Deception Detection and Remote Physiological Monitoring: A Dataset and Baseline Experimental Results

Nathan Vance, *Graduate Student Member, IEEE*, Jeremy Speth, *Graduate Student Member, IEEE*, Siamul Khan, Adam Czajka, *Senior Member, IEEE*, Kevin W. Bowyer, *Life Fellow, IEEE*, Diane Wright, and Patrick Flynn, *Fellow, IEEE*

***Abstract*—We present the Deception Detection and Physiological Monitoring (DDPM) dataset and initial baseline results of deception detection on this dataset. Our application context is an interview scenario in which the interviewee attempts to deceive the interviewer on selected responses. The interviewee is recorded in visible, near-infrared, and long-wave infrared light, along with cardiac pulse, blood oxygenation, and audio. After collection, data were annotated for interviewer/interviewee, curated, ground-truthed, and organized into train / test parts for a set of canonical deception detection experiments. For the collected data, we estimated subject heart rates (remotely from face videos with a mean absolute error lower than two beats per minute), pupil size, and eye gaze. Fusion experiments discovered that a combination of remote plethysmography, pupil size, and thermal data yields the best deception detection results, with an equal error rate of 0.357. The database contains almost 13 hours of recordings of 70 subjects, and over 8 million visible-light, near-infrared, and thermal video frames, along with appropriate meta, audio and pulse oximeter data. To our knowledge, this is the only collection offering recordings of five modalities in an interview scenario that can be used in both deception detection and remote photoplethysmography research.**

***Index Terms*—Photoplethysmography, deep learning, affective computing.**

I. Introduction

NEW DIGITAL sensors and algorithms offer the potential to address challenges in human monitoring. Two interesting problems in this domain are remote physiological monitoring (*e.g.,* via remote photoplethysmography (rPPG) [1]) and deception detection (via remote analysis of various signals, such as pulse rate, blinking, or EEG that attempts to predict anxiety and/or cognitive load [2], [3], [4], [5], [6], [7], [8], [9], [10]). In this paper, we present the *Deception Detection and Physiological Monitoring (DDPM)* dataset and baseline experiments with this dataset. DDPM is collected in an interview context, in which the interviewee attempts to deceive the interviewer with selected responses. DDPM supports analysis of video and pulse data for facial features including pulse, gaze, blinking, pupillometry, face temperature, and micro-expressions. The dataset comprises over 8 million high resolution RGB, near-infrared (NIR) and thermal (LWIR) frames from face videos, along with cardiac pulse, blood oxygenation, audio, and deception-oriented interview data. We provide this dataset with evaluation protocols to help researchers assess automated deception detection techniques.[1] The **main contributions** of this work are:

a) the largest **deception detection dataset** in terms of total truthful and deceptive responses, recording length, and raw data size;
b) the first dataset for both deception detection and **remote pulse monitoring** with RGB, NIR, and thermal imaging modalities, synchronized in time;
c) the first rPPG dataset with **facial movement and expressions** in a natural conversational setting;
d) **baseline results for deception detection** using pupillometry, heart rate estimation, dynamics of face temperature, and fusion of these measurements.

This work is an extension of our conference paper [11], with the following significant additions: (a) results from experiments probing the robustness of the proposed rPPG model (RPNet), (b) a pupillometry method for working with low resolution visible-light videos and accommodating off-angle gaze; (c) a feature fusion analysis utilizing rPPG, pupillometry, and thermal data for deception detection.

II. Background

A. Databases for Deception Detection

Most research in deception detection has been designed and evaluated on private datasets, typically using a single sensing modality. The proposed DDPM dataset addresses these drawbacks. The top of Table I compares sensor modalities and acquisition characteristics for existing datasets and DDPM. Early researchers, inspired by the polygraph, believed that

Manuscript received 28 February 2022; revised 23 June 2022 and 25 August 2022; accepted 26 August 2022. Date of publication 3 November 2022; date of current version 5 December 2022. This article was recommended for publication by Associate Editor S. Yu upon evaluation of the reviewers' comments. *(Corresponding author: Nathan Vance.)*

Approval of all ethical and experimental procedures and protocols was granted by the Human Subjects Institutional Review Board under Application No. 19-08-5484.

The authors are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: nvance1@nd.edu; jspeth@nd.edu; skhan22@nd.edu; aczajka@nd.edu; kwb@nd.edu; dwright2@nd.edu; flynn@nd.edu).

Digital Object Identifier 10.1109/TBIOM.2022.3218956

[1] https://cvrl.nd.edu/projects/data/#deception-detection-and-physiological-monitoringddpm

2637-6407 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISON OF THE DIFFERENT MODALITIES AND ENVIRONMENTS FOR EXISTING DATABASES FOR DECEPTION DETECTION AND rPPG

	Dataset	Subject Count	Length (Minutes)	Head Motion	Talking	RGB	NIR	Thermal	Physio-logical	Audio	Train/Test Splits	Raw Data
Deception	Silesian [8]	101	186	✓	✓	✓						
	Multimodal [15]	30	-		✓	✓		✓	✓	✓		-
	Real Trials [16]	56	56	✓	✓	✓				✓		
	EEG-P300 [12]	11	-		✓				✓			✓
	Box-of-Lies [14]	26	144	✓	✓	✓				✓		
	Bag-of-Lies [7]	35	<241		✓	✓			✓	✓	✓	
rPPG	MAHNOB-HCI [17]	27	264			✓			✓			
	PURE [18]	10	60	✓	✓	✓			✓			✓
	AFRL [19]	25	1500	✓		✓			✓			✓
	MMSE-HR [20]	40	<102	✓		✓		✓	✓			
	COHFACE [21]	40	160			✓			✓		✓	
	VIPL-HR [22]	107	1150/380*	✓	✓	✓	✓		✓		✓	
	UBFC-RPPG [23]	43	70	✓		✓			✓			
	DDPM (Ours)	**70**	**776**	**✓**	**✓**	**✓**	**✓**	**✓**	**✓**	**✓**	**✓**	**✓**

– Certain features could not be acquired. <The length is estimated using the maximum length of a recording described in literature. * Some NIR videos in VIPL-HR were discarded, so the length of the RGB/NIR videos are given individually.

information from the nervous system would likely give the best signals for deceit. Along these lines, the EEG-P300 dataset [12] was proposed, which consists solely of EEG data. Early work by Ekman [13] asserted that humans could be trained to detect deception with high accuracy, using micro-expressions. Inspired by human visual capabilities, the Silesian dataset [8] contains high frame-rate RGB video for more than 100 subjects. The Box-of-Lies dataset [14] was released with RGB video and audio from a game show, and presents preliminary findings using linguistic, dialog, and visual features. Multiple modalities have been introduced in the hope of enabling more robust detection. Pérez-Rosas et al. [15] introduced a dataset for deception including RGB and thermal imaging, as well as physiological and audio recordings. DDPM includes these modalities but adds NIR imaging, higher temporal and spatial resolution in RGB, and twice as many interviews. Gupta et al. [7] proposed Bag-of-Lies, a multimodal dataset with gaze data for detecting deception in casual settings. Concerns about the authenticity of deception in constrained environments spurred the creation of the Real-life Trial dataset [16]. Transcripts and video from the courtroom were obtained from public multimedia sources to construct nearly an hour of authentic deception footage. While the environment for "high-stakes" behavior is more difficult to achieve in the lab setting, the number of free variables involved in retrospectively assembling a real-world dataset (*e.g.,* camera resolution, angle, lighting, distance) may make algorithm design difficult.

