International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

# Multi-modal Emotion Detection System

**Akanksha Unde[1], Shashank Kulkarni[2], Harshada Sutar[3], Pradnya Suryanwanshi[4]**

*Department Of Information Technology*
*Marathwada Mitra Mandal's College of Engineering, Karvenagar, Savitribai Phule University, Pune*

---

**Abstract –**

Emotions are fundamental aspects of human communication and play a crucial role in our daily lives. Accurately detecting and understanding emotions can have profound implications in various domains, including human- computer interaction, mental health analysis, and personalized user experiences. However, existing emotion detection systems often focus on individual modalities, such as facial expressions or textual cues, which limit their effectiveness and real-world applicability. To address this, The Multi-Modal Emotion Detection System presented in this project integrates multi-modal data to effectively detect and classify emotions. By leveraging Convolutional Neural Networks (CNN) for real- time emotion recognition from webcam input and utilizing the Natural Language Toolkit (NLTK) for text processing and analysis, the system aims to capture a more comprehensive understanding of emotional states. The motivation behind this project stems from the need for a reliable and versatile emotion. This project encompasses several key components, including data collection, pre-processing, feature extraction, and fusion of multi-modal information. A diverse dataset comprising facial expression images and corresponding textual data is collected and annotated. The collected data is then pre-processed to ensure its quality and consistency. Feature extraction techniques tailored to each modality are employed to extract meaningful representations from the input data. To effectively combine the modalities, fusion strategies are employed, enabling the integration of facial expressions and textual features. Convolutional Neural Networks (CNNs) are trained to learn the underlying patterns and relationships within the fused data, facilitating accuracy.

Keywords: Deep learning, Image processing, Feature Extraction, CNN Model, Face Detection, Regression, Real Time

## 1. INTRODUCTION

Emotions play a significant role in human communication and have a profound impact on our decision-making, behaviour, and overall well-being. Understanding and accurately detecting emotions are crucial in various domains, including human- computer interaction, mental health analysis, and personalized user experiences. However, traditional emotion detection systems often rely on single-modal approaches, such as analysing facial expressions or processing textual data, which limits their effectiveness.

and real-world applicability. To address this limitation, this project introduces a Multi-Modal Emotion Detection System that combines facial expressions and textual data to enhance the accuracy and holistic understanding of emotions.

The motivation behind this project stems from the need for a more comprehensive and versatile emotion detection system capable of integrating multiple modalities. By leveraging Convolutional Neural Networks (CNN) for real-time emotion recognition from webcam input and utilizing the Natural Language Toolkit (NLTK) for text processing, our system aims to capture a more nuanced and context-aware understanding of emotional states. By fusing visual and textual cues, the Multi-Modal Emotion Detection System
overcomes the limitations of single-modality approaches and offers a more robust and accurate emotion detection capability.

The project involves several key components, including data collection, pre-processing, feature extraction, and fusion of multi-modal information. A diverse dataset comprising facial expression images and corresponding textual data is collected and annotated. To ensure data quality and consistency, the collected data undergoes pre-processing steps. Feature extraction techniques tailored to each modality are then applied to extract meaningful representations from the input data. The fusion of facial expressions and textual features is achieved through carefully designed fusion strategies, enabling a holistic understanding of emotions.

The proposed Multi-Modal Emotion Detection System holds great potential for real-time emotion monitoring, sentiment analysis, and adaptive user interfaces. By combining real-time webcam-based facial expression analysis with textual data processing, the system can accurately recognize and categorize emotions, providing valuable insights into users' emotional states. Furthermore, this project serves as a foundation for future research and development in the field of multi-modal emotion detection.

A. **Problem Statement:**

Current emotion detection systems struggle to accurately detect emotions from multiple modalities. A more sophisticated multi-modal system is needed to better interpret complex emotional cues and enable accurate classification, with potential applications in psychology, healthcare, and human-computer interaction.

B. **Motivation:**

• Overall, the motivation for this project lies in the need for a more advanced and versatile emotion detection system that can accurately interpret emotions from multiple modalities.

• By leveraging the power of multi-modal data analysis, we aim to overcome the limitations of existing approaches and unlock the potential for significant advancements in psychology, healthcare, and human- computer interaction.

• The current work is especially undertaken to seek out the presence of depression in faculty students by learning their countenance. This technique chiefly uses totally different image process techniques for face detection, feature extraction and classification of those options as depressed or non-depressed. The system is going to be trained with options of depression.

C**. Objective:**

• Collect and curate a diverse dataset comprising facial expression images and corresponding textual data.

• Pre-process the collected data to ensure consistency, quality, and compatibility for subsequent analysis.

• Develop feature extraction techniques tailored to each modality (facial expressions and textual cues) to extract meaningful representations.

• Implement fusion strategies to integrate and combine the modalities effectively, enabling a more comprehensive understanding of emotions.

