EmoSense, a mobile therapy application

Lucia Fang
Dietrich College of Humanities and Social Sciences
Carnegie Mellon University
yufang@andrew.cmu.edu

1 Introduction

Depression impacts 9.2% of the population in the United States in 2020 [1]. Therapy sessions are useful, however, such uncomfortable/high arousal settings often times lead to false memory recall [2]. In addition, researchers have also found that the recall accuracy scales with emotional valence, leading to a emotional state bias [3]. To allow an accurate record of how people feel whenever and wherever, I propose a mobile solution where I implement state-of-the-art machine learning technologies to record and automate emotional states analyses. The summary statistics can then be retrieved by mental health professionals to keep track and provide more targeted interventions swiftly. Recent computer vision advances have allowed human-level performance in face recognition [4]. In addition, openly available facial expression data sets have facilitated development of emotional recognition algorithms [5]. Beyond vision, researchers have demonstrated that audio can support emotional recognition [6]. While speech-to-text can sometimes directly identify the user's emotional state, recent development of large language modeling with GPT can decode implicit emotion. Taken together, I developed a multi-modal emotional recognition application that can allow users to record, identify, and eventually present their emotional portfolio to better receive help from professionals.

2 Implementation Details

Once the user uploads a video, the video is being split into static frames versus sound file. The video frames will be analyzed prior to the audio file simply due to the openly available facial expression dataset label constrain.

2.1 Facial expression analysis

Every half a second, each frame will be preprocessed by the VGG face recognition model within DeepFace `https://github.com/serengil/deepface` [4, 7, 8, 9]. VGG face model identifies if one or more faces are in the frame. Upon identifying face(s) in a frame, there may be differences in face size. To standardize the faces across all various client inputs, dynamic cropping and rescaling based on face location is performed by OpenCV package `https://opencv.org/` within DeepFace. One of the most famous facial expression datasets is the FEC2013 dataset `https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data`. The model I used is trained on 28K facial images consisting of 7 unique expressions (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral). I then run the cropped faces into the model for an emotion output. The output is probability values for each of the 7 emotions.

Final Project Report for 05-499. Do not distribute.

2.2 Speech expression analysis

Upon converting the audio component of the video into a .wav file, I fed this audio file to a speech-to-text algorithm `https://github.com/Uberi/speech_recognition`. Currently, I have enabled the user to choose from 4 different languages (English, Chinese, Spanish and French). Upon transforming speech to text, OpenAI chatGPT will then determine the client's emotion. To do that, I found a command line wrapper for OpenAI's chatGPT `https://github.com/acheong08/ChatGPT` and submitted the user's text to chatGPT. In order to combine the two modality, I combined the top three emotions decoded from the face and have the large language model decide based on implicit language cues.

3 Results

I developed a web application "EmoSense" that automates emotional analyses. The basic pipeline takes client's video and splits it into visual and audio components (Fig. 1a). With the help of DeepFace, the visual component will be processed independently of the audio file. Upon decoding facial expression using purely video frames, the audio file will be transcribed and submitted to OpenAI ChatGPT through a command line wrapper. The client's message will then be tagged with a the question "Of the emotions my face expressed, which am I most likely feeling?".