B. Databases for rPPG

The middle part of Table I collects various properties of the existing rPPG datasets. The first widely used and publicly available rPPG dataset was MAHNOB-HCI [17], in which subjects' faces are relatively stationary (a significant limitation). Stricker et al. [18] introduced PURE, the first public dataset with stationary and moving faces. Estepp et al. [19] simultaneously collected raw video from 9 imagers of 25 subjects at different orientations to form the AFRL dataset. Later, MMSE-HR [20] was used for rPPG during elicited emotion. The dataset consisted of more subjects than MAHNOB-HCI with more facial motion. The COHFACE dataset and open source implementations of three rPPG algorithms were introduced in [21]. To accommodate data requirements for deep learning-based solutions, the VIPL-HR dataset [22] was created. Aside from being the largest publicly available dataset for rPPG, they released preliminary results from a CNN that outperformed then-existing techniques. Recently, the UBFC-RPPG [23] dataset (containing rigid motion and a skin segmentation algorithm for rPPG) was released.

To our knowledge, no rPPG datasets other than DDPM contains *natural* conversational behavior with unconstrained facial movement.

C. Deception Detection Methods

Ekman et al. used his Facial Action Coding System (FACS) [24] to detect *microexpressions* and make inferences about deception [13]. Zhang et al. [2] leveraged FACS to detect deceitful facial expressions, which do not reproduce the full set of action units exhibited by a genuinely felt emotion. Wang et al. [25] used pupil diameter to distinguish liars from truth tellers, finding that pupils dilate when a subject is attempting to mislead. We investigate this claim on the data collected in this study. Bhaskaran et al. [3] leveraged eye movements to detect deceit. A study comparing non-visual saccades between planned lies, spontaneous lies, and truth telling found that the eye movement rate (in saccades per second) was greater in 68% of subjects when telling a spontaneous lie versus telling the truth [6]. Caso et al. discovered that a deceptive person utilizes relatively fewer *deictic* (pointing) gestures and more *metaphoric* gestures (*e.g.,* forming a fist as a metaphor for strength) when compared to a truthful person [26]. Michael et al. take the subject's full body posture into account when detecting deceit [27]. Nonverbal clues to deceit remain a controversial topic [28]. A meta analysis conducted by Luke [29] suggests that all of the nonverbal clues to deceit existing in psychological literature could plausibly

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.

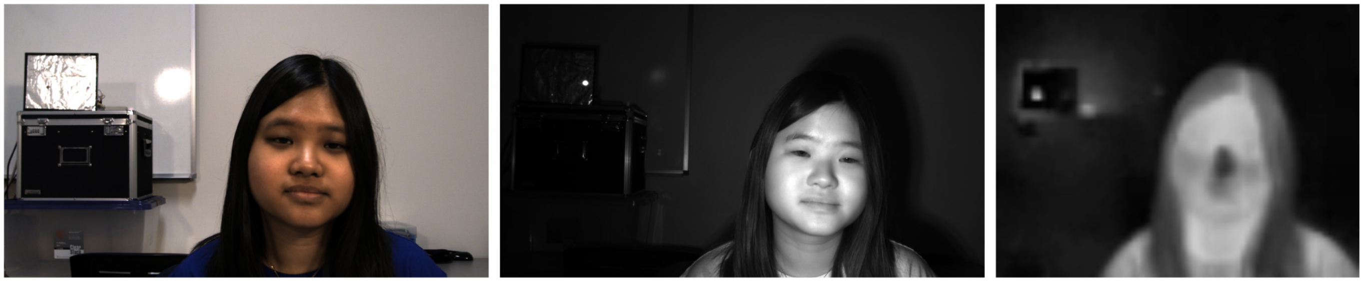

Fig. 1. Sample images from the RGB, NIR, and thermal cameras (left to right) from the collected DDPM dataset.

have arisen due to random chance as a result of small sample sizes and selective reporting. Our dataset enables researchers to more rigorously test claims regarding nonverbal clues to deceit as well as assertions to the contrary.

III. Apparatus

A. Sensors

Detection of facial movements requires high spatial and temporal resolution. Analyzing images collected in different spectra as in Fig. 1 may provide deeper insight into facial cues associated with deception. Additionally, changes observed in the cardiac pulse rate as in Fig. 4 may elucidate one's emotional state [5]. Speech dynamics such as tone changes provide another mode for detecting deception [30]. We assembled an acquisition arrangement composed of three cameras, a pulse oximeter, and a microphone to address these needs. All data were captured by a workstation designed to accommodate the continuous streaming of data from the three cameras (750 Mbps), operating a GUI that contained subject registration and interview progression components.

The **entire sensing apparatus** consisted of:

a) a DFK 33UX290 **RGB camera** from The Imaging Source (TIS) operating at 90 FPS with a resolution of 1920 × 1080 px;

b) a DMK 33UX290 monochrome camera from TIS with a bandpass filter to capture **near-infrared images** (730 to 1100 nm) at 90 FPS and 1920 × 1080 px;

c) a FLIR C2 compact **thermal camera** that yielded 80 × 60 px images at 9 FPS;

d) a FDA-certified Contec CMS50EA **pulse oximeter** that provides a 60 samples/second SpO2 and heart rate profile;

e) a Jabra SPEAK 410 omni-directional **microphone** recording both interviewer and interviewee at 44.1 kHz with 16-bit audio measurements.

B. Synchronization Device

The NIR, RGB and LWIR sensors were time-synchronized using visible and thermal artifacts generated by an Arduino-controlled device, illustrated in Fig. 2. The synchronization device generates signals in visible, near infrared and thermal spectra with pre-defined duty cycle. While the duty cycle is kept constant (50% of the period), we gradually increase the period from 5 seconds to 13 seconds, as shown in Fig. 3, to create a pattern that repeats every 81 seconds. Signals modulated in this way allow for cameras to be unambiguously synchronized over longer synchronization times, *i.e.,* up to 81 seconds, with multiple signal edges in between for finer alignment. In the current setup, a 50 ms delay is added to RGB and NIR signals due to inertia of the shutter used to generate the LWIR signal. Assuming that the position of the synchronization target is known within the video frame, the signal reconstruction is straightforward and based on reading local image intensity, as shown in Fig. 3.

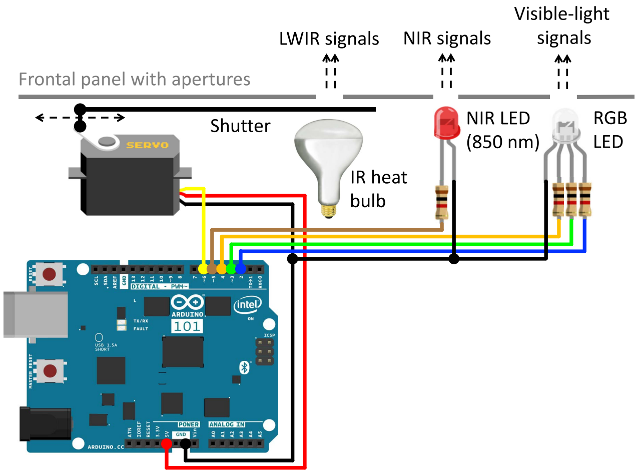

Fig. 2. Synchronization target generating signals in visible, near infrared and thermal spectra.

IV. Procedure

The collection protocol ensures that each subject gave both honest and deceptive answers. Each session consisted of a brief preparatory meeting followed by a 10 - 20 minute interview in which 24 questions were asked, nine of which the subject was instructed to answer deceptively. All data were collected under a protocol approved by the authors' institution's Human Subjects Institutional Review Board.

A. The Mock Interview Setting

A professional actor was hired to conduct the interviews and to provide their judgment of truthful or deceptive answers to each question they asked. The actor was instructed to be stoic and non-reactionary during the interviews, and to wear clothing consistent with that of a security official. They were

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.

