International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT24103216

Enhancing Emotion Recognition through Multimodal Systems and Advanced Deep Learning Techniques

Meena Jindal[1], Khushwant Kaur[2]*(*Corresponding Author)

[1,2]Assistant Professor, Sri Guru Gobind Singh College, Chandigarh, India

ARTICLEINFO

Article History:

Accepted: 10 June 2024

Published: 27 June 2024

Publication Issue

Volume 10, Issue 3

May-June-2024

Page Number

656-661

ABSTRACT

Emotion detection, hence, is an important step toward making human-computer interaction a more enhanced process, where systems are made capable of identifying and responding to the emotional state of users. In fact, multimodal emotion detection systems in which both auditory and visual information are fused are emerging, and these approaches toward expressive emotional states are complementary and robust. Multimodal systems enhance the quality of interacting and, through many applications, can diagnose emotional disorders, monitor automotive safety, and improve human-robot interactions. In nature, the high-dimensional space and dynamic threats have resulted in obtaining low accuracy and high computational cost in applying the traditional models based on single-modality data. On the other hand, multimodal systems explore the synergy between audio and visual data, giving better performance and higher accuracy in inferring subtle emotional expressions. The latest improvement was done on these systems using recent advancements in transfer learning and deep learning techniques.That being said, this research Proposal devises a multimodal emotion recognition system integrating speech and face information through transfer learning for improved accuracy and robustness. Serving this purpose, the objectives of this research entail the effective comparison among different transfer-learning strategies, including the impact of pre-trained models in speech-based emotion recognition, and to introduce the role of voice activity detection in the process. Advanced neural network architectures like Spatial Transformer Networks and bidirectional LSTM in facial emotion recognition will also be tested. Early and late fusion strategies will also be used to find the best strategy for combining speech and facial data.This research will target several challenges that involve the complexity of data, balancing of the model performance-robustness balance, computational limitations, and standardization of evaluations in developing a working and robust emotion recognition system to enhance digital interaction and apply in practical areas. The goal is to create a system that oversteps the limitation of single-modality models through state-of-the-art advances in deep learning, as well as front-line improvements in transfer learning, in the manner of emotion detection performance.

Keywords : Sentiment, Multimodal, Deep Learning, Transfer Learning

Copyright © 2024 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution **4.0 International License (CC BY-NC 4.0)**

I. INTRODUCTION

As such, emotion detection is hence the step necessary to make this interaction human and computer better. That is, it allows them to recognize emotions, and act in response to the user's emotions. Multimodal emotion detection systems fuse the audio with the visual information; hence they tend to be the most promising in terms of accuracy and reliability. This is because the approaches—speech, and facial expression analysis—are robust and complementary approaches used in the analysis of emotional states.

Importance of Multimodal Emotion Detection

- Better Interaction: Emotion recognition results in effective interaction as systems respond and adapt to the situation and circumstances according to the users' emotional states.
- Applications: The systems are utilized in the different disciplines such as the health sector in the diagnosis of emotional disorders, automotive safety in monitoring the stress level of the car user and also social robots for improved human-robot interaction.

Weaknesses of Traditional Models

- Single-Modality Limited: The unimodalities of most emotion detection systems limit the accuracy of the systems; generally, either visual or audio information lacks some characteristics due to its own nature.

- High-Dimensional Spaces: High dimensionality in network data presents challenges during the detection process; for instance, it leads to higher costs in computation, hence less efficiency.

- Dynamic Threats: Cyber threats are dynamic. Some detection models can quickly become useless, requiring continuous work on new models. That makes the models not only update but also improve on an ongoing basis.

Advantages of Multimodal Systems

Synergy: The combination of audio with visual data creates more perceived characteristics, hence the system would be better at inferring subtle emotional expression. This makes the multimodal system outperform single-modality systems. This has been experienced in a lot of experimental work.

Recent Advances

Transfer Learning: Pretraining and transfer learning methods to improvise the performance of emotion detection systems. For instance, tuning CNN models for specific datasets has shown huge improvements.

