NON-INVASIVE STRESS MONITORING FROM VIDEO

Akshata Tiwari[1] *Brian Matejek*[2] *Daniel Haehn*[3]

[1] Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[2]Computer Science Laboratory, SRI International, Arlington, Virginia, USA
[3]Department of Computer Science, University of Massachusetts Boston, Boston, Massachusetts, USA

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, non-invasive 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 on a dataset of video clips with individuals labeled as undergoing low, moderate, or high-stress levels. All code and data are openly available.

***Index Terms*—** stress detection, emotion, heartrate, facial features

1. INTRODUCTION

A recent study reported that over four out of five adult Americans experience medium to high-stress levels [1, 2]. Experts have identified correlations between elevated stress levels, decreased job performance, reduced quality of life, and cognitive decline [3]. In 2019 the Occupational Safety and Health Administration (OSHA), the American federal agency tasked with protecting the health and safety of workers, expanded its role to include mental health considerations [4]. A recent survey found that heavy workloads, deadlines, high expectations from bosses, and the pressure of maintaining a healthy work-life balance are some leading causes of stress in workers [5].

Students have also reported an increase in stress levels [8]. In the past year alone, over 88% of students reported moderate to severe stress levels [9]. Traditionally, clinicians have measured stress in dedicated medical settings through retrospective user questionnaires or a life-events checklist that is subjectively assessed [10]. However, these current methods are time-intensive and subjective, rendering them unfit for rapid stress detection in many environments.

We propose a framework to conduct rapid stress detection using video input from a webcam (Figure 1). Without expensive equipment, this architecture is usable in various situations, particularly on Zoom and other video conferencing platforms. Our method consists of three modules that identify common characteristics of stressed individuals. First, we classify the subject's emotional state by extracting frames from video and labeling the expressions. Second, we measure the user's heart rate with an algorithm that amplifies the slight changes in skin hue as blood pumps through the region of interest. Third, we calculate the distances of specific facial landmarks, such as the eyebrows and lips, in successive frames. We integrate these three measurements and devise mapping scores to predict the candidate's overall stress level.

Many researchers have focused on developing automatic methods to identify stress. Some prior work focuses on unimodal stress identification by using wearable sensors. Can *et al.* use contact-based devices to measure skin conductance to classify individual stress levels [11]. Increasingly, researchers use machine learning models and computer vision techniques to perform stress detection. Almeida *et al.* also use emotion-recognition software to identify individuals undergoing stress [12]. Other research considers alternative intrinsic features, such as heart rate, for automatic stress detection. However, these unimodal approaches can yield inconsistent or false results. In other fields, researchers demonstrate higher efficiency using many modalities [13]. To our knowledge, no studies have yet implemented a multifaceted approach toward contactless stress identification like ours.

2. METHODS

2.1. 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 In this stage, we input