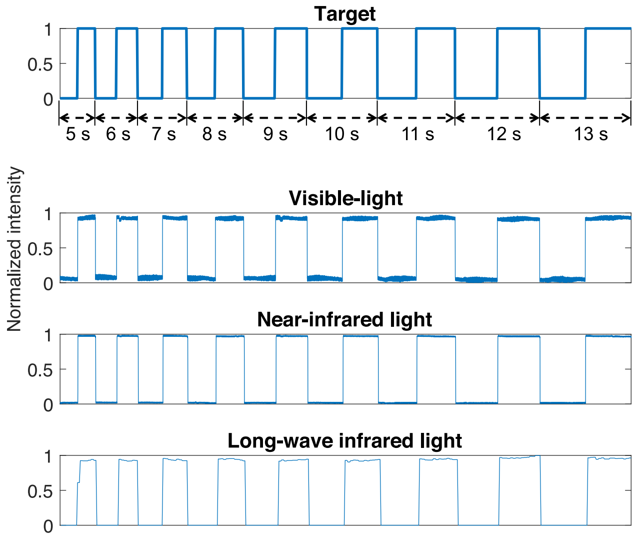

Fig. 3. Synchronization signal generated by the target (top) and its example reconstructions done by three sensors used in this study. The reconstructed intensities were normalized to ⟨0, 1⟩ as the absolute signal value does not play a role in synchronization.

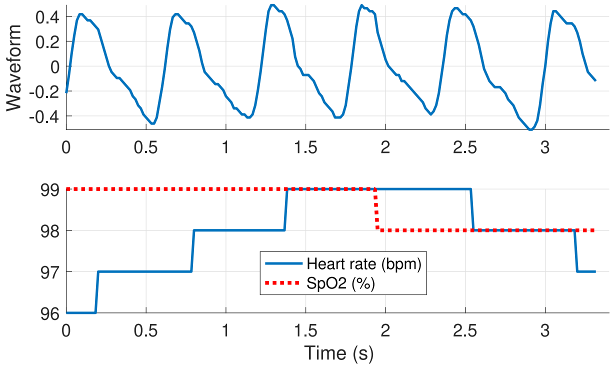

Fig. 4. Sample data recorded from the pulse oximeter, included into the collected DDPM dataset.

supplied with an interview script to follow. Subjects were motivated to deceive successfully through two levels of bonus compensation: if they were able to deceive the interviewer in five or six of the nine deceptive responses, they were given a 150 percent of a base incentive payment; the base payment was doubled if they were successfully deceptive in seven or more questions.

B. Interview Content

Interview questions comprised three categories: experiential, travel screening, and "superlative" questions. The first group inquires whether or not the subject had a particular experience within the prior year, *e.g.,* "Have you traveled outside of the United States in the last 12 months?" The second relates to an exercise conducted before the interview wherein the subject was given or asked to pack a suitcase with predetermined props, *e.g.,* "Do you currently have any foreign currency in your possession?" The final set of questions solicits the subject's opinion on their best/worst or favorite/least favorite experience, *e.g.,* "What is your favorite flavor of ice cream?" When answering such questions deceptively, subjects were to assert the opposite of the truth. The first three "warm up" questions were always to be answered honestly. They allowed the subject to get settled, and gave the interviewer an idea of the subject's demeanor when answering a question honestly. The order of the remaining questions and those selected for deception were randomly assigned for each subject.

Subjects were prepared for the interview by the Data Collection Coordinator (DCC). After signing a consent form, subjects were given a brief description of the recording equipment and an explanation of the experiment. The categories of questions were described, but not specific questions, with the exception of the superlative questions that they were to answer deceptively. The DCC emphasized that the more convincing the subject was vis-à-vis the interviewer, the more the subject would be compensated. At this point in interview preparation, either the subject or the DCC would pack the suitcase. The subject was then given a survey to indicate which questions to answer deceptively and verify that they had answered the question according to the assignment.

V. Collected Data

A. Deception Metadata

Age, gender, ethnicity, and race were recorded for all participants. Each of the 70 interviews consisted of 24 responses, 9 of which were deceptive. Overall, we collected 630 deceptive and 1050 honest responses. To our knowledge, the 1,680 annotated responses is the most ever recorded in a deception detection dataset.

The interviewee recorded whether they had answered as instructed for each question. For deceptive responses, they also rated how convincing they felt they were, on a 5-point Likert scale ranging from "I was not convincing at all" to "I was certainly convincing". The interviewer recorded their belief about each response, on a 5-point scale from "certainly the answer was deceptive" to "certainly the answer was honest". The data was additionally annotated to indicate which person (interviewer or interviewee) was speaking and the interval in time when they were speaking.

B. Data Post-Processing

The RGB and NIR videos were losslessly compressed. The interviews' average, minimum and maximum durations were 11 minutes, 8.9 minutes, and 19.9 minutes, respectively. In total, our dataset consists of 776 minutes of recording from all sensor modalities. The oximeter recorded SpO2, heart rate, and pulse waveform at 60 Hz giving average heart rates for the whole interview ranging from 40 bpm to 161 bpm.

To encourage reproducible research, we defined subject-disjoint training, validation, and testing sets, with stratified random sampling across demographic features. Table II shows the demographics for each set.

VI. Pulse Detection Experiments

Five pulse detection techniques were evaluated for Remote Photoplethysmography (rPPG) performance on the DDPM dataset, relying on **blind-source separation** [31], [32],

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.

TABLE II
NUMBER OF SUBJECTS IN VARIOUS DEMOGRAPHIC CATEGORIES ACROSS THE TRAINING, VALIDATION, AND TEST SETS

	Race						Gender			Age					
	White	**African American**	**Chinese**	**Asian Indian**	**Filipino**	**Other**	**Female**	**Male**	**Nonbinary**	**18-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60+**
Train	33	3	3	3	1	5	31	16	1	4	32	2	4	5	1
Val.	6	1	2	0	1	1	6	5	0	0	9	1	0	1	0
Test	6	1	1	1	0	2	7	4	0	0	7	2	1	1	0

TABLE III
COMPARISON BETWEEN BASELINE PULSE ESTIMATORS AND RPNET. MAE IS MEAN ABSOLUTE ERROR IN ESTIMATED HEART RATE RELATIVE TO GROUND TRUTH FORM THE PULSE OXIMETER, RMSE IS THE RMS ERROR RELATIVE TO THE SAME GROUND TRUTH, AND R_WAVE IS THE PEARSON CORRELATION COEFFICIENT BETWEEN THE ESTIMATED PULSE WAVEFORM AND THE GROUND TRUTH DATA OBTAINED FROM THE PULSE OXIMETER. ERROR ESTIMATES ARE 95% CONFIDENCE INTERVALS

Method	MAE	RMSE	r_wave
CHROM [33]	3.591	7.536	0.527
POS [34]	7.697	15.902	0.311
POH10 [31]	24.050	33.113	0.057
POH11 [32]	15.777	24.823	0.126
RPNet (ours)	**1.830** ± 0.556	**3.972** ± 1.325	**0.665** ± 0.052

chrominance and color space transformations [33], [34], and **deep learning** [35], [36]. All methods are the authors' implementations based on the published descriptions.

A. Methods

The general pipeline for pulse detection contains region selection, spatial averaging, a transformation or signal decomposition, and frequency analysis. For region selection, we used OpenFace [37] to detect 68 facial landmarks used to define a face bounding box. The bounding box was extended horizontally by 5% on each side, and by 30% above and 5% below, and then converted to a square with a side length that was the larger of the expanded horizontal and vertical sizes, to ensure that the cheeks, forehead and jaw were contained. For the chrominance-based approaches, we select the skin pixels within the face with the method of Heusch et al. [21].

