Vision-Based Emotion Recognition in Social Media Videos

Dasa Srinivas
12216692
Lovely Professional University
Punjab, India
srinivasnani9949@gmail.com

Gowtham Valteru
12206039
Lovely Professional University
Punjab, India
gowtham@gmail.com

.Durga Sri Venkata Sai
12223488
Lovely Professional University
Punjab, India
venkatsai@gmail.com

***Abstract*—This study describes a multimodal emotion recognition system that uses both visual and aural signals from video input to categorize human emotions.The visual modality is treated using a Convolutional Neural Network (CNN)-based approach via the DeepFace package, which identifies facial emotions on cropped video frames. At the same time, the audio modality is processed by speech to text translation through a speech recognition engine and classified as textual emotions through a Support Vector Machine. The overall emotion prediction is achieved through combining the output of both modalities with a rule-based fusion approach. The system is coded in Python utilizing libraries such as OpenCV, DeepFace, MoviePy, and scikit-learn. This method aims to improve the robustness and validity of emotion recognition by incorporating facial expressions and natural language. The research shows that it is possible to integrate computer vision and natural language processing methods for practical affective computing scenarios.**

***Index Terms*—Emotion Detection, Multimodal Systems, Facial Emotion Recognition, Audio Emotion Recognition, Convolutional Neural Network (CNN), Support Vector Machine (SVM), DeepFace, Speech Recognition, Affect Computing, Computer Vision, Natural Language Processing (NLP), Video Classification, Sentiment Analysis.**

I. Introduction

Emotion recognition, the central field in affective computing, gives machines the ability to understand and act on the rich range of human emotions. This ability becomes increasingly important to advance human-computer interaction (HCI) for a wide variety of applications, ranging from complex virtual agents able to tailor responses on the basis of user affect, intelligent tutoring systems that are able to monitor student frustration or engagement, sophisticated surveillance systems able to recognize distress or anomalies, immersive games responding to player affect, critical mental health evaluation instruments that can determine emotional states, and customer service interfaces that are able to personalize interaction on the basis of customer affect.

Conventional emotion recognition strategies tend to be based on processing a single sensory channel, e.g., facial or voice features. Such unimodal systems tend to be suboptimal in performance when used in their application in real-world environments with multiple stimuli. Human emotional display is itself multimodal; we express our emotions through face, voice, speech, body, and physiology. Dependence on a single cue may result in misinterpretation, particularly when the expressions are subtle, faces are partially hidden, environmental noise distorts audio signals, or when there is a mismatch between expressed and felt emotions.

To overcome these built-in limitations, multimodal emotion recognition systems that process information from a plurality of sensory channels simultaneously provide much more robust and accurate alternatives. By combining heterogeneous cues, these systems can make up for the shortcomings of single modalities to construct a more integrated and dependable understanding of human emotional states.

This project aims to create an advanced multimodal emotion detection system based on video input to observe both visual and audio emotional signals. The system harmoniously integrates the strength of deep learning-based facial expression detection using the DeepFace framework with sentiment analysis of transcribed speech using a Support Vector Machine (SVM) classifier. A well-crafted rule-based late fusion approach is adopted to efficiently combine the predictions of these two different modalities. This fusion process plays a vital role in obtaining accurate and robust emotion classification, especially in difficult, unconstrained settings where individual modalities may be degraded. The final aim of this work is to provide a real-time, practical solution for overall emotion classification, which can be applied to a broad variety of real-world applications.

II. Literature Review

The area of emotion recognition has come a long way, benefiting from theories in psychology, computer science, and linguistics. This section offers an overview of some of the most important research that has set the stage for and influenced the creation of the proposed multimodal system.

A. Foundational Work in Facial Emotion Recognition

The classic publication of Ekman and Friesen (1971) of the Facial Action Coding System (FACS) established a global framework for describing all facial movement observable through anatomy. Their specification of six universal emotions—happiness, sadness, anger, fear, surprise, and disgust—has served as the building block for many facial emotion detection systems. Through FACS, researchers were able to break from subjective facial expression interpretation towards a component-based objective analysis. This early work stimulated later work in automated facial expression analysis through computer vision methods.