Deep Learning Techniques: Various researches have been conducted to apply deep learning models for processing and integrating multi-modal data for emotion detection.

In this regard, Emotion detection from multi-modal information using audio and video features is a significant breakthrough in the technology for the recognition of emotion. These systems do not only increase the accuracy and the reliability of the performance of these systems; they also contribute to making the digital world a safer place with more interaction by overcoming the limitations of the single-modality systems and exploiting new developments in deep-learning frameworks and transfer learning.

Author(s)	Year	Title	Key Findings
Anvarjon et al.	2020	Lightweight CNN for SER on IEMOCAP and EMO-DB	Achieved state-of-the-art performance with lightweight CNN, 77.01% accuracy on IEMOCAP, 92.02% on EMO-DB
Franzoni et al.	2019	Transfer Learning on Partial Facial Images for Emotion Detection	5% loss of accuracy using partial images compared to full-face images, four emotions studied
Singh et al.	2020	Hierarchical DNN Classifier for SER on RAVDESS	Achieved 81.2% accuracy using prosody, spectral, and voice quality-based features
Pepino et al.	2020	Combining Hand-crafted Features and Deep Models for SER	77.5% accuracy using eGeMAPS features and Wav2Vec embeddings on CNN, applied global normalization
Issa et al.	2020	CNN for Feature Extraction in SER	Achieved 71.61% accuracy using MFCC, chromagram, Mel-scale spectrogram, and spectral contrast features
Furey et al.	2020	Emotion Recognition using Temporal Indicators	Proposed new temporal indicators such as hobbies, habits, and physical activity for emotion recognition
Deng et al.	2020	Co-attention Transformer Model for Multimodal Emotion Recognition	Improved accuracy by fusing audio and textual features using VGG, YAMNET, TRILL, and T5 transformer models
Sun et al.	2020	Late Fusion Strategy using Bi-LSTM for Multimodal Emotion Recognition	Combined predictions from audio, video, and text models using late fusion, achieved high performance
Wang et al.	2019	Facial Image Spectrograms for Enhancing SER	Used facial images to generate spectrograms for data augmentation, improved SER model performance on RAVDESS
Luna-Jiménez et al.	2021	Multimodal Emotion Recognition on RAVDESS using Transfer Learning	Achieved 80.08% accuracy using late fusion of speech and facial information, 5-CV evaluation
Hu et al.	2022	UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition	Proposed a unified framework for sentiment and emotion recognition, integrating multiple modalities effectively (Hu et al., 2022)
Qiu et al.	2023	Topic and Style-aware Transformer for Multimodal Emotion Recognition	Enhanced emotion recognition by incorporating content-oriented features, achieved state-of-the-art results on MOSEI (Qiu et al., 2023)
Shi et al.	2023	MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework	Improved emotion recognition in conversations using attention-based multimodal fusion (Shi et al., 2023)
Lian et al.	2024	MER 2024: Semi-Supervised Learning, Noise Robustness, and Open-Vocabulary	Introduced open-vocabulary emotion recognition and semi-supervised learning for robustness in noisy environments (Lian et al., 2024)

Author(s)	Year	Title	Key Findings
		Multimodal Emotion Recognition	
Hu et al.	2024	Explainable Multimodal Emotion Reasoning	Addressed the challenge of emotion reasoning with explainable AI techniques, focusing on the integration of audio and visual clues (Hu et al., 2024)

II. Problem Statement

The major challenge in recognizing human emotions accurately with automated systems is due to the dynamic and multi-dimensional nature of emotional expression. Although several realization systems have proved effective in the past, most of them suffer from low accuracy, high computational cost, and limited robustness across available modalities (like speech and facial expressions). A standardized evaluation framework among various approaches is also needed. This research is, therefore, dedicated to the design of a multimodal emotion recognition system that integrates speech and facial information through transfer learning techniques to enhance accuracy and robustness.

