Evaluating Visual Photoplethysmography Method

Debjyoti Talukdar [1] , Luis Felipe de Deus [2] , Nikhil Sehgal [2]

1. Medical Research, Mkhitar Gosh Armenian-Russian International University, Yerevan, ARM 2. AI Research, Vastmindz Limited, London, GBR

Corresponding author: Debjyoti Talukdar, drtalukdard@aol.com

Abstract

Regular monitoring of common physiological signs, including heart rate, blood pressure, and oxygen saturation, can be an effective way to either prevent or detect many kinds of chronic conditions. In particular, cardiovascular diseases (CVDs) are a worldwide concern. According to the World Health Organization, 32% of all deaths worldwide are from CVDs. In addition, stress-related illnesses cost $190 billion in healthcare costs per year. Currently, contact devices are required to extract most of an individual's physiological information, which can be uncomfortable for users and can cause discomfort. However, in recent years, remote photoplethysmography (rPPG) technology is gaining interest, which enables contactless monitoring of the blood volume pulse signal using a regular camera, and ultimately can provide the same physiological information as a contact device. In this paper, we propose a benchmark comparison using a new multimodal database consisting of 56 subjects where each subject was submitted to three different tasks. Each subject wore a wearable device capable of extracting photoplethysmography signals and was filmed to allow simultaneous rPPG signal extraction. Several experiments were conducted, including a comparison between information from contact and remote signals and stress state recognition. Results have shown that in this dataset, rPPG signals were capable of dealing with motion artifacts better than contact PPG sensors and overall had better quality if compared to the signals from the contact sensor. Moreover, the statistical analysis of the variance method had shown that at least two heart-rate variability (HRV) features, NNi 20 and SAMPEN, were capable of differentiating between stress and non-stress states. In addition, three features, inter-beat interval (IBI), NNi 20, and SAMPEN, were capable of differentiating between tasks relating to different levels of difficulty. Furthermore, using machine learning to classify a "stressed" or "unstressed" state, the models were able to achieve an accuracy score of 83.11%.

Categories: Cardiology, Healthcare Technology
Keywords: blood pressure, cardiovascular disease, heart rate variability, stress recognition, remote photoplethysmography

Introduction

There is growing interest in technologies related to remote patient monitoring (RPM) solutions, an interest that has largely piqued as of late amid the COVID-19 pandemic. Furthermore, these technologies can be used to monitor several disorders. Cardiovascular disease (CVD) is a group of disorders related to the heart and blood vessels. According to the World Health Organization, 32% of all deaths worldwide are attributed to CVDs [1]. The most common health problems include the risk of a heart attack, stroke, and heart failure.

In addition, stress-related illnesses place a significant burden on the global healthcare system. Data from the American Institute of Stress show that 83% of US workers suffer from work-related stress [2]. As a result of work stress, US businesses lose up to $300 billion yearly, and work-related stress issues cause up to 120,000 deaths and result in $190 billion in healthcare costs per year.

RPM solutions are being developed for a multitude of customers of varying ages and varying health conditions of patients are efficiently being monitored outside traditional settings with increased accessibility to regular physiological assessments. Regular physical evaluations of a patient's health allow the capture of longitudinal data that can be used to provide a highly personalized level of care and has the potential to predict the onset of severe health problems.

Well-established methods for capturing physiological data include the use of electrocardiograms (ECG) or photoplethysmography (PPG), both of which require the use of contact sensors. These methods, through their physiological signals, have the ability and the potential to measure several physiological parameters such as heart rate (HR), respiration rate, oxygenation, and mental stress.

One of the main parameters used in these studies is heart-rate variability (HRV), which can be accessed through the time difference between two peaks of the signal, the ECG with the R part and the PPG with the systolic peak, which in both cases can simply be named the R-R interval.

Since HRV can be deduced from both methods, it is well established that HRV is an accurate noninvasive proxy for measuring changes in the autonomic nervous system and thus accesses the patient's psychological

information. Kim et al. [3] investigated the assumption of correlating HRV and stress levels. Their findings show that HRV features changed in response to stress induced by different methods. HRV characteristics change in association with low parasympathetic activity, which is characterized by a decrease in the high-frequency band (0.16 Hz-0.4 Hz) and an increase in the low-frequency band (0.04 Hz-0.15 Hz).

HRV can also be used to measure potential cardiovascular issues. HRV has the potential to assess autonomic imbalances, diseases, and mortality as it is associated with CVD risk factors like hypertension, obesity, work-related stress, and family history [4]. Although contact-based methods are noninvasive and can capture valuable information about an individual's health and wellness, they can irritate those with sensitive skin, and such physical devices that capture ECG or PPG signals may not be easily accessible.

Researchers have recently introduced the remote photoplethysmography (rPPG) technique, which is a low-cost, non-contact method and an alternative solution for measuring the same parameters as the PPG signal does in a contactless way. Since it is a method that can be performed on any consumer technology device with an embedded camera, its ease of use makes it an attractive addition to the suite of RPM solutions.

The information acquired through rPPG reflects the variations of blood volume in skin tissue which is modified by cardiac activity. The reflection of light is influenced by the change in the volume of blood and the movement of the wall of blood vessels; this phenomenon is visible within frame-to-frame changes of a red, green, and blue (RGB) camera. There are, however, several challenges when attempting to retrieve an optimal rPPG signal. Distortion to clean rPPG signals mostly arises from low illumination, significant head movement, a camera's frame rate, and its resolution [4].

rPPG methods are usually carried out using a four-step methodology, which can be summarized as a frame-to-frame extraction, region of interest (ROI) detection, signal processing, and vitals estimation. First, the video files are usually separated into several frames, and the amount of frames in a certain period is denoted as frame rate, measured in frames per second (FPS). ROI detection is performed by detecting face regions in each video frame, and this process is commonly used with face tracking and landmark detection algorithms such as the Viola-Jones method [4]. Once the ROIs are selected, pixel intensity components are extracted; those components are in the RGB color space. In addition, the RGB components are spatially averaged over all pixels in the ROI to yield an RGB component for each frame and form the raw signals.

