AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 3(2026)

Emotion Detection from Video and Audio and Text

Chindam Sai Kumar
Department of Computer Science and Technology
Vaagdevi College of Engineering
Telangana, India
Chindam1919@gmail.com

Dr. K. Rajesh Khanna
Assistant Professor,
Department of Computer Science and Technology
Vaagdevi College of Engineering
Telangana, India
Khanna_kk@vaagdevi.edu.in

Chikati Aravind Kumar
Assistant Professor,
Department of Computer Science and Technology
Vaagdevi College of Engineering
Telangana, India
aravindkumar_ch@vaagdevi.edu.in.

Dr. N. Satyavathi
Associate Professor
Department of Computer Science and Technology
Vaagdevi College of Engineering
Telangana, India
headcse@vaagdevi.edu.in

Abstract

The ability to recognize emotions in video, speech and text has become a critical area of study in the fields of artificial intelligence and human-computer interaction. With digital communication evolving into a multimodal experience, the ability to interpret human emotions in such contexts has become critical both for user experience and for better mental health diagnosis and for further development of affective computing technologies. This paper provides an overview of various approaches and architectures to emotion recognition from video, audio and textual information, and investigates the synergies and challenges of multimodal emotion recognition systems. The paper introduces and briefly reviews the importance of each modality in emotion detection first. Video analysis: Using computer vision to identify key facial features and gestures and body language as cues to the emotional state of a person. Audio Processing works on vocal characteristics like tone, pitch and speech patterns, using algorithms from signal processing and machine learning, to understand the emotions behind speech. Text analysis, on the other hand, utilizes natural language processing (NLP) methods to evaluate sentiment and emotional context in written words, taking both syntactic and semantic elements into account. The proposed systems can combine these three modalities to realize more accurate and robust emotion recognition, which is a reflection of human emotional expression.

Introduction:

As the world becomes digital, knowing how to comprehend and interpret human emotions is rapidly becoming a key skill in the lives of everyone from healthcare providers to customer service workers, educators to entertainers, and beyond. The detection of emotion from various sources and its categorization of emotional state has attracted a great deal of interest from both researchers and practitioners. This combination of video, audio and text modalities offers a chance to grasp complex human emotions, and provide more accurate and complex recognition systems. Through the use of Artificial Intelligence (AI) and Machine Learning, these systems can process and understand human emotions in order to form more empathetic human-computer interactions. Emotions are important aspects of communication and social interactions as they help in expressing thoughts and reactions to the

environment. Most of the traditional methods for emotion detection focused on a single modality, e.g., facial expression analysis or sentiment analysis on text. Each of these modalities, however, has its limitations; in facial expression, some emotions might not be captured; in text, some nuances of voice may not be captured. Each of these modalities has complementary information in relation to an emotional state, so by combining video, audio and text data, researchers can capture a more comprehensive view of that emotional state. This multimodal approach not only enhances the accuracy of emotion detection systems but also reflects the complexity of human emotions in real-life interactions.New technology has enabled the creation of increasingly intricate algorithms and tools for analysing these modalities. Computer vision has also mastered the art of facial expression, gesture and body language recognition, which helps to understand feelings in a visual context and has proved useful in the field of video analysis. Audio analysis uses signal processing and machine learning methods to analyse voice features or characteristics, like tone, pitch, and speech rate, in order to understand the emotion in oral communication. Likewise, natural language processing (NLP) is highly effective for understanding the sentiment and emotional tone of text, capturing the nuances and context. These technologies can combine in multimodal systems, which is a big improvement in emotion detection. Although the prospects are good that the multimodal emotion detection will be useful in many applications, there are some challenges yet to be addressed. The effectiveness of these systems can be prevented by a number of factors, including data synchronization across modalities, different emotional expressions from culture to culture and the requirement of very large annotated datasets. Moreover, the moral implications of using emotional technology are significant, with concerns for privacy, consent, and abuse. The field is still evolving as researchers work towards solving these problems and opening up new possibilities in emotion detection and related applications