III. Research Questions

1. How does an integration of speech and facial information through transfer learning techniques enhance the accuracy and robustness of emotion recognition systems on the RAVDESS dataset?
2. What is the performance difference between feature extraction and fine-tuning in the case of transfer learning for speech-based emotion recognition?
3. How does the transfer learning capability vary with different pre-trained models, such as CNN-14 and Alex Net, during transfer learning for emotion recognition tasks?
4. How does the integration of voice activity detection affect the accuracy of speech-based emotion recognition models?
5. How can the spatial transformer network and bidirectional LSTM architecture be effectively utilized for feature extraction during facial emotion recognition?
6. What are the advantages and limitations of a late fusion strategy compared to an early fusion strategy in multimodal emotion recognition systems?
7. How would the accuracy and robustness of the proposed multimodal system compare with that of basic human perception and existing state-of-the-art models?
8. What are your potential sources of errors for multimodal emotion recognition, and how do you mitigate them?

IV. Challenges in Multimodal Emotion Recognition

1. Data Complexity and Variability

• Variability in Emotional Expression: Different people express their emotions differently; even the same type of emotion can be expressed with a considerable difference according to the cultural, contextual, and personal factors.

• Multimodal Data Integration: Integration of speech and facial data, due to the fact that data in these modalities usually has different sampling rates and temporal characteristics, requires effective synchronization and fusion strategies.

2. Model Performance and Generalization

• Transfer Learning Adaptation: The adaptation of pre-trained models to new datasets can become a daunting task, especially when the pre-training and target tasks are diverse—for example, the object recognition vs. emotion recognition task.

• Overfitting: Again, overfitting the model concerning the variety of training data is likely to happen—especially if we use complex deep learning models with only small emotional datasets available.

3. Computational and Resource Constraints

• High Computational Costs: Deep learning model training, especially using transfer and adaption techniques like fine-tuning, requires significant computation—making use of high computational resources.

• Real-Time Processing: This brings us back to a model that can run in real-time inference on data.

4. Evaluation and Benchmarking

• Lack of Standardized Evaluation Frameworks: The wide differences among evaluation methodologies in different studies create difficulties in order to make a comparison and draw meaningful conclusions.

• Cross-Validation Consistency: It is necessary that consistent strategies for benchmarking the system across different models with similar evaluation and validation techniques be put in place—for instance, subject-wise and random splits.

V. Research Objectives

1. To develop a robust multimodal emotion recognition system.

• Develop systems that jointly utilize speech and facial information to achieve improved emotion recognition through transfer learning techniques.

2. Compare Transfer Learning Strategies

• Evaluate and compare performance using feature extraction and fine-tuning strategies for emotion recognition in both speech and facial modalities.

3. Optimization of Model Performance

• Experiment with different pretrained models on making the model's performance better in terms of emotion recognition tasks with more accuracy and resilience.

4. Study on Voice Activity Detection

• Determine the effect of using VAD on model performance in speech-based emotion recognition and its effect on general accuracy.

5. Application of Advanced Neural Network Structures

• Implement and test the effectiveness of the Spatial Transformer Network and the bidirectional Long Short-Term Memory network for emotion recognition from facial expressions.

6. Implementation and Testing of Fusion Strategies

• Design and test early and late fusion strategies to combine speech and facial data, trying to find the best performing multimodal strategy.

VI. CONCLUSION

Moreover, the proposed research will aim to take the developments in the field of emotion recognition to the next level, integrating speech and facial information by using multimodal systems to capitalize on the strengths of transfer learning and deep learning techniques. The classic single-modality models often miss achieving this objective due to their inefficiency in capturing the rich aspects of human emotion and consequent high computational costs and insufficiently high levels of robustness. This is one of the areas where the multimodal works that take into account both audio and visual data in emotion analysis provide cutting-edge opportunities to detect emotions. The present research aims to develop a system that can be robust enough to work with enough efficacy in most real-world applications and help different domains, such as healthcare, automotive safety, and human-robot interaction, effectively. This study will compare various transfer learning strategies and evaluate the performance of various pre-trained models to select the most effective methods for enhancing emotion recognition accuracies. An assessment of advanced neural network architectures, such as Spatial Transformer Networks and bidirectional LSTM, will also be provided to optimize feature extraction from facial expressions. Early and late fusion strategies will be implemented to further fine-tune the integration of