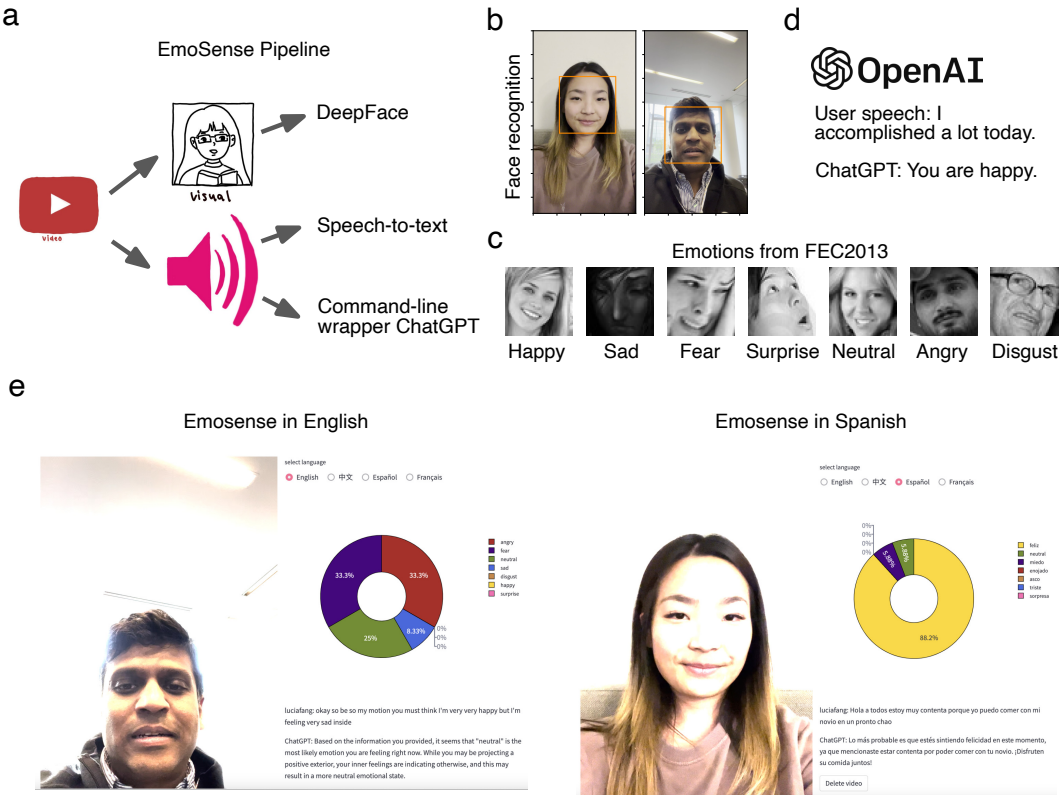

Figure 1: EmoSense pipeline and web application user interface (UI). a) A schematic of the EmoSense pipeline. b) Face recognition and dynamic cropping demonstrated using DeepFace. c) Single example image from each emotional state in the FEC2013 dataset. d) Example conversation with OpenAI ChatGPT. e) EmoSense UI demonstrated in English (left) and Spanish (right). Note that it can currently perform in two additional languages (Chinese and French)

4 Anticipated Ethical Issues

Since I do not have access to "VGG" facial dataset, I cannot provide any insights regarding potential biases in gender, race, or other demographic attributes. In addition, privacy can be a significant concern when dealing with personal data. In order to address the issue, I have implemented a login screen to store individuals recordings. However, despite taking necessary precautions, data breaches can still occur. The danger of having personalized video record of their emotional well-beings is the potential misuse of the information as exploiting people and cause more psychological harm [10].

References

[1] Renee D. Goodwin et al. "Trends in U.S. Depression Prevalence From 2015 to 2020: The Widening Treatment Gap". In: *American Journal of Preventive Medicine* 63.5 (Nov. 2022), p. 726. ISSN: 18732607. DOI: 10.1016/J.AMEPRE.2022.05.014. URL: /pmc/articles/PMC9483000/%20/pmc/articles/PMC9483000/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9483000/.

[2] Yves Corson and Nadège Verrier. "Emotions and false memories: valence or arousal?" In: *Psychological science* 18.3 (Mar. 2007), pp. 208–211. ISSN: 0956-7976. DOI: 10.1111/J.1467-9280.2007.01874.X. URL: https://pubmed.ncbi.nlm.nih.gov/17444912/.

[3] C. J. Brainerd et al. "How Does Negative Emotion Cause False Memories?" In: *https://doi.org/10.1111/j.1467-9280.2008.02177.x* 19.9 (Sept. 2008), pp. 919–925. ISSN: 09567976. DOI: 10.1111/J.1467-9280.2008.02177.X. URL: https://journals.sagepub.com/doi/abs/10.1111/j.1467-9280.2008.02177.x.

[4] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition". In: ().

[5] Yousif Khaireddin and Zhuofa Chen. "Facial Emotion Recognition: State of the Art Performance on FER2013". In: ().

[6] Sung Woo Byun and Seok Pil Lee. "Human emotion recognition based on the weighted integration method using image sequences and acoustic features". In: *Multimedia Tools and Applications* 80.28-29 (Nov. 2021), pp. 35871–35885. ISSN: 15737721. DOI: 10.1007/S11042-020-09842-1/TABLES/1. URL: https://link.springer.com/article/10.1007/s11042-020-09842-1.

[7] Sefik Ilkin Serengil and Alper Ozpinar. "LightFace: A Hybrid Deep Face Recognition Framework". In: *Proceedings - 2020 Innovations in Intelligent Systems and Applications Conference, ASYU 2020* (Oct. 2020). DOI: 10.1109/ASYU50717.2020.9259802.

[8] Sefik Ilkin Serengil and Alper Ozpinar. "HyperExtended LightFace: A Facial Attribute Analysis Framework". In: *7th International Conference on Engineering and Emerging Technologies, ICEET 2021* (2021). DOI: 10.1109/ICEET53442.2021.9659697.

[9] Sefik Ilkin Serengil and Alper Ozpinar. "An Evaluation of SQL and NoSQL Databases for Facial Recognition Pipelines". In: (Feb. 2023). DOI: 10.33774/COE-2023-18RCN. URL: https://www.cambridge.org/engage/coe/article-details/63f3e5541d2d184063d4f569.

[10] Javier Hernandez et al. "Guidelines for Assessing and Minimizing Risks of Emotion Recognition Applications". In: *2021 9th International Conference on Affective Computing and Intelligent Interaction, ACII 2021* (2021). DOI: 10.1109/ACII52823.2021.9597452.