StressVision: Non-Invasive Stress Detection from Thermal Videos

StressVision: Non-Invasive Stress Detection from Thermal Videos

Calvin Xia[1], Vikram Bhagavatula[1], Jason Moraes[1],
William Peng[1], Ryan Murakawa-Rubin[1], Tom Bullock[2], Satish Kumar[1], B.S. Manjunath[1]
[1]Department of Electrical & Computer Engineering
[2]Department of Psychological & Brain Sciences
University of California, Santa Barbara
{calvin_xia, vikrambhagavatula, jmoraes, williampeng, rtr, tombullock, satishkumar, manj}@ucsb.edu

September 1, 2023

ABSTRACT

Timely and accurate stress detection is crucial for effective healthcare monitoring and intervention. Existing methods for stress detection often rely on invasive or subjective measures, limiting their use. Here, we propose StressVision; a non-invasive and automated transformer-based deep learning approach that uses thermal video analysis to capture and analyze facial thermal patterns, and enables objective and continuous stress detection. We validate our approach by applying StressVision to two datasets comprised of healthy human adult participants who were exposed to an acute stressor (ice-cold water) while thermal video of their faces and electrocardiography were recorded. One of these datasets was collected specifically for the purpose of this work (*n*=36) and the other dataset was acquired from a previous study (*n*=42). With StressVision we were able to achieve state-of-the-art stress detection performance, such that stress state could be classified (i.e., stress, no-stress) with accuracy = 0.8748. We make the StressVision source code available on GitHub along with our new dataset, which will serve as a valuable resource for stress-detection research and allow for bench-marking against other methods.

INTRODUCTION

Physical stress in humans is a common occurrence and a precursor of many health conditions – thus, it is imperative we develop methods for monitoring it easily. However, most methods for stress detection require monitoring physiological signals which requires medical expertise to use and are intrusive such as collecting electrocardiography (ECG) or impedance cardiography (ICG) signals. This is the primary motivation for developing an easy, non-intrusive approach to stress detection especially during/after the COVID pandemic where the need for such health monitoring system has become imperative. Recent work in remote photoplethysmography (rPPG) [31], which concerns itself with non-intrusive measurement of physiological signals, makes us of computer vision techniques to monitor one's health [8, 33, 26, 2]. rPPG shows that it is possible to estimate physiological signals and state of physiological stress from video, particularly that of the face.

However, the majority of prior research in remote photoplethysmography (rPPG) has predominantly focused on analyzing RGB facial videos. RGB videos have many limitations such as most of the subtle information is present in the green channel, which is heavily influenced by surrounding noise. Additionally, privacy concerns arise as RGB videos capture the true likeness of individuals. Hence, exploring alternative video modalities is necessary for improved performance and privacy preservation. Thermal video presents several compelling benefits over RGB video analysis. Notably, it exhibits greater robustness to environmental variations, including changes in illumination intensities, light directions, and multiple light sources [34]. Additionally, thermal video affords enhanced privacy protection, as it does not reveal the true likeness of the face [13]. In this paper, we investigate a deep learning-based computer vision approach tailored specifically for thermal facial video processing.

StressVision builds upon the foundation of StressNet, developed by the UCSB Vision Research Lab [21]. StressNet leveraged thermal video processing to predict stress levels and additionally estimated the initial systolic time interval (ISTI) as a physiological signal to enhance stress prediction accuracy. In light of the simplification of StressNet, we strive to move away from explicitly estimating physiological signals and make a direct prediction of stress. We propose a novel transformer-based neural network architecture that directly estimates the physiological state of stress. By leveraging the power of transformer models, we aim to capture the intricate spatio-temporal correlations present in thermal video data, enabling more accurate and reliable stress detection without relying on the estimation of specific physiological signals. Furthermore, to improve the effectiveness of our model, we have meticulously curated a novel dataset comprising thermal video recordings of subjects experiencing various levels of stress induced, alongside their corresponding ECG signals. This comprehensive dataset, which is made publicly available, provides a valuable resource for researchers and facilitates benchmarking against other stress detection methods.