Given the region of interest, we performed channel-wise spatial averaging to produce a 1D temporal signal for each channel. The blind source separation approaches apply independent component analysis (ICA) to separate the pulse from motion and noise. The chrominance-based approaches linearly combine the channels to define a robust pulse signal. The heart rate is then found over a time window by converting the signal to the frequency domain and selecting the peak frequency f_p as the cardiac pulse. The heart rate is computed as $\widehat{HR} = 60 \times f_p$ beats per minute (bpm).

For training the deep learning-based approach, we employed RPNet [36], a 3D Convolutional Neural Network (3DCNN) [35] that is fed with the face cropped at the bounding box and downsized to 64 × 64 pixels with bicubic interpolation. When training and inferring using RPNet, the model is given clips of the video consisting of 136 frames (1.5 seconds) and outputs a 136 element pulse waveform. We selected 136 frames as it is the minimum time for an entire heartbeat to occur, considering 40 bpm as a lower bound for average subjects. RPNet was trained to minimize the negative Pearson correlation between predicted and normalized ground truth pulse waveforms. We use the Adam optimizer with a learning rate of $\alpha = 0.0001$, and parameter values of $\beta_1 = 0.99$ and $\beta_2 = 0.999$ to train the model for 50 epochs, then select the model with the lowest loss on the validation set as our final model.

The oximeter recorded ground truth waveform and heart rate estimates at 60 Hz, which we upsampled to 90 Hz to match the RGB camera frame rate. One of the difficulties in defining an oximeter waveform as a target arises from the phase difference observed at the face and finger, coupled with time lags from the acquisition apparatus [38]. To mitigate the phase shift, we use the output waveform predicted by CHROM [33] (chosen because it does not require supervised training) to shift the ground truth waveform such that the cross-correlation between them is maximized. Our ground truth waveforms contain infrequent noisy segments caused by subjects moving their fingers inside the pulse oximeter. We detect these patches as jumps in heart rate over 7 bpm in a second, as calculated using a FFT with bandpass bounds of 40 and 180 bpm and a sliding window of 10 seconds. If such a jump occurs, that 10 second FFT window is marked as invalid and masked from the dataset.

For videos longer than the clip length of 136 frames it is necessary to perform predictions in sliding window fashion over the full video. Similar to [33], we use a stride of half the clip length to slide across the video. The windowed outputs are standardized, a Hann function is applied to mitigate edge effects from convolution, and they are added together to produce a single value per frame.

B. Evaluation of rPPG

Pulse detection performance is analyzed by calculating the error between predicted and ground truth heart rates. The heart rate is calculated by applying a 10 second wide Hamming window to the signal and converting to the frequency domain, from which the index of the maximum spectral peak between $0.6\overline{6}$ Hz and 3 Hz (40 bpm to 180 bpm) is selected as the heart rate. Since the frequency domain suffers quantization effects, we dequantize spectral peaks by taking the weighted average of spectral readings between adjacent valleys. We used metrics from the rPPG literature to evaluate performance, such as mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient for the pulse waveform, r_wave, as shown in Table III. We found that while masking out noisy sections from the ground truth improved evaluation metrics for CHROM and RPNet, it degraded results for the other methods. As such, we only apply masking to CHROM and RPNet.

The original blind-source separation approach, POH10 [31], is outperformed by POH11 [32] due to signal detrending and filtering, which removes noise from motion. Both chrominance-based approaches perform similarly, although

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.

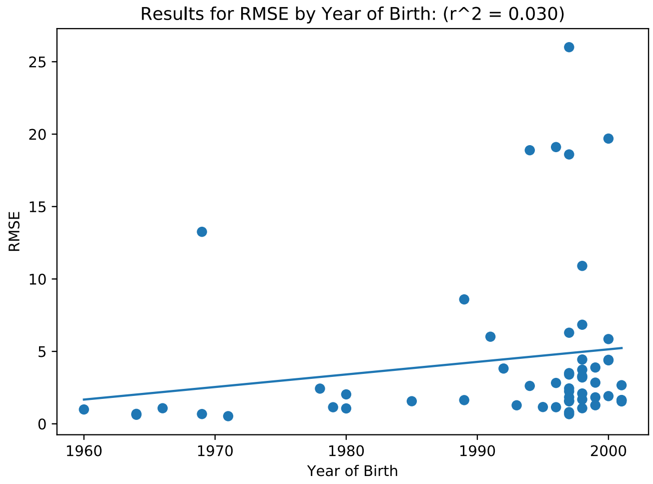

Fig. 5. Older subjects do not appear to have degraded performance compared to younger subjects (P-value = 0.23).

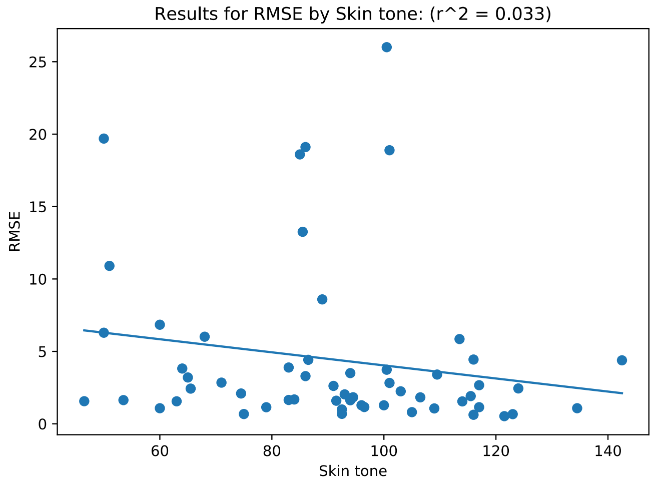

Fig. 6. RPNet does not appear to be significantly affected by skin tone (P-value = 0.13).

POS [34] gives good accuracy without filtering. We evaluate RPNet using 5-fold cross validation, reporting the mean and 95% confidence interval for each performance metric. We observe that RPNet outperforms the non-deep learning baselines. We investigate characteristics of its performance in the following sections.

C. Demographics Study

In machine learning fields, algorithmic bias, in which the model behaves differently on different types of input, is an important issue [39]. Certain demographics such as skin tone and age affect light's interaction and reflection from the tissue. Recent studies have shown that rPPG algorithms perform worse on dark-skinned subjects [40]. A performance drop is likely influenced by two main factors: 1) insufficient diversity in model training data and 2) stronger light absorption from melanin, which decreases the intensity of the reflected light. Consequently, we were interested to see if RPNet was similarly effected by skin tone and age.

To investigate this issue, we evaluated RPNet performance on the demographics in the test sets for each of the 5 folds in our 5-fold cross validation. We did not collect ground truth skin tone measurements, so we estimate the skin tone as the average grayscale pixel value of a subject's face as suggested in [39]. The results for the age study are shown in Fig. 5, and the results for the study on skin tone are shown in Fig. 6.

In both demographics scenarios (*i.e.,* age and skin tone), we were unable to find evidence showing that the performance of RPNet was affected by demographic variables (obtaining P-values > 0.05), nor did we prove that RPNet is indifferent to these variables. Since these studies had a small sample size we intend to perform a followup study after collecting from a more diverse pool of subjects. A further limitation is that our skin tone estimation based on the brightness of the subject's face is confounded by factors such as lighting, which while controlled in a laboratory setting, has been shown to add noise to skin tone estimates [39]. Therefore, we plan to collect ground truth skin tone data using a dedicated dermatological device in order to provide more reliable skin tone measurements.

