rPPG BASED HEART RATE ESTIMATION USING DEEP LEARNING

by

Nur Deniz ÇAYLI

Minel SAYGISEVER

Esra POLAT

CSE497 / CSE498 Engineering Project report submitted to Faculty of Engineering
in partial fulfilment of the requirements for the degree of

BACHELOR OF SCIENCE

Supervised by:
Prof. Dr. Çiğdem EROĞLU ERDEM

Marmara University, Faculty of Engineering

Computer Engineering Department

2021

rPPG BASED HEART RATE ESTIMATION USING DEEP LEARNING

by

Nur Deniz ÇAYLI

Minel SAYGISEVER

Esra POLAT

CSE497 / CSE498 Engineering Project report submitted to Faculty of Engineering
in partial fulfilment of the requirements for the degree of

BACHELOR OF SCIENCE

Supervised by:
Prof. Dr. Çiğdem EROĞLU ERDEM

Marmara University, Faculty of Engineering

Computer Engineering Department

2021

ABSTRACT

Remote heart rate estimation is the measurement of heart rate without any physical contact with the patients. This is accomplished using remote photoplethysmography (rPPG). rPPG signals are usually collected using a video camera with a limitation of being sensitive to multiple contributing factors, such as different skin tones, lighting condition of environment and facial structure. There are multiple studies and generally two basic approaches in the literature to process and make sense of these signals: Firstly, we examined the traditional methods as CHROM DEHAN [1], ICA POH [2], GREEN VERCRUYSSE [3] and POS WANG [4]. Secondly, we examined MTTS-CAN [5], one of the deep learning methods. While we tried traditional methods with the UBFC [6] dataset, we ran deep learning methods with UBFC and PURE [7] datasets. When we used SNR [8] to calculate heart rate based on the Blood Volume Pulse (BVM) signal resulting from deep learning-based methods, we observed a significant improvement in some results. In summary, we concluded that deep learning-based methods play an important role in the development of rPPG technologies and their introduction into our daily lives.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Problem Description and Motivation

Heart rate estimation has great importance in determining a person's mental and physiological state. In some cases, it is not possible to use medical devices such as the finger pulse oximeter with photoplethysmography (PPG) technology due to the patient's delicate health and skin conditions. For example, such a technology is needed for continuous monitoring of premature infants. In such a case, it is necessary to measure the heart rate remotely. Remote PPG studies (rPPG) bring us a solution in such sensitive situations, allowing us to estimate the heart rate through a face video obtained with a standard webcam. We can see an illustration of the main principles of rPPG in Figure 1. With every heartbeat, there are changes in the light and hence colour reflected from our skin caused by the cardiac cycle. We can not see these changes with our eyes, but we can analyze the intensity of these colours with image processing techniques. If we can get the RGB values of the skin pixels in the frames of a video, we have 3 colour signals. By processing these signals, we can estimate the heart rate.