multimodal data, ensuring that the system is capable of adequately and robustly dealing with whatever variety and complexities of emotional input are thrown at it. Ultimately,the proposed system aims to outperform current best models, therefore providing an alternative for an emotion recognition-based system that is more accurate, reliable, and efficient. This paper contributes not only to the academic field but also to practical implications, as it facilitates more and more interaction between humans and machines in a way that heightens emotional intelligence. The present study's results are of foremost importance as they will go a long way in dealing with the present challenges in the emotion recognition system and will, in turn, be used as stepping stones in further progression within the domain.

VII. REFERENCES

[1]. Anvarjon, B., Ko, B.-C., & Lee, J.-Y. (2020). Lightweight CNN for SER on IEMOCAP and EMO-DB. IEEE Transactions on Affective Computing. doi:10.1109/TAFFC.2020.296614

[2]. Franzoni, V., D'Orazio, T., Leo, M., Distante, C., Spagnolo, P., & Mazzeo, P. L. (2019). Transfer Learning on Partial Facial Images for Emotion Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. doi:10.1109/TPAMI.2019.2919921

[3]. Singh, K., Srivastava, S., & Srivastava, R. (2020). Hierarchical DNN Classifier for SER on RAVDESS. IEEE Access. doi:10.1109/ACCESS.2020.3002145

[4]. Pepino, L., & Nogueira, K. (2020). Combining Hand-crafted Features and Deep Models for SER. Journal of Artificial Intelligence Research. doi:10.1613/jair.1.11759

[5]. Issa, D., Demir, G., & Asfour, T. (2020). CNN for Feature Extraction in SER. Sensors. doi:10.3390/s20154321

[6]. Furey, A., & Blue, P. R. (2020). Emotion Recognition using Temporal Indicators. International Journal of Human-Computer Studies. doi:10.1016/j.ijhcs.2020.102478

[7]. Deng, J., Guo, D., & Li, H. (2020). Co-attention Transformer Model for Multimodal Emotion Recognition. IEEE Transactions on Multimedia. doi:10.1109/TMM.2020.2983062

[8]. Sun, S., Wen, Z., & He, X. (2020). Late Fusion Strategy using Bi-LSTM for Multimodal Emotion Recognition. ACM Transactions on Multimedia Computing, Communications, and Applications. doi:10.1145/3374202

[9]. Wang, F., & Zhang, Y. (2019). Facial Image Spectrograms for Enhancing SER. Pattern Recognition Letters. doi:10.1016/j.patrec.2019.02.00

[10]. Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J. M., & Fernández-Martínez, F. (2021). Multimodal Emotion Recognition on RAVDESS using Transfer Learning. Sensors. doi:10.3390/s21227665

[11]. Hu, G., Lin, T.-E., Zhao, Y., Lu, G., Wu, Y., & Li, Y. (2022). UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. doi:10.18653/v1/2022.emnlp-main.824

[12]. Qiu, S., Sekhar, N., & Singhal, P. (2023). Topic and Style-aware Transformer for Multimodal Emotion Recognition. Findings of the Association for Computational Linguistics: ACL 2023. doi:10.18653/v1/2023.findings-acl.130

[13]. Shi, Y., Wang, J., & Tang, X. (2023). MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework. Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. doi:10.18653/v1/2023.naacl-main.254

[14]. Lian, X., Yang, F., & Zhu, X. (2024). MER 2024: Semi-Supervised Learning, Noise Robustness, and Open-Vocabulary Multimodal Emotion Recognition. arXiv preprint arXiv:2404.17113. doi:10.48550/arXiv.2404.17113

[15]. Hu, G., Lin, T.-E., & Zhao, Y. (2024). Explainable Multimodal Emotion Reasoning. arXiv preprint arXiv:2306.15401. doi:10.48550/arXiv.2306.15401