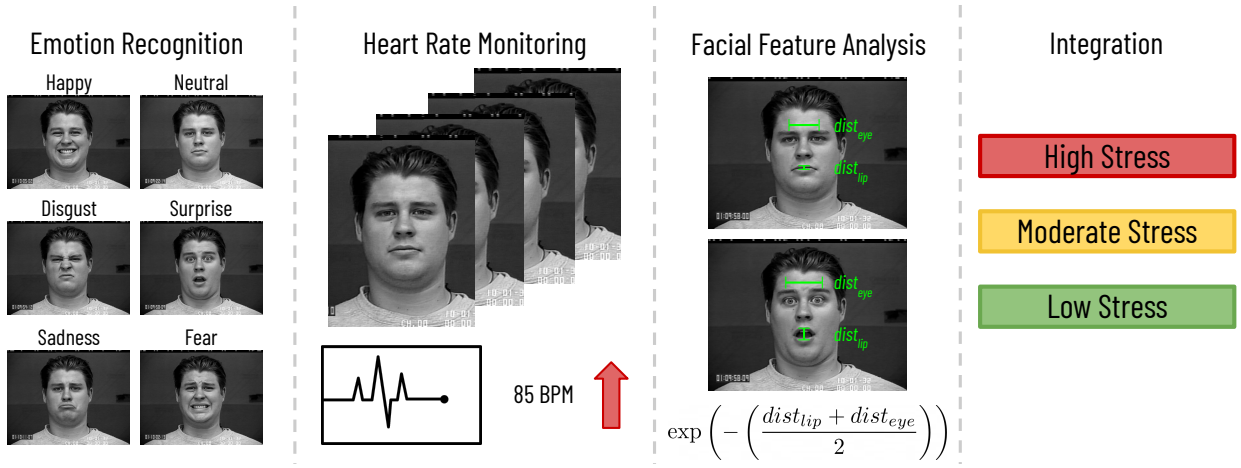

Fig. 1. Our framework combines three modules to analyze the signs of a stressed individual. First, we classify the subject's expressions into various emotional states. We use Eulerian Video Magnification to detect the individual's heart rate, and then measure the distances of certain facial landmarks per frame. We integrate these three module outputs to produce a final stress score. Subject images taken from the Cohn-Kanade (CK) [6] and Extended Cohn-Kanade (CK+) [7] datasets.

our video into each module at full resolution, since each subcomponent requires a different level of cropping and resizing.

2.2. 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 as our region of interest. We resize each frame around the user's face into 48×48 pixels and normalize each pixel between 0 and 1. Our emotion recognition software predicts seven expressions: neutral, anger, disgust, fear, happiness, sadness, and surprise. We use a VGG-19 network with batch normalization [14],[1] after evaluating against two other neural network and traditional ML models [15, 16, 17].

2.3. Heart Rate Detection

Our heart rate 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 [18]. Oftentimes, these small fluctuations are imperceptible to the naked eye.[2]

Preprocessing. We convert our video into a sequence of image frames, resizing each to 320×240 pixels. We identify facial regions via a "region-of-interest" selection procedure.

Spatial Filtering. We perform pyramid multiresolution decomposition of the input video sequence and extract features of interest. Then, we decompose the video sequence into multiple spatial frequency bands and magnify these bands differently to account for differences in the signal-to-noise ratios. By performing spatial processing, we increase the temporal signal-to-noise ratio by pooling pixels. These frames are decomposed with a Gaussian pyramid by downsampling layers. We chose three levels for our pyramid.

Time Domain Filtering. We performed time domain filtering, and temporal processing performed on each spatial band in the Gaussian pyramid. We obtain several frequency bands through a Fourier transform. We extract a frequency band of interest by passing pixels through a bandpass filter. The range chosen for selecting frequencies was between 1.0 to 3.0 Hz, the range of viable heart rates ($60 - 180$ beats per minute).

Amplification. We approximate the signal of each frequency by a Taylor series. After computing a Gaussian pyramid, the first level is set to 0. This ends up boosting the pulse signal that we have specified. The amplification factor, α, is set to 170. To reduce noise, we prefer narrow temporal bandpass filters, as opposed to the broader ones used for color amplification. The temporal bandpass filter pulls out motions in our range of amplification. Then, the filter is applied to $I(x, t)$ at every position. To convert this to beats per minute, we establish heart rate calculation variables that include beat-buffer-size and frequency. The beats per minute are finally then obtained by multiplying the obtained frequency by 60 Hz.

2.4. 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 select the eyebrows, lips, and mouth because of their strong correlation with stress. Eyebrow tensing and lip parting (significant in teeth-baring or an agape mouth) correlate with higher stress levels.