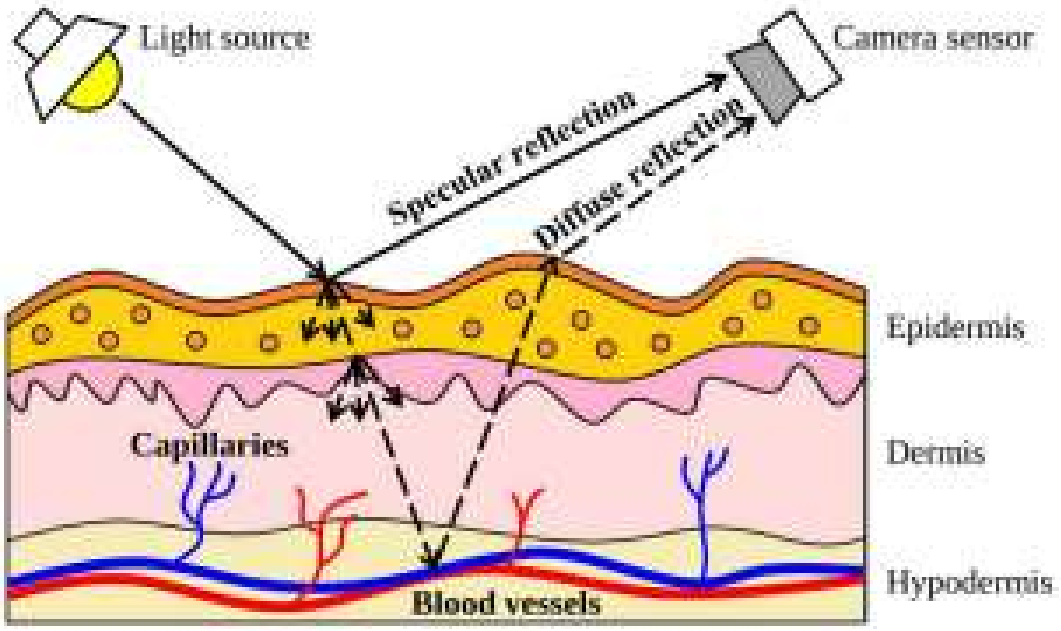

Figure1: Skin reflection model illustration [4]

1.2 Aims of the Project

Project Aim 1: The heart rate that we could get from the video of the face of a person standing still in a well-lit environment should have been measured close to the reference heart rate. We aimed to achieve an RMSE [9] error of less than 5 bpm on PURE [7] and UBFC [6] datasets.

Project Aim 2: In a well-lit environment, the heart rate that we could get from the video of the face of a person who moves his / her head right and left up and down without rotating his / her head should have been measured close to the reference heart rate.

Project Aim 3: In a well-lit environment, a result close to the reference heart rate should have been obtained by taking into account the person's speech or facial expressions such as laughing or being surprised.

Project Aim 4: In a well-lit environment, the heart rate of a person wearing glasses should have been measured close to the reference heart rate.

2. DEFINITION OF THE PROJECT

2.1 Scope of the Project

There were some requirements for the project to give an accurate result. We processed the signals from the skin pixels of the face captured by the camera and these skin pixels must have been visible to the camera.

Our assumptions and constraints were:

- The environment must have been well-lit.
- The subject's head movements should have been minimal.
- The face of the subject should have been visible, so the subject should have not turned his/her head.
- The subject's facial expressions should have been minimal.
- The subject should have been minimal facial hair.

We implemented a deep learning-based algorithm for heart rate estimation using rPPG and compared it with signal processing based methods in the literature. We implemented the deep learning-based method based on the MTTS-CAN [5] framework. MTTS-CAN offers a video-based and on-device optical cardiopulmonary vital sign measure approach. It leverages a novel Multi-Task Temporal Shift Convolutional Attention Network (MTTS-CAN) and enables real-time cardiovascular and respiratory measurements on mobile platforms. MTTS-CAN runs the system on an ARM-CPU and it reaches state-of-the-art accuracy while running at over 150 frames per second, enabling real-time applications. Systematic trial on large benchmark datasets makes it more clear that the MTTS-CAN approach leads to important (20%-50%) reductions in error and generalizes greatly against datasets. While we worked on the framework, we used two public datasets which are PURE [7] and UBFC [6]. Our main goal was to measure the heart rate most accurately. If the RMSE [9], between the estimated heart rate and the actual heart rate, were 5 beats per minute or less, we considered the result as successful.

2.2 Success Factors and Benefits

2.2.1 Measurability / Measuring Success

The main objective of the project was to develop a contactless photoplethysmography method that can monitor heart rate using a webcam using deep learning methods. The project considered successful if our requirements listed below are fulfilled:

- We determined the region of interest, shortly ROI [10] area, such as. forehead area and cheeks area, on the face correctly.
- We tracked the detected ROI region throughout the video.
- The heart rate RMSE [9] measured by our method is less than 5 beats per minute.
- The heart rate means absolute error, shortly MAD [11], measured by our method.
- The Pearson correlation measured by our method.

2.2.2 Benefits / Implications

The benefits of our project are:

- It minimized the changes in heart rate due to the stress of being connected to a device.
- Remotely measuring heart rate for sensitive individuals and babies in intensive care units.
- Measuring the remote heart rate of a person who has a burn because he or she can't wear pulse oximeters.
- Other applications: measuring the heart rate of a person at a gym, and drivers.