Next, a signal-processing stage is applied, also known as the "rPPG Core." This has been the object of various studies in the last decade, resulting in multiple methods that seek to extract reliable rPPG signals from RGB components. Some rely on blind source separation methods, which can retrieve information by de-mixing raw signals into different sources. Principal component analysis (PCA) and independent component analysis (ICA), which use different criteria to separate temporal RGB traces into uncorrelated or independent signal sources, are some of the used techniques.

ICA separates the pulsatile signal from noise by minimizing the Gaussianity within the de-mixed signal. Mc Duff et al. [5] have used the JADE implementation of ICA to recover source signals from the observations, maximizing the non-Gaussianity within the sources. However, the experiment uses a novel digital single-lens reflex camera, capable of capturing five color channels: red, green, blue, cyan, and orange (RGBCO). On the other hand, Lewandowska et al. [6] obtained the rPPG signal using a 640 x 480 pixels RGB camera, and the chosen method was through PCA. The authors then compared the PCA results with FastICA as well as the pulse rate obtained with the ECG ground truth (GT).

Other authors have tried to improve the quality of the signal by changing the color space to a chrominance-based domain. Haan and Jeanne [7], in their work, have proposed a chrominance-based method (CHROM) to extract the rPPG signal by assuming a standardized fixed skin-color tone, where it is assumed to be the same for everyone under white light.

In a recent study, Wang et al. [8] introduced a new alternative to process RGB components into an rPPG signal, called the "plane-orthogonal-to-skin" (POS) algorithm. The main idea of the algorithm is to filter out intensity variations by projecting RGB components on a plane orthogonal to a determined normalized skin-tone vector. As a result, a two-dimensional (2D) signal referent to the projections is obtained and then combined into a 1D signal which is one of the input signal dimensions that is weighted by an alpha parameter. The alpha parameter is the quotient of the standard deviations of each signal.

In this article, we aim to evaluate the efficacy of our rPPG method using the UBFC-Phys public database [9]. The advantage of this dataset over other datasets that are suitable for rPPG evaluation is that UBFC-Phys contains additional data that can be used for stress and emotion analysis. The dataset also contains videos in which subjects are exhibiting significant head movements, which are a good test of how well our rPPG method can handle such conditions.

Materials And Methods

The methodology proposed by this work is divided into different stages that go from data acquisition and

---

signal processing to the evaluation method. The flowchart in Figure *1* shows the proposed method.
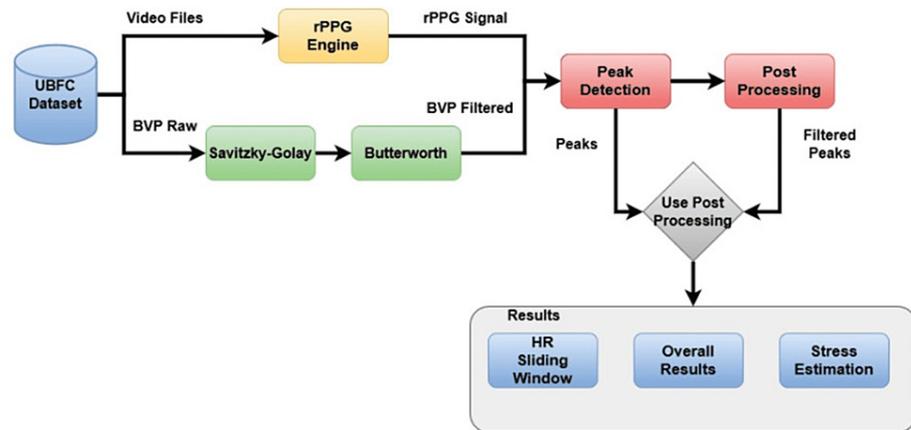

FIGURE 1: Proposed method - flowchart.

rPPG, remote photoplethysmography; HR, heart rate.

Dataset

To evaluate the efficacy of the proposed rPPG methodology, the method was benchmarked against the UBFC-Phys dataset [9]. This dataset is a public multimodal dataset and while it is dedicated to psychophysiological studies, it contains information that can be used to benchmark general rPPG methods. The dataset contains 56 participants of ages between 19 and 38, with 46 females and 10 males. Each of them followed a three-step experience that involved a resting task "T1," a speech task "T2," and an arithmetic task "T3." The participants were filmed during each of the three tasks with an EO-23121C camera by Edmund Optics at 35 FPS and a resolution of 1024 x 1024 pixels. An artificial light source was used to ensure adequate lighting conditions. The participants wore a wristband (Empatica E4) that measured their blood volume pulse signal (BVP), sampled at 64 Hz as well as their electrodermal activity signal, sampled at 4 Hz. Both the wristband signals and the video file are exactly 3 minutes long.

During the rest of the task, participants were asked to be quiet and not to talk. Therefore, most of the clean signals can be found in the T1 samples. T2 and T3 were interactive tasks and the subjects were randomly assigned to a "test" or "ctrl" group to denote the varying levels of difficulty for each of the speech and arithmetic tasks. Twenty-six subjects were assigned to the "test" group; in this group, they had to endure higher levels of difficulty in comparison to the "ctrl" group, which consisted of 30 individuals, for both tasks. The researcher intended to collect physiological responses that could be well indicative of stress and emotion for the "test" group. Further details of the dataset can be found in the UBFC-Phys original paper [10].

