School of Electronic Engineering and Computer Science
Queen Mary University of London

MSc PROJECT DEFINITION
2024-25

Project Title: Multimodal Emotion Recognition from video, audio and text

Supervisor: Dr. Juntao Yu

Student name: Sarvagya Rastogi

Student e-mail: ec24781@qmul.ac.uk

PROJECT AIMS:
This project aims to develop a Multimodal Emotion Recognition system that integrates facial expressions, vocal tones, and text sentiment to improve emotional intelligence in AI systems. The system will leverage deep learning, OpenCV, Librosa, and Hugging Face Transformers to process multimodal data efficiently. The primary goal is to enhance emotion detection accuracy by combining visual, auditory, and textual cues.
The model will be trained and evaluated using CMU-MOSEI and CREMA-D datasets, ensuring robustness across diverse emotional expressions. The system has potential applications in customer service, virtual therapy, and interactive gaming.

PROJECT OBJECTIVES:
- To investigate existing AI tools and techniques used in multimodal emotion recognition.
- To collect and preprocess datasets (CMU-MOSEI and CREMA-D) for emotion analysis.
- To develop machine learning models to integrate and analyse video, audio, and text modalities.
- To extract and represent multimodal features using OpenCV (facial expressions), Librosa (speech), and Hugging Face Transformers (text sentiment).
- To experiment with multimodal fusion techniques, including concatenation, attention-based fusion, and cross-modal transformers.
- To assess model performance using accuracy, F1-score, confusion matrix, and AUC-ROC metrics.
- To optimise model for real-time inference, ensuring efficient emotion recognition.
- To develop a prototype (if feasible) to demonstrate real-world application of the emotion recognition system.

State how your project will be aligned with the learning outcomes of your programme of study.

- Applies deep learning techniques to process multimodal data, aligning with AI model development objectives.
- Integrates Natural Language Processing (NLP), Computer Vision, and Speech Processing, demonstrating interdisciplinary AI applications.
- Uses state-of-the-art machine learning frameworks (Transformers, CNNs, LSTMs) to enhance multimodal emotion recognition.
- Emphasises dataset handling, feature engineering, and model optimisation, reflecting key AI research skills.
- Develops a real-world application, supporting AI deployment and real-time inference strategies.

METHODOLOGY:

- **Data Collection:** Utilise publicly available benchmark datasets such as CMU-MOSEI and CREMA-D for emotion analysis, ensuring a diverse range of emotions across multiple modalities.
- **Model Selection:** Leverage state-of-the-art multimodal deep learning architectures, including transformers, CNNs, and LSTMs, to effectively process video, audio, and text inputs.
- **Preprocessing:** Format and clean dataset inputs by extracting facial expression features (OpenCV), speech emotion features (Librosa), and text sentiment features (Hugging Face Transformers) for optimal model performance.
- **Model Training and Fine-Tuning:** Compare different multimodal models and fusion techniques (concatenation, attention-based fusion, cross-modal transformers) to identify the most effective approach for emotion recognition.
- **System Integration:** Develop a pipeline that seamlessly integrates multiple modalities into a single, unified model for real-time emotion detection.
- **Testing and Validation:** Evaluate system performance using standard metrics such as accuracy, F1-score, confusion matrix, and AUC-ROC, ensuring robustness and reliability.
- **Iterative Improvement:** Implement a feedback-driven approach to continuously fine-tune the model, optimising hyper-parameters and refining feature extraction techniques for enhanced accuracy.

PROJECT MILESTONES

- Conduct a literature review on existing multimodal emotion recognition techniques, datasets, and models.
- Collect and preprocess publicly available datasets (CMU-MOSEI and CREMA-D) for multimodal learning.
- Extract and analyse features from video (facial expressions using OpenCV), audio (speech signals using Librosa), and text (sentiment using Hugging Face Transformers).

- Implement and compare different deep learning architectures (e.g., CNN, LSTM, Transformer-based models) for multimodal fusion.
- Experiment with various fusion techniques, such as concatenation, attention-based fusion, and cross-modal transformers.
- Evaluate the system performance using benchmark metrics such as accuracy, F1-score, confusion matrix, and AUC-ROC.
- Optimise the model for real-time inference, ensuring computational efficiency and usability.
- Develop an interactive prototype (if feasible) to demonstrate the practical application of multimodal emotion recognition.
- Conduct testing and validation using real-world emotion datasets and user feedback.
- Finalise system deployment and prepare documentation for submission, including technical reports and model evaluation results.