2.3 Professional Considerations

2.3.1 Methodological Considerations / Engineering Standards

- We used GitHub to manage version control
- We used Python to develop the software.
- We used UBFC and PURE datasets.
- We used Gantt charts for our management plan.

2.3.2 Societal / Ethical Considerations

Economical: Since our studies did not require the use of devices such as pulse oximetry used in hospitals, it was an economically viable study as we could measure it using a simple webcam.

Environmental: The equipment we used in our work does not contain any substances that can pollute the environment.

Health and Safety: Since the main purpose of rPPG was to calculate heart rate for non-contact patients, it was a method that can be used for every patient. In addition, it did not pose any threat to patients in terms of health and safety.

Legal Considerations: Since the datasets and libraries we used are open sources, they did not pose a legal problem. Python source code and installers were available for download for all versions and we had a free license for students on GitHub.

2.4 Literature Survey

The main subject of our project was to measure the heart rate of the person without touching the person. We did this with rPPG technology, so we measured heart rate with an RGB camera. We estimated the heart rate by selecting the region of interest on the skin and inferring the rPPG signal from the colour changes.
Below we give a brief overview of the methods in the literature. We group the methods as contact and remote (non-contact) methods.

2.4.1 Contact Methods

Photoplethysmography (PPG): Photoplethysmography (PPG) is a technique used to measure the volumetric changes in the blood affected by the heartbeat. PPG is usually obtained using pulse oximetry to measure the heart rate. A normal pulse oximeter monitors the circulation of blood in the dermis layer under the skin. With each cardiac cycle, the heart pumps blood. Even though this pressure pulse is somewhat damped by the time it reaches the skin, it is enough to

distend the arteries and arterioles in the subcutaneous tissue. The change in volume caused by the pressure pulse is detected by illuminating the skin with the light from a light-emitting diode (LED) and then measuring the amount of light either transmitted or reflected by a photodiode. Each cardiac cycle appears as a peak.

Figure 2: Finger pulse oximeter [14]

PPG technology is also used in other applications. For example, blood oxygen saturation, blood pressure, cardiac output, respiration, vascular assessment, arterial disease. Additionally, the shape of the PPG waveform differs from subject to subject and varies with the location and manner in which the pulse oximeter is attached as we can see in Figure 2.

2.4.2 Remote Methods

Independent Component Analysis (ICA POH) Method: The ICA POH [2] tries to separate a multivariate signal into independent non-Gaussian signals.

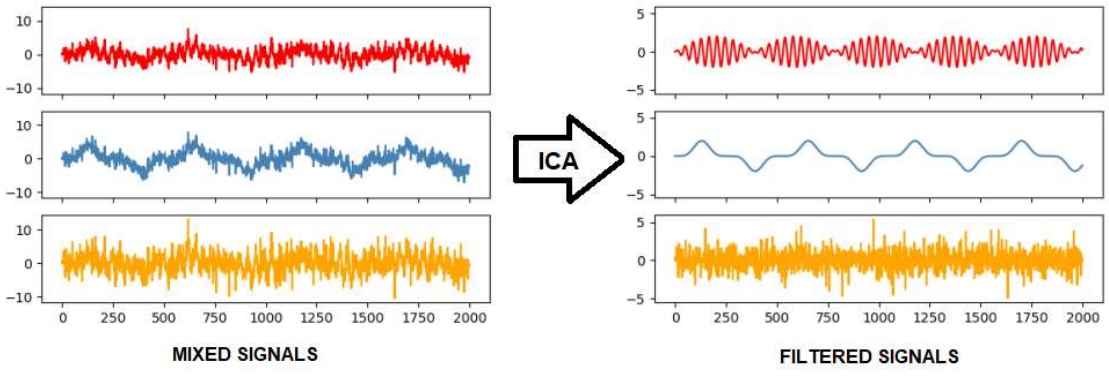

Figure 3: ICA working example

For example, an audio signal occurs the numerical addition, at each time *t*, of signals from different sound sources. In this signal, the problem is whether it is possible to separate these subscripts to sources from the observed entire signal. If the statistical independence assumption is right, blind ICA POH separation of a mixed-signal gives very good outcomes. Also, ICA POH can be used for signals that are not needed to be generated by mixing for analysis purposes. We can see these processes in Figure 3.