A. Technical Contributions

- We propose a novel transformer-based spatio-temporal network that effectively processes thermal videos of human faces to capture both the spatial and temporal dynamics inherent in thermal video data, allowing for accurate stress classification.
- Our temporal network design plays a crucial role in balancing computational efficiency and decision accuracy. Our network does this by making stress classification decisions based on a single token instead of processing the entire vector.
- We collected a novel dataset of thermal videos of participants under stress and no-stress conditions, along with their ECG signals.
- We achieve state of the art performance on both the previous dataset [4] that we used to develop StressNet and the new dataset we collected for this paper.

B. Related Works:

Heart rate variability has been explored as a way for binary classification of stress [33, 8, 15, 28]. Earlier works on heart rate estimation depended exclusively on traditional signal processing techniques. The approach documented in Li et al. [23] located a particular region of interest (ROI), took the mean of the green channel from RGB video, and applied a band-pass filter resulting in an estimation of the individual's heart-beat. At the time of its publication in 2014, it accomplished the impressive feat of attaining a mean-squared error of 7.62 bpm on the MAHNOB-HCI dataset [23]. However, ROI tracking is necessary for this approach to work, thus results are highly susceptible to head movement.

Later approaches employed the use of RGB sensors to estimate heart-rate variability [26, 2, 27, 1]. The first end-to-end neural network model trained to perform rPPG was DeepPhys [8]. Using a deep convolutional neural network (CNN), it can estimate heart and breathing rates. It also uses a skin reflection model and a deep learning attention mechanism to learn motion representation of a person's head. A more recent model that builds upon the works of DeepPhys is PhysNet [33]. This model used a long short term memory (LSTM) network to distinguish temporal relations. For pulse detection, the new network made mild improvements over its predecessor. However, the introduction of LSTM allowed for the model to be trained to detect atrial fibrillation.

ISTI is another signal proposed as a reliable way to quantify both physical and psychological stress [17, 29, 12, 14, 3]. It requires the measurement of the ECG and impedance cardiography (ICG) signals to compute. StressNet, the foundation for this work, was the first approach to estimate ISTI from thermal video of humans [21]. It uses deep neural network estimate the ISTI and makes a binary classification based off the estimated signal[21]. It propagated a weighted sum of the loss from the estimated ISTI and the loss from the prediction to train the model. Like PhysNet, it utilized a CNN-LSTM configuration for feature extraction.

Transformer networks over the past couple of years have sought to replace LSTM when it comes to finding sequential dependencies in the field of computer vision. Through self-attention, they're better at capturing long-term dependencies and enable parallelization which improves the efficiency of training [30]. Several works have adapted this network for the purpose of spatial feature extraction when performing image segmentation tasks for various cross-domain applications [11, 19, 5, 16, 20]. These networks have also been implemented for feature extraction in the time domain [32, 25]. The input to these models is electrical brain activity recorded at individual's scalps via electroencephalography (EEG). In Yan et al. [32], they used signal processing techniques to extract local features of the EEG signal while Liu et al. [25] used a CNN to do so. The commonality between the two works is the use of transformer networks to capture long-term dependencies in the signal, resulting in high accuracy for epileptic seizure predictions.

APPROACH

Our approach uses an emission representation model to pre-process raw thermal videos and generate a more refined version of our data to input into our spatio-temporal network [21]. The final prediction yielded by our model is a binary classification (stress, no-stress) based on raw input thermal videos. Figure 1 depicts the proposed model architecture.

The interest in shifting away from explicitly estimating physiological signals like ISTI led to

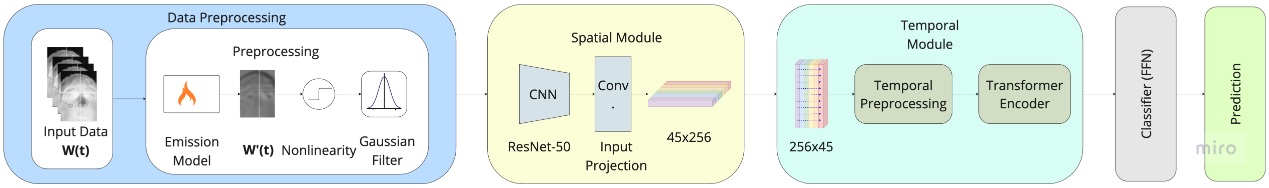

Figure 1: *Model architecture overview. The Preprocessing module applies the emission representation model, yielding the sequence fed into the spatial and temporal modules. The Spatial module yields feature vectors of length 256, with 45 of such vectors in this diagram (one for each frame in the video). These are subsequently processed by the Temporal module. The Classifier draws upon the output of the Temporal module, enabling for a final stress classification.*

the development of a new architecture for our spatio-temporal network. To this end, we developed an architecture inspired by both the Detection Transformer (DETR) [5] and the Vision Transformer [9, 10] for video classification.

