Automatic Multi-Modal Emotion Recognition for AI Interaction

Interim Report

Student: Li Shiyi 3035844405

Department of Computer Science University of Hong Kong

Supervisor: Wu Chuan

2024/01/26

Abstract

Emotion recognition is a critical part of human-computer interaction, with applications across various fields such as marketing, security, and mental healthcare. However, existing ER models often require high computational resources due to their complex structures and large parameter sets, limiting their practicality for real-time applications on smart devices. To address these issues, this project develops a lightweight, multi-modal model for efficient real-time emotion recognition. The model utilizes data inputs including whole images, cropped facial images, and audio data. End-to-end training is conducted using shallow Convolutional Neural Networks and Long Short-Term Memory networks for spatial and temporal feature extraction, with the simple concatenation-based fusion approach further reducing complexity. Initial experiments on the MELD dataset, enhanced with data augmentation strategies, have demonstrated promising results. Future work will focus on several key areas like the integration of additional datasets and the evaluation of advanced feature extraction and fusion methods.

Table of Contents

List of Figures

1. INTRODUCTION

As a crucial part of human-computer interaction, emotion recognition (ER) is drawing increasing attention because of its vital importance in several application fields, such as marketing, security, and mental healthcare [1]. Especially, in the development of interactive chatbots, the integration of human emotion status into AI models can help generate more natural and intuitive responses, thereby enhancing user experience and improving response efficiency.

Although some models have shown promising accuracy in emotion recognition tasks, most of them feature complex structures and large parameter sets, which hinder their ability to produce results in real time. For example, the M2FNet (Multi-modal Fusion Network) model achieves high accuracy by employing a hierarchical framework that processes features at both the utterance and dialog levels. It first uses the RoBERTa-based text feature extractor and custom visual and audio feature extractors to obtain utterance-level information, then combines these features through a multi-head attention fusion mechanism [2]. All these feature extractors and the attention mechanism significantly contribute to the model's high parameter count, making it computationally intensive and challenging for daily usage.

Significant demand exists for developing smaller models that can run effectively on smart devices (e.g., mobile phones or embedded systems) while 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.

This paper makes two contributions. First, it designs a lightweight, multi-modal model capable of efficient automatic emotion recognition and easy integration into existing systems. The model will be trained and validated using up-to-date datasets to ensure robustness and accuracy. Second, this paper develops a web interface that provides real-time emotion recognition for demonstration purposes. The interface leverages audio and visual data collected from the smart

devices to generate emotion labels. A server will also be established to host this web page, ensuring efficient handling and processing of requests.

The remainder of this paper proceeds as follows. Sections 2.1 to 2.3 outline the three main components of our emotion recognition model. Section 2.1 introduces the preprocessing procedures implemented. Section 2.2 discusses the feature extraction methods used to efficiently capture temporal and spatial information. Section 2.3 describes the feature fusion approach to fuse embeddings derived from different modalities. The front-end and server-side design of the website is shown in Section 2.4. Finally, Section 3 discusses the progress being made so far, and Section 4 presents the project schedules and milestones.