We decompose each input video into a series of frames and resize each frame to 500×500 pixels. From each frame,

[1] https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch
[2] Modified from https://github.com/giladoved/webcam-heart-rate-monitor

we detect certain landmarks: the right eyebrow, left eyebrow, and mouth. We then extract the shape of the aforementioned landmarks (i.e., eyebrows and lips) (Figure 1, center right). We generate the convex hull for each landmark to quantify their proper location and shape. We calculate the Euclidean distance between the left and right eyebrows and the top and bottom lips. These values are normalized by the size of the input (here, 500×500 pixels). To create a uniform point, we take the farthest distance between two points in the respective convex hulls. We then generate a stress score using the following equation, which includes a clamping value of 0.85:

$$\mathcal{S}_{FF} = \max\left(\exp\left(-\frac{dist_{lip} + dist_{eye}}{2}\right), 0.85\right) \quad (1)$$

2.5. Integration

Emotion Stress Mapping. We use a weighted correspondence between the recognized emotions and their stress levels. Gao *et al.* demonstrate that anger and disgust are among the highest stress indicators [19]. Likewise, Arslan *et al.* find that random outbursts, irritation, and aggression are direct indication of stress, which correspond to emotions of anger, fear, and disgust [20]. Therefore, we rank the emotions anger and surprise as the highest stress indicators. Directly below that, we evaluate fear, with sadness and disgust as medium stress indicators [21]. Neutral and happy facial expressions indicate lower to no amounts of stress.

$$\mathcal{S}_{ER} = \begin{cases} 0 & \text{if ER } \in \text{ happiness} \\ 0.6 & \text{if ER } \in \text{ neutral} \\ 0.7 & \text{if ER } \in \text{ disgust, sadness} \\ 0.8 & \text{if ER } \in \text{ fear} \\ 1.0 & \text{if ER } \in \text{ anger, surprise} \end{cases} \quad (2)$$

Heart Rate Stress Mapping. Like emotion recognition, we map heart rate values to a stress threshold. The human heart rate spikes up 38 BPM from resting rate in a stressed state [22]. We convert our heart rate values into stress scores: values below 80 correspond to low stress, values above 120 correspond to high stress, and moderate stress is in between.

$$\mathcal{S}_{HR} = \begin{cases} 0.4 & \text{if HRV} < 80 \\ 0.7 & \text{if } 80 \leq \text{ HRV } < 100 \\ 0.8 & \text{if } 100 \leq \text{ HRV } < 120 \\ 1.0 & \text{if HRV} \geq 120 \end{cases} \quad (3)$$

Facial Feature Analysis Stress Mapping. For calculating the stress scores from the facial feature analysis module, we simply take the value from Equation 1.

Combination. We add these three weights together to produce a final score:

$$\mathcal{S} = \begin{cases} \text{low} & \mathcal{S}_{ER} + \mathcal{S}_{HR} + \mathcal{S}_{FF} < 1.2 \\ \text{moderate} & 1.2 \leq \mathcal{S}_{ER} + \mathcal{S}_{HR} + \mathcal{S}_{FF} < 1.8 \\ \text{high} & \mathcal{S}_{ER} + \mathcal{S}_{HR} + \mathcal{S}_{FF} \geq 1.8 \end{cases} \quad (4)$$

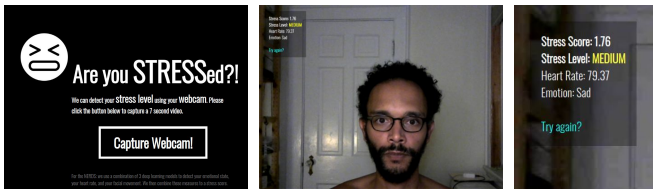

Fig. 2. Our web application tracks a user's stress level and reports back their estimated heart rate and emotional state. (Photo courtesy of David Degras/Bougdou).

2.6. Web Application

We present a web application that enables a user to evaluate their stress levels by allowing video capture from their webcam (Figure 2).[3] We use Flask for our web server and connect the HTML5 MediaStream Recorder API[4] to capture the user's webcam in the browser. For normalization of different webcam models, we process the videos using the `ffmpeg` library. Our server-side NVIDIA GeForce RTX 3080 GPU allows for fast inference throughput, and we additionally reduce the number of frames for emotion recognition and facial analysis by a factor of 6. The complete analysis of a 7-second video capture takes less than 9 seconds.