C. Data Preprocessing

Our initial preprocessing step is inspired by StressNet which involved modeling the reflection of light in skin and blood vessels with Shafer's dichromatic reflection model [21]. The model defines the total energy of a single pixel as the weighted sum of total heat emission from blood movement, absorption of radiation from skin tissue and blood vessel, and absorption of radiation from the atmosphere.

By eliminating atmospheric absorption due to experimental conditions and treating human skin as a black-body radiator, the model is able to narrow down changes in thermal energy to head movement and blood flow in the face. Taking the first-order derivative with respect to time, the model filters out the time-independent components of each of the sources of thermal variability.

$$\boldsymbol{W}'(t) = p'(t)\ .\ E_o\ .\ (\epsilon_b + \epsilon_b\ .\ \frac{\partial f_1}{\partial p} + \epsilon_s\ .\ \frac{\partial f_2}{\partial p}) + m'(t)\ .\ E_o\ .\ (\epsilon_b\ .\ \frac{\partial f_1}{\partial m} + \epsilon_s\ .\ \frac{\partial f_2}{\partial m}) \tag{1}$$

$\mathbf{W}(t)$ represents the energy of a single pixel of a frame. The equation takes into consideration the thermal variation observed by both the thermal camera and skin tissue denoted by f_1 and f_2 respectively. This observed variation is further separated into its time dependent components $m(t)$, which are changes in energy due to non-physiological causes such as head movement and facial expressions, and $p(t)$, which are changes in energy due to blood volume in skin tissue and blood vessels. E_o is the energy of a black-body radiator at constant temperature which is modulated by ϵ_s and ϵ_b to correctly scale each of the time-dependent terms by the emissivity of skin and blood vessels respectively.

We then apply $\log$ non-linearity to suppress outliers [21]. This non-linearity is as follows

$$\boldsymbol{X}(t) = sign(\boldsymbol{W}'(t))\ .\ \log(1 + \mod \boldsymbol{W}'(t)). \tag{2}$$

It also applies a Gaussian filter with $\sigma = 3$ in the spatial domain and $\sigma = 4$ in the temporal domain to remove high frequency components.

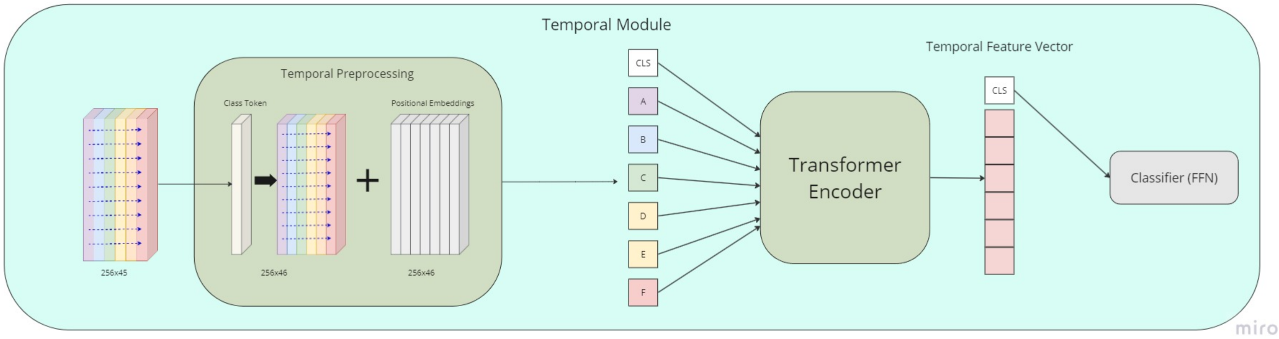

Figure 2: *Temporal and Classifier Module. A class token is prepended to the sequence of feature vectors output by the Spatial module; then a learnable positional embedding is added. The Classifier draws upon the very first token in the transformer encoder's output, similar to the Vision Transformer.*

$X(t)$ is the input into our network after Gaussian smothing. With classical signal processing techniques, the model transforms our input data into a form that is more information-rich.

