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 (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 [17].

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 (ANN) 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 [1]. 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 3 details the most used datasets to train, validate, and evaluate ER and CA algorithms. Section 4 introduces and discusses the core AI algorithms used nowadays to build ER systems and CAs, while Section 5 presents the architecture, features, and characteristics of our proposed solution for HCI aided with ER and an empathic CA. Lastly, Section 6 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. This is no longer confined to basic hardware-based communication but now encompasses more sophisticated techniques that are gradually becoming a part of everyday life. These include voice recognition, face recognition, and gesture recognition, which are essential for facilitating more natural and intuitive interactions between humans and computers [2]. Furthermore, the ability of machines to perceive and interact with humans, whether in physical or virtual environments, needs an understanding of human motion. This understanding must account for physical constraints, such as muscle torque and gravity, as well as the intentions behind movements, making motion modeling a highly complex [14].

Deep learning, a subset of machine learning, has been crucial in advancing the capabilities of HCI. By utilizing computational models with multiple processing layers, deep learning makes easier the learning of data representations at several levels of abstraction. 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 deep learning. Supervised learning

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. This process involves a high number of adjustable parameters, or weights, which define the system's input-output function [13] [15].

The implementation of deep learning in supervised learning models follows a structured process. This encompasses phases like data collection, where the importance of data quality cannot be overstated, data preprocessing to enhance data quality, training the model, optimization based on validation, and testing [18]. In recent years, Facial Emotion Recognition (FER) systems have exemplified the efficacy of Artificial Neural Networks (ANNs) over traditional machine learning methods. ANNs have demonstrated superior performance in detecting and recognizing emotions in a subject-independent manner, analyzing training data from various individuals. This approach has opened new opportunities in fields like healthcare, security, business, education, and manufacturing [5] [4] [21].

In the context of computer vision, neural networks have been particularly successful in image classification tasks, including face identification and facial emotion recognition. These technologies are not only used in surveillance systems but also in medical diagnostics and user-interactive applications. Different neural network architectures have been employed to meet the specific requirements of these tasks, including the use of pre-trained networks for classification, feature extraction, and transfer learning. Transfer learning, in particular, involves adjusting and reusing layers of a neural network trained on one dataset to work with a new dataset, demonstrating the versatility and adaptability of neural networks in various applications [4] [21].

Deep learning processes rely on data collection and processing, notably 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 [10].

The evolution of deep learning has been marked by a shift from hand-engineered features to trainable multilayer networks, a concept realized through the implementation of backpropagation. This method allows the calculation of gradients in multilayer structures by working backward from the output. Convolutional Neural Networks (CNNs) architecture, shown in Figure 1, are an excellent example of this approach, they are 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 [13].