BVP processing

Since the BVP signals were obtained using a wristband, the body movement performed by the individual can lead to poor signal quality. In an attempt to prove it, a processing stage was performed. First, the BVP signal was re-sampled from its original sampling frequency (64 Hz) to the sampling frequency of our proposed method (30.30 Hz).

Once the BVP signal was resampled, digital filtering techniques were applied to improve the signal quality. First, a fourth-order and a 19-frame Savitzky-Golay filter [11] was used to smooth the time series through a moving average. In addition, a second-order Butterworth band-pass filter [12], with low cut and high cut of 0.7 Hz-7 Hz, respectively, was applied to remove noise artifacts.

rPPG processing

In this section, we will describe our method in more detail. The purpose of the method is to find an optimal rPPG signal capable of holding the same physiological information as a PPG signal from a contact sensor. The method is described in Figure *2*.
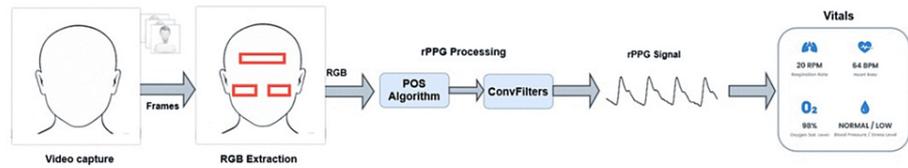
FIGURE 2: rPPG signal extraction methodology.

rPPG, remote photoplethysmography; RGB, red, green, and blue; POS, plane-orthogonal-to-skin.

The landmark detection algorithm from the OpenCV library [13] was used to extract RGB components from ROIs. This work proposes the use of three ROIs from the forehead, left cheek, and right cheek. Once the raw signal was collected, the rPPG Core, which is composed of the POS algorithm and a filtering stage based on convolutional filters, was applied to extract a clean rPPG signal.

POS algorithm

Originally proposed by Wang et al. [8], the POS algorithm seeks to mix RGB channels into a single-channel rPPG signal. According to the authors, the input RGB signal channels are mixed on the time interval t as follows:

1. $U(t) = G_n(t) - B_n(t)$

2. $V(t) = G_n(t) - R_n(t) + B_n(t) - R_n(t)$

3. $rPPG(t, t_i) = U(t) + \alpha(t_i)V(t)$

4. $s\_heart[n] = \frac{1}{2K+1} \sum\limits_{k=-K}^{K} s\_orig[n-k]t[-k]$

5. $s\_heart[n] = \frac{1}{2K+1} \sum\limits_{k=-K}^{K} s\_orig[n+k]t[k]$

6. $RR(i) = \frac{peak(i) - peak(i+1)}{sampling\,frequency} \times 1000.0$

7. $HR = \frac{60,000}{mean(RR)}$

8. $IBI = mean(RR)$

9. $RMSSD = \sqrt{\frac{1}{N-1} \times \sum\limits_{j=1}^{N} (RR(i) - RR(i+1))^2}$

10. $SDNN = \sqrt{\frac{1}{N-1} \times \sum\limits_{j=1}^{N} (RR(i) - mean(RR))^2}$

11. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

12. $Precision = \frac{TP}{TP+FP}$

13. $Recall = \frac{TP}{TP+FN}$

14. $Specificity = \frac{TN}{TN+FP}$

15. $F1Score = 2 * \frac{Precision*Recall}{Precision+Recall}$

16. $y = A * x + B$

The subscript n represents normalization, representing the instant color values divided by the mean value of the color channel. The rPPG signal on this interval is constructed as denoted in Eq. 3, where α is the ratio between the standard deviation of U(t) and V(t) calculated in the interval.

Convolutional filter

To clean the signal, this work proposes the use of a convolutional filter (ConvFilter). The ConvFilter will apply the convolution operation of the input single-channel signal "s orig," which is extracted after the POS algorithm, with the template that represents a single heartbeat peak of the same signal. The template is built by averaging segments of the "s orig" signal around the detected peaks in the signal. Since the signal could be a bit noisy, to make peak detection easier, it is recommended to apply a bandpass filter with the pass bandwidth from 0.7 Hz to 7.0 Hz. The cleaner "heart" signal is obtained via convolution Eq. 4 or the equivalent correlation Eq. 5 with this template t[k].

An example of a resulting rPPG signal is shown along with the corresponding BVP GT signal from the dataset in Figure *3*.
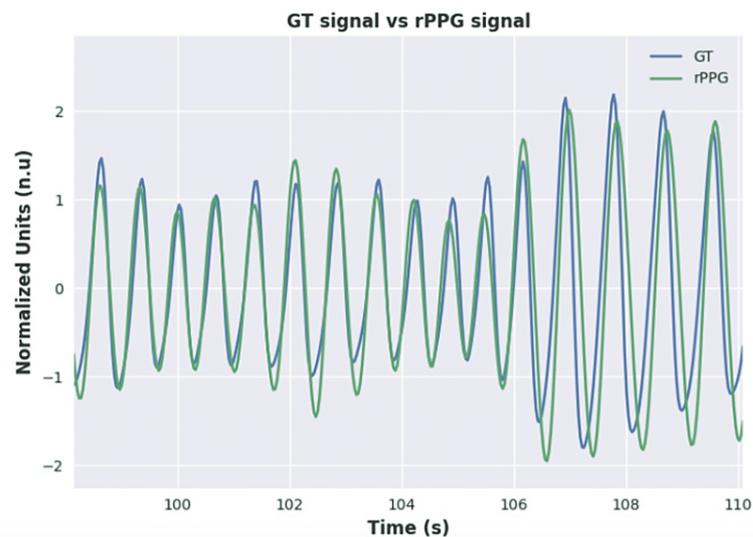

FIGURE 3: GT Signal vs. rPPG signal.

rPPG, remote photoplethysmography; GT, ground truth.