D. Spatio-Temporal Module

Intuitively, we want an architecture which would perform spatial feature extraction on the preprocessed video frames, followed by sequence processing which would pick up on temporal patterns and relations between the video frames, ultimately allowing us to converge upon a classification result. We found a CNN-Transformer structure (similar to that in DETR [5]) suitable for performing this task. The CNN would help obtain a compact feature representation for each frame in the video, and these feature vectors would then be fed into a transformer structure. The output of this structure would then be used for the final classification (many to one).

Specifically, our model initially uses a CNN spatial backbone for feature extraction from each preprocessed frame in the video. This spatial backbone is applied to each frame $\boldsymbol{X}(t) \in \mathbb{R}^{1 \times H_0 \times W_0}$, yielding an output activation map $f \in \mathbb{R}^{C \times H \times W}$ for each frame. Values of $C = 2048$, $H = \frac{H_0}{32}$, and $W = \frac{W_0}{32}$ are used. The channel dimension for each of these activation maps is reduced using 1x1 convolutions, bringing the channel count to 256. Finally, adaptive average pooling is applied to each of these activation maps, collapsing each of them to a 256 length feature vector – this helps us preserve information while enabling for dimensionality reduction. Each of these feature vectors make up the sequence that is to be fed into the transformer encoder.

Our classifier module is inspired by the Vision Transformer [10, 19, 18, 22, 16]. A learnable class token of size 256 is prepended to the sequence of aforementioned feature vectors, and a learnable positional embedding is added to the tokens. This sequence is then fed into the transformer encoder. Only the class token is used because there is less information to process and it contains nominal sequential information about all the other tokens. Using other tokens for decision making adds an implicit bias, as they are all strongly correlated towards themselves. Because the class token was initially "empty," it will accumulate relevant information about all the other tokens without adding additional biased information. This de-biasing the class token provides, while only containing relevant time dependent information, is what makes it inherently useful for classification [10]. A classifier head at the end of the transformer encoder draws upon the class token's final representation at the output, ultimately yielding a binary classification. The process of prepending the class token and obtaining the final classification is illustrated in Figure 2.

E. Loss Function:

Since our model purely performs binary classification, we used the binary cross entropy (BCE) loss function, which is as follows:

$$BCE = \frac{1}{N}\sum_{i=1}^{k} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \tag{3}$$

EXPERIMENTS

In this section we discuss the various design choices of StressVision and experiments that were carried out to validate those design choices. We provide details on our experimental methods and compare the performance of StressVision with that of current state-of-the-art methods.

F. Existing Dataset

We acquired an existing dataset from the UC Santa Barbara Biomarkers of Stress States (BOSS) [4]. This dataset includes concurrent thermal video and ECG recordings of 42 healthy human adult participants repeatedly exposed to acute cold stress five times within a single laboratory based testing session. The original BOSS study was designed to investigate how different types of stress impact the human brain, physiology and behavior. Participants were considered ineligible if any of the following criteria applied: history of heart condition or joint issues, recent surgeries that would inhibit movement, BMI of more than 30, and currently taking blood pressure medication or any psychostimulants or antidepressants. The procedures were approved by Western IRB and The U.S. Army Human Research Protection Office. All processes adhered to the policies of the UC Santa Barbara Human Subjects Committee. At the start of each session, participants were given all details regarding the experiment and provided informed consent.

G. Current Dataset

The methodological approach for our new experiment is inspired by the BOSS protocols for cold-exposure [4]. Since the BOSS data were collected in a controlled environment, to make the detection more robust and adapt to a natural environment settings, we collected a new dataset. We collected the data with multiple and varying light source and multiple people present around the subject. In collecting our new dataset we applied a similar methodological approach to BOSS so that the two datasets could be either merged or compared easily. The experiment was approved by the Institutional Review Board (IRB) at UC Santa Barbara and conducted on 36 healthy adult

