Enhancing Facial Emotion Recognition Through Deep Learning: Integrating CNN and RNN-LSTM Models

Jianhui Xu[a]
Khoury College, Northeastern University, 360 Huntington Ave, Boston, U.S.A.

Keywords: Facial Emotion Recognition (FER), Deep Learning, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNNs).

Abstract: This article investigates the application of deep learning techniques in Facial Emotion Recognition (FER) to advance psychological research and practical applications. Given its increasing relevance for improving human-computer interaction, mental health assessment, and accessibility for individuals with disabilities, FER is a field of growing importance. The proposed method combines Convolutional Neural Networks (CNN) for extracting spatial features from facial images with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) units for analyzing temporal evolution, particularly in video sequences. CNNs are employed to discern subtle variations in facial expressions, while RNN-LSTM models capture the progression of emotions over time. Experiments conducted on the FER-2013 and AffectNet datasets demonstrate that the CNN model outperforms other models, achieving accuracy levels that exceed those of human recognition on the FER-2013 dataset. This integration of CNN and RNN-LSTM models holds significant promise for enhancing the accuracy and efficiency of FER systems. Future research will focus on mitigating cultural biases, optimizing real-time application performance, and addressing privacy and ethical considerations in FER technology deployment.

1 INTRODUCTION

The rapid progress of artificial intelligence has opened up new possibilities for understanding human emotions through its intersection with psychology. This makes deep learning more valuable for research and development in facial emotion recognition (FER). FER involves using algorithms to analyze facial expressions and accurately identify displayed emotions. This technology not only enhances the ability to interpret emotions, but also creates new opportunities for practical applications in psychological research and clinical practice.

Emotion is one of the core factors driving human behavior, profoundly influencing key psychological mechanisms such as perception, attention, decision-making, and learning. Therefore, emotional regulation plays a crucial role in all aspects of human life. For many years, psychologists have firmly believed that a deep understanding of emotional states is crucial for a comprehensive understanding and interpretation of human behavior patterns, ways of thinking, and levels of intellectual development (Bota, et.al., 2019). Mastering the dynamic changes of emotions not only helps to explain individual psychological processes, but also facilitates research in broader social behavioral and cognitive contexts. In society, the use of machines to perform different tasks is constantly increasing. Providing machines with perceptual abilities can guide them in performing various tasks, while machine perception requires machines to understand their environment and the intentions of their interlocutors. Therefore, machine perception may help identify facial emotions. Inevitably, deep learning techniques will be widely applied in various fields, including displaying images of facial emotions. Although the results obtained are not state-of-the-art, the evidence collected suggests that deep learning may be suitable for classifying facial emotional expressions (Kumar, et.al., 2024). Therefore, deep learning has the potential to improve human-computer interaction, as its ability to learn features will enable machines to develop perception (Kumar, et.al., 2024). By possessing perceptual

[a] https://orcid.org/0009-0002-8875-5215

Xu and J.
Enhancing Facial Emotion Recognition Through Deep Learning: Integrating CNN and RNN-LSTM Models.
DOI: 10.5220/0013510700004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 131-136
ISBN: 978-989-758-754-2

abilities, machines can provide smoother responses, greatly improving the user experience.

Deep learning, a subset of machine learning, focuses on training artificial neural networks with large datasets to identify patterns and generate predictions.(Cowie, et.al., 2001). In the context of FER, deep learning models are trained on thousands of facial expression images to learn subtle features that distinguish different emotions. Once trained, these models can analyze new images and accurately classify expressed emotions (Lebovics, 1999). The strength of deep learning comes from its capacity to learn from vast datasets, making it especially effective for tasks that demand high precision and consistency.

