International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Design and Implementation of a Framework for Emotion Detection System Utilizing Audio and Video-Based Expressions Using AI

Shaik Kareemulla Sha[1], Ch Santhi[2], Pinisetti Devi Sri Satya Sai[3]

[1]M.Tech Scholar, Computer Science And Engineering, Bonam Venkata Chalamayya Institute Of Technology And Science
[2]Assistant Professor, Computer Science And Engineering, Bonam Venkata Chalamayya Institute Of Technology And Science
[3]B.Tech Scholar, Computer Science And Engineering, Bonam Venkata Chalamayya Institute Of Technology And Science

Abstract
Emotion recognition is an important area in artificial intelligence and human-computer interaction. Traditional emotion detection systems mainly depend on a single source such as facial expressions or speech signals, which may reduce accuracy in real-time conditions. This paper proposes a multimodal emotion detection framework using audio, video, and text-based expressions with artificial intelligence techniques.
The proposed system uses Convolutional Neural Networks (CNN) for facial emotion recognition, Long Short-Term Memory (LSTM) networks for speech emotion analysis, and Natural Language Processing (NLP) for understanding text queries. The system processes multimedia inputs and predicts emotions such as happiness, sadness, anger, fear, surprise, and neutral emotions.
The framework is implemented using Python, TensorFlow, Keras, OpenCV, and Flask. Experimental results show that combining multiple modalities improves emotion recognition accuracy and system performance compared to traditional single-modality methods.

Keywords: Emotion Detection, Artificial Intelligence, Deep Learning, CNN, LSTM, NLP, Multimodal Emotion Recognition, Audio and Video Analysis.

I. Introduction

Emotion recognition is widely used in applications such as healthcare, surveillance, education, robotics, and human-computer interaction. Artificial intelligence helps systems understand human emotions through facial expressions, speech signals, and text analysis.
Most existing systems focus on only one type of input such as image or audio data. These single-modality systems may produce inaccurate results because human emotions are expressed through multiple forms like face, voice, and language together.
To overcome these limitations, the proposed system combines image processing, speech analysis, and natural language processing into one framework. CNN models are used for facial emotion detection from

images and videos, while LSTM models analyse emotions from speech signals. NLP techniques are used to understand user text queries.

The system allows users to upload multimedia inputs and retrieve emotion-based predictions through a web interface. By combining multiple modalities, the proposed framework improves prediction accuracy and emotional understanding.

II. Related Work

Several researchers have explored emotion recognition techniques using facial expressions, speech analysis, multimodal learning, and deep learning models.

- Ekman and Friesen (1971) established that emotions such as happiness, sadness, anger, fear, surprise, and disgust are universally recognized through facial expressions across different cultures. Their work laid the foundation for modern facial emotion recognition systems.
- Russell (1980) proposed the Circumplex Model of Affect, which represented emotions using dimensions such as valence and arousal instead of fixed categories. This model helped improve emotional understanding and sentiment analysis.
- Kosti et al. (2017) emphasized the importance of contextual understanding in emotion recognition. Their work demonstrated that emotions cannot be accurately identified using facial expressions alone and that surrounding environmental context also contributes to emotional perception.
- You et al. (2016) introduced a large-scale image emotion recognition dataset developed through crowd-sourced annotations. Their dataset significantly improved the performance of deep learning-based emotion classification systems.
- Wang et al. explored multimodal sentiment analysis by integrating text, audio, and visual features using deep learning approaches. Their work demonstrated that combining multiple modalities improves prediction accuracy compared to single-modality systems.
- Radford et al. introduced the CLIP model for learning visual representations from natural language supervision. This work enabled semantic understanding between text and images and influenced multimodal AI research.
- Vivek Jain and Erik Learned-Miller proposed neural network-based facial emotion recognition systems that automatically learn features from image data rather than relying on handcrafted methods.
- Xiaohua Zhang et al. explored deep neural networks for facial emotion recognition and demonstrated the effectiveness of advanced architectures in capturing subtle emotional variations.
- Wang and Ji proposed a unified probabilistic framework for multimodal emotion recognition by combining facial expressions, speech, and contextual information.
- Sun et al. reviewed video-based emotion recognition techniques using deep learning and highlighted the importance of temporal feature extraction for analysing emotional dynamics in videos.

These studies collectively show that multimodal deep learning approaches provide better emotional understanding, improved prediction accuracy, and greater adaptability in real-world emotion recognition applications.