In the ICA POH model formulas $x_1(t)$, $x_2(t)$ and $x_3(t)$ red, green and blue signals. Source signals are represented by $s_1(t)$, $s_2(t)$ and $s_3(t)$.

$$x_i(t) = \Sigma a_{ij} s_j(t) \qquad \text{for each} \qquad i = 1, 2, 3 \tag{1}$$

$$x(t) = As(t)$$

the column vectors $x(t) = [x_1(t), x_2(t), x_3(t)]$, $s(t) = [s(t), s(t), s(t)]T$.

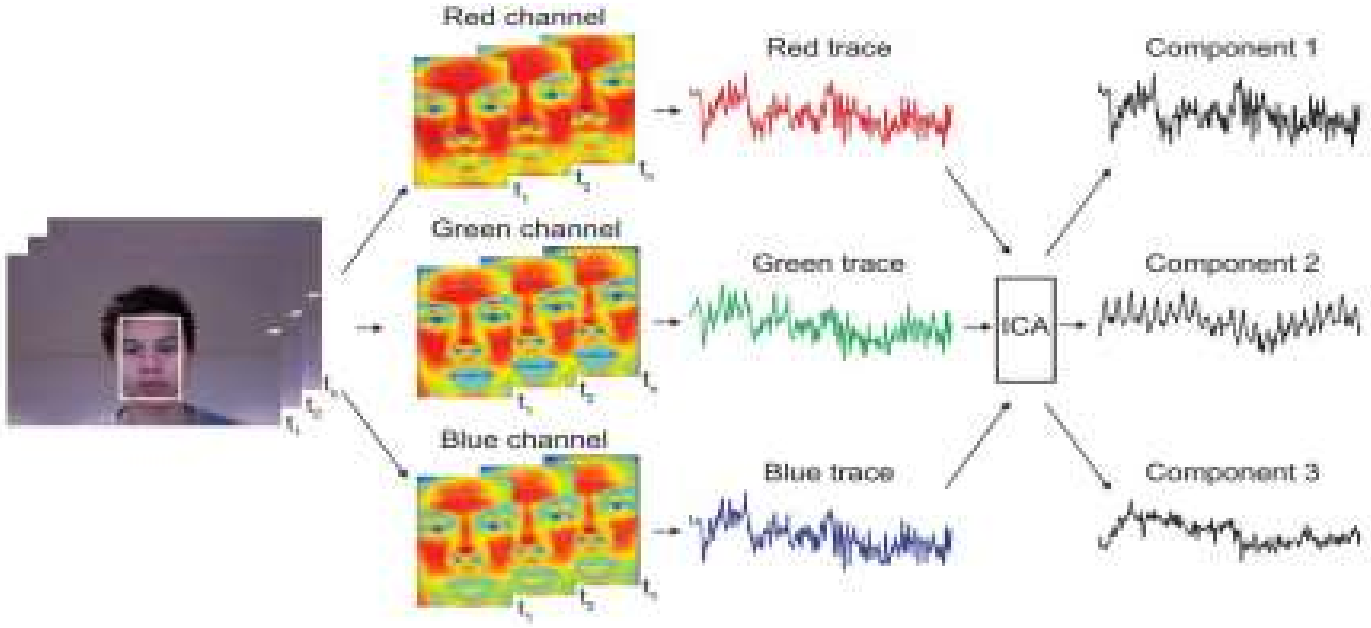

Figure 4: A diagram showing the processing of two source signals with ICA POH [2]

Chrominance-Based (CHROM DEHAN) Method: CHROM DEHAN [1] signal processing method allows obtaining the pulse signal in case of specular and motion artefacts. RGB channels are reflected in a chrominance subspace. Here the movement component is largely eliminated. The CHROM DEHAN method creates a vector using a standard skin tone. It obtains the pulse signal using an alpha setting. However, these settings sometimes may not match the actual situations and as a result, the method may fail. We can see the schema of CHROM DEHAN in Figure 5.