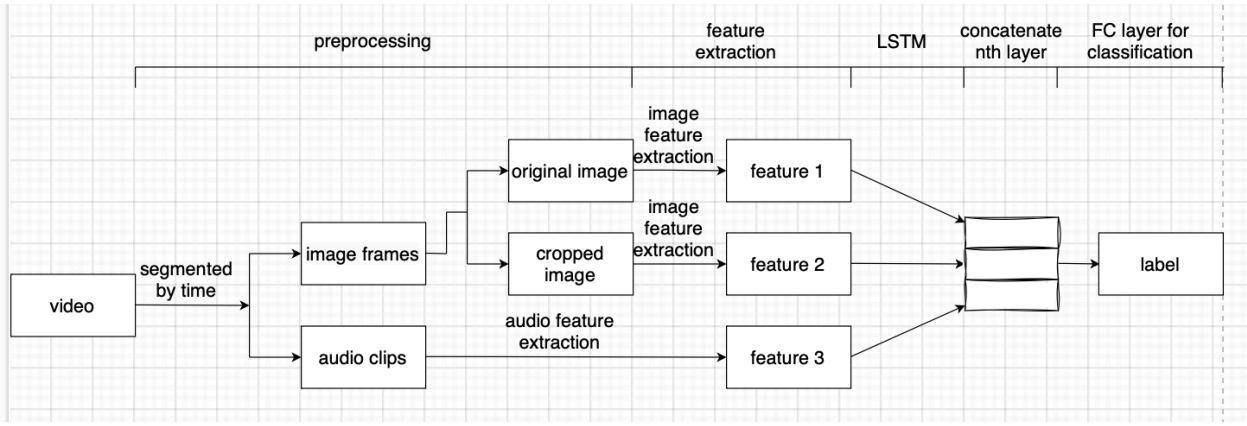

Figure 1. Model Structure

2. METHODOLOGY

Multi-modal emotion recognition models, compared to uni-modal models, utilize multiple data sources rather than a single source to detect human emotions. They have been proven to outperform uni-modal classifiers in detection accuracy[4]. 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. Therefore, we will employ a multi-modal model for emotion recognition to leverage these advantages.

As shown in Figure 1, the model structure can be divided into three main sections: data preprocessing, feature extraction, and feature fusion across modalities. The process is facilitated using the End2You toolkit, which supports the integration of these components efficiently.

2.1 Data Preprocessing

Three types of input data will be used: the whole image, the cropped facial region, and the corresponding audio data. The original video will first be segmented into audio clips and image frames based on time intervals. The image frames will then be cropped to extract the facial region using face detection techniques. This method is inspired by the approaches in [2] and [3], where both the entire image and cropped facial regions are incorporated as input to the model. The integration of facial expressions and contextual information is likely to enable higher accuracy. To mitigate data scarcity issues, data augmentation techniques such as pitch shifting for audio data and random cropping for visual frames were employed, thus increasing data variability and improving model robustness.

2.2 Feature Extraction

It has been proven in [5] that when training end-to-end, the shallow CNN + RNN architecture is highly effective. Under the same situation, this approach surpasses the performance of self-attention and cross-attention methods while using fewer parameters. Therefore, considering the constraints on our model size, we will implement end-to-end training. Shallow CNN models will be employed for spatial feature extraction. Additionally, the LSTM method will be used for processing temporal information, as LSTM retains temporal information more effectively than standard RNNs. According to [5], MobileFaceNet was utilized for visual feature extraction, while a 1D CNN proposed by [6] was used for audio extraction. Further experimentation is required to determine whether alternative models should also be considered.

2.3 Feature Fusion

Considering the constraints on model size, we did not adopt complex methods like cross-attentional audio-visual fusion [7]. Instead, we opted for a straightforward approach. Feature embeddings derived from different modalities are simply concatenated and processed through several fully connected layers to generate the final label.

2.4 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.

2.5 End2You Toolkit as the Model Framework

The End2You toolkit[8] was adopted to facilitate the development of our model, offering an end-to-end framework that supports both regression and classification outputs. It offers flexible customization options, allowing for the adaptation to meet specific project requirements. To implement our model, modules such as CustomProvider, CustomModel, CustomLoss, and CustomMetric were developed for data preprocessing, training, and evaluation respectively. This approach simplified the setup process, eliminating the need for manual construction of the underlying architecture. Additionally, its built-in conda environment ensures compatibility across different systems.

3. PROGRESS TO DATE AND FUTURE PLAN

The preliminary results demonstrate successful adaptation of the MELD dataset to the End2You toolkit and the construction of the intended model. Despite challenges such as limited storage capacity for large .hdf5 files and the integration of additional contextual information, initial experiments show promising results in multi-modal emotion recognition. Future plan will focus on integrating additional datasets, refining feature extraction and fusion methods, and replacing the End2You toolkit with a customized implementation to further optimize performance.

3.1 Preliminary Results and Discussion

MELD dataset[9] is a widely recognized benchmark dataset for emotion recognition, featuring multi-modal conversational data annotated with emotion labels. In order to adapt the raw MELD data to the End2You toolkit, several preprocessing steps, including data segmentation and label handling were taken to meet the toolkit's input requirements.

The audiovisual modality of the End2You toolkit was initially designed to handle only two input channels: visual facial data and audio data. However, the inclusion of additional contextual background information, which is crucial for our model's performance, was not supported by the toolkit out-of-the-box. To address this limitation, the toolkit's source code was carefully modified, and these changes have been documented and uploaded to GitHub to ensure traceability.