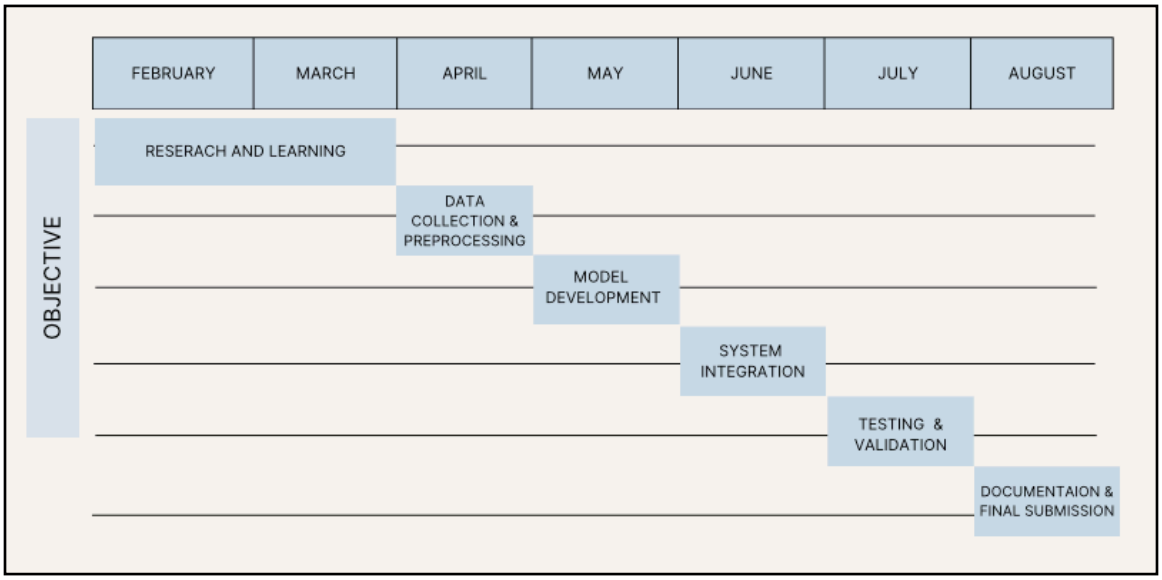
REQUIRED KNOWLEDGE/ SKILLS/TOOLS/RESOURCES:

- Familiarity with multimodal deep learning architectures, including CNNs, LSTMs, and Transformers.
- Proficiency in Python and machine learning frameworks, such as PyTorch and TensorFlow.
- Experience with computer vision (OpenCV), speech processing (Librosa), and NLP (Hugging Face Transformers).
- Understanding of dataset handling and preprocessing techniques for video, audio, and text modalities.
- Access to computational resources, such as cloud-based GPUs (Google Colab, Kaggle, or university computing clusters) for model training and evaluation.
- Knowledge of multimodal fusion techniques, including concatenation, attention-based fusion, and cross-modal transformers.
- Experience in real-time model optimisation to enhance inference speed and efficiency.
- Project management tools, such as JIRA or Trello, for organising tasks and tracking progress.

BACKGROUND MATERIAL:

- Joe Dhanith P R, Shravan Venkatraman, Vigya Sharma, Santhosh Malarvannan, Modigari Narendra."Multimodal Emotion Recognition using Audio-Video Transformer Fusion with Cross Attention." arXiv preprint arXiv:2407.18552 (2024).
- Gagah Dwiki Putra Aryono, Dede Ferawati, Sigit Auliana. "Advanced Deep Learning Models For Emotion Detection In Speech: Applying The RAVDESS Dataset." Jurasik (Jurnal Riset Sistem Informasi dan Teknik Informatika), vol. 9, no. 2, 2024.

TIMEPLAN:



Objective	Activities
Research and Learning	• Literature Review • Choose frameworks • Define Multimodal Dataset Requirements
Data Collection and Preprocessing	• Dataset Sourcing • Preprocess video, audio and text separately • Feature Extraction
Model Development	• Train individual modality models • Multimodal Model Training • Model Fine-tuning
System Integration	• Create and API or prototype application (if possible) • Multimodal Feature Fusion Implementation • Performance Optimisation
Testing and Validation	• Evaluate Model Performance • User Testing • Iterate and Improve
Documentation and Final Submission	• Write Final Report • Prepare Presentation • Submit Code and Documentation