Automatic Multi-Modal Emotion Recognition for AI Interaction

Detailed Project Plan

Student: Li Shiyi

Department of Computer Science University of Hong Kong
Final Year Project

Supervisor: Wu Chuan

2024/09

Table of Content **Page**

1. Project Background

With the increasing integration of smart devices into daily life—from smart homes to intelligent vehicles and healthcare systems—human-computer interaction (HCI) has become one of the most important research directions in modern technology. One of the key challenges for HCI is to create more natural and intuitive interactions in these diverse environments, and emotion recognition (ER) plays a crucial role in achieving this goal. Multi-modal emotion recognition, which combines facial expressions, voice, and other cues, has proven to be superior to single-modality approaches in enhancing interaction quality [1]. By utilizing complementary data from multiple sources, multi-modal systems offer higher accuracy and robustness, especially in complex, real-world scenarios where noise or ambiguity can degrade single-modality performance.

However, many smart devices—especially mobile and embedded systems—face limitations in computational power, making it challenging to deploy large multi-modal models with billions of parameters for real-time interactions. These constraints highlight the demand for developing smaller, more efficient multi-modal models that can run effectively on these devices while still maintaining a reasonable level of accuracy for emotion recognition. Adjusting the balance between model complexity and efficiency is essential for enabling real-time emotion detection across a wide range of smart devices.

2. Project Objectives

a. Design the structure of the lightweight multi-modal model, making it capable of performing automatic ER tasks

b. Employ various multi-modal ER datasets for model training and validation

c. Build a webpage that utilizes audio and visual data collected from the smart device to return real-time emotion labels for specific scenarios (for illustrative purposes, as in practical applications, the model is intended to function as a functional module).

d. A server for the web page which stores the trained model and reply to requests

3. Methodology

a. Input Data Types

Inspired by the approaches in [2] and [3], which incorporate both the entire image and cropped facial regions as input to the model, three types of input data will be used: the whole image, the cropped facial region, and corresponding audio data. This enables the integration of facial expressions and contextual information.

b. Feature extraction models & LSTM Method

It has been proven in [4] that when training end-to-end, the shallow CNN + RNN architecture is highly effective, surpassing the performance of self-attention and cross-attention methods with fewer parameters. Therefore, considering the constraints on our model size, we'll employ shallow CNN models for visual feature extraction, implement end-to-end training, and take the LSTM method for temporal information processing. According to [4], MobileFaceNet is utilized for visual feature extraction, while a 1D CNN proposed by [5] is used for audio extraction. Further experimentation is required to determine whether alternative models should also be considered.

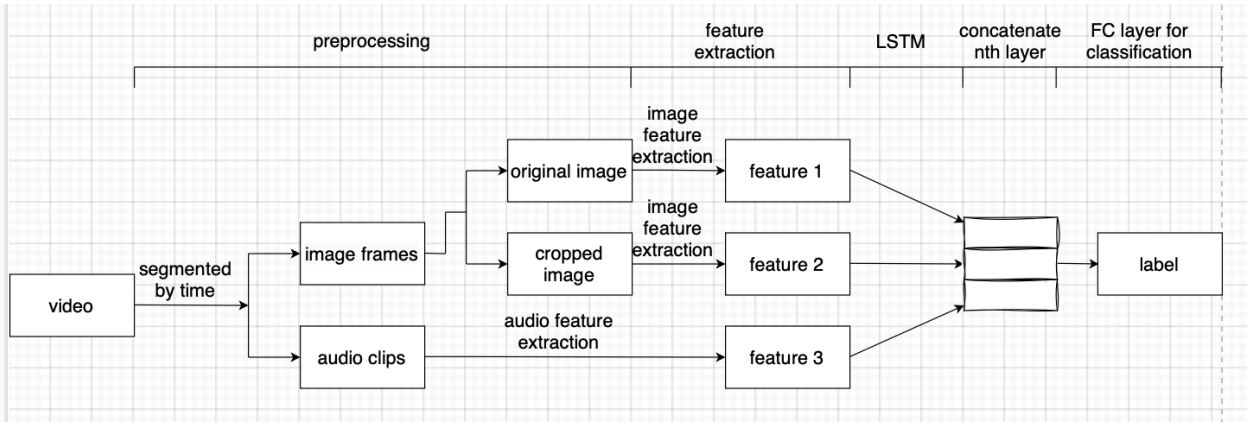

Figure 1. Overview of the model

c. Website Front End & Server

The front end will display the real-time video captured by the device, as well as the detected emotion label. Meanwhile, on the server side, upon receiving the video, the server will apply the model for object detection, process the data, and return the label back to the device.

4. Schedules and Milestones

Tasks	Due Date
Explore related papers and derive a draft model framework	10/01
Determine model structure, dataset, and loss function to utilize	10/15
Code writing & Model Training	12/31
Prepare interim report and first presentation	01/31
Implement front-end functions and further exploration	03/31
Prepare final report, presentation, and project exhibition	04/30

5. References

[1] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

[2] Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. (2017). Emotion recognition in context. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1960–1968). IEEE. https://doi.org/10.1109/CVPR.2017.212

[3] Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., & Onoe, N. (2022). M2FNet: Multi-modal fusion network for emotion recognition in conversation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4651–4660. https://doi.org/10.48550/arXiv.2206.02187

[4] Vincent, K., Kumar, T. M., Mesaros, A., Valstar, M., & Schuller, B. W. (2022). Time-continuous audiovisual fusion with recurrence vs. attention for in-the-wild affect recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2022.XXXX

[5] Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control, 47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035