FER has great potential in promoting the achievement of the Sustainable Development Goals and can contribute to them (Kumar, et.al., 2024; 6. LeCun, 2015). FER offers numerous practical applications across various domains. Firstly, FER can play a crucial role in mental health monitoring by detecting individual emotional changes and identifying potential mental health issues. By recognizing these changes early, timely intervention measures can be implemented, leading to improved mental health outcomes. Secondly, in public spaces, FER can be utilized to monitor people's emotions and behavioral patterns, which can enhance public safety and crowd management. Thirdly, FER provides significant benefits for individuals with hearing or language impairments by improving communication accessibility. Furthermore, FER has valuable applications in the education sector, where it can be used to assess and understand students' emotions. By analyzing emotional data, educators can tailor their teaching methods to better support students who may be facing emotional challenges (Mellouk, et.al., 2020). For example, when the system detects that certain students exhibit anxiety or frustration in the classroom, teachers can quickly take measures, such as changing the teaching pace or introducing more interactive learning activities, to alleviate students' emotional pressure, ensure that they maintain a positive emotional state during the learning process, and achieve better learning outcomes.

Although FER has hope, it is important to recognize the challenges and limitations associated with this technology. An important issue is the possibility of bias in deep learning models. If the training data cannot represent the diversity of human facial expressions, the model may produce biased results, especially for individuals from underrepresented groups. This may lead to inaccurate emotion recognition and exacerbate inequalities in psychological assessment and intervention. When developing and deploying FER systems, addressing these biases requires careful consideration to ensure that they are trained on diverse and inclusive datasets.

Another challenge is the ethical impact of using FER technology, particularly in terms of privacy and consent. The collection and analysis of facial data have raised concerns about how to store, share, and use this information. Establishing ethical guidelines and regulations is essential to safeguard individual rights and prevent the potential misuse of FER technology.

This paper aims to investigate the integration of deep learning techniques with FER to advance both psychological research and practical applications. It delves into the fundamental concepts and methodologies of FER, offering a comprehensive analysis of the primary models employed in this domain. The paper evaluates the performance of these models and explores potential future developments in FER technology. It also addresses the implications for enhancing human-computer interaction, improving mental health interventions, and considering the ethical aspects of deploying FER technologies.

2 METHODOLOGIES

2.1 Dataset Description and Preprocessing

Selecting a comprehensive and diverse dataset is crucial for model performance. The main datasets used in this study include the 2013 Facial Emotion Recognition (FER-2013) and AffectNet (Goodfellow, et.al., 2013). The dataset comprises 35,887 facial images, each meticulously labeled to reflect a specific emotional state (Giannopoulos, et.al., 2018). These emotional categories cover a broad spectrum of human expressions, capturing both positive and negative emotions, as well as more neutral states. It comprehensive labeling allows for detailed analysis and accurate recognition of a wide range of facial expressions, making the dataset a valuable resource for training deep learning models in emotion detection. This dataset is highly challenging due to significant differences in facial features such as age, posture, and occlusion. In addition, in the FER-2013 dataset, the accuracy of human facial emotion recognition is approximately 65%, with an error range of plus or minus 5% (Giannopoulos, et.al., 2018). In contrast, the AffectNet dataset has a larger scale and wider coverage. It used 1250 emotion related keywords in three major search engines to

search, including six languages, and finally collected more than 1 million facial images from the Internet (Mollahosseini, et.al., 2017). These datasets serve as a robust foundation for training and evaluating deep learning models in FER. Overall, the high utilization rate of neural network baselines indirectly proves its superiority over traditional machine learning methods and existing FER systems.

2.2 Proposed Approach

This study aims to enhance the ability of FER by combining deep learning techniques, thereby improving its efficiency and effectiveness in psychological research and practical applications. The goal of the research is to leverage the advantages of deep learning to optimize the performance of FER, in order to better support research progress in the field of psychology and provide more efficient solutions for practical application scenarios. The main objective of the study is to evaluate and compare several models that can accurately classify human emotions based on facial expressions. Specifically, the research focuses on two main models: CNN and RNN-LSTM. By comparing and analyzing these two types of models, the research aims to determine which model performs the best in recognizing facial expressions and analyzing emotions. This will not only provide more accurate technical support for psychological research, but also provide more effective solutions for practical application scenarios such as human-computer interaction and emotional computing. This study not only focuses on theoretical exploration, but also emphasizes the feasibility and effectiveness of the model in practical applications.