Signal post-processing

Even though the set of digital filtering techniques, which were applied to the signals, can well remove artifacts and noise issues, it may not be able to improve the signal quality to an acceptable level. Meziati et al. [10], in the original paper, have chosen a strategy to cope with those signals by simply removing them. In our work, we propose a post-processing stage, which seeks to find and remove specific segments of the signal that are damaged in an attempt to avoid removing the entire sample. The procedure is entirely based on detecting corrupted peaks, either peaks detected when they should not be or missing peaks. The peak validation process is a set of test rules that are used to find reliable and unreliable peaks either in the BVP or rPPG signals. First, using the original peaks, the inter-beat interval (IBI) signal is calculated. In the first rule, we propose the use of an IBI threshold, in which the IBI signal should be within the range from min IBI to max IBI values that correspond to 210 bpm and 42 bpm of HR, respectively. Second, since each peak has its left and right IBI components, for valid peaks, the absolute IBI sequential difference should be less than 0.5 seconds.

The valid peaks are related to the left and right IBI components. Therefore, the first detected peak in the signal is invalid because it has no left IBI, and the last peak is invalid because it has no right IBI.

Finally, a peak that is just before or after the invalid peak is assumed to be invalid. This processing is applied only once, as a last validation test. In this way, there are two invalid peaks at the beginning and the end of the signal. Moreover, four invalid peaks can be expected to appear around one invalid IBI in the signal. This rule aims to eliminate isolated peaks surrounded by invalid peaks. Figure *4* presents an example of a post-processing algorithm in use.
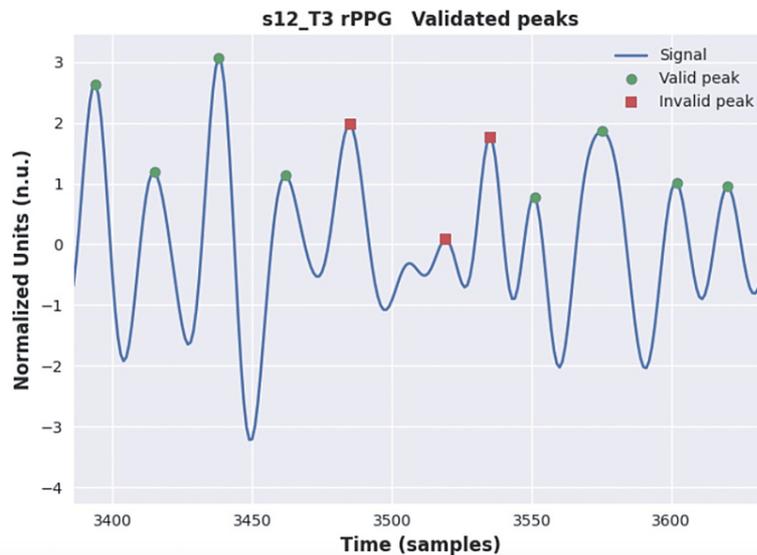
FIGURE 4: Post-processing algorithm over the rPPG signal.

rPPG, remote photoplethysmography.

Metrics

As stated previously, this work seeks to evaluate our proposed method of remote health screening using rPPG signals extracted from video files. To compare the performance of our method, the BVP signal was used as the "ground truth." Thus, physiological features were taken from both contact and remote signals and then compared.

The main tool used to calculate those features is the RR Interval, also known as the pulse-to-pulse interval, which is the time difference between two peaks in terms of milliseconds (ms) (Eq. 6). HR, measured in beats per minute (bpm), can be calculated from the RR intervals, and Eq. 7 shows the definition. In addition, we also compile the following features in the time domain: IBI (Eq.8), root mean square of successive differences between normal heartbeats (RMSSD; Eq. 9), and the standard deviation of normal to normal heartbeats (SDNN; Eq. 10). In the frequency domain, we measured the power from the low-frequency band (LF) [0.04; 0.15]Hz and the high-frequency band (HF) [0.15; 0.4]Hz.

To evaluate the accuracy of our rPPG method against the contact sensor, we measured the mean absolute error (MAE) for each comparative feature described above. For the stress classification task, statistical metrics were used, as described in Eqs. 11-15. We also use the F1-score Eq. 15 to evaluate how well the rPPG method can identify pulsatile peaks in a signal.

The F1 score is the weighted average of precision Eq. 12 and recall Eq. 13, in which the score takes both, false positives and false negatives, into account. The concepts of true positives, false positives, and false negatives are entirely based on the position of the pulsatile peaks in the signal.

• True Positive (TP): Event in which there was a peak and the algorithm found it correctly.

• False Positive (FP): Event in which there was not a peak and the algorithm incorrectly points.

• False Negative (FN): Event in which there was a peak and the algorithm has not found it.

Stress experiments

This work also seeks to evaluate how well we can predict whether an individual is under stress or not, just by looking at the features of the rPPG signal. In the UBFC-Phys dataset, there are at least three different ways to measure this: stress classification between stress and unstress, task differentiation between subjects in T1, T2, and T3, and finally group differentiation between test and control.

To conduct the analysis, only HRV features in the time, frequency, and nonlinear domains were used. Moreover, this work does not use the self-reported scores assigned to subjects before and after the experiment. The group separation between stress and non-stress was created in the following way: All T1 samples were labeled as non-stress, T2 and T3 samples were labeled as stress if the sample belongs to the

test set otherwise, if the subject belongs to the control set, then this sample is labeled as non-stress.

In our work we used two different approaches for each experiment, using a statistical one-way analysis of variance (ANOVA) test and machine learning (ML) models. ML models are often used for classification tasks, and in this work, we propose the use of HRV features as inputs to those models.