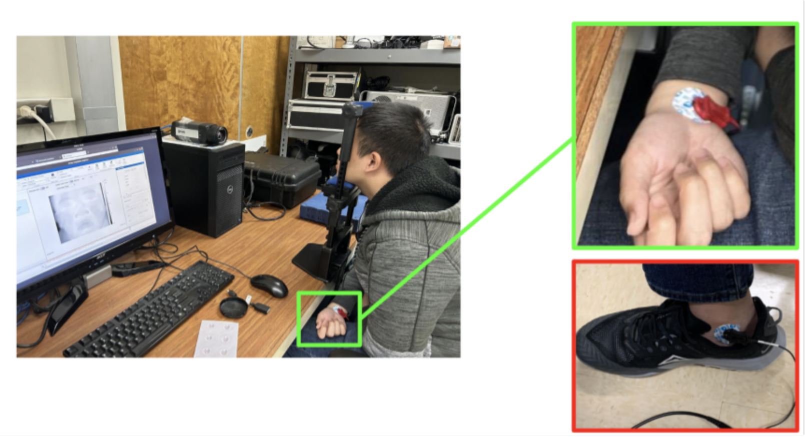

Figure 3: *Experimental Setup. The left image shows the entire setup, the images on the right show placement of electrodes on subject for recording ECG. The red box illustrates electrode placement on the subject's foot (out of view in the left picture).*

volunteers who were members of the UC Santa Barbara campus community. All previous exclusion criteria were applied. Additional criteria include: history of toe tip blanching, or diabetes, diagnosed with peripheral nerve injury, Raynaud's Disease, or hand-arm vibration syndrome, and history of neurological conditions or neck pain. Participants were briefed about the study and provided fully informed consent prior to participation. An image of the experimental setup is shown in Figure 3.

Each individual completed two tests within a single session. The Cold Pressor Test (CPT) involved immersing the left hand in cold water ($\sim 4°C$) for two minutes, and a warm water control (Warm Pressor Test; WPT) involved immersing the left hand in warm water ($\sim 34°C$) for two minutes. The goal of these tests was to induce stressed and non-stressed states, respectively. Each test consisted of a three minute session: a one minute pre-exposure baseline recording followed by a two minute exposure to either cold or warm water. The order that the CPT and WPT were completed in was counterbalanced between participants.

We trained and evaluated our model on a combination of thermal videos and ECG data acquired from the BOSS dataset (59 videos in total) and our novel dataset (35 videos).

H. *Evaluation Metrics*

The metric for evaluating the performance of StressVision are: the probability metrics, precision, recall, and true negative rate (tnr) represented with equation 4.

$$Precision = \frac{tp}{tp + fp}, \qquad Recall = \frac{tp}{tp + fn}, \qquad TNR = \frac{tn}{tn + fp}, \tag{4}$$

where, tp is true positive, tn is true negative, fp is false positive, and fn is false negative. It is worth noting that recall is also referred to as the true positive rate (tpr). These quantities are then used to calculate both the accuracy and balanced accuracy of the predictions made by the model.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}, \qquad Balanced\ Accuracy = \frac{tpr + tnr}{2} \tag{5}$$

Considering the training data's skewed distribution, with more instances of "no stress" cases than "stress" cases, it becomes crucial to incorporate balanced accuracy as a significant metric.

I. Implementation Details

Transformers are known to take large amounts of data to converge [24]. Training a transformer from scratch given our limited data is difficult. We initialized our spatial backbone and temporal transformer with the pretrained weights of DETR [5]. Our spatial module is a ResNet-50 backbone which extracts spatial features frame by frame of the thermal video. The spatial feature extractor is followed by average pooling layer to pool relevant features. The adopted average pooling instead of max pooling because the signatures we are looking for are very subtle and pooling max values only might overlook those weak features. This is followed by input projection layer to collapse the output activation map (which was $256 \times H \times W$ in volume) into a 256 length feature vector for each frame. We used only the encoder from transformer model inspired from Vision Transformer [10, 30]. The encoder is initialized with pre-trained weights [5], it has 6 layers and 8 attention heads with feature size 256 [6]. The output of average pooling layer is appended with a class token of size 256 and fed to transformer encoder. The first encoded feature (corresponding to class token) of length 256 from encoder is passed through a 2 fully connected layer classifier to make a binary decision of Stress or no-stress. This classifier consists of two linear layers with 256 neurons in the input layer and 128 in the hidden layer, along with ReLU activation functions.