3. EXPERIMENTS

3.1. Datasets

We use three publicly available datasets for our experiments. The extended **Cohn-Kanade dataset (CK+)** dataset contains facial action units and images of facial expressions [6, 7]. The **PURE Pulse Rate Detection** dataset comprises of 60 video sequences involving 10 subjects [23], each recorded under six different environments and conditions with varying degrees of movement. The **UBFC-Phys** dataset is a multimodal dataset that includes videos of 56 participants [24].

We repurpose the UBFC-Phys dataset [24] to fit our stress labeling framework. We first divide each of their three-minute-long testing videos into 18 ten-second clips. We filter each clip to make sure that each clip only includes the individual's face and select 60 clips for a reserved testing set. To obtain ground truth values, three individuals labeled the selected clips as "high", "moderate", and "low" stress. To encourage reproducibility and further research in the field, our labeling is publicly available.[5]

3.2. Emotion Recognition

We divide the CK and CK+ datasets into training and testing splits of 70% and 30%, respectively. We augment the data by adding variations of contrast, positioning, and lighting. We train the CNN model for 100 epochs with stochastic gradient

[3] https://github.com/mpsych/stress
[4] https://w3c.github.io/mediacapture-record/#mediarecorder-api.
[5] https://github.com/mpsych/stress

Machine Learning Model	Accuracy (↑)
Linear SVM	84.29
2D CNN	85.50
Decision Tree	80.05
EfficientNet	**98.48**
VGG-19	96.46

Table 1. Model performance on the emotion recognition task.

descent (SGD) and a categorical crossentropy loss function. Similarly, after training on the ImageNet challenge, we fine-tune the EfficientNet [25]. Our fine-tuning uses the Adam optimizer and categorical crossentropy loss function.

3.3. Heart Rate Detection

We use the PURE Pulse Rate Dataset to evaluate our heart rate detection module. We consider 60 ten-second videos from this dataset, and evaluate our predictions to the ground truth provided by the authors. We evaluate our results using the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE).

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 (Table 4). 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 struggled with other video inputs. Therefore, for our web application, 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 [26].

Ablation Studies. We conduct a series of additional studies to identify the importance of each module to our overall framework. We remove one component in each study and evaluate the results on our hand-labeled dataset. Note that the threshold values from Equation 4 are adjusted to account for the missing component, as they are rescaled by the new maximum score. Table 2 shows our results. We see that each component contributes to the overall accuracies that we achieve.

Modules Included	Accuracy (↑)
Heart Rate + Facial Features	61.67%
Emotion Recognition + Facial Features	71.67%
Emotion Recognition + Heart Rate	80.00%

Table 2. Removing any component from our framework significantly reduces our accuracy in labeling videos as high, moderate, or low stress. The emotion recognition module in particular significantly contributes to our overall performance.

However, removing the emotion recognition module significantly reduces performance.

Limitations. Our software only detects *perceivable stress*, which manifests itself in detectable heart rate, and emotional, and facial feature changes. Since the software operates on visual means, it is not able to detect any psychological stress without outward manifestations.

5. CONCLUSIONS

In this study, we propose a method that takes into account multimodal data while performing real-time stress detection. Integrating emotion recognition, heart rate detection, and facial feature analysis, our system generates a holistic score that produces an evaluation of an individual's stress level. We package our three modules into an easy-to-use web application that enables a user to identify their stress levels in real time. Current state-of-art stress detection systems frequently rely on invasive measures from contact sensors, techniques that are limited in remote situations or during intensive tasks like driving. We are currently testing an embedded, low-cost device that runs the software on an NVIDIA Jetson Nano and connects to a miniature camera. Integration of this software with semi-autonomous vehicles could play a role in driver settings. A future iteration of our facial-feature analysis software could track the movement of the irises of the eyes and analyze high-frequency movement.