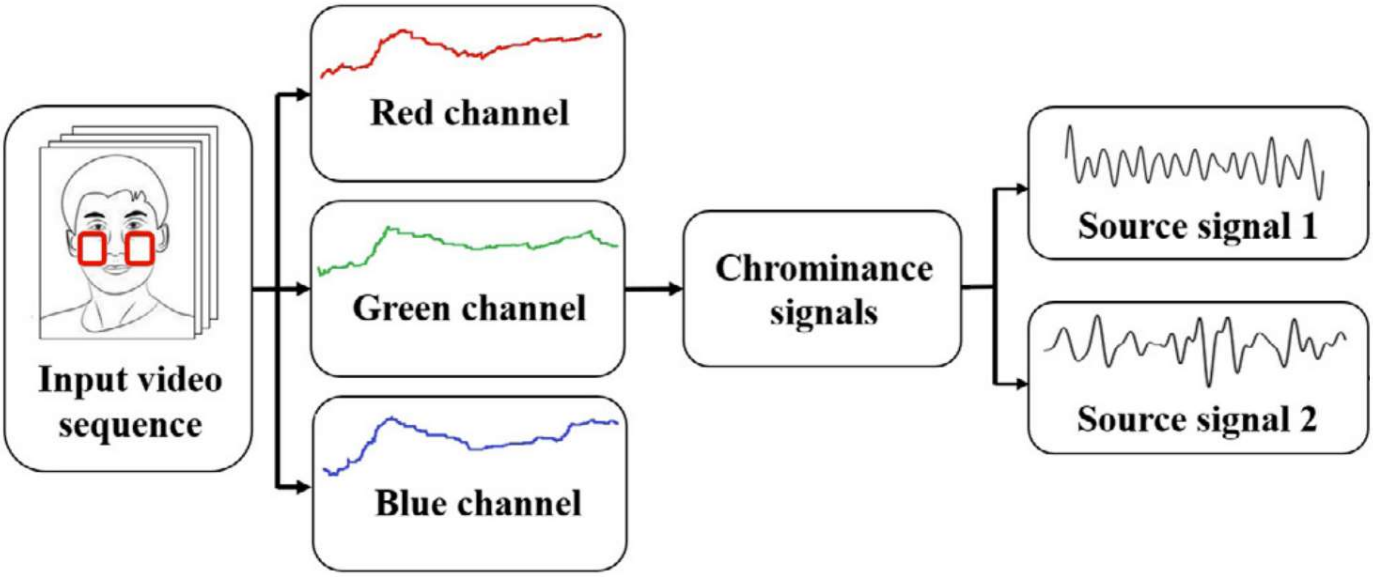

Figure 5: CHROM schema [15]

Green - Vercruysse Method: According to the premise of the GREEN VERCRUYSSE [3] method, the green channel contains the powerful plethysmographic signal, consistent with the fact that haemoglobin absorbs green light better than red and on the other hand passes through sufficiently deeper into the skin as compared to blue light to study the vasculature. This method uses Fourier transforms for filtering. For steps, we can look at Figure 6.