III. Existing System

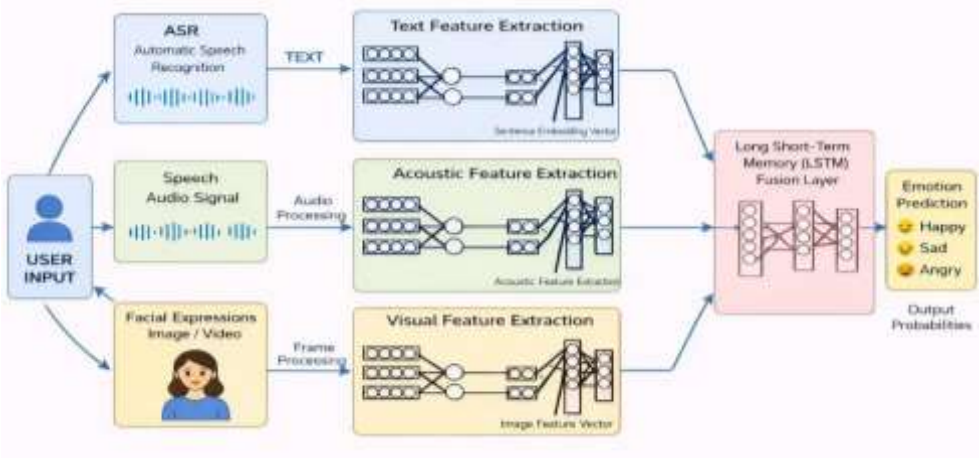
Existing emotion recognition systems mainly focus on single-modality analysis such as facial expression recognition, speech emotion detection, or text-based sentiment analysis. These systems use traditional machine learning and image processing techniques to identify emotions.

Image-based systems analyse facial expressions, while speech-based systems use voice features such as pitch and tone. Text-based systems classify emotions using sentiment analysis methods.

Although these systems provide reasonable results, they have several limitations:

1. They depend on only one type of input.
2. Accuracy decreases in noisy or real-time environments.
3. Complex emotional patterns are difficult to identify.
4. Contextual understanding is limited.
5. Real-time multimodal emotion analysis is not supported.

Due to these limitations, a multimodal emotion recognition framework is required for better accuracy and performance.



IV. Proposed Methodology

The proposed system uses a multimodal deep learning framework to detect emotions from images, videos, audio, and text inputs. The framework combines Computer Vision, Speech Emotion Recognition, and Natural Language Processing techniques.

Initially, users upload multimedia files and enter text queries through a web interface. The input data is pre-processed by resizing images, extracting video frames, cleaning audio signals, and processing text.

CNN models such as VGG16 or ResNet are used for facial emotion recognition from images and video frames. LSTM models analyse speech signals and identify emotional patterns from audio data. NLP techniques process text queries and understand emotional meaning.

The extracted features from different modalities are combined using a fusion layer. The final output predicts emotions such as happy, sad, angry, fear, surprise, and neutral. The predicted results are displayed through the web interface.

V. System Architecture

The proposed system consists of the following layers:

- Input Layer – Accepts images, videos, audio files, and text queries.
- Preprocessing Layer – Performs data cleaning, resizing, normalization, and tokenization.
- CNN Layer – Extracts visual emotion features.
- LSTM Layer – Analyses emotional speech patterns.
- NLP Layer – Understands emotion-based textual queries.
- Fusion Layer – Combines multimodal features.

- Classification Layer – Predicts final emotions.
- Output Layer – Displays emotion detection results and performance metrics.

VI. Implementation

The proposed emotion detection system is implemented using Python and AI libraries for processing image, audio, video, and text data.

The system uses a Flask-based web interface where users can upload multimedia inputs and enter text queries. Images and video frames are processed using OpenCV, while audio signals are analysed using LSTM models. CNN models are used for facial emotion recognition, and NLP libraries process text-based emotional queries.

The outputs from all modalities are combined to improve prediction accuracy and provide reliable emotion detection results.

Technologies Used

- Python – Overall system development
- TensorFlow – Deep learning framework
- Keras – Neural network implementation
- OpenCV – Image and video processing
- NumPy – Numerical computations
- Pandas – Data handling and analysis
- Flask – Web application development
- NLTK / SpaCy – Natural Language Processing

Implementation Advantages

1. Supports multimodal emotion recognition.
2. Improves prediction accuracy using deep learning.
3. Provides real-time multimedia analysis.
4. User-friendly web interface.
5. Flexible and scalable architecture.
6. Supports future integration with advanced AI models.