CNNs are employed to extract essential features from facial images, emphasizing spatial details that convey emotions. In contrast, RNN-LSTM models analyze the temporal dynamics of these features, which is crucial for interpreting emotions in video sequences where expressions evolve over time. By comparing these methodologies, the study seeks to determine the most effective approach for integrating deep learning into FER, assessing its impact on psychological assessment and human-computer interaction. Figure 1 illustrates the workflow for using deep learning in FER research. It begins with dataset description and preprocessing, followed by feature extraction and model selection. The models are then evaluated and compared to determine the most effective method. The study concludes with a discussion on future developments and ethical considerations, focusing on enhancing the accuracy of FER applications and safeguarding privacy.

2.2.1 CNN

As a variant of deep neural networks, significant breakthroughs have been made in many tasks related to computer vision. In other words, as a highly favored algorithm, it performs exceptionally well in object detection and medical image analysis (Girshick, et.al., 2014; 12. Dondeti, et.al., 2020). From this, it can be seen that CNN is an efficient feature extractor. Convolution and pooling greatly help it extract features from deeper images (Bodapati, et.al., 2022). Initially, researchers input images through convolutional layers and apply filters in the convolutional layers to extract relevant features from the images. In convolutional layers, multiple filters can be used to capture various features required for the task. These filters can effectively extract different types of information from images, thereby helping models better understand and process specific tasks. Following the convolution operation is a pooling layer, whose main function is to reduce data redundancy, minimize duplicate information while preserving important features. By processing through the pooling layer, not only does it reduce the computational cost of the model, but it also effectively increases the depth and complexity of the

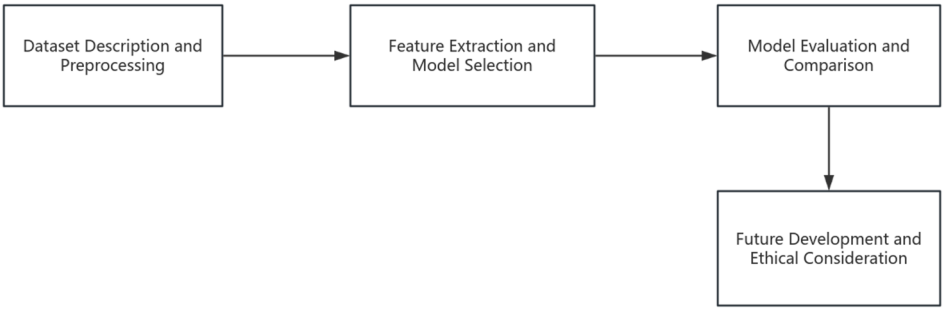

Figure 1: The pipeline of this study (Picture credit: Original).

network, enabling the model to handle more complex input data without adding too much computational burden. This process is crucial for improving the overall efficiency and performance of the model, and helps to analyze and understand data at a deeper level. Deep functionality is becoming increasingly popular and significantly improving the performance of models developed for FER (Georgescu, et.al., 2019). Given the widespread use of CNN in facial emotion recognition, some research teams have leveraged existing studies and conducted thorough analyses of different CNN architectures to propose a novel, relatively simple, and straightforward CNN design. (Bodapati, et.al., 2022). Their structure showed significant performance in FER-2013 testing, and through testing and experimentation, their CNN structure achieved an accuracy of approximately 69.57% in FER-2013, which is higher than the accuracy of humans on this dataset, about 65% (Bodapati, et.al., 2022). Through research and continuous improvement of the CNN structure, the team has concluded that the CNN model currently has significant efficiency and versatility in facial emotion recognition.