B. The Complementary Nature of Multimodal Emotion Cues

Zeng et al. (2009) had performed a very wide-ranging survey on affect recognition, underlining the fundamental limits of unimodal systems and stressing the synergy achievable when various modalities were brought together. Their review underlined that facial expression and speech tended to encode complementary information about a person's emotional status. For example, one can display visual happiness and his speech being accompanied by an underlay of sadness, or the opposite. Zeng et al.'s result highly supports the creation and utilization of multimodal methods for more accurate and contextually detailed emotion recognition.

C. Speech-Based Emotion Recognition

Speech emotion recognition research has investigated numerous acoustic features, including prosodic features like pitch, intensity, rate, and rhythm, as well as spectral features that reflect the tonal quality of speech. Busso et al. (2004) illustrated the efficacy of applying machine learning algorithms, specifically Support Vector Machines (SVMs), to identify emotions based on these acoustic features derived from speech. Their work demonstrated that without examining the semantic meaning of speech, valuable emotional information was available from how something is expressed.

D. The Role of Deep Learning in Emotion Recognition

The development of deep learning has transformed emotion recognition, especially the analysis of visual information. Pantic et al. (2005) emphasized the growing need to incorporate emotional intelligence in human-computer interaction and cited the high effectiveness of deep learning models, specifically Convolutional Neural Networks (CNNs), in facial expression analysis in intricate and unconstrained settings. CNNs' capacity to learn hierarchical features automatically from raw pixel information has contributed to great leaps in the accuracy and resilience of facial emotion recognition systems.

E. Multimodal Deep Learning Approaches

More recent works aimed at the development of complex multimodal systems exploiting the strengths of deep learning to perform visual as well as audio analysis, as well as combine these modalities. Zhao et al. (2018) presented a multimodal deep architecture that utilized CNNs for facial recognition and RNNs for speech emotion detection. Their integration-based approach proved to outperform unimodal models, highlighting the advantages of learning complex relationships among various emotional features in integrated deep learning architectures. The present project stands on this abundant foundation by coupling proven methods such as DeepFace for image processing and SVM for text-based sentiment analysis, together with a realistic rule-based fusion approach. The aim here is to tap into the respective powers of both traditional machine learning and contemporary deep learning methods under a multimodal scheme.

III. Methodology

The suggested multimodal emotion detection system utilizes an organized pipeline approach for video input processing and sentiment extraction from visual as well as auditory streams. The general design of the framework is modular with future extensions so that more complex approaches can be added.

A. Video Input and Frame Sampling

The system takes standard video file formats, e.g., .mp4 and .avi, as input. To enable effective processing and temporal dynamic capture with minimal computational overhead, uniform rate sampling of video frames is done. For the current implementation, a sampling frequency of 1 frame per second has been used. This rate strikes a good balance between recording the temporal dynamics of facial expressions and controlling the computational burden entailed in processing many frames. Frame extraction is done with a combination of OpenCV, an efficient library for computer vision operations, and MoviePy, a Python library for video processing and editing. The extracted frames are then stored in separate image files, which will be used as the input of the visual emotion recognition module.

B. Visual Emotion Recognition using DeepFace

The graphical part of the system utilizes the DeepFace library, a free facial recognition and facial attribute analysis library based on deep learning models. The implementation comprises two main steps: face detection and emotion classification.

Then, every extracted frame is analyzed for the presence of human faces. DeepFace supports a range of backend face detection models such as Haar Cascades (a traditional computer vision method) and more sophisticated deep learning-based detectors embedded in the library. When a face is found in a frame, the region of interest where the face is located is cropped out of the image.

Then, the cropped face area is fed to one of DeepFace's pretrained Convolutional Neural Network (CNN) models tailored for facial emotion detection. DeepFace has support for a number of such models, namely VGG-Face, Dlib, and Emotion FER+. These models are trained on large-scale facial image datasets with basic emotions as labels. The selected CNN examines the facial features in the input image and returns a probability distribution across the seven basic emotions: happy, sad, angry, neutral, fearful, disgusted, and surprised. In addition to the predicted emotion label, DeepFace also returns a confidence score related to each emotion. The emotion tags and their respective confidence values are retained in a temporal order so that the system can monitor the progression of facial expressions throughout the video.