TABLE IV
5-Fold Testing Results for RPNet on Lower Frame Rates. FPS Means Frames Per Second; MAE Is Mean Absolute Error (MAE); RMSE Is Root Mean Squared Error, and r_wave Is Pearson Correlation Coefficient Between Predicted and Ground-Truth Pulse Waveforms. Error Estimates Are 95% Confidence Intervals

FPS	MAE	RMSE	r_wave
90	1.830 ± 0.556	3.972 ± 1.325	0.665 ± 0.052
45	2.947 ± 1.509	5.660 ± 2.255	0.652 ± 0.048
30	3.335 ± 1.949	5.860 ± 2.451	0.647 ± 0.051
22.5	3.899 ± 3.074	6.407 ± 3.861	0.639 ± 0.066
18	3.546 ± 1.198	6.280 ± 1.565	0.643 ± 0.055
15	4.785 ± 2.731	8.008 ± 3.426	0.605 ± 0.067

D. Frame Rate Study

We are interested in applying RPNet to videos of various sources. As a prerequisite, we retrained RPNet on the following frame rates: 45, 30, 22.5, 18, and 15 fps. The data for this purpose was obtained by downsampling the DDPM videos by averaging pixel values across adjacent frames.

We performed 5-fold cross-validation, using a validation set size of 17 videos and a test set size of 17 videos. We selected models based on validation loss. The evaluation results are given in Table IV.

From this study we conclude that RPNet has somewhat degraded yet comparable performance on lower frame rates. This is corroborated by prior findings as published in [36].

E. Spatial Analysis

We worked to identify which regions of the face produce the best rPPG results. While we have utilized Grad-CAM [41] in the past [42], we have found that it reveals the regions of interest as learned by the network rather than the regions of interest generally suitable for rPPG, which was our goal in this experiment. The heatmap in Fig. 7 was created by performing

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.

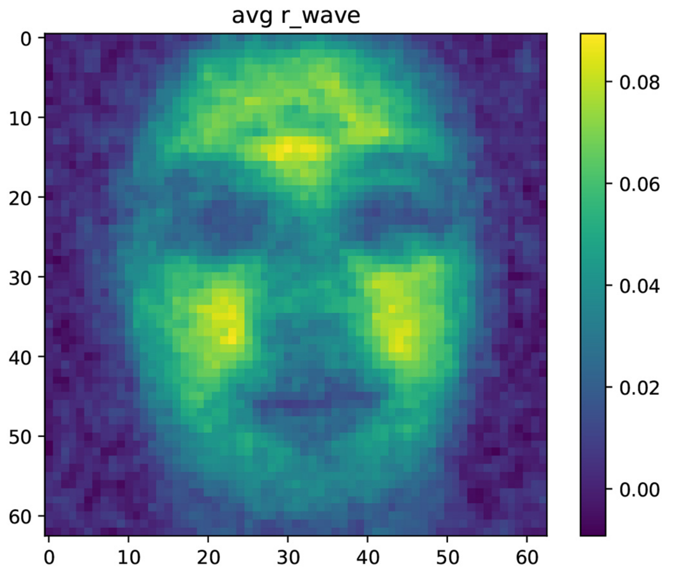

Fig. 7. The correlation between inferred and ground truth rPPG signals at each facial region. The cheeks and forehead give a rPPG signal that is somewhat more correlated with the ground truth than other parts of the face.

an evaluation using (for each subject) a 2 × 2 pixel region from every location across the 64 × 64 pixel video, which were then scaled to the 64 × 64 input for RPNet. For each of the 63^2 regions we obtained the Pearson r correlation between the waveform and the ground truth, then averaged this result across subjects to obtain the intensity of a single pixel in the heatmap.

From the image, we see that the cheeks and forehead produce a better rPPG wave than other facial skin, which is plausible since those regions are more highly vascularized than other parts of the face. We found in this study that the two subjects wearing lipstick in the set did not appear to have as strong of an rPPG signal on the cheeks and forehead. Unfortunately, we did not collect ground truth on which subjects wore makeup, so while we believe it is likely that topical cosmetics will have a detrimental effect on rPPG, more study is needed in this area.

We investigated whether RPNet model, proposed by these authors, could be improved by focusing it on regions with a stronger signal, *i.e.,* the forehead and cheeks. To this end we divided the face into the three regions (forehead, right cheek, left cheek) as shown in Fig. 8. Using the models trained over the full face, we inferred an rPPG wave over these regions. The results for the individual regions and for combining them by averaging wave values is given in Table V. We found that the forehead obtained the most accurate results of the subregions, although even when the three regions are combined, RPNet utilizing the full frame still outperforms these more focused regions. By this we conclude that RPNet may have already learned from the data to pay attention to the most important facial regions.

F. Landmarker Study

We had chosen the OpenFace landmarker in our image processing pipeline for the purpose of generating bounding boxes because it exhibits superior landmark stability, resulting in a low amount of jitter in the landmarks. However, we have also investigated MediaPipe [43] as an alternative due to its applicability in real time systems.

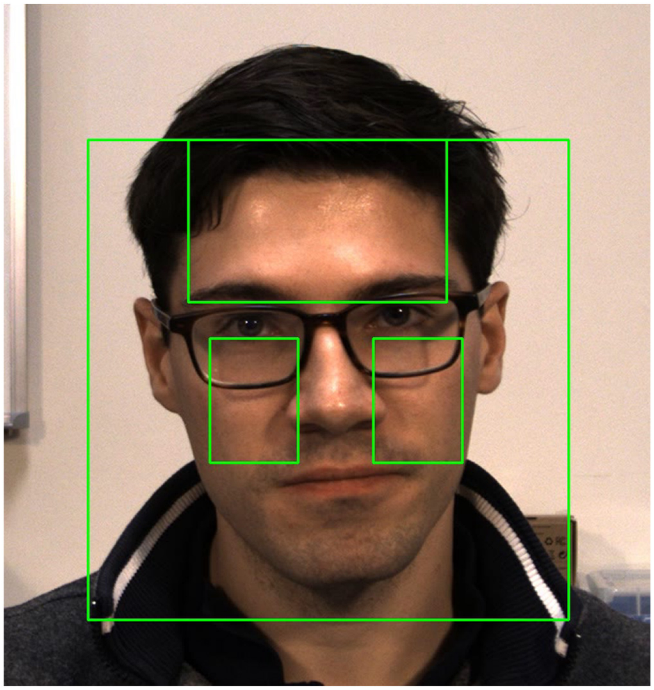

Fig. 8. Bounding boxes around regions of interest: forehead and cheeks.

TABLE V
Error Metrics Obtained From the RPNet Model on Various Face Subregions. MAE Is Mean Absolute Error (MAE); RMSE Is Root Mean Squared Error, and r_wave Is Pearson Correlation Coefficient Between Predicted and Ground-Truth Pulse Waveforms. Error Estimates Are 95% Confidence Intervals

Region	MAE	RMSE	r_wave
Forehead	2.597 ± 1.272	5.571 ± 1.828	0.631 ± 0.063
Left Cheek	5.203 ± 1.714	10.352 ± 2.040	0.542 ± 0.064
Right Cheek	6.597 ± 2.520	11.747 ± 2.997	0.525 ± 0.074
Combined	2.270 ± 1.034	5.262 ± 1.993	0.579 ± 0.090

We calculated bounding boxes using MediaPipe for the videos in the test set and performed an evaluation comparing the OpenFace and MediaPipe data. The evaluation results are given in Table VI. We notice that MediaPipe performs nearly as well as OpenFace as a bounding box method for rPPG, despite its poorer bounding box stability (as measured by average displacement in pixels between successive frames), exhibiting an increase in MAE by only 33% despite a degradation in bounding box stability by 260%. As it is a lighter weight solution than OpenFace, we expect MediaPipe to be useful for real time rPPG systems.