Moreover, ANOVA was used to verify the hypothesis that the groups are separable. ANOVA is a statistical tool that compares the means of several samples; this is done by analyzing the variances between the data and within these groups, and it is also known as an extension of the t-test for two independent samples to more than two groups [14]. ANOVA test of the hypothesis is based on a comparison of two independent estimates of the population variance [15].

The null hypothesis of the ANOVA test represents that there is no difference between the groups; to reject this hypothesis, the resulting p-value is observed and compared with a certain threshold, usually a significance level of 0.05. In other words, if the resulting p-value is below 0.05 (p-value < 0.05), it rejects the hypothesis and it is possible to conclude with 95% confidence that there are differences between the means of the group, and thus the groups are separable.

However, to use the ANOVA test, a normally distributed dataset is required. To test whether a feature follows a normal distribution or not, the Shapiro-Wilk test [16] was applied. The Shapiro-Wilk test is based on the correlation between the data and the corresponding normal scores [17]. The null hypothesis of the Shapiro-Wilk test is that the feature is normally distributed, which is represented by a p-value above 0.05 (p-value > 0.05).

Even though ANOVA is a powerful tool that indicates that at least one group differs from the other groups, the method does not show which particular group differs or if there is more than one. Thus, ANOVA is often equipped with a specific test that compares the two means between the pairs or groups, also known as pairwise comparisons. In this work, Tukey's test was applied [18].

Results

In this study, video files from the UBFC dataset [9] were used to extract physiological information from individuals. Each one of the videos was processed using our proposed method described in the section related to rPPG processing, which returns the rPPG signal.

The information obtained through the rPPG signal was then compared to that obtained from the BVP signal, which was taken as GT.

First, after the rPPG signals were obtained, the post-processing algorithm described previously was used to better understand the relationship between the quality of the contact and remote signals.

To extract the HR values over time from each subject, a sliding window approach, with a size of 30 seconds and a stride of 1 second, was used.

To understand the real spatial difference between estimated and GT HR, two plots were created.

It is clear to see that the UBFC dataset has damaged signals, and these signals can be either one of BVP GT or rPPG signals or both. In contrast with the original article, which proposes exclusion criteria based on a comparative assessment between the rPPG and BVP signals, this work aims to create a more robust approach to identifying unreliable signals. This condition must be under the assumption that four scenarios are possible:

These scenarios are apparent in any of the three tasks.

(a) Acceptable GT and acceptable rPPG.

(b) Acceptable GT and unacceptable rPPG.

(c) Unacceptable GT and acceptable rPPG.

(d) Unacceptable GT and unacceptable rPPG.

This work looked at several parameters to formulate a fair and robust exclusion criterion solely based on removing damaged GT signals. To reach this goal, the following parameters were investigated:

• percentage of GT valid peaks;

• F1 score of valid peaks (uses GT and rPPG peaks);

• signal-to-noise ratio (SNR);

• standard deviation of GT HR.

Experiments have shown that the most efficient and reliable parameter was the percentage of GT valid peaks. However, the threshold used as the minimum required value to validate the sample has its limitations. Table _1_ shows the results for different thresholds of validated GT peaks.

	GT Validated Peaks' Score			
T1	TH > 60%	TH > 70%	TH > 80%	TH > 90%
No. files	56	52	49	39
Mean value	3.38	3.23	3.42	2.94
Min value	0.01	0.01	0.01	0.01
Max value	33.05	33.05	33.05	33.05
T2	TH > 60%	TH > 70%	TH > 80%	TH > 90%
No. files	34	24	11	4
Mean value	5.58	4.99	3.72	3.81
Min value	0.00	0.00	0.37	1.28
Max value	26.34	26.34	8.87	5.82
T3	TH > 60%	TH > 70%	TH > 80%	TH > 90%
No. files	42	30	16	9
Mean value	4.02	4.15	3.12	4.18
Min value	0.06	0.06	0.06	0.39
Max value	22.31	13.69	13.69	13.69

TABLE 1: Overall absolute heart rate error using different exclusion criteria.

Validated peaks were obtained from GT. GT, ground truth; TH, threshold.

Evaluation of exclusion criteria - absolute error HR (bpm)

Although using an exclusion criterion based on GT valid peaks can increase the reliability of the results, it is possible to see that the maximum value of all three tasks was considerable. Those large errors can be accepted if the problem was due to suboptimal rPPG signals, which can be understood as those samples in which the rPPG method failed; however, this is not always the case.

Using the samples that survived from the exclusion criterion of GT valid peaks >60%, each sample that could be a potential outlier was investigated by evaluating the SNR of the GT signal. Using a threshold of 4 dB, we removed GT signals that fell below this threshold, indicating that this signal was not of sufficient quality for comparative evaluation. Table _2_ shows the results using these two exclusion criteria.

T1	This work
No. files	55
Mean value	3.33
Min value	0.01
Max value	33.05
T2	**This work**
No. files	22
Mean value	3.08
Min value	0.00
Max value	16.26
T3	**This work**
No. files	34
Mean value	2.77
Min value	0.06
Max value	22.31

TABLE 2: Results after exclusion criteria of GT valid peaks >60% and SNR >4 dB. MAE HR (bpm)

SNR, signal-to-noise ratio; MAE HR, mean absolute error of heart rate; bpm, beats per minute; GT, ground truth.

Despite applying these techniques to avoid including damaged GT signals in our analysis, it is still evident that damaged GT signals were still left. Therefore, a manual check on the GT signal was performed to avoid damaged signals that neither of the proposed materials was able to filter. We only found additional damaged signals in Task 1 during our manual check.

Table 3 presents the results after applying our exclusion criteria which are based on valid GT valid peaks >60%, SNR >4 dB, and a manual GT signal check.