2.2.2 RNN-LSTM

Artificial neural networks can also be applied to FER, where RNNs can generate sequences with additional weights and maintain internal states. This enables it to make very calm and stable predictions when facing sequence related problems, including those involving sequence or time components (Antonio, et.al., 2018). Although RNN has shown encouraging performance on some tasks, it is not an easy task to train on long-term sequences, mainly because of the disappearance and explosion of gradient problems (Graves, et.al., 2014). However, these issues can be addressed by introducing memory mechanisms that enable the network to effectively remember and selectively forget early states (Mikolov, et.al., 2010). This memory mechanism makes the network more stable when processing long time series, not only retaining important information but also avoiding unnecessary interference, thereby improving the learning ability and prediction accuracy of the model. Through this approach, networks can better understand complex temporal dependencies, especially playing a crucial role in tasks that require long-term dependencies such as emotion recognition. LSTM networks equip RNNs with this capability, enabling them to retain information over extended periods (Salih, et.al., 2019). Consequently, employing LSTM-RNN for extracting various types of facial expressions in feature detection leads to improved performance and more accurate results. Compared to traditional neural networks, efficiency evaluations of LSTM-RNN on image and video frame sequences demonstrate a performance improvement of over 5% (Salih, et.al., 2019). Additionally, the network is capable of capturing both the presence and dynamics of partial and complete geometric shapes, while also effectively processing stationary data. It advanced technology has proven to be successful in this cutting-edge application (Salih, et.al., 2019).

3 RESULT AND DISCUSSION

3.1 CNN Model Architecture

According to Table 1 (Bodapati, et.al., 2022), this comparison graph demonstrates the excellent performance of the CNN model architecture, as shown in the figure. Bodapati's team compared the performance of the CNN model structure with the state-of-the-art FER-2013 facial recognition task model. In addition to evaluating the performance of the model, the model parameters were also compared to ensure that they could more easily compare the advantages of the CNN model. This comparison underscores the model's efficiency while maintaining competitive accuracy. The provided results and parameters indicate that the CNN model structure is superior to several models currently created for the FER-2013 dataset. Compared to shallow models such as SVM, deep neural architectures are much better. In addition, the model structure of CNN is superior to many existing models based on VGGNet and AlexNet architectures. The model's accuracy on the FER-2013 dataset exceeds human performance, achieving a rate of approximately 65%. From this, it can be seen that the CNN model structure is more efficient in FER compared to other model structures, and can also be improved based on the original model to achieve higher accuracy. This proves the enormous potential of CNN.

Table 1: The result of different models.

Model	Baseline	Accuracy	Parameters
Mishra	SVM	63.03	–
Gan et al.	VGGNet	64.24	–
Manual	–	65.00	–
Liu et al.	VGGNet	65.03	84M
Wan et al.	AlexNet+VGG	65.34	14M

Agarwal et al.	–	65.77	0.93M
Mollahosseini et al.	AlexNet	66.4	25M
Tang	AlexNet	69.30	7.17M
Bodapati	CNN	69.57	2.3M

3.2 Discussion

The focus is on the advantages and disadvantages of the methods used in FER. CNN can effectively extract spatial features from images, making it suitable for recognizing subtle differences in facial expressions. However, they lack the ability to capture temporal changes in emotions, which limits their performance in analyzing video sequences. On the other hand, RNN-LSTM excels at processing data sequences, allowing it to capture the evolution of emotions over time, although it is more complex to train and requires more computational resources.

Since CNN already has great potential, future research can explore combining CNN and RNN-LSTM, leveraging the advantages of these two models to create a more accurate system for FER, especially in videos. In addition, by using different datasets to address cultural differences in emotional expression, it can help reduce biases in these models. Exploring the real-time application of FER in low-power devices through edge computing is another promising direction. Ethical issues, such as privacy, should also be a focus, with research aimed at developing FER systems for security and privacy protection. Finally, further research can examine the application of FER in monitoring mental health, assisting in early detection and prevention of psychological problems.