VII. Pupil Size Estimation

The general pipeline for pupil detection contains eye region selection, and the estimation of the pupil and iris ellipse parameters. For selecting the eye region, we utilized OpenFace facial landmarks (as in pulse detection pipeline) and used the

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.

TABLE VI

Error Metrics for Two Landmarkers Used in This Study. MAE Is Mean Absolute Error (MAE); RMSE Is Root Mean Squared Error, and r_wave Is Pearson Correlation Coefficient Between Predicted and Ground-Truth Pulse Waveforms; Stability Is Bounding Box Movement in Pixels Per Frame. Error Estimates Are 95% Confidence Intervals

Landmarker	MAE	RMSE	r_wave	stability
OpenFace	1.830 ± 0.556	3.972 ± 1.325	0.665 ± 0.052	0.689
MediaPipe	2.435 ± 1.002	5.219 ± 2.185	0.631 ± 0.053	2.480

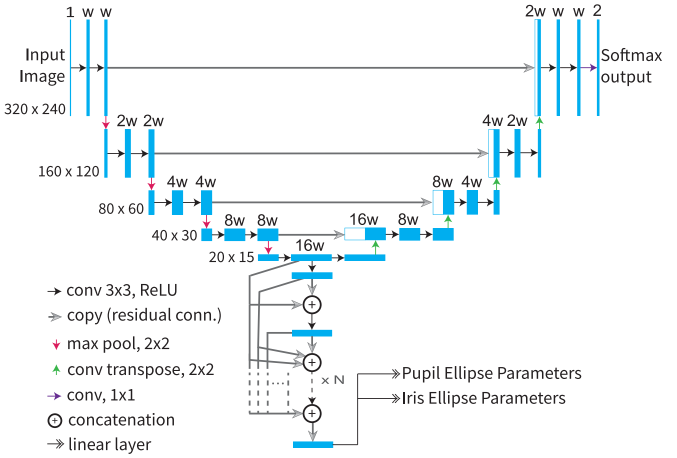

Fig. 9. Modified CC-Net architecture for estimating iris and pupil ellipeses and masks.

points around the eyelids to define a eye bounding box. We convert the bounding box to have a 4:3 aspect ratio (which corresponds to the aspect ratio of iris images recommended by ISO/IEC 19794-6) by lengthening the shorter side (which is usually the vertical side).

To detect the pupil and iris ellipse parameters, we utilize a modified CC-Net architecture [44]. In particular, we use the encodings from the CC-Net to train a Convolutional Neural Network-based regressor with dense residual connections to approximate the parameters of ellipses fitting the iris and pupil, as illustrated in Fig. 9. In other words, the left part of the CC-Net model, along with the extra regression layers, serve as a one-shot end-to-end approximator of the iris and pupil ellipses, directly from the input image. In addition, the right part of the CC-Net model predicts the iris segmentation mask. Hence this model, apart from its general applications to pupil size estimation, may be a convenient image pre-processing tool for iris recognition, where both the segmentation mask and locations/sizes of pupil and iris are needed.

To train our modified CC-Net architecture, we utilize the combination of multiple publicly-available iris recognition datasets (Biosec, BATH database, ND0405, UBIRIS and CASIA-V4-Iris-Interval) previously used in [45]. This combined dataset comes with manual ground-truth segmentations, but does not contain any annotations for the pupil and iris ellipse parameters. Also, off-axis eye images are very rare in these data sets. To find ellipse parameters, we applied Hough transform to find iris and pupil ellipsoidal approximations from the segmentation masks. To augment the training data with off-axis samples, we applied perspective transformation to randomly deform the eye image, mask and recalculate the ellipse parameters in the 3D space, accordingly. Finally, we jointly train the modified CC-Net architecture to predict both the mask, and the pupil and iris ellipsoidal approximations, through a loss function combining both tasks (segmentation and regression).

TABLE VII

Comparison of Different CNN-Based Regressors in the Pupil/Iris Detection Model, Shown in Fig. 9. ‘Conv × N, Indicates That There Are N Convolutional Layers. ‘DenseConv × N’ Indicates That There Are N Convolutional Layers Connected via Dense Residual Connections. We Have ‘W’ as the Number of Channels After the Input Goes Through the First Convolution (Same as in Fig. 9). ‘IoU’ Means Intersection Over Union. Avg. IoU Is Simply (Pupil IoU + Iris IoU)/2. For Avg. IoU, Both Mean and Standard Deviation Are Provided

Model	Pupil IoU	Iris IoU	Avg. IoU (Pupil and Iris)
Conv ×1 w=4	0.4436	0.6719	0.5578 ± 0.0038
Conv ×2 w=4	0.5408	0.6877	0.6143 ± 0.0058
Conv ×3 w=4	0.5510	0.7428	0.6469 ± 0.0047
Conv ×4 w=4	0.5797	0.7283	0.6540 ± 0.0050
Conv ×10 w=4	0.6015	0.7692	0.6853 ± 0.0047
DenseConv ×1 w=4	0.5767	0.6942	0.6355 ± 0.0059
DenseConv ×2 w=4	0.5943	0.7410	0.6677 ± 0.0040
DenseConv ×3 w=4	0.5844	0.7809	0.6826 ± 0.0063
DenseConv ×4 w=4	0.6221	0.7700	0.6961 ± 0.0038
DenseConv ×10 w=4	0.6500	0.7900	0.7200 ± 0.0052
DenseConv ×10 w=8	**0.6664**	**0.7982**	**0.7323 ± 0.0041**

To validate the performance of the model on the DDPM dataset, we extract eye regions, ensuring that the eye is open, from the DDPM dataset and manually annotate ellipses for pupil and iris. To evaluate the ellipse detection on this dataset, we randomly warp the images and annotated ellipses as before. As the augmentation on the validation dataset is done randomly, we repeat the validation step 10 times and average the results.

We experimented with different architectures for the CNN-based regressor in the modified CC-Net model. Table VII shows *Intersection over Union* results on the eye regions extracted from the DDPM dataset for different architectures. Dense residual connections improve the results as we use deeper networks. While making the regressor deeper could potentially further increase the IoU, our goal was also to let this model work in real time on a CPU, hence we stopped after building a model with 10 convolutional layers incorporating dense residual connections.

We utilize the ‘DenseConv × 10 w = 8’ model, which consists of 10 convolutional layers with dense residual connections and starts with a channel width of 8 after the first convolution (see Fig. 9), as the CNN regressor for our deception detection experiments.

VIII. Fusion

We investigated using fusion between several techniques for deception detection. In particular, we investigated rPPG as

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
ABLATION STUDY DATA WITH EQUAL ERROR RATE (EER) AND AREA UNDER THE ROC CURVE

Features	EER	AUC
rPPG	0.3853	0.6557
Pupil	0.4498	0.5698
Saccades	0.4882	0.5284
Blink_EAR	0.4735	0.5528
Thermal	0.4380	0.5825
Fusion (all)	0.3586 ± 0.0012	0.6918 ± 0.0012
Fusion (rPPG, Pupil, Thermal, Blink)	**0.3571 ± 0.0012**	0.6928 ± 0.0013
Fusion (rPPG, Pupil, Thermal)	0.3572 ± 9.6773e-04	**0.6945 ± 9.5610e-04**
Fusion (rPPG, Pupil)	0.3748 ± 7.7375e-04	0.6679 ± 8.6410e-04
Fusion (rPPG, Thermal)	0.3581 ± 8.8042e-04	0.6894 ± 6.1589e-04

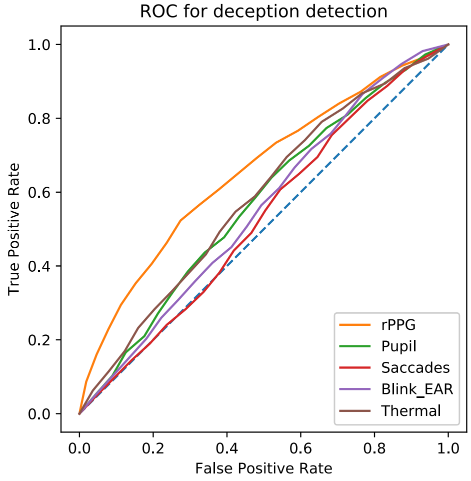

Fig. 10. ROC curves for the deception detection obtained for all individual features considered in this study.