2022 Talukdar et al. Cureus 14(7): e26871. DOI 10.7759/cureus.26871

9 of 15

T1	This work
No. files	47
Mean value	1.10
Min value	0.01
Max value	4.70
T2	**This work**
No. files	22
Mean value	3.08
Min value	0.00
Max value	16.26
T3	**This work**
No. files	34
Mean value	2.77
Min value	0.06
Max value	22.31

TABLE 3: Results of this work after proposed exclusion criteria, GT valid peaks >60%, SNR >4 dB, and manual GT signal check. MAE HR (bpm)

SNR, signal-to-noise ratio; MAE HR, mean absolute error of heart rate; bpm, beats per minute; GT, ground truth.

Using the exclusion criteria based on the percentage of valid GT peaks >60%, SNR >4 dB, and a manual check, features from time, frequency, and non-linear domains were extracted from each one of the remaining samples using contact and remote signals after the post-processing algorithm.

From those HRV features, the most important ones, according to this work, are shown in Table 4, which demonstrates the average of each feature per task as well as the MAE in the specific feature over all the remaining samples.

T1	Contact	Remote	MAE
HR (bpm)	83.56	84.13	1.10
IBI (ms)	745.18	728.51	22.09
RMSSD (ms)	153.98	75.22	93.98
SDNN (ms)	127.53	77.40	62.10
LF (s2)	7.42	3.89	6.31
HF (s2)	6.29	1.85	5.18
LF/HF	1.40	2.13	0.98
T2	**Contact**	**Remote**	**MAE**
HR (bpm)	80.51	80.53	3.08
IBI (ms)	810.57	778.81	57.82
RMSSD (ms)	352.00	266.76	120.37
SDNN (ms)	267.14	212.93	88.64
LF (s2)	22.61	29.88	28.14
HF (s2)	23.77	14.39	13.07
LF/HF	0.97	1.40	0.73
T3	**Contact**	**Remote**	**MAE**
HR (bpm)	82.70	82.26	2.77
IBI (ms)	777.15	756.81	58.34
RMSSD (ms)	296.97	208.75	125.45
SDNN (ms)	224.09	165.73	89.28
LF (s2)	15.58	9.25	10.04
HF (s2)	17.46	9.69	11.76
LF/HF	1.09	1.05	0.52

TABLE 4: Comparison of contact/remote features per task using proposed exclusion criteria

HR, heart rate; IBI, inter-beat interval; RMSSD, root mean square of successive differences between normal heartbeats; SDNN, standard deviation of normal-to-normal heartbeats; LF, low-frequency band; HF, high-frequency band; MAE, mean absolute error

Since the main purpose of the UBFC dataset was to create a database for social stress studies, in this work, we have investigated the ability of the proposed rPPG method to recognize stress states. First, statistical methods were used to explore the assumption that HRV tissues can differentiate between stress and non-stress states.

To investigate whether this assumption may be valid, we can visualize the distribution of certain HRV features across each task.

In addition, we investigated the use of ML models to correctly identify stress states and distinguish between tasks. Four classifiers were considered: Random Forest (RF) with 20 trees, Support Vector Machine (SVM) with a linear kernel, K-Nearest Neighbors (KNN) with k = 5, and Light Gradient Boosting Machine (Light-GBM).

Using metrics such as Accuracy, Precision, Sensitivity, Specificity, and F1 score, Table *5* shows the results for stress state recognition.

Model	Accuracy	Precision	Sensitivity	Specificity	F1 score
RF	83.11%	77.08%	94.87%	71.05%	85.05%
SVM	72.72%	68.00%	87.17%	57.89%	76.40%
KNN	71.42%	68.08%	82.05%	60.52%	74.41%
Light-GBM	77.92%	70.37%	97.43%	57.89%	81.72%

TABLE 5: Statistical analysis of this work, stress recognition.

RF, Random Forest; SVM, Support Vector Machine; KNN, K-Nearest Neighbors; Light-GBM, Light Gradient Boosting Machine.

Discussion

In this work, we discovered that to accurately and fairly benchmark PPG signals against remote methods, there is a need to develop an exclusion criterion to evaluate which signals should be discarded and which signals should be included for comparative purposes. Across the UBFC dataset, we found that, on average, the quality of our rPPG signals was better than that of PPG signals from a contact sensor when applying the same evaluation metric to both [9,10].

After synchronizing the PPG and rPPG signals in time, we could analyze how well our rPPG method was able to measure GT pulsatile peaks in the signals and we used the F1 score as the metric for this evaluation. The rPPG and PPG signals were well synchronized in time, with a mean F1 score of above 90%; however, the lower F1 score for tasks T2 and T3 can be explained by some phase shift in the signals when motion is introduced.

Even though our post-processing algorithm can handle some damaged signals, as shown previously, in some samples, it is not possible to correct them to an acceptable level; therefore, the most appropriate approach is to filter out those files.

Table 1 details different thresholds that we can use for our exclusion criteria, which are based on the number of valid peaks in a signal as a percentage of all peaks detected in that signal. A threshold of 60% would mean that the signal must have at least 60% of its peaks marked as valid for comparative purposes. Table 1 shows how many files survived for each threshold and the comparative mean HR MAE as well as the minimum and maximum HR MAE.

However, despite applying this filter to the GT signals, we could still observe some outliers in terms of MAE values >10 bpm; therefore, to ensure that we are making a fair comparison of these samples, there was a need to include second exclusion criteria. Table 2 shows that while using a threshold of 60% in our first criteria and applying an additional SNR filter to the GT signal, there was a significant improvement in MAE values in both tasks T2 and T3.