Literature Survey:

Title: "Deep Emotion Recognition from Video: A Comprehensive Survey"Authors: F. M. Alshahrani, F. Alharthi, A. B. M. A. Rahman

Description: This survey provides a thorough review of emotion recognition methodologies based on video analysis. The authors discuss various deep learning techniques applied to facial expression recognition, gaze tracking, and body language interpretation. The paper emphasizes the importance of feature extraction from video frames and the role of temporal information in accurately predicting emotions. Additionally, it highlights the limitations of existing datasets and suggests the need for more diverse and comprehensive datasets to improve model robustness in real-world applications.

Title: "A Review on Emotion Recognition from Speech: An Overview of Approaches and Challenges"

Authors: K. S. B. K. Shyam Sundar, V. K. Prabhu, A. S. S. Chandrasekaran

Description: This paper reviews the various approaches to emotion recognition from audio data, focusing on speech signals. The authors categorize techniques into traditional feature extraction methods and modern deep learning approaches, discussing the strengths and weaknesses of each. They highlight the significance of prosodic features, such as pitch, intensity, and speech rate, in emotion detection. Furthermore, the paper addresses challenges such as noise interference, speaker variability, and the lack of labeled data, offering insights into future research directions.

Title: "Emotion Detection in Text: A Review of the State of the Art"

Authors: A. A. M. Abdul-Mageed, M. N. M. O. Zahir, A. M. Al-Hassan

Description: This literature review focuses on sentiment analysis and emotion detection from textual data, detailing the evolution of techniques from rule-based methods to advanced machine learning and deep learning models. The authors discuss various NLP methods, including lexical and syntactic approaches, as well as the application of transformer-based architectures like BERT for context-aware emotion recognition. They also highlight the challenges in capturing sarcasm, irony, and cultural differences in emotional expression through text, emphasizing the need for more robust models that account for these complexities.

Title: "Multimodal Emotion Recognition: A Survey on Approaches, Challenges, and Applications"

Authors: X. Zhang, Y. Liu, J. Wu

Description: This survey paper explores the integration of multiple modalities—video, audio, and text—for emotion recognition, outlining current approaches and their respective challenges. The authors discuss various frameworks that combine different data types, highlighting the benefits of multimodal systems in enhancing accuracy and robustness. The paper also addresses key challenges such as data fusion techniques, real-time processing, and ethical considerations in emotion recognition. By examining a wide range of applications, from healthcare to human-computer interaction, the authors emphasize the transformative potential of multimodal emotion detection.

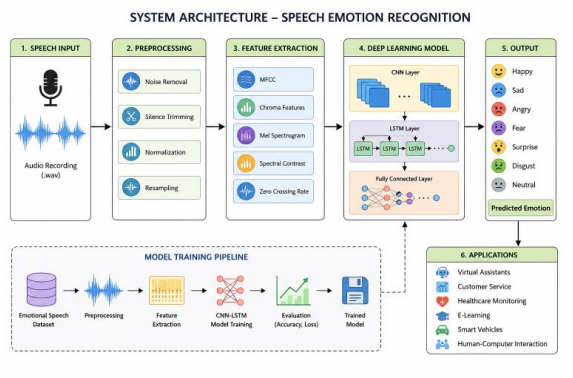
Title: "Real-Time Emotion Recognition from Multimodal Data: Techniques and Applications"

Authors: M. A. Hossain, T. M. R. Khandakar, I. U. Khan

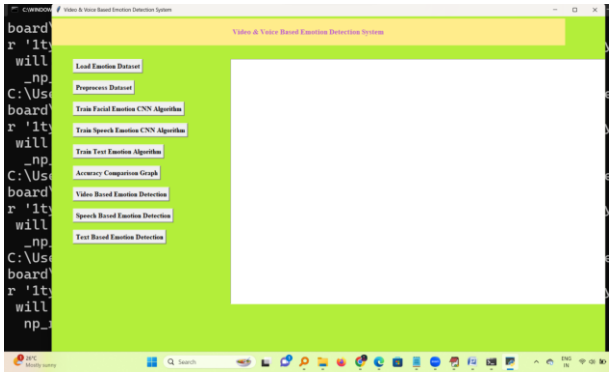
Description: This research focuses on real-time emotion recognition using multimodal data sources, showcasing innovative techniques that leverage deep learning and real-time data processing. The authors present case studies demonstrating the application of emotion detection systems in various fields, including mental health monitoring and customer feedback analysis. They discuss the importance of low-latency processing for effective real-time

