10. 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
11. 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
12. 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
13. 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
14. 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
15. 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
16. 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
17. 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

18. 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
19. 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
20. 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
21. 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