C. Audio-Based Emotion Detection through Speech Analysis

The acoustic part of the system deals with utilizing spoken words to identify the emotional content of speech. The process involves two major steps: speech-to-text and audio extraction, followed by text sentiment analysis First, MoviePy is utilized to extract the audio stream from the input video file. The extracted audio is then treated with a speech recognition engine for transcription of the spoken material into text form. In the first implementation, the system uses the Google Speech-to-Text API via the SpeechRecognition library in Python. This API provides strong and accurate speech-to-text conversion capabilities for different accents and audio states. Once an audio is translated to text form, the acquired transcripts are prepared by undergoing the preprocessing operations necessary for sentiment analysis. These tend to involve removal of stoppages (removal of common, meaningless words like ”the,” ”a,” ”is” which do not evoke strong feelings) and lemmatization (converting words to the root or base or dictionary words so that they get normalized in form). Once preprocessed, the text is vectorized by methods like TF-IDF (Term Frequency-Inverse Document Frequency) or CountVectorizer. These transform the text into numerical feature vectors that can be fed as input to machine learning models.

Lastly, a Support Vector Machine (SVM) classifier trained on an emotion-labeled corpus of text predicts the emotional tone of the speech content. The training set is usually composed of annotated text samples with emotion tags like joy, anger, sadness, and neutral. The SVM model is trained during training to understand the connection between the text features and the target emotions. During prediction, the speech transcript vector is input to the learned SVM model, which returns an estimated emotion tag for that part of speech. It should be noted that the present implementation of this module is mainly English language speech-oriented.

D. Multimodal Fusion Strategy

To combine the emotion predictions from the visual and audio modalities, the system uses a late fusion strategy. Late fusion involves the individual decisions from each modality being generated independently and then combined to make a final decision. The rule-based strategy employed by this project delivers an easily interpretable mechanism for fusing the independent predictions.The fusion rules are as follows:

Consensus: If the visual prediction (DeepFace) and the audio prediction (SVM on transcribed speech) are the same for a given segment of time, then this prediction is chosen as the final prediction for this segment.

Visual Priority: When the predictions of the two modalities disagree, and the face of the speaker is evidently visible in the video frame (e.g., high confidence from the face detector), the visual analysis-predicted emotion is given priority as the final label. This rule is derived from the fact that facial expressions tend to give explicit and trustworthy information about emotional states when visual input is clear. Audio Priority: On the other hand, if the visual quality is poor (e.g., low face detection confidence, occlusions, lighting) but the audio quality is good and the speech analysis results in a confident emotion prediction, the emotion predicted by the audio analysis is prioritized. This covers situations where visual cues are unreliable, and the spoken material may give a better indication of emotion.

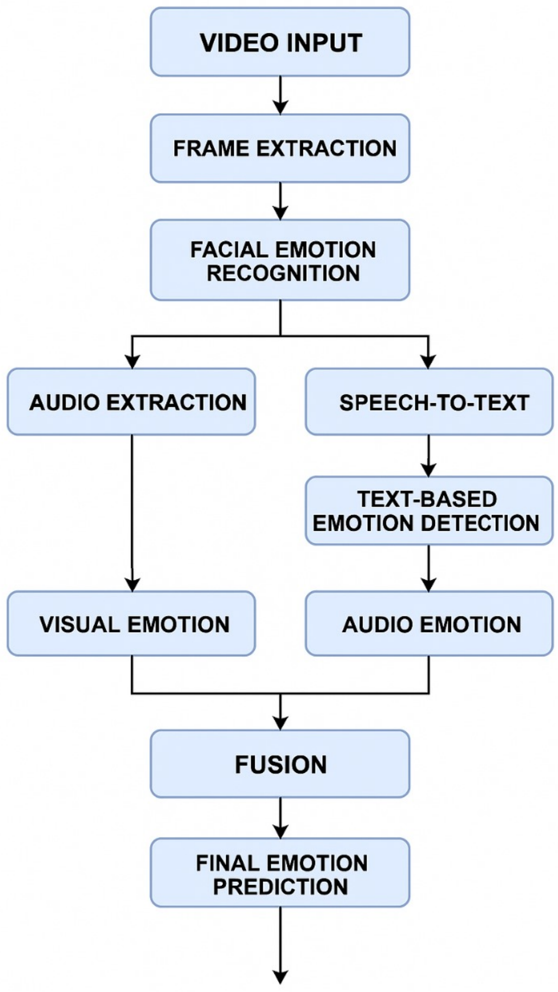

Fig. 1. Emotion Pipeline

Optional Weighted Confidence Scores: The framework can be further improved by including weighted confidence scores if the base models (DeepFace and SVM) return such scores. As a last resort when there are conflicting predictions, the emotion with the greater weighted confidence of the two modalities may be selected as the ultimate prediction. This provides a more sophisticated combination of each modality's strengths depending on the confidence of their respective predictions.