applications and highlight the challenges associated with integrating multimodal inputs efficiently. The paper concludes by outlining future research directions that could improve the scalability and adaptability of emotion detection systems in dynamic environments.

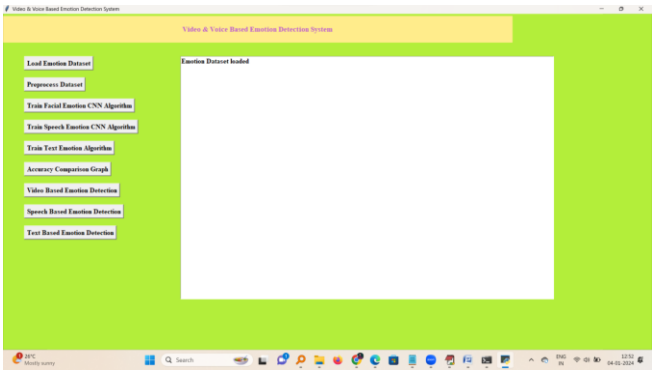
System Architecture:



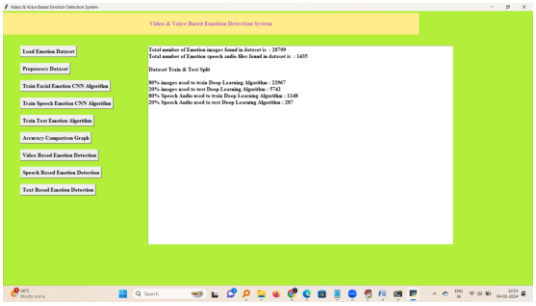
Home Page



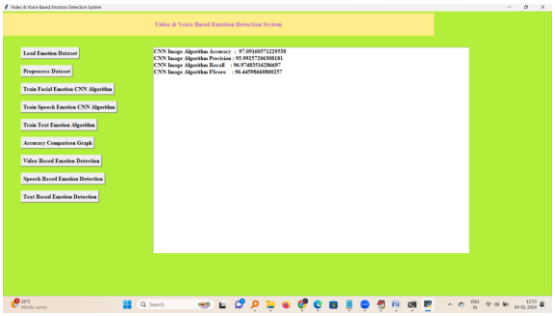
Load Emotion Dataset



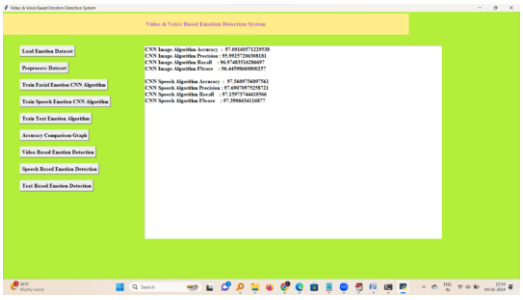
Preprocess Dataset



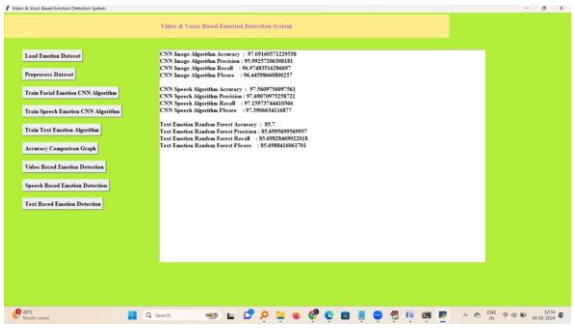
Train Facial Emotion using CNN



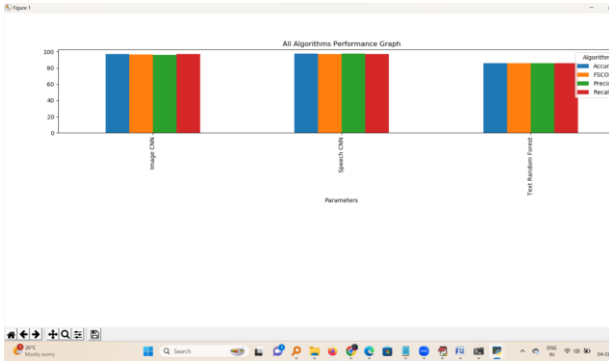
Train Speech Emotion using CNN algorithm



Train Text Emotion algorithm

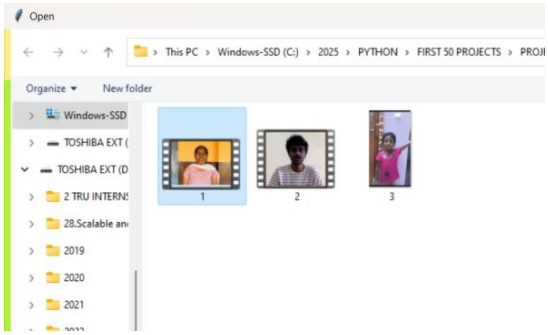


Accuracy Comparison Graph

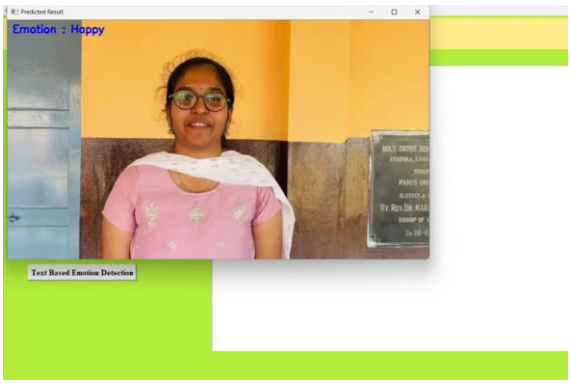