For the model training process, we made usage of the Adam optimizer with a learning rate of 10^{-5} for all parts of the model: spatial backbone, input projection layer, transformer encoder, and classifier. Additionally, a learning rate scheduler was used. For the spatial backbone and input projection layer, we used a learning rate multiplier of 1 for the first 10 epochs. Afterward, we reduced the learning rate by a factor of 0.1 every epoch to enable the model to make fine adjustments and converge more precisely. Similarly, for the transformer encoder and classifier, we applied a learning rate multiplier of 1 for the first 20 epochs and the scaling it down by a factor of 0.1 every epoch for subsequent training.

We trained StressVision on 2 NVIDIA RTX A4000 GPUs with 16GB of VRAM. The system was equipped with 64GB of system memory, and had a Intel(R) Core(TM) i9-10900X CPU with a clock speed of 3.70GHz and 20 cores.

J. Dataset Collection Details:

We recorded our dataset containing thermal videos of human faces and concurrent ECG under stressed non-stressed (WPT) and (CPT) conditions. We also collected subjective pain ratings at the end of each test using a visual analogue scale (VAS). The participants were 36 UCSB students (24 male, 12 female) between the ages of 18-35 (mean age=23). Adjacent frames taken from each of the two videos for the same subject during both the stress and no-stress conditions are shown in Figure 4.

The thermal videos are taken with a FLIR thermal camera (Model A655sc, Flir Systems, Wilsonville, OR, USA). Recorded thermal videos have a resolution of 640 x 420 pixels at 15 frames per second. Our experiments were performed in a natural setting with various light sources and people walking around in the background which better simulates real world noise. The participant's rested their chin on a chin-rest and the camera was positioned 45 cm from the face.

ECG was recorded with a Biopac ECG device (Model MP36, Biopac, Goleta, CA, USA) at a sampling rate of 2,000 Hz. The ECG electrodes were placed on the right wrist and each ankle of the participant. By using an Arduino serial port, we were able to start the ECG recording and the thermal video recording within 2 ms of each other.

Arm and leg movement during the experiment introduced low frequency noise in the ECG recording, commonly referred to as baseline wander. To compensate for this, we filtered out the noise using a high-pass filter. We will provide the both the unfiltered and filtered versions of each ECG signal in the dataset.

Finally, after completing the CPT and WPT, participants completed a VAS pain rating. Participants were handed a piece of paper with a scale numbered 0-10 labeled "No pain at all" and "Worst pain imaginable" at 0 and 10, respectively and asked to rate their pain. These pain ratings will also be made available for each participant along with their corresponding thermal video and ECG data.

Our data were split using an 80-20 training to validation split. Due to GPU memory constraints, videos are split into segments of 45 frames each (3 seconds duration with a 15 fps sampling rate), each to be classified separately. This also had the added benefit of generating more samples from our dataset, rendering our measured performance more reflective of potential real world performance (had we had access to more data). Ultimately, this generated 3646 training samples and 607 test samples.

RESULTS

Performance: StressVision is evaluated along the lines of balanced classification accuracy. After training for 52 epochs, a balanced classification accuracy of 0.8748 was obtained. StressVision outperforms StressNet [21] by a noticeable margin, which had a classification accuracy of 0.843 as shown in Table 1. These evaluations are performed on both videos from BOSS and our novel dataset. As shown in Table 1 our implementation of DeepPhys model [8] did not perform well in detecting stress from thermal videos. This poor performance mainly stems from main reason that DeepPhys model is not suited for thermal videos. For baseline methods, ECG signal used is extracted by simple statistical filtering methods hence very noisy. We tracked regions from nose tip and forehead for our baseline. Additional performance metrics like precision and recall can be seen in Table 1.

Ablations: Due to the infeasibility of training the transformer from scratch, it was difficult to carry out experiments varying different aspects of the transformer (i.e. the number of encoder layers). It was desirable to keep the same structure as DETR so we could use the pretrained weights, and speed up convergence. However, one important experiment we did (and could feasibly carry out) was the removal of the class token. The model with no class token was able to achieve a balanced test classification accuracy of 0.8778 within $\sim$ 52 epochs, which is comparable to that of the class token model. However, its precision (0.8494) and true negative rate (0.8517) were