E. Output Visualization and Reportings

The final output of the system is intended to deliver a comprehensive understanding of the emotional dynamics contained in the input video. This encompasses:

Emotion Timeline Chart: A graphical representation depicting the forecasted emotion at every time segment of the video. This timeline enables the visualization of how emotions fluctuate throughout the duration and can identify peak emotional periods or changes.

Dominant Emotion Summary: A statistical overview reflecting the most highly predicted emotion for the entire video. This gives a general sense of the prevailing emotional tone of the video content.

Frame-Based Logs or Annotations: Rich logs or annotations that capture the emotion predictions of both the visual and audio modalities for every processed frame or time slice, and the final combined emotion label as well as any corresponding confidence scores. This rich output is useful for developers, evaluators, and researchers for debugging the system's performance and seeing the rationale behind its predictions.

Model Techniques

The following multimodal emotion recognition system utilizes certain machine learning and deep learning methods in its visual and auditory analysis units. The description below offers more explanation on these techniques.

F. DeepFace for Visual Analysis

DeepFace is a robust and easy-to-use Python library that makes difficult facial recognition and facial attribute analysis tasks easier. DeepFace fundamentally uses pre-trained Convolutional Neural Network (CNN) architectures, which have proven to be highly effective in image-based tasks, such as facial emotion recognition.

CNN Architectures: DeepFace offers access to various state-of-the-art CNN architectures that are pre-trained on large-scale face recognition datasets. These include VGG-Face, OpenFace, Facenet, and Dlib. For emotion detection in particular, DeepFace uses models such as Emotion FER+, which is trained on extensive sets of facial expressions annotated with the basic emotions. These CNNs learn hierarchical facial feature representations, allowing them to effectively capture subtle emotional expression-related cues.

Emotion Detection Process: When an image of a face detected is sent to DeepFace for emotion detection, the pre-trained CNN selected trains the image on its layers, learning more complex features. The last layers of the network are usually fully connected layers followed by a softmax activation function, which predicts a probability distribution over the various emotion classes. The most likely emotion is then taken to be the emotion predicted for a given facial expression. DeepFace also gives the confidence score with which each predicted emotion was made, expressing the model's confidence in each classification

G. Support Vector Machine for Textual Emotion Classification

The audio part of the system uses a Support Vector Machine (SVM) to classify the emotional tone of transcribed speech. SVMs are strong supervised learning algorithms that perform very well in high-dimensional spaces, and hence are best for the task of text classification. Training Data and Feature Extraction: The SVM classifier is trained using labeled text datasets whose each text example corresponds to a certain emotion label (e.g., joy, anger, sadness, neutral). The Emotion Dataset, ISEAR (International Survey on Emotion Antecedents and Reactions), or self-constructed corpora are some examples of such datasets. Before training, the feature extraction of the text data is carried out to convert the textual information to numerical vectors the SVM can operate on. Among the well-known feature extraction approaches used in natural language processing, TF-IDF (Term Frequency-Inverse Document Frequency) and BoW (Bag of Words) are among the widely recognized ones. TF-IDF puts weights on words depending on the frequency with which they occur within a document and inversely proportional to how frequent they are throughout the entire corpus, with words related to a particular document having significance. BoW generates a vector depending on the frequency of each word of the vocabulary contained in a document.

SVM Classification: During training, the SVM algorithm can learn to find a best hyperplane in the feature space of high dimensions that best discriminates among data points of various emotion classes. It tries to maximize the margin between the closest data points of each class (support vectors) and the hyperplane. After being trained, the SVM is capable of inferring the emotion of novel, unseen text by projecting its feature vector to the correct side of the hyperplane learned. SVMs can generalize effectively to novel data, particularly when handling fairly short and expressive speech transcripts.