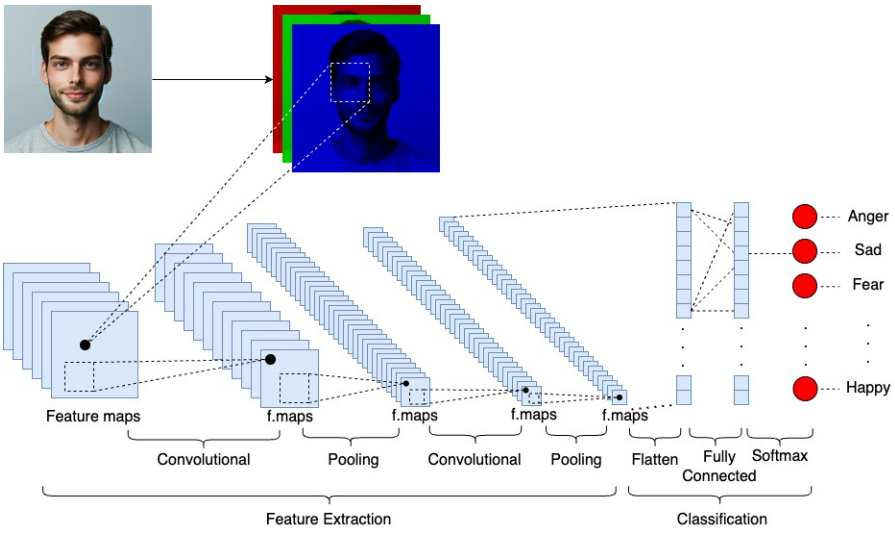

Fig. 1. An illustrative diagram of a CNN for emotion detection. The process begins with an input image of a smiling person and progresses through successive convolution and pooling layers for feature extraction. After flattening the feature maps, a fully connected network follows, leading to a final 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 computer vision. 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, transfer learning 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 [11].

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

Emotion detection encompasses the analysis of various human expressions gathered through different modalities. In computer vision (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, 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 [3] [20] [9].

The topic of emotion detection using computer vision had significant improvements in facial and pose detection using CNNs. Which had proven capabilities detecting emotions from facial expressions and body language [7] [12] [6] [16]. Despite over a decade of research, the field of emotion recognition using computer vision, specifically in the context of CNNs, remains a subject of active investigation, highlighting the ongoing evolution and potential for future breakthroughs in this area [3].

For speech emotion recognition, 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. This combination is particularly adept at capturing both temporal and spectral features from raw speech signals, enhancing the accuracy of emotion detection. 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. The process involves preprocessing text data, followed by the application of LSTM networks to recognize emotions. This method is effective in identifying emotional states from textual data, showcasing the ability of deep learning models to comprehend and process human emotions in written form [9].

3 Datasets

4 Methods

5 Proposed solution

6 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. 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/
2. 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
3. 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
4. Cîrneanu, A.L., Popescu, D., Iordache, D.: New trends in emotion recognition using image analysis by neural networks, a systematic review. Sensors 2023, Vol. 23, Page 7092 **23**, 7092 (8 2023). https://doi.org/10.3390/S23167092, https://www.mdpi.com/1424-8220/23/16/7092/htm https://www.mdpi.com/1424-8220/23/16/7092
5. Giannopoulos, P., Perikos, I., Hatzilygeroudis, I.: Deep learning approaches for facial emotion recognition: A case study on fer-2013. Smart Innovation, Systems and Technologies **85**, 1–16 (2018). https://doi.org/10.1007/978-3-319-66790-4_1/COVER, https://link.springer.com/chapter/10.1007/978-3-319-66790-4_1
6. Gorbova, J., Colovic, M., Marjanovic, M., Njegus, A., Anbarjafari, G.: Going deeper in hidden sadness recognition using spontaneous micro expressions database. Multimedia Tools and Applications **78**, 23161–23178 (8 2019). https://doi.org/10.1007/S11042-019-7658-5
7. Hayale, W., Negi, P.S., Mahoor, M.H.: Deep siamese neural networks for facial expression recognition in the wild. IEEE Transactions on Affective Computing **14**, 1148–1158 (4 2023). https://doi.org/10.1109/TAFFC.2021.3077248
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. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Context based emotion recognition using emotic dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 2755–2766 (11 2020). https://doi.org/10.1109/TPAMI.2019.2916866

13. 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
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. 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. Romeo, M., García, D.H., Han, T., Cangelosi, A., Jokinen, K.: Predicting apparent personality from body language: benchmarking deep learning architectures for adaptive social human–robot interaction. Advanced Robotics **35**, 1167–1179 (2021). https://doi.org/10.1080/01691864.2021.1974941
17. 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
18. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks **61**, 85–117 (1 2015). https://doi.org/10.1016/J.NEUNET.2014.09.003
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. Zhao, X., Shi, X., Zhang, S.: Facial expression recognition via deep learning. IETE Technical Review **32**, 347–355 (2015). https://doi.org/10.1080/02564602.2015.1017542, https://www.tandfonline.com/doi/abs/10.1080/02564602.2015.1017542