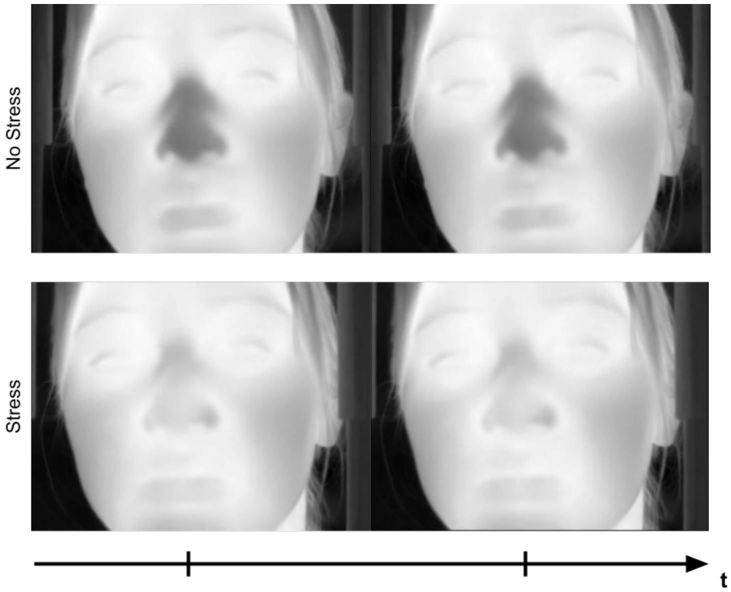

Figure 4: *Two adjacent frames taken from two thermal videos in the assembled dataset. The video frames show a participant under no-stress (WPT) and stress (CPT) conditions, respectively. The individual shown here is a member of the research team, used for illustration purposes only.*

Model	Accuracy	Precision	Recall	TNR
Baseline	0.170	–	–	–
DeepPhys [8]	0.575	–	–	–
I3D [7] + Detection Network	0.84	–	–	–
StressNet [21]	0.843	–	–	–
StressVision (No Class Token)	**0.8778**	0.8494	**0.9138**	0.8517
StressVision (Class Token)	0.8748	**0.8967**	0.8379	**0.9117**

Table 1: *StressVision's performance in classifying stress states, as compared to previous works, as well as variants of StressVision. "–" represents metrics not available.*

lower, while its recall was better (0.9138). This indicates that the class token model is potentially more conservative in its predictions, but makes a more accurate positive prediction when it does. Depending on the desired application, one may be willing to tolerate a higher precision in exchange for lower recall (or vice versa).

CONCLUSION

This paper presents a novel modeling approach to the classification of acute stress states from thermal videos of the human face. Our modeling approach achieves superior performance when compared to the current state-of-the-art modeling approaches to stress classification, without requiring the estimation of complicated physiological signals. It builds upon a vision-transformer model pretrained for object detection, ultimately achieving a stress vs. no-stress classification accuracy of 87.48%.

Our new method does have its limitations. The model is trained and tested on physical stress related signals, so its generalization to detecting psychological stress needs to be validated. Furthermore, placing a hand in warm water ($\sim 34°C$) does not truly represent a non-stressed state since the body has to adjust to the external stimulus. It is possible that the network is learning to identify a state that is somewhere in between stressed and non-stressed.

In summary, in this paper we outline and validate a novel modeling approach that achieves best-in-class stress detection performance. Furthermore, we contribute a novel open-source dataset consisting of thermal videos, ECG data and pain-ratings collected from 36 healthy adults. This dataset will be valuable for other researchers applying computer vision approaches to stress detection.

ACKNOWLEDGEMENT

This research is partially supported by the following grants: NSF award $SI2-SSI$ $\#1664172$ and International Telemetry Conference Grant 2023.

REFERENCES

[1] Timon Blöcher, Johannes Schneider, Markus Schinle, and Wilhelm Stork. An online ppgi approach for camera based heart rate monitoring using beat-to-beat detection. In *2017 IEEE Sensors Applications Symposium (SAS)*, pages 1–6. IEEE, 2017.

[2] Frédéric Bousefsaf, Choubeila Maaoui, and Alain Pruski. Remote assessment of the heart rate variability to detect mental stress. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 348–351. IEEE, 2013.

[3] Sharon L Brenner and Theodore P Beauchaine. Pre-ejection period reactivity and psychiatric comorbidity prospectively predict substance use initiation among middle-schoolers: A pilot study. *Psychophysiology*, 48(11):1588–1596, 2011.