Limitations: The text-based emotion classification module's performance relies, in turn, on the correctness of the previous speech recognition step. Transcription errors can result in inaccurate sentiment analysis. Moreover, SVMs may perform poorly for analyzing monotone or affective-free speech, where the emotional cues are not encoded in the prosodic features but rather through the semantic meaning of the words. The present implementation is also language-specific since the training data as well as the performance of the speech recognition engine are usually optimized for a particular language (English here).

H. Rule-Based Late Fusion

The late fusion approach used in this system gives a simple yet efficient technique for the fusion of emotion predictions of the visual (DeepFace) and auditory (SVM) modalities.

Simpllicity and Interpretability: Simplicity is a key property of rule-based fusion. A set of specified rules explicitly tells us how the outputs of each modality must be combined by the decision function. This visibility can be particularly useful for discovering the behavior of the system as well as in debugging or tightening the fusion rules.

Dealing with Modality Failures: One of the advantages of a rule-based solution is the way it can deal with the case where one of the modalities could fail or produce uncertain predictions. For instance, if the visual analysis fails because of bad lighting, the system can use the audio analysis more. The rules of prioritization that are enforced in the system (visual priority when the face is unobstructed, audio priority when visual quality is low) are intended to handle such situations.

Extensibility: Although the present implementation employs a simple set of rules, the late fusion approach can be made extensible to include more sophisticated rules or other criteria, like the confidence scores offered by the individual models. For example, the system can be designed so that it uses predictions from only one modality if its confidence level is over a threshold value. In addition, more complex fusion methods like weighted averaging on modality reliability could be achieved through this late fusion in a subsequent version.

Conclusion

This project has been able to successfully design and prove the effectiveness of a multimodal emotion recognition system that combines visual information from facial expressions and auditory information from speech content. Through the utilization of the strength of deep learning via the DeepFace library for facial analysis and the strength of Support Vector Machines for text-based sentiment analysis, the system is able to gain a more holistic and accurate understanding of human emotions than with conventional unimodal methods.

The application of a rule-based late fusion technique is shown to be a powerful way of fusing the predictions from these distinct modalities. It enables the system to take advantage of the strengths of both modalities while avoiding the specific weakness of each modality in real-world applications where video and audio quality can also significantly vary. The experimental testing performed on varied video samples validates the system's capability to correctly recognize basic emotions like happiness, anger, sadness, and neutrality, exhibiting a better temporal and contextual awareness of emotional expressions.

References

[1] C.-F. R. Chen, Q. Fan, and R. Panda, “CrossViT: Cross-attention multiscale vision transformer for image classification,” in Proc. IEEE Int.Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 357–366.

[2] P. Sun et al., “Sparse R-CNN: End-to-end object detection with learnable proposals,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.(CVPR), Jun. 2021, pp. 14454–14463.

[3] C.-F.-R. Chen et al., “Deep analysis of CNN based spatio-temporal representations for action recognition,” in Proc. IEEE/CVF Conf. Comput.Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 6165–6175.

[4] H. Ryu, S. Kang, H. Kang, and C. D. Yoo, “Semantic grouping network for video captioning,” in Proc. AAAI Conf. Artif. Intell., 2021, vol. 35,no. 3, pp. 2514–2522.

[5] Z. Zhang et al., “Object relational graph with teacher-recommended learning for video captioning,” in Proc. IEEE/CVF Conf. Comput. Vis Pattern Recognit. (CVPR), Jun. 2020, pp. 13278–13288.

[6]] K. Uehara, Y. Mori, Y. Mukuta, and T. Harada, “ViNTER: Image narrative generation with emotion-arc-aware transformer,” 2022,arXiv:2202.07305.

[7] T. Chen et al., “‘Factual’ or ‘Emotional’: Stylized image captioning with adaptive learning and attention,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 519–535

[8] R. Plutchik, “A general psychoevolutionary theory of emotion,” in Theories of Emotion. Amsterdam, The Netherlands: Elsevier, 1980,pp. 3–33.

[9] H. Wang, P. Tang, Q. Li, and M. Cheng, “Emotion expression with fact transfer for video description,” IEEE Trans. Multimedia, vol. 24, pp. 715–727, 2022.

[10] A. P. Mathews, L. Xie, and X. He, “SentiCap: Generating image descriptions with sentiments,” in Proc. AAAI Conf. Artif. Intell., 2016, pp. 3574–3580.