There were still some outliers present in T1; therefore, to avoid comparing our rPPG signals against unacceptable GT signals, we performed a manual check on some of the GT samples. Eight GT signal samples from T1 were discarded due to their poor quality and none were further discarded from T2 and T3.

Overall, after applying various exclusion criteria to filter out poor quality GT signals, the HR MAE for T1, T2, and T3 was 1.10 bpm, 3.08 bpm, and 2.77 bpm, respectively.

Stress recognition

To evaluate our ability to detect stress states, this work proposed the use of two different approaches, a statistical ANOVA test and ML.

First, a one-way ANOVA was performed between subjects to compare the effect of each HRV characteristic sequentially on stress recognition. A probability p-value and an F score are defined to decide the significance of the variability. The p-value was compared to a significance level of 0.05. Hence, for p values lesser than the significance level, it is possible to conclude that the group means are different, and the groups can be separable. However, before running the ANOVA test, we first need to check that the features are normally distributed using the Shapiro-Wilk test.

Based on the RR intervals from each signal, 32 HRV features were extracted. For those, there are 168 rows, which represent 56 subjects of three tasks. However, only four of those features are normally distributed

according to the Shapiro-Wilk test: IBI, NNi 20, MEAN HR, and SAMPEN. The following conclusions can be made for those features.

The IBI and MEAN HR feature were not shown to be statistically significant for stress state at the p value <0.05 level for two conditions. IBI: $[F_{(1, 166)} = 1.75, p = 0.187]$, mean HR: $[F_{(1, 166)} = 1.00, p = 0.317]$. However both NNi 20 and SAMPEN were shown to be statistically significant for stress state recognition at the p value <0.05. NNi 20: $[F_{(1, 166)} = 8.65, p = 0.003]$, SAMPEN: $[F_{(1, 166)} = 7.34, p = 0.007]$. We can conclude that both NNi 20 and SAMPEN were able to detect stress states with a 95% of confidence level.

These results suggest that in the time domain only the number of pairs of successive RR intervals that differ by more than 20 ms can differentiate between stress and non-stress. However, it should be noted that other HRV features could potentially accomplish this result.

In addition, the same experiment was conducted to find features that could differentiate between tasks.

Once again there was not a significant effect of mean HR on task recognition at the p-value <0.05 level for three conditions. Mean HR: $[F_{(2, 165)} = 1.52, p = 0.221]$. Mean HR failed to separate tasks T1, T2, and T3.

However, there was a significant effect of IBI, NNi 20, and SAMPEN on task recognition at the p-value <0.05 level for three conditions IBI: $[F_{(2, 165)} = 6.47, p = 0.001]$, NNi 20: $[F_{(2, 165)} = 26.97, p = 7.286\mathrm{e}{-11}]$, and SAM- PEN: $[F_{(2, 165)} = 8.27, p = 0.0003]$. In fact, IBI, NNi 20, and SAMPEN succeeded a 95% confidence interval in separating at least two of three tasks T1, T2, and T3.

Even though ANOVA can statistically show whether the framework is able or not to separate the target, ML approaches are an easier way to classify whether the subject is under stress or not. The HRV features previously mentioned were extracted without any kind of windowing function, which means that there are only 168 entries that would be divided into train and test sets.

Besides that, the number of stress and non-stress samples were heavily imbalanced, which might represent unreliable results. To deal with this problem, the SMOTE algorithm was applied to provide an evenly distributed dataset [19]. This algorithm creates synthetic samples based on the nearest neighbors of the minority class sample. Hence, the number of entries increased to 232, equally distributed between stress and non-stress classes.

Furthermore, as stated previously, the dataset has 32 HRV features; however, only 20 of them were used to build the models. A grid search analysis between globally assigned features was performed to define the best set of features.

Table *5* shows that all four ML models performed reasonably well. However, SVM, KNN, and LightGBM had problems with false positives, represented by the specificity metric. The specificity of around 60% can be explained by the situations where the subject was not under stress; however, the model classifies it as stress. In real-world situations, where a stress recognition system is applied to decide if a driver should rest, wrong classifications in the non-stress class represent a more tolerable situation than wrong predictions in the stress class. If the subject was under stress and the system allows him to drive, this wrong decision may contribute to a severe consequence such as a car accident.

Overall, the Random Forest model performed the best in most of the metrics and was able to achieve 83% accuracy with an F1 score of 85%.

The RF model was able to achieve 95% of accuracy in the stress class and only 71% in classifying the non-stress state.

Furthermore, in task classification the results were similar between the classifiers; however, the overall results were worse than in stress classification. The best result shows 77% accuracy and 66.51% F1 score in the Random Forest model.

These results show that the model struggled to classify between T2 and T3, corroborating the hypothesis that T2 and T3 are similar tasks that should be within a single group, as activity.

Conclusions

In this work, we proposed a benchmark comparison using the novel multimodal UBFC dataset. The dataset was built to allow studies on social stress since each of the 56 subjects was assigned to three different tasks, the first (T1) being a resting task and the second and third (T2 and T3) activity tasks. However, the main contribution of this dataset was to allow the comparison between contact and remote extraction of physiological information through BVP and rPPG signals.

The BVP signal had fewer filtering stages, which typically refer to a band-pass filter to remove unwanted frequencies as well as a moving average filter to smooth the signal, while the rPPG signal was obtained using the method proposed by this work. First, the RGB components were obtained from three different ROIs and averaged. Hence, a POS algorithm was used to mix the RGB channels into the rPPG signal, which ultimately was passed through a convolutional filter to improve the quality of the signal and make the peak detection more robust.

Furthermore, this work proposed a post-processing stage that seeks to filter out unreliable peaks. This method excludes peaks that do not abide by the IBI threshold rules. Although, the post-processing stage was possible to extract the percentage of validated peaks in each one of the signals as well as synchronize the contact and remote signals to measure the F1 score to evaluate how accurately our proposed method can detect pulsatile peaks.