[4] Tom Bullock, Mary H. MacLean, Tyler Santander, Alexander P. Boone, Viktoriya Babenko, Neil M. Dundon, Alexander Stuber, Liann Jimmons, Jamie Raymer, Gold N. Okafor, Michael B. Miller, Barry Giesbrecht, and Scott T. Grafton. Habituation of the stress response multiplex to repeated cold pressor exposure. *Frontiers in Physiology*, 13, Jan. 2023.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[8] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[11] Ruiming Du, Zhihong Ma, Pengyao Xie, Yong He, and Haiyan Cen. Pst: Plant segmentation transformer for 3d point clouds of rapeseed plants at the podding stage, 2022.

[12] Mohamad Forouzanfar, Fiona C Baker, Ian M Colrain, Aimée Goldstone, and Massimiliano de Zambotti. Automatic analysis of pre-ejection period during sleep using impedance cardiogram. *Psychophysiology*, 56(7):e13355, 2019.

[13] Jacob Gunther and Nathan E Ruben. Remote heart rate estimation, Dec. 26 2017. US Patent 9,852,507.

[14] J Benjamin Hinnant, Lori Elmore-Staton, and Mona El-Sheikh. Developmental trajectories of respiratory sinus arrhythmia and preejection period in middle childhood. *Developmental psychobiology*, 53(1):59–68, 2011.

[15] Gee-Sern Hsu, ArulMurugan Ambikapathi, and Ming-Shiang Chen. Deep learning with time-frequency representation for pulse estimation from facial videos. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 383–389. IEEE, 2017.

[16] ASM Iftekhar, Satish Kumar, R Austin McEver, Suya You, and BS Manjunath. Gtnet: Guided transformer network for detecting human-object interactions. In *Pattern Recognition and Tracking XXXIV*, volume 12527, pages 192–205. SPIE, 2023.

[17] Robert M Kelsey. Beta-adrenergic cardiovascular reactivity and adaptation to stress: The cardiac pre-ejection period as an index of effort. 2012.

[18] Satish Kumar, Ivan Arevalo, ASM Iftekhar, and BS Manjunath. Methanemapper: Spectral absorption aware hyperspectral transformer for methane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17609–17618, 2023.

[19] Satish Kumar, Ivan Arevalo, ASM Iftekhar, and B S Manjunath. Methanemapper: Spectral absorption aware hyperspectral transformer for methane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17609–17618, June 2023.

[20] Satish Kumar, ASM Iftekhar, Ekta Prashnani, and BS Manjunath. Locl: Learning object-attribute composition using localization. *arXiv preprint arXiv:2210.03780*, 2022.

[21] Satish Kumar, A S M Iftekhar, Michael Goebel, Tom Bullock, Mary H. MacLean, Michael B. Miller, Tyler Santander, Barry Giesbrecht, Scott T. Grafton, and B. S. Manjunath. Stressnet: Detecting stress in thermal videos, 2020.

[22] Satish Kumar, William Kingwill, Rozanne Mouton, Wojciech Adamczyk, Robert Huppertz, and Evan D Sherwin. Guided transformer network for detecting methane emissions in sentinel-2 satellite imagery. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*, 2022.

[23] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4264–4271, 2014.

[24] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers, 2021.

[25] Chenyu Liu, Xinliang Zhou, and Yang Liu. Eened: End-to-end neural epilepsy detection based on convolutional transformer, 2023.

[26] Daniel McDuff, Sarah Gontarek, and Rosalind Picard. Remote measurement of cognitive stress via heart rate variability. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2957–2960. IEEE, 2014.

[27] Hamed Monkaresi, Nigel Bosch, Rafael A Calvo, and Sidney K D'Mello. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28, 2016.

[28] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3580–3585. IEEE, 2018.

[29] K Purushotham Prasad and Dr B Anuradha. Detection of abnormalities in fetal electrocardiogram. *International Journal of Applied Engineering Research*, 12(1):2017, 2017.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[31] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

[32] Jianzhuo Yan, Jinnan Li, Hongxia Xu, Yongchuan Yu, and Tianyu Xu. Seizure prediction based on transformer using scalp electroencephalogram. *Applied Sciences*, 12(9), 2022.

[33] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *Proc. BMVC*, 2019.

[34] Qi Zhang, Yimin Zhou, Shuang Song, Guoyuan Liang, and Haiyang Ni. Heart rate extraction based on near-infrared camera: Towards driver state monitoring. *IEEE Access*, 6:33076–33087, 2018.