**Multimodal Emotion Recognition System Using Image and Audio**

**Introduction**

In the evolving field of artificial intelligence, understanding and interpreting human emotions has become a critical capability. Emotion recognition has numerous real-world applications—ranging from enhancing virtual assistants and tutoring systems to improving user experience in entertainment and telehealth services. Traditionally, emotion recognition systems have focused on a single modality such as text, facial expressions, or audio. However, such unimodal systems often struggle in real-world settings where one source of information may be distorted, occluded, or ambiguous.

This project introduces a **multimodal emotion recognition system** that integrates two powerful sources of emotional information: **facial expressions** (from static images) and **vocal tone and speech features** (from audio or video). By analyzing both, the system is capable of providing a more accurate and robust emotional assessment.

**Problem Statement**

Human emotional expression is complex, dynamic, and often multimodal. Relying solely on one form of input (e.g., facial image or speech) can lead to misinterpretation, especially in challenging environments. Common issues in unimodal systems include:

* Poor facial detection in low light or occluded images.
* Misclassification of emotions in noisy or accented speech.
* Inability to assess emotional nuance when relying only on transcribed text.

Therefore, there is a pressing need for a system that can **combine multiple cues**—visual and auditory—to detect emotions more effectively. This project addresses that need by creating a hybrid pipeline that integrates image-based and audio-based emotion recognition.

**Existing System Limitations**

Several unimodal systems have been proposed and deployed in various domains:

**Facial Emotion Recognition**

* Uses convolutional neural networks trained on datasets like FER-2013 or CK+.
* Performs well under ideal lighting and frontal face images.
* **Limitations**:
  + Sensitive to facial occlusions (e.g., glasses, masks).
  + Ineffective in poorly lit or dynamic environments.

**Audio Emotion Recognition**

* Uses acoustic features like MFCCs and pitch for emotion classification.
* Models include SVMs and RNNs trained on emotion-labeled datasets.
* **Limitations**:
  + Easily affected by background noise.
  + Struggles with multiple speakers or emotional neutrality.

**Text Emotion Recognition (optional)**

* Uses transformers like BERT for sentiment/emotion classification.
* Relies entirely on the transcribed speech.
* **Limitations**:
  + Ignores vocal tone and visual emotion.
  + Prone to misinterpretation without prosody or facial cues.

These systems fail to capture the **rich, complementary nature** of multimodal emotional expression.

**Proposed System**

The proposed solution is a **hybrid multimodal emotion detection system**. It accepts user-uploaded image and audio/video files via Google Colab, detects facial emotion from images using a trained CNN model, and detects vocal emotion by analyzing audio features such as MFCC and pitch.

**System Highlights:**

* Facial emotion recognition from images using a trained Keras CNN model.
* Whisper ASR model for speech-to-text and language detection.
* Audio emotion recognition using extracted MFCC + pitch features and a simple SVM classifier.
* Real-time feedback using Colab interface and OpenCV display.
* Modular design for future integration of text-based emotion classification.

**System Pipeline**

The system follows a structured step-by-step pipeline:

1. **File Upload**:
   * Image files: .jpg, .png
   * Audio/video files: .mp3, .wav, .mp4
2. **Facial Emotion Recognition**:
   * Detect face using Haar cascades.
   * Preprocess region of interest and pass through CNN.
   * Output predicted emotion with annotation on the image.
3. **Audio Emotion Recognition**:
   * Convert audio/video to WAV using MoviePy.
   * Use Whisper to transcribe and detect spoken language.
   * Extract MFCC and pitch features using Librosa.
   * Classify emotion using a trained SVM classifier.
4. **Result Display**:
   * Facial emotion label.
   * Vocal emotion label.
   * Transcript of the speech.
   * Detected language.

**Architecture Overview**

**System Components:**

* **Whisper (ASR)**: Transcribes speech and detects language from uploaded audio/video.
* **Librosa**: Extracts MFCC and pitch for emotion classification.
* **SVM Classifier**: Predicts emotion based on audio features.
* **CNN (Keras)**: Recognizes facial emotion from preprocessed grayscale images.
* **OpenCV**: Detects face regions and annotates emotion.
* **Google Colab Interface**: For interactive file upload and results display.

This modular architecture allows for scalability, where additional components like text-based emotion recognition or real-time video analysis can be plugged in later.

**Future Enhancements**

Several enhancements can be added to improve accuracy, usability, and flexibility:

* **Textual Emotion Analysis**: Use transformer-based models on transcribed text for deeper emotion inference.
* **Multimodal Fusion**: Combine visual, audio, and text features into a unified deep learning model using attention mechanisms.
* **Dataset Integration**: Replace randomly generated audio training data with real emotion datasets like RAVDESS or EmoDB.
* **Real-Time Processing**: Add webcam/video stream processing to detect temporal emotion changes.
* **Mobile/Edge Support**: Deploy models using TensorFlow Lite for low-power devices like smartphones or Raspberry Pi.
* **Cross-Cultural Emotion Support**: Integrate culturally diverse datasets and customize emotion labeling across demographics.

**Limitations of the Current System**

Despite its advantages, the proposed system has some limitations:

* **Audio Classifier Training**: The current SVM is trained on synthetic data and lacks generalization.
* **No True Fusion**: The facial and audio models run in parallel but do not interact or resolve conflicts.
* **Environment Sensitive**: Facial model may fail in dim light; audio may fail with strong background noise.
* **Emotion Granularity**: Detects only a fixed set of emotion classes (7 facial, 4 vocal).
* **Static Input**: Does not support continuous or video frame-by-frame emotion detection.

**Conclusion**

This project demonstrates the development of a **multimodal emotion recognition system** capable of analyzing both facial and vocal cues from user inputs. By combining speech recognition (via Whisper), feature extraction (via Librosa), facial emotion detection (via CNN), and audio emotion classification (via SVM), the system presents a powerful prototype for emotion-aware applications.

While the model currently runs on static inputs and basic classifiers, its **modular and scalable design** paves the way for a more robust, dynamic, and context-aware emotion recognition system. Future iterations can integrate deep fusion models, support real-time inputs, and operate across languages and cultures, ultimately bringing machines closer to empathetic human interaction.