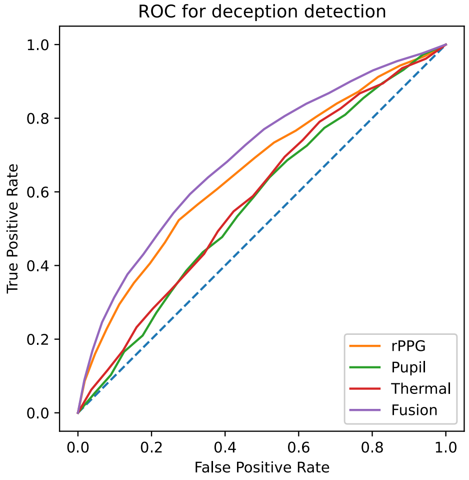

Fig. 11. Fusion results for rPPG, Pupil, and Thermal data using a linear SVM.

outlined in Section VI, Pupillometry as outlined in Section VII, Saccadic eye movement data as reported in [11], Blink rate as in [46], and the Temperature data collected using our thermal camera. In a past work we also investigated the use of Microexpressions [11], however, we found state-of-the-art microexpression detection techniques to be unreliable for deception detection, so those are not included here. Furthermore, we performed feature fusion using a linear Support Vector Machine (SVM), and performed an ablation study. An ROC curve depicting the utility for these methods as used on their own is shown in Fig. 10.

To perform our ablation study, we trained a linear SVM with the default parameters supplied by the scikit-learn python package [47] (*i.e.,* C = 1.0, gamma = 1/(N*variance) where applicable) using features derived from a combination of the five techniques (*i.e.,* rPPG, Pupillometry, Saccades, Blinks, and Thermal). We define these features as follows:

- *rPPG:* The mean relative heart rate.
- *Pupillometry:* The mean relative pupil diameter.
- *Saccades:* The mean relative saccade rate, with saccades detected based on the angular velocity of the eye as in [11].
- *Blinks:* The mean relative blink rate, with blinks detected using the Eye Aspect Ratio technique as in [46].
- *Thermal:* The mean relative facial temperature as detected using the FLIR C2 thermal camera.

Each of the above features are mean values calculated over the intervals when the subject is responding deceptively, relative to the subject's mean response over the entire interview. These features are further normalized by subtracting the mean and dividing by the standard deviation across the training set.

We trained the SVM on half of the data (partitioning by subject and determined randomly) and obtained class probabilities for the other half, then switched, thus obtaining class probabilities for the full dataset. These class probabilities were then used to generate an ROC curve for the fusion technique. This was repeated 100 times for the purpose of calculating the 95% confidence intervals.

In our ablation study we determined that the best combination of features for deception detection is rPPG, Pupil, and Thermal data, as this combination maximizes the Area under the ROC Curve (AUC), as shown in Table VIII. The ROC curves for this subset of features plus an ROC curve averaged over the 100 fusion trials is shown in Fig. 11.

Of the features, rPPG performs the best for deception detection. This is likely because the heart rate is tightly correlated to stress, is difficult for subjects to control, and changes relatively quickly when the subject enters the stressful situation of deception (as opposed to, *e.g.,* Thermal). Other features such as the blink and saccade rates have a near random correlation because they are easier for a subject to control. We believe that adding such additional features, all of which are weakly

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.

TABLE IX
FUSION UNDER VARIED SVM KERNELS

Kernel	EER	AUC
Linear	**0.3572 ± 9.6773e-04**	**0.6945 ± 9.5610e-04**
Poly (2nd Degree)	0.5016 ± 0.0036	0.5006 ± 0.0042
Poly (3rd Degree)	0.4821 ± 0.0114	0.5222 ± 0.0150
Poly (4th Degree)	0.4932 ± 0.0031	0.5105 ± 0.0040
Sigmoid	0.4263 ± 0.0023	0.5962 ± 0.0030
RBF	0.3611 ± 0.0015	0.6798 ± 0.0014

correlated with deceit, causes the model to overfit to the noise in the features.

We investigated varying the SVM kernel to optimize fusion performance, with results in Table IX. We found that the linear kernel yields the best performance for the deception detection task. We believe that, as was the case with feature selection, more complex SVM kernels are more able to fit to the noise in the features, thus leading to poor results when compared to the more simple linear kernel.

In conclusion, we have determined that feature fusion combining rPPG, Pupil, and Thermal data with a linear SVM yields improved deception detection results as compared to any of these features alone; feature fusion obtains an EER of 0.357, whereas the best solo feature, rPPG, obtains an EER of 0.385.

While additional techniques may be applied such as deep learning, we believe that due to the weak and near random correlation between our features and deception that further analysis invites an optimistic interpretation of features that may be correlated only due to random chance, as argued by [29]. Rather, we believe that effort is better spent developing and refining the individual feature extractors (for measuring rPPG, pupil size, saccades, blinking and spatial distribution of face temperature), which are applicable in domains beyond deception detection. Therefore, in addition to analysis of deception detection as the main application, we also provided detailed experiments to assess the value of several individual feature extractors.

IX. Conclusion

We present the Deception Detection and Physiological Monitoring (DDPM) dataset, the most comprehensive dataset to date in terms of number of different modalities and volume of raw video, to support exploration of deception detection and remote physiological monitoring in a natural conversation setup. The sensors are temporally synchronized, and imaging across visible, near-infrared and thermal spectra provides more than 8 million high-resolution images from almost 13 hours of recordings in a deception-focused interview scenario.

Along with this dataset, we provide baseline results for heart rate detection, and the feasibility of deception detection using pupillometry, heart rate, and thermal data. We have determined that feature fusion using these three features obtains an equal error rate of 0.357, exceeding any of these features on their own. This new dataset and evaluation protocols are made publicly available to further advance research in the areas of remote physiological monitoring and deception detection.

Acknowledgment

The authors would like to thank Marybeth Saunders for conducting the interviews during data collection. This research was sponsored by the Securiport Global Innovation Cell, a division of Securiport LLC. Commercial equipment is identified in this work in order to adequately specify or describe the subject matter. In no case does such identification imply recommendation or endorsement by Securiport LLC, nor does it imply that the equipment identified is necessarily the best available for this purpose. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of our sponsors.

References

[1] Y. Sun and N. Thakor, “Photoplethysmography revisited: From contact to noncontact, from point to imaging,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 463–477, Mar. 2016.

[2] Z. Zhang, V. Singh, T. E. Slowe, S. Tulyakov, and V. Govindaraju, “Real-time automatic deceit detection from involuntary facial expressions,” in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, 2007, pp. 1–6.

[3] N. Bhaskaran, I. Nwogu, M. G. Frank, and V. Govindaraju, “Lie to me: Deceit detection via online behavioral learning,” in *Proc. Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2011, pp. 24–29.