4 CONCLUSIONS

This study investigates the integration of deep learning techniques into FER, with the goal of enhancing both psychological research and practical applications. The proposed approach employs CNNs to extract spatial features from facial images and RNNs with LSTM units to analyze the temporal evolution of these features, particularly in video sequences. Through extensive experimentation, this study demonstrates that the CNN significantly outperforms other models on the FER-2013 dataset, achieving accuracy levels that surpass human recognition. This finding underscores the effectiveness of deep learning in capturing and interpreting complex emotional expressions. Looking ahead, future research will aim to refine the FER system by further integrating CNN and RNN-LSTM models to improve accuracy and robustness. Key areas of focus will include addressing cultural biases to ensure the system's applicability across diverse populations, developing real-time FER applications for deployment on low-power devices, and addressing privacy and ethical considerations. These efforts will help in creating more accurate, inclusive, and practical FER solutions that can be effectively used in a variety of settings.

REFERENCES

Antonio, V. A. A., Ono, N., Saito, A., Sato, T., Altaf-Ul-Amin, M., & Kanaya, S. 2018. Classification of lung adenocarcinoma transcriptome subtypes from pathological images using deep convolutional networks. International journal of computer assisted radiology and surgery, 13, 1905-1913.

Bodapati, J. D., Srilakshmi, U., & Veeranjaneyulu, N. 2022. FERNet: a deep CNN architecture for facial expression recognition in the wild. Journal of The institution of engineers (India): series B, 103(2), 439-448.

Bota, P. J., Wang, C., Fred, A. L., & Da Silva, H. P. 2019. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. IEEE access, 7, 140990-141020. (2)

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. 2001. Emotion recognition in human-computer interaction. IEEE Signal processing magazine, 18(1), 32-80. (7)

Dondeti, V., Bodapati, J. D., Shareef, S. N., & Veeranjaneyulu, N. 2020. Deep Convolution Features in Non-linear Embedding Space for Fundus Image Classification. Rev. d'Intelligence Artif., 34(3), 307-313.

Georgescu, M. I., Ionescu, R. T., & Popescu, M. 2019. Local learning with deep and handcrafted features for facial expression recognition. IEEE Access, 7, 64827-64836.

Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I. 2018. Deep learning approaches for facial emotion recognition: A case study on FER-2013. Advances in hybridization of intelligent methods: Models, systems and applications, 1-16.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. 2013. Challenges in representation learning: A report on three machine learning contests. In Neural information processing: 20th international conference, ICONIP

2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20 (pp. 117-124). Springer berlin heidelberg.
Graves, A., & Jaitly, N. 2014. Towards end-to-end speech recognition with recurrent neural networks. In International conference on machine learning (pp. 1764-1772). PMLR.
Kumar, A., Sindhwani, M., & Sachdeva, S. 2024. Facial Emotion Recognition (FER) with Deep Learning Algorithm for Sustainable Development. In Sustainable Engineering: Concepts and Practices (pp. 415-434). Cham: Springer International Publishing. (1)
Kumar, A., Sindhwani, M., & Sachdeva, S. 2024. Facial Emotion Recognition (FER) with Deep Learning Algorithm for Sustainable Development. In Sustainable Engineering: Concepts and Practices (pp. 415-434). Cham: Springer International Publishing. (9)
Lebovics, H. 1999. Mona Lisa's escort: André Malraux and the reinvention of French culture. Cornell University Press. (8)
LeCun, Y., Bengio, Y., & Hinton, G. 2015. Deep learning. nature, 521(7553), 436-444. (5)
Mellouk, W., & Handouzi, W. 2020. Facial emotion recognition using deep learning: review and insights. Procedia Computer Science, 175, 689-694. (6)
Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. 2010. Recurrent neural network based language model. In Interspeech (Vol. 2, No. 3, pp. 1045-1048).
Mollahosseini, A., Hasani, B., & Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31.
Salih, W. M., Nadher, I., & Tariq, A. 2019. Deep learning for face expressions detection: Enhanced recurrent neural network with long short term memory. In International Conference on Applied Computing to Support Industry: Innovation and Technology (pp. 237-247). Cham: Springer International Publishing.