## 2. PROPOSED SYSTEM

A. **System Architecture:**
The system is a multi-modal emotion detection system that uses a combination of computer vision and natural language processing techniques to detect emotions in real-time from video and text inputs. The system consists of a convolutional neural network (CNN) for analysing video input, and Text2Emotion algorithm for analysing text input. The output of these models is then combined and processed to provide a comprehensive emotional analysis of the input. The system can be deployed on a standard computer with a graphics card and sufficient processing power to handle real-time video analysis.
This project aims to develop a multi-modal emotion detection system using machine learning techniques to accurately classify emotions from facial expressions, voice intonation, and body language. The system will have potential applications in psychology, healthcare, and human- computer interaction. The project will involve data collection, model development and training, and system evaluation and testing to develop a proof-of-concept system.

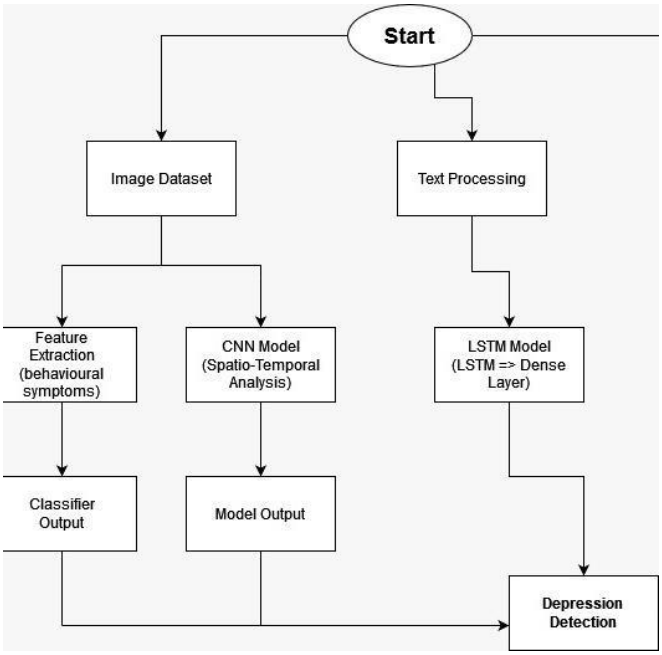


**Figure-1:** System Architecture

**B. Dataset:**
It is a dataset available freely on Kaggle. It contains around 36000 grayscale images. It was collected for easier recognition of the most common facial expressions portrayed by humans. These emotions include “Anger”, “Disgust”, “Fear”, “Happy”, “Neutral”, “Sad”, “Surprise”. Dataset is already divided between around 28000 Training Set and the rest being the validation set. This dataset was chosen considering its overall size and availability of all the basic and common facial emotions that are portrayed by humans.

C. **SYSTEM REQUIREMENT**

I. Hardware Requirement:
• Processor: - Intel Core 5 / AMD Ryzen 5 or more

• RAM: - 4 GB or Higher

• Storage: - Space 256 GB or above

• I/O Devices: - Mouse, Keyboard, Camera

II. **Software Requirement:**
• Operating System: - Windows 10/11

• Libraries: - TensorFlow, Keras, Matplotlib, NumPy, OpenCV, pandas

• Front end & Back end: - Python 3.6

## 3. RESULT ANALYSIS

A. **Text Classification: -**
For the text classification task, the model achieved the following results:

Emotion: Happy
Precision: 0.69
Recall: 0.74
F1-score: 0.71

Emotion: Sad
Precision: 0.66
Recall: 0.66
F1-score: 0.66

Emotion: Angry
Precision: 0.64
Recall: 0.62
F1-score: 0.63

Emotion: Surprise
Precision: 0.71
Recall: 0.79
F1-score: 0.75

Emotion: Fear
Precision: 0.80
Recall: 0.73
F1-score: 0.76

The model demonstrates good performance in detecting emotions such as "Happy" and "Surprise," with relatively high precision, recall, and F1-score values. However, it exhibits slightly lower performance in identifying emotions like "Sad" and "Angry."
These results provide insights into the strengths and weaknesses of the Text Classification Model, highlighting areas for potential improvement to enhance its accuracy and performance across all emotion categories.

**B. Live Model Results: -**
The Live Emotion Detection Model, integrated into our Multi-Modal Emotion Detection System, analyzes real-time video input to predict emotions. We evaluated the model's performance, obtaining an accuracy of 72% and a data loss of 0.53.

The accuracy metric measures the overall correctness of the model's predictions, indicating that it accurately classifies emotions in approximately 72% of cases. This signifies the model's ability to make reasonably accurate predictions in real-time scenarios.
The data loss of 0.53 represents the discrepancy between the actual emotions expressed in the video input and the emotions predicted by the model. A lower data loss indicates a better alignment between the model's predictions and the ground truth emotions.