Upload Video

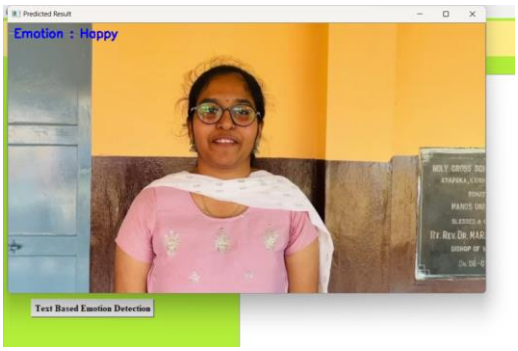
Train Facial Emotion using CNN



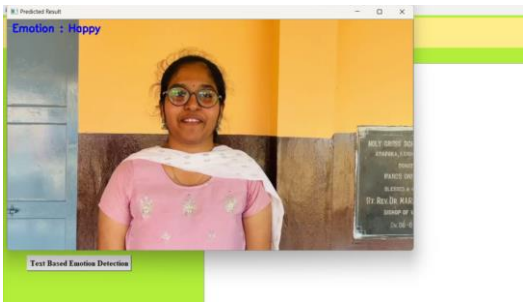
Detect Emotion



Detect Audio Emotion



Detect Text Emotion



Conclusion

Overall, the use of video, audio and text modalities to detect emotions is a major step forward in the domain of affective computing. The proposed system combines the strengths of each modality to provide a more comprehensive understanding of

human emotions with a more nuanced and accurate approach. The multimodal architecture allows detecting the complexities of emotional expression, making it a more rich and more powerful analysis than a system working with a single data source. This approach leads to more precise emotion recognition, and to better user-to-technology relationships.