The storage space limitation emerged due to the large size of the .hdf5 files. For instance, the size of only 43 samples(even after compression using the LZF method) reached 12GB. A rough estimate indicates that with a total of 9,000 training samples, over 1TB of storage would be required, far exceeding the available capacity of the GPU farm. As a result, efforts are currently focused on splitting the dataset into smaller, more manageable batches for model training. Although the model design has been completed, overall training and validation are still in progress. Initial experiments conducted on a small subset of the dataset have yielded promising results, demonstrating the model's potential to accurately recognize emotional expressions across modalities.

3.2. Future Plan

To enhance the model's performance, the CREMA-D dataset[10] will be integrated into our model in future phases. This dataset offers great variability in terms of race, ethnicity, and age, which is expected to improve the model's generalization across diverse demographic groups. Meanwhile, the potential for more advanced feature extraction and fusion methods to improve model accuracy will be assessed. This evaluation will carefully balance the trade-off between

potential performance improvements and the added complexity, guiding our decisions on further model refinement.

A possible limitation of the End2You toolkit is that the built-in implementations may not represent the most advanced or optimal solutions available, which could hinder performance optimization. To address this issue, the toolkit will be used only during the initial stages for debugging and model structure refinement. Once optimized, we will replace the toolkit with our own implementation to achieve better performance.

4. CONCLUSIONS

This project focuses on the development of a lightweight, multi-modal model for efficient real-time emotion recognition, addressing the challenges of existing models that require high computational resources. By leveraging a CNN+LSTM architecture for feature extraction and employing a simple concatenation-based fusion approach, the model is expected to achieve reduced computational complexity while delivering promising performance.

The adaptation of the MELD dataset to the End2You toolkit and the model construction were completed successfully, overcoming challenges such as limited storage for large .hdf5 files and the integration of additional contextual information. Initial experiments on a small subset of the dataset showed promising results, highlighting the model's potential in multi-modal emotion recognition.

Despite these promising outcomes, several limitations can be identified. The model was only evaluated on a small subset of the dataset, which may lead to an overestimation of its performance. The End2You toolkit’s underlying code could constrain further optimization and limit the model’s overall performance. Additionally, the simplified feature extraction and fusion methods may not fully capture the rich interactions and the detailed information hidden in the multi-modal data.

To address these limitations, an important avenue for future work will be the integration of additional datasets, such as the CREMA-D dataset, to enhance robustness and generalization. Further exploration of advanced feature extraction and fusion methods will also be conducted to guide the model structure refinement, ensuring the balance between accuracy and computational efficiency. Once the model structure is finalized, the End2You toolkit will be replaced with our own implementation to achieve better performance.

5. REFERENCES

[1] K. Kamble and J. Sengupta, "A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals," *Multimedia Tools and Applications*, Feb. 2023, doi: https://doi.org/10.1007/s11042-023-14489-9.

[2] V. Chudasama, P. Kar, Ashish Gudmalwar, N. J. Shah, Pankaj Wasnik, and Naoyuki Onoe, "M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2022, doi: https://doi.org/10.1109/cvprw56347.2022.00511.

[3] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion Recognition in Context," *IEEE Xplore*, Jul. 01, 2017. https://ieeexplore.ieee.org/document/8099695

[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, no. 37, pp. 98–125, Sep. 2017, doi: https://doi.org/10.1016/j.inffus.2017.02.003.

[5] V. Karas, Mani Kumar Tellamekala, A. Mallol-Ragolta, M. Valstar, and B. Schuller, "Time-Continuous Audiovisual Fusion with Recurrence vs Attention for In-The-Wild Affect Recognition," Jun. 2022, doi: https://doi.org/10.1109/cvprw56347.2022.00266.

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

[7] R, Gnana Praveen, E. Granger, and P. Cardinal, “Cross Attentional Audio-Visual Fusion for Dimensional Emotion Recognition,” *arXiv (Cornell University)*, Nov. 2021, doi: https://doi.org/10.48550/arxiv.2111.05222.

[8]P. Tzirakis, S. Zafeiriou, and B. W. Schuller, “End2You -- The Imperial Toolkit for Multimodal Profiling by End-to-End Learning,” arXiv (Cornell University), Jan. 2018, doi: https://doi.org/10.48550/arxiv.1802.01115.

[9]S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations,” *arXiv:1810.02508 [cs]*, Jun. 2019, Available: https://arxiv.org/abs/1810.02508

[10]H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset,” *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, Oct. 2014, doi: https://doi.org/10.1109/TAFFC.2014.2336244.

[11] OpenAI, ChatGPT, Oct. 2023. [Online]. Available: https://chat.openai.com/chat