These results demonstrate the effectiveness of the Live Emotion Detection Model in capturing and interpreting emotional cues from real-time video inputs. The accuracy achieved and the data loss value provide a quantitative assessment of the model's performance,

contributing to its validation and potential applications in fields such as psychology, healthcare, and human-computer interaction.

## 3. Future Scope

• The successful development of our Multi-Modal Emotion Detection System opens several promising avenues for future exploration and enhancements. Here are some potential future scopes for the project:

• Integration of additional modalities: While our system currently combines video and text inputs, future research can explore the integration of other modalities, such as audio and physiological signals. Incorporating audio analysis techniques can capture vocal intonations and speech patterns, further enriching the emotion detection process. Additionally, incorporating physiological sensors, such as heart rate monitors or galvanic skin response sensors, can provide valuable insights into the user's physiological responses and emotional arousal levels.

• Continuous emotion tracking: Enhancing the system to enable continuous emotion tracking would provide a more detailed understanding of emotional dynamics over time. By analysing emotions across multiple frames or text passages, our system could capture the temporal aspects of emotions, allowing for more comprehensive emotion monitoring and analysis.

• Deep learning architectures: Further exploration can be done to investigate advanced deep learning architectures for emotion detection. For instance, recurrent neural networks (RNNs) or long short-term memory (LSTM) networks can capture sequential dependencies in video or textual data, enabling better modelling of temporal dynamics in emotions. Transformer-based architectures, such as the popular BERT model, can also be explored for text analysis, considering their effectiveness in capturing contextual information and semantic relationships.

• Real-world deployment and validation: Conducting extensive validation studies and deploying the system in real-world scenarios would provide valuable insights into its performance and practical usability. This includes testing the system with diverse datasets, including different cultural contexts and demographic groups, to ensure its effectiveness across various populations.

• User interface and interaction improvements: Enhancing the user interface and interaction design of the system can improve its usability and user experience. Implementing intuitive visualizations of emotional analysis, providing real-time feedback, and integrating interactive features can make the system more engaging and accessible to users.

• Overall, the future scope of the project lies in advancing the system's capabilities, exploring new modalities, refining the algorithms, conducting rigorous validations, and improving user interaction. By continuously refining and expanding the system, we can unlock its full potential and further contribute to the field of emotion detection and its applications.

### 3. Conclusion

Our project focused on developing a Multi-Modal Emotion Detection System that combines computer vision and natural language processing techniques for real-time emotion analysis. By integrating video and text inputs, our system provides a comprehensive understanding of emotional cues, overcoming the limitations of existing single-modal approaches. Through the implementation of a Convolutional Neural Network (CNN) model, we successfully analyzed facial expressions from video inputs, capturing intricate details of facial features to predict emotions. Additionally, the integration of the "Text2Emotion" algorithm allowed us to extract emotions from textual data using sentiment analysis and keyword matching techniques. The results of our project highlight the effectiveness of combining visual and textual cues for accurate emotion detection and classification. Our Multi- Modal Emotion Detection System holds significant potential in psychology, healthcare, and human-computer interaction domains, enabling better understanding of emotional states, supporting mental health monitoring, and enhancing interactive experiences. In conclusion, our project showcases the successful implementation of a Multi-Modal Emotion Detection System that leverages computer vision and natural language processing techniques. As further advancements are made, our system has the potential to contribute to various fields.
and open new avenues for emotion analysis and its applications in real-world scenarios.

## REFERENCES

*1. Girard, Jeffrey M., Jeffrey F. Cohn, Mohammad H. Mahoor, SeyedmohammadMavadati, and Dean P. Rosenwald. "Social risk and depression: Evidence from manual and automatic facial expression analysis." Multimodal Deep Learning Framework for Mental Disorder Recognition on, pp. 1-8. IEEE, 2013.*

*2. W. C. de Melo, E. Granger and A. Hadid, "Depression Detection Based on Deep Distribution Learning," 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 4544- 4548, doi: 10.1109/ICIP.2019.8803467.*

*3. Thati, R.P., Dhadwal, A.S., Kumar, P. et al. A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms. Multimed Tools Appl (2022). (Springer)*

*4. Victor, Ezekiel, M. Aghajan, Zahra Sewart, Amy Christian, Ray. (2019). Detecting Depression Using a Framework Combining Deep Multimodal Neural Networks With a Purpose-Built Automated Evaluation. Psychological Assessment. 31. 10.1037/pas0000724.*

*5. S. Alghowinem et al., "Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors," IEEE Transactions on Affective Computing, vol. 9, no. 4, pp. 478–490, Oct. 2018.*

*6. Indonesian Journal of Electrical Engineering and Computer Science- A computer vision based image processing system for depression detection among students for counseling Vol. 14, No. 1, April 2019, ISSN: 2502-4752, DOI: 10.11591/ijeecs.v14.*