6. COMPLIANCE WITH ETHICAL STANDARDS

All datasets used in this study were publicly available or obtained via a signed agreement form. Our own dataset used videos from a publicly available dataset and was hand-labeled by select individuals. Our data is publicly available.

7. REFERENCES

[1] Anjana Bhattacharjee and Tatini Ghosh, "Covid-19 pandemic and stress: Coping with the new normal," *Journal of Prevention and Health Promotion*, vol. 3, no. 1, pp. 30–52, 2022.

[2] Michelle Lambright Black, "Americans' stress levels - and financial anxiety - on the rise," Aug 2022.

[3] Roqayeh Parsaei, Hamidreza Roohafza, Awat Feizi, Masoumeh Sadeghi, and Nizal Sarrafzadegan, "How dif-

ferent stressors affect quality of life: an application of multilevel latent class analysis on a large sample of industrial employees," *Risk Management and Healthcare Policy*, vol. 13, pp. 1261, 2020.

[4] Phillip Montgomery, "Osha role expands to mental health protections," *CBIA*.

[5] Kamaldeep Bhui, Sokratis Dinos, Magdalena Galant-Miecznikowska, Bertine de Jongh, and Stephen Stansfeld, "Perceptions of work stress causes and effective interventions in employees working in public, private and non-governmental organisations: a qualitative study," *BJPsych bulletin*, vol. 40, no. 6, pp. 318–325, 2016.

[6] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian, "Comprehensive database for facial expression analysis," in *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)*. IEEE, 2000, pp. 46–53.

[7] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[8] "Stress in america™ 2020: A national mental health crisis," Oct 2020.

[9] Jungmin Lee, Hyun Ju Jeong, and Sujin Kim, "Stress, anxiety, and depression among undergraduate students during the covid-19 pandemic and their use of mental health services," *Innovative higher education*, vol. 46, no. 5, pp. 519–538, 2021.

[10] Anna Butjosa, Juana Gómez-Benito, Inez Myin-Germeys, Ana Barajas, Iris Baños, Judith Usall, Norma Grau, Luis Granell, Andrea Sola, Janina Carlson, et al., "Development and validation of the questionnaire of stressful life events (qsle)," *Journal of psychiatric research*, vol. 95, pp. 213–223, 2017.

[11] Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, pp. 1849, 2019.

[12] José Almeida and Fátima Rodrigues, "Facial expression recognition system for stress detection with deep learning," 2021.

[13] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang, "What makes multimodal learning better than single (provably)," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[14] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[15] Wei-Yin Loh, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[16] John Platt et al., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[17] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[18] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–8, 2012.

[19] Hua Gao, Anil Yüce, and Jean-Philippe Thiran, "Detecting emotional stress from facial expressions for driving safety," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 5961–5965.

[20] Coskun Arslan, "An investigation of anger and anger expression in terms of coping with stress and interpersonal problem-solving.," *Educational Sciences: Theory and Practice*, vol. 10, no. 1, pp. 25–43, 2010.

[21] Amie M Gordon and Wendy Berry Mendes, "A large-scale study of stress, emotions, and blood pressure in daily life using a digital platform," *Proceedings of the National Academy of Sciences*, vol. 118, no. 31, 2021.

[22] Second Opinion, "Stress and its adverse effect on the human heart," May 2012.

[23] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jedrzej Nowak, "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity," in *2011 federated conference on computer science and information systems (FedCSIS)*. IEEE, 2011, pp. 405–410.

[24] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang, "Ubfc-phys: A multimodal database for psychophysiological studies of social stress," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[26] Ze Yang, Haofei Wang, and Feng Lu, "Assessment of deep learning-based heart rate estimation using remote photoplethysmography under different illuminations," *arXiv preprint arXiv:2107.13193*, 2021.