This innovative system's capabilities in real-time processing allow for instant feedback in different applications, such as client support, psychological monitoring, and so on. In situations where users' emotional responses matter a great deal to their experiences and results, this immediacy is vital. The system's ability to handle environmental variability demonstrates robustness and reliability, ensuring it performs consistently even in less than ideal conditions. Its versatility allows it to be used in a variety of settings, making it more practical and effective.

Moreover, because of the emphasis placed on context and expression in context, the system will be able to catch nuances that aren't normally picked up by traditional systems. The system proposed here can use sophisticated natural language processing (NLP) algorithms to understand complex emotions in text, helping to achieve a more comprehensive evaluation of emotions. It is also very useful in mental health, where the intensity of emotional responses is critical to effective interventions and support.

Another important component of the proposed system is the focus on transparency and interpretability, as the integration of explainable AI features is essential for ensuring that the system is understandable and reliable. The system gives information about the decision making process, which helps users gain trust and confidence in the system, particularly in sensitive applications. In an era of growing adoption of emotion detection technologies, the importance of their correct and responsible use is paramount—and this focus on explainability is appropriate for those objectives.

In conclusion, the proposed emotion detection system is a cutting-edge innovation that has the potential to significantly improve HCI. It brings together cutting-edge methods from video, audio and text analysis, enabling more empathetic and responsive systems to better understand and interpret human emotions. With ongoing advancements in research and technology, these systems could revolutionize several industries and foster more meaningful and impactful interaction between people and digital platforms. As these technologies continue to mature and focus on ethical issues and user-centered design, they will play a key role in defining the future of emotion detection and its uses.

References

1. Deep Learning, I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
2. Alex Krizhevsky, I. Sutskever, and Geoffrey E. Hinton, “ImageNet classification with deep

convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.
3. Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
4. Kaiming He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
5. Ashish Vaswani *et al*., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
6. François Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. IEEE CVPR*, 2017.
7. Joseph Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018.
8. Karen Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2015.
9. Yann LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
10. Paul Ekman, "An Argument for Basic Emotions," *Cognition & Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.
11. Florian Schuller *et al*., "The INTERSPEECH 2013 Computational Paralinguistics Challenge," *Proc. INTERSPEECH*, 2013.
12. Björn Schuller, S. Steidl, and A. Batliner, *The INTERSPEECH Computational Paralinguistics Challenge*, 2014.
13. S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PLoS ONE*, vol. 13, no. 5, 2018.
14. Steven R. Livingstone and F. Russo, "RAVDESS: A Dynamic Multimodal Set of Emotional Expressions," 2018.
15. F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," *Proc. ACM Multimedia*, 2010.
16. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
17. Louis-Philippe Morency, R. Zadeh, and P. Liang, "Tensor Fusion Network for Multimodal Sentiment Analysis," *Proc. EMNLP*, 2017.
18. D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Proc. ICLR*, 2015.
19. C. Busso *et al*., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
20. A. Krizhevsky, "Learning Multiple Layers of Features from Tiny

Images," University of Toronto, Tech. Rep., 2009.
21. H. Fayek, M. Lech, and L. Cavedon, "Evaluating Deep Learning Architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.
22. S. Latif *et al*., "Deep Representation Learning in Speech Processing: Challenges, Recent Advances, and Future Trends," *IEEE Access*, vol. 9, pp. 15209–15266, 2021.
23. P. Tzirakis, G. Trigeorgis, M. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.
24. S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.
25. Z. Zhao, Y. Bao, Z. Zhang, and H. Wang, "Speech Emotion Recognition Based on CNN-LSTM and Attention Mechanism," *IEEE Access*, vol. 7, pp. 171499–171510, 2019.
26. Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proc. EMNLP*, 2014.
27. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. NAACL-HLT*, 2019.
28. A. Graves, A.-R. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *Proc. IEEE ICASSP*, 2013.
29. S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, *Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies*. IEEE Press, 2001.
30. B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," *Proc. INTERSPEECH*, pp. 2794–2797, 2010.