The results showed that, according to the post-processing algorithm, the rPPG signal had fewer corrupted peaks than the BVP signal, which represents a cleaner signal. Furthermore, by comparing contact and remote HR extracted using a 30-second sliding window approach, there was a high correlation between the majority of the people. However, there were highly compromised samples, which correspond to signals with a very high level of noise that even the post-processing algorithm was unable to deal with.

This work performed an exclusion criterion to create a fair comparison using only reliable GT signals. Ultimately the exclusion criteria were based on the percentage of validated GT peaks above 60%, SNR of GT signal above 4 dB, and a manual GT signal quality check. The results have shown a considerable improvement, with 1.10 bpm of mean error in T1 in 47 samples, 3.08 bpm in T2 in 22 samples, and finally 2.77 bpm in T3 in 34 samples.

In terms of stress recognition, this work proposed two different approaches. The ANOVA test was conducted over HRV features extracted from the rPPG signal, and the results show that two features were able to statistically differentiate between stress and non-stress. Moreover, three HRV features were able to differentiate between tasks. Using Tukey's test, we discovered that those three characteristics were able to differentiate between T1 (rest task) and T2 or T3, which represents the tasks for which there was a higher level of activity.

Ultimately, four ML models were evaluated to investigate if any of them were able to classify between stress and non-stress as well as between T1, T2, and T3. Results have shown that the Random Forest model reached 83.11% of accuracy, 77.08% of precision, and 94.87% of recall in the binary classification, which denotes the effectiveness of the model in classifying the stress status.

In task recognition problems, once again the Random Forest model obtained the best score with 77.38% of accuracy, 66.91% of precision, and 66.13% of recall. Although these results highlighted the difficulty of distinguishing between T2 and T3, they also showed that the model was able to differentiate between T1 and T2 or T3.

Further work will include the enhancement of the rPPG core algorithm to better deal with motion robustness; once the method can improve the results in those noisy signals it will considerably decrease the mean error. Moreover, further studies in stress recognition using the remote signal will be performed; through the results of this work it can be concluded that there is a correlation between HRV features and stress tasks; however, with additional data, it will be possible to conclude which features are more likely to carry the stress information.

Additional Information

Disclosures

Acknowledgements

References

1. World Health Organization. Cardiovascular diseases (CVDs) . (2021). Accessed: June 20, 2021:

https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(CVDS).

2. American Institute of Stress, 42 Worrying Workplace Stress Statistics . (2021). Accessed: June 20, 2021: https://www.stress.org/42-worrying-workplace-stress-statistics.

3. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH: Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig. 2018, 15:235-245. 10.30773/pi.2017.08.17

4. Viola P, Jones M: Robust real-time object detection. Int J Comput Vis. 2001,

5. McDuff D, Gontarek S, Picard RW: Improvements in remote cardiopulmonary measurement using a five band digital camera. IEEE Trans Biomed Eng. 2014, 61:2593-601. 10.1109/TBME.2014.2323695

6. Lewandowska M, Ruminski J, Kocejko T, Nowak J: Measuring pulse rate with a webcam - a non-contact method for evaluating cardiac activity. 2011 Federated Conference on Computer Science and Information Systems (FedCSIS). 2011. 405-410.

7. de Haan G, Jeanne V: Robust pulse rate from chrominance-based rPPG . IEEE Trans Biomed Eng. 2013, 60:2878-86. 10.1109/TBME.2013.2266196

8. Wang W, den Brinker AC, Stuijk S, de Haan G: Algorithmic principles of remote PPG . IEEE Trans Biomed Eng. 2017, 64:1479-91. 10.1109/TBME.2016.2609282

9. Meziati R, Benezeth Y, Oliveira P, Chapp´e J, Yang F: UBFC-PHYS. IEEE Dataport . 2021. 10.21227/5da0-7344

10. Meziati Sabour R, Benezeth Y, De Oliveira P, Chappe J, Yang F: A multimodal database for psychophysiological studies of social stress. IEEE Trans Affect Comput. 2021, 10.1109/TAFFC.2021.3056960

11. Savitzky A, Golay MJE: Smoothing and differentiation of data by simplified least squares procedures . Anal Chem. 1964, 36:1627-39. 10.1021/ac60214a047

12. Selesnick IW, Burrus CS: Generalized digital butter- worth filter design . IEEE Trans Signal Process. 1998, 46:1688-94. 10.1109/78.678493

13. Bradski G: The OpenCV Library, Dr. Dobb's Journal of Software Tools . 2000. https://www.bibsonomy.org/bibtex/25fa6cccd386170d95ae17fd29a42303d/ross_mck.

14. Ostertagova E, Ostertag O: Methodology and application of one-way ANOVA . Am J Mech Eng. 2013, 1:256-61. 10.12691/ajme-1-7-21

15. Montgomery DC, Runger GC: Applied Statistics and Probability for Engineers . John Wiley and Sons, New York; 2013.

16. Shapiro SS, Wilk MB: An analysis of variance test for normality (complete samples). Biometrika. 1965, 52:591-611. 10.2307/2333709

17. Barton B, Peat J: Medical Statistics: A Guide to SPSS, Data Analysis and Critical Appraisal, 2nd ed . John Wiley and Sons, New Jersey, USA; 2014.

18. Tukey JW: Exploratory Data Analysis. Pearson, London, UK; 1977. http://www.ru.ac.bd/wp-content/uploads/sites/25/2019/03/102_05_01_Tukey-Exploratory-Data-Analysis-1977.pdf.

19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: Synthetic minority over-sampling technique . J Artif Int Res. 2002, 16:321-57. 10.48550/arXiv.1106.1813