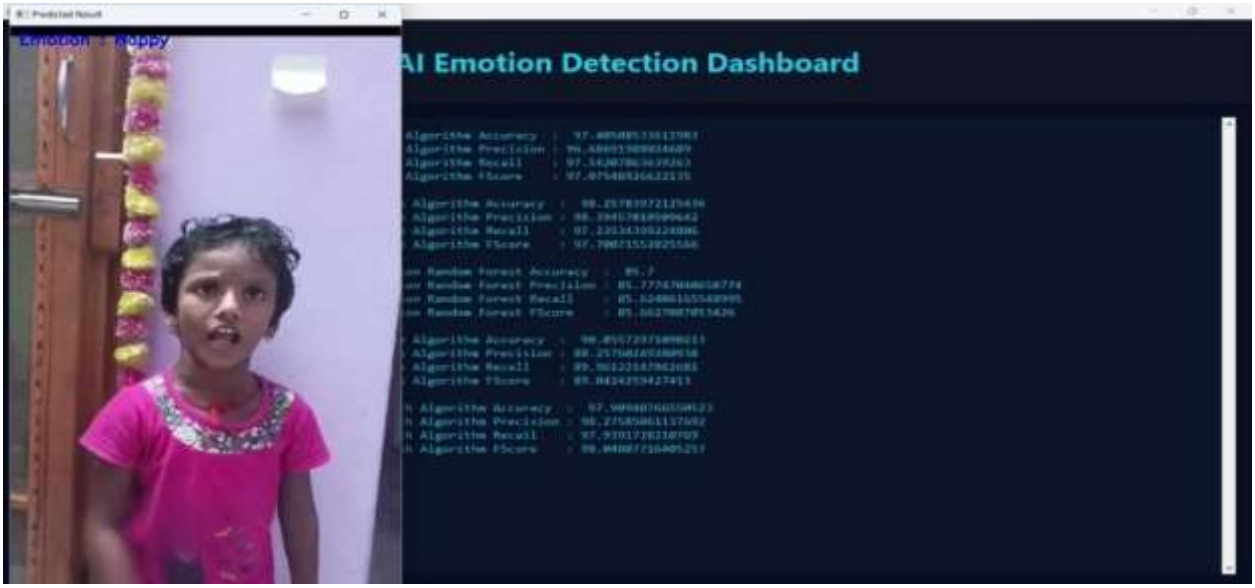
VII. Results and Discussion

The proposed emotion recognition framework was tested using image, audio, and text inputs. The system successfully detected emotions such as happy, sad, angry, fear, surprise, and neutral emotions.

CNN models provided effective facial emotion recognition from images and video frames, while LSTM models accurately analysed speech emotions from audio signals. NLP techniques improved the understanding of user text queries.

The multimodal approach improved overall prediction accuracy compared to single-modality systems. The system also provided reliable performance and smooth user interaction through the web interface.


AI Emotion Detection Dashboard
All Algorithms Performance Graph
Algorithms
Accuracy
FSCORE
Precision
Recall
Parameters


AI Emotion Detection Dashboard


AI Emotion Detection Dashboard

VIII. Conclusion

The proposed emotion detection system successfully combines image, audio, and text analysis using artificial intelligence techniques. CNN, LSTM, and NLP models are used to identify emotions from multimedia inputs.

The system improves emotion recognition accuracy compared to traditional single-modality systems and provides better emotional understanding through multimodal analysis. The framework is flexible, efficient, and suitable for real-world applications such as healthcare, surveillance, and human-computer interaction.

IX. Future Work

The proposed system can be further enhanced in several ways:

1. Support real-time emotion detection using live video streams.
2. Improve audio emotion recognition accuracy using advanced transformers.
3. Integrate advanced multimodal fusion techniques.
4. Support multilingual text-based emotion understanding.
5. Develop mobile and cloud-based deployment solutions.
6. Add emotion intensity and psychological state analysis.
7. Implement explainable AI visualization methods.
8. Enhance system scalability for large multimedia datasets.
9. Improve performance under noisy and low-light conditions.
10. Integrate wearable sensors for physiological emotion analysis.

References

1. Ekman, P., and Friesen, W. V., “Constants across Cultures in the Face and Emotion,” Journal of Personality and Social Psychology, 1971.
2. Russell, J. A., “A Circumplex Model of Affect,” Journal of Personality and Social Psychology, 1980.
3. You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J., “Building a Large-Scale Dataset for Image Emotion Recognition,” IEEE Transactions on Multimedia, 2016.
4. Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A., “Emotion Recognition in Context,” Proceedings of CVPR, 2017.
5. Busso, C., et al., “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database,” Language Resources and Evaluation, 2008.
6. Soleymani, M., et al., “A Survey of Multimodal Sentiment Analysis,” Image and Vision Computing, 2017.
7. Zhang, Z., et al., “Facial Emotion Recognition using Deep Learning: A Survey,” IEEE Access, 2018.
8. Wang, S., and Ji, Q., “A Unified Probabilistic Framework for Multimodal Emotion Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.
9. Poria, S., Cambria, E., Bajpai, R., and Hussain, A., “A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion,” Information Fusion, 2017.
10. Sun, R., Wu, J., and Zhao, Y., “Video-Based Emotion Recognition using Deep Learning: A Review,” Information Fusion, 2020.
11. Goodfellow, I., Bengio, Y., and Courville, A., “Deep Learning,” MIT Press, 2016.
12. LeCun, Y., Bengio, Y., and Hinton, G. E., “Deep Learning,” Nature, 2015.

13. Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," NIPS, 2012.
14. Hochreiter, S., and Schmidhuber, J., "Long Short-Term Memory," Neural Computation, 1997.
15. Simonyan, K., and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," ICLR, 2015.