Non-Invasive Stress Monitoring From Video

***Abstract*—Identifying stress is crucial for maintaining a healthy lifestyle. Current stress-detection methods are relatively slow and subjective and often take place through invasive measurements via medical devices. We instead propose end-to-end, noninvasive detection of stress through video. We incorporate several modalities to perform holistic detection of a user's stress level. Our framework employs an emotion recognition model to detect expressions through facial recordings. Then, we evaluate a user's heart rate by amplifying changes in skin coloration through Eulerian Video Magnification. Finally, we analyze differentials in eyebrow and lip movements. We combine these three measurements to output a final stress score per unit of time. Our chosen emotion recognition model achieves an accuracy of 96.46%, and our remote heart rate detection module has a mean absolute error of 5.79 BPM. We provide a web-based application that allows for rapid, contactless stress detection through a webcam. We achieve over 84% accuracy, with 90% in detecting the presence of stress.**

1. Introduction

A recent study reported that over four out of five adult Americans experience medium to high-stress levels, with national stress levels rising. Traditionally, clinicians have measured stress in medical settings through user questionnaires or a life-events checklist, methods which are time-intensive and subjective, unfit for rapid stress detection.

We propose a framework to conduct rapid stress detection using video input from a webcam. Without expensive equipment, this architecture is usable in various situations, particularly on Zoom and other video conferencing platforms. First, we classify the subject's emotional state by extracting frames from video, then measure the HR using an algorithm that identifies changes in skin hue, and third calculate distances of specific facial landmarks to predict their overall stress level in real time.

2. Methods

Input: Accepting a video of any length as input, we downsample the frames per second (fps) to 30 for inputs with higher frequencies using the ffmpeg library.

Emotion Recognition: Our first module identifies the emotional expressions of the subject in the input video. We use the Haar-cascade feature selection technique to localize and capture the individual's face. It predicts seven expressions: neutral, anger, disgust, fear, happiness, sadness, and surprise. After experimenting with 6 models, we use a VGG-19 network with batch normalization.

Heart Rate Detection: Our HR detection module uses Eulerian Video Magnification, a computational approach for visualizing small perturbations in color and motion in video, to estimate the changes in heart rate. Oftentimes, these small fluctuations are imperceptible to the naked eye. We developed and tested all steps in the EVM method.

Facial Feature Analysis: In our third module, we measure the variation of the eyebrows and lips from their neutral location, categorizing the results into a movement-based stress score. We decompose each input video into a series of frames and resize each frame to 500×500 pixels. From each frame, we detect certain landmarks, generate the convex hull, and calculate the Euclidean distance between the left and right eyebrows and the top and bottom lips.

Fusion: After testing + linear regression analysis, we determine high/moderate/low thresholds for the three values. Each score is classified as the appropriate stress level.

3. Experiments

We use three publicly available datasets:

- The PURE Pulse Rate Detection Dataset.
- The UBFC-Phys Dataset.
- Hand-Labeled Stress Dataset.

Emotion Recognition: We divide the CK and CK+ datasets into training and testing splits of 70% and 30% and augment the data. We train the CNN model for 100 epochs with stochastic gradient descent (SGD) and a categorical crossentropy loss function. Similarly, After training on Imagement challenge, We fine-tune the EfficientNet. Our fine-tuning uses the Adam optimizer and categorical crossentropy loss function.

Heart Rate Detection: We use the PURE Pulse Rate Dataset, using 60 ten-second videos from this dataset. We evaluate our results using the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). We also generate a Bland-Altman plot, which shows the consistency between two signals.

4. Results

Stress Detection. We evaluate our end-to-end framework on our hand-labeled dataset, and it correctly identifies the level of stress nearly 84% of the time, achieving a 90% accuracy when combining moderate to high-stress categories. The most incorrectly predicted combination is guessing moderate stress when the individual experiences low stress.

Emotion Recognition. The deep learning approaches to emotion recognition outperformed the others on the CK and CK+ datasets. The fine-tuned EfficientNet model achieves the highest accuracy at 98.48%, followed closely by the VGG-19 architecture at 96.46%. We find that the EfficientNet classifier overfits the CK/CK+ dataset and and struggled with other video inputs. For our webapp, we use the VGG-19 model to predict a user's emotional state.

Heart Rate Detection. After testing the EVM framework on the PURE dataset, the framework achieved near state-of-the-art results. The Mean Absolute Error determined was 5.79 BPM, with some of the best predictions ranging between 0 BPM and 2.8 BPM. The Root Mean Squared Error (RMSE) is 10.13. We found that this performance ranks better compared to existing work.

Ablation Studies. We conduct a series of additional studies where we remove a single component and re-did the experiment. We found emotions and HR to be the most important indicators.