[4] A. Vrij, K. Edward, K. P. Roberts, and R. Bull, “Detecting deceit via analysis of verbal and nonverbal behavior,” *J. Nonverbal Behav.*, vol. 24, no. 4, pp. 239–263, 2000.

[5] G. Duran, I. Tapiero, and G. A. Michael, “Resting heart rate: A physiological predicator of lie detection ability,” *Physiol. Behav.*, vol. 186, pp. 10–15, Mar. 2018. [Online]. Available: https://doi.org/10.1016/j.physbeh.2018.01.002

[6] A. Vrij, J. Oliveira, A. Hammond, and H. Ehrlichman, “Saccadic eye movement rate as a cue to deceit,” *J. Appl. Res. Memory Cogn.*, vol. 4, no. 1, pp. 15–19, 2015.

[7] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, “Bag-of-lies: A multimodal dataset for deception detection,” in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit. Workshops*, 2019, pp. 83–90.

[8] K. Radlak, M. Bozek, and B. Smolka, “Silesian deception database: Presentation and analysis,” in *Proc. ACM Workshop Multimodal Deception Detection*, 2015, pp. 29–35. [Online]. Available: https://doi.org/10.1145/2823465.2823469

[9] B. A. Rajoub and R. Zwiggelaar, “Thermal facial analysis for deception detection,” *IEEE Trans. Inf. Forensics Security*, vol. 9, pp. 1015–1023, 2014.

[10] M. Owayjan, A. Kashour, N. Al Haddad, M. Fadel, and G. Al Souki, “The design and development of a lie detection system using facial micro-expressions,” in *Proc. Int. Conf. Adv. Comput. Tools Eng. Appl.*, 2012, pp. 33–38.

[11] J. Speth, N. Vance, A. Czajka, K. W. Bowyer, D. Wright, and P. Flynn, “Deception detection and remote physiological monitoring: A dataset and baseline experimental results,” in *Proc. Int. Joint Conf. Biometrics (IJCB)*, 2021, pp. 4264–4271.

[12] A. Turnip, M. F. Amri, H. Fakrurroja, A. I. Simbolon, M. A. Suhendra, and D. E. Kusumandari, “Deception detection of EEG-P300 component classified by SVM method,” in *Proc. Int. Conf. Softw. Comput. Appl.*, New York, NY, USA, 2017, pp. 299–303. [Online]. Available: https://doi.org/10.1145/3056662.3056709

[13] P. Ekman, “Lie catching and microexpressions,” *Philos. Deception*, vol. 1, no. 2, pp. 118–136, 2009.

[14] F. Soldner, V. Pérez-Rosas, and R. Mihalcea, “Box of lies: Multimodal deception detection in dialogues,” in *Proc. NAACL-HLT*, 2019, pp. 1768–1777.

[15] V. Pérez-Rosas, R. Mihalcea, A. Narvaez, and M. Burzo, “A multimodal Dataset for deception detection,” in *Proc. Int. Conf. Lang. Resour. Eval.*, Reykjavik, Iceland, 2014, pp. 3118–3122. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/869_Paper.pdf

[16] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, “Deception detection using real-life trial data,” in *Proc. Int. Conf. Multimodal Interact.*, 2015, pp. 59–66. [Online]. Available: https://doi.org/10.1145/2818346.2820758

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.

[17] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan.–Mar. 2012.

[18] R. Stricker, S. Muller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *Proc. Int. Symp. Robot Human Interact. Commun.*, 2014, pp. 1056–1062.

[19] J. R. Estepp, E. B. Blackford, and C. M. Meier, "Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography," in *Proc. IEEE Int. Conf. Syst. Man Cybern. (SMC)*, 2014, pp. 1462–1469.

[20] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2396–2404.

[21] G. Heusch, A. Anjos, and S. Marcel, "A reproducible study on remote heart rate measurement," 2017, *arXiv:1709.00962*.

[22] X. Niu, H. Han, S. Shan, and X. Chen, "VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 562–576.

[23] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognit. Lett.*, vol. 124, pp. 82–90, Jun. 2019. [Online]. Available: https://doi.org/10.1016/j.patrec.2017.10.017

[24] P. Ekman and E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. New York, NY, USA: Oxford Univ. Press, 1997.

[25] J. T.-Y. Wang, M. Spezio, and C. F. Camerer, "Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games," *Amer. Econ. Rev.*, vol. 100, no. 3, pp. 984–1007, 2010.

[26] L. Caso, F. Maricchiolo, M. Bonaiuto, A. Vrij, and S. Mann, "The impact of deception and suspicion on different hand movements," *J. Nonverbal Behav.*, vol. 30, no. 1, pp. 1–19, 2006.

[27] N. Michael, M. Dilsizian, D. Metaxas, and J. K. Burgoon, "Motion profiles for deception detection using visual cues," in *Proc. ECCV*, 2010, pp. 462–475.

[28] T. Brennen and S. Magnussen, "Research on non-verbal signs of lies and deceit: A blind alley," *Front. Psychol.*, vol. 11, Dec. 2020, Art. no. 613410.

[29] T. J. Luke, "Lessons from pinocchio: Cues to deception may be highly exaggerated," *Perspect. Psychol. Sci.*, vol. 14, no. 4, pp. 646–671, 2019.

[30] H. Nasri, W. Ouarda, and A. M. Alimi, "ReLiDSS: Novel lie detection system from speech signal," in *Proc. Int. Conf. Comput. Syst. Appl.*, 2016, pp. 1–8.

[31] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Opt. Exp.*, vol. 18, no. 10, pp. 10762–10774, May 2010. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-18-10-10762

[32] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a Webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[33] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Trans. Biom. Eng.*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.

[34] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote PPG," *IEEE Trans. Biom. Eng.*, vol. 64, no. 7, pp. 1479–1491, Jul. 2017.

[35] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–12.

[36] J. Speth, N. Vance, A. Czajka, K. Bowyer, and P. Flynn, "Unifying frame rate and temporal dilations for improved remote pulse detection," *Comput. Vis. Image Understand.*, vol. 210, pp. 1056–1062, Sep. 2021.

[37] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.

[38] Q. Zhan, W. Wang, and G. de Haan, "Analysis of CNN-based remote-PPG to understand limitations and sensitivities," *Biomed. Opt. Exp.*, vol. 11, no. 3, pp. 1268–1283, 2020.

[39] J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Reliability and validity of image-based and self-reported skin phenotype metrics," 2021, *arXiv:2106.11240*.

[40] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "A meta-analysis of the impact of skin tone and gender on non-contact Photoplethysmography measurements," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2020, pp. 284–285.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[42] J. Speth, N. Vance, P. Flynn, K. Bowyer, and A. Czajka, "Remote pulse estimation in the presence of face masks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2086–2095.

[43] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.

[44] S. Mishra, P. Liang, A. Czajka, D. Z. Chen, and X. S. Hu, "CC-NET: Image complexity guided network compression for biomedical image segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, 2019, pp. 57–60.

[45] M. Trokielewicz, A. Czajka, and P. Maciejewicz, "Post-mortem iris recognition resistant to biological eye decay processes," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Snowmass, CO, USA, Mar. 2020, pp. 2296–2304. [Online]. Available: https://doi.org/10.1109/WACV45572.2020.9093281

[46] J. Cech and T. Soukupova, "Real-time eye blink detection using facial landmarks," in *Proc. 21st Comput. Vis. Winter Workshop*, 2016, pp. 1–8.

[47] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

Authorized licensed use limited to: Pontificia Universidad Javeriana. Downloaded on July 25,2024 at 20:45:57 UTC from IEEE Xplore. Restrictions apply.