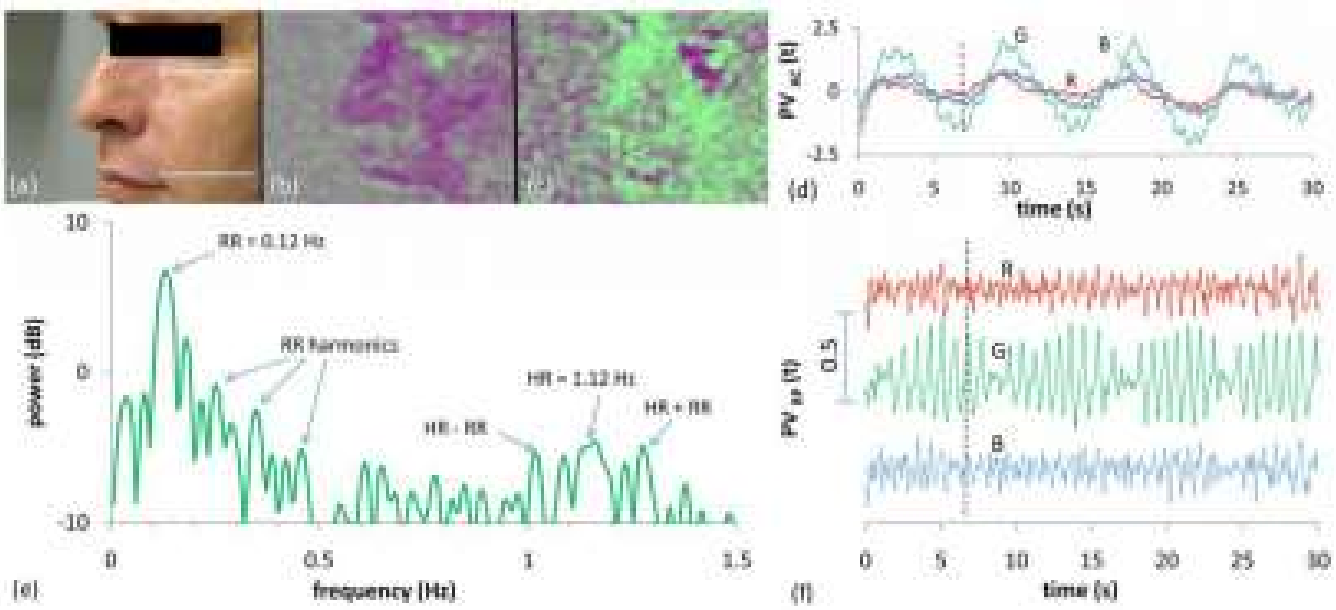

Figure 6: An example of pulse amplitude modulation [3]

Plane-Orthogonal-to-Skin (POS WANG) Method: POS WANG [4] primarily makes skin detection and only takes signals from the skin. The POS WANG algorithm suggests adding the 2SR property to the model. 2SR or data-driven method is a new development. It creates a subject-dependent skin-colour area and tracks the colour-change over time to measure the pulse, also the sudden colour is determined depending on the statistics of the skin pixels.

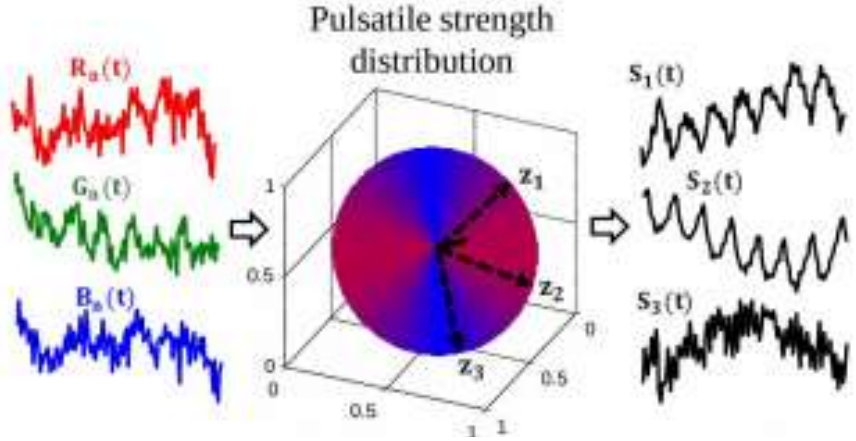

Figure 7: The distribution of the pulsatile strength on the plane orthogonal to 1 as a function of z [4]

From Figure 7, we can see that the projection direction is highly related to the pulsatility that determines the signal quality; different z may give very different projected signals. As we can see in Figure 8, POS WANG and CHROM DEHAN have different volume and reflective variations. In this context, the solid black line shows the primary normal vector and projection axes in both. So, we can say that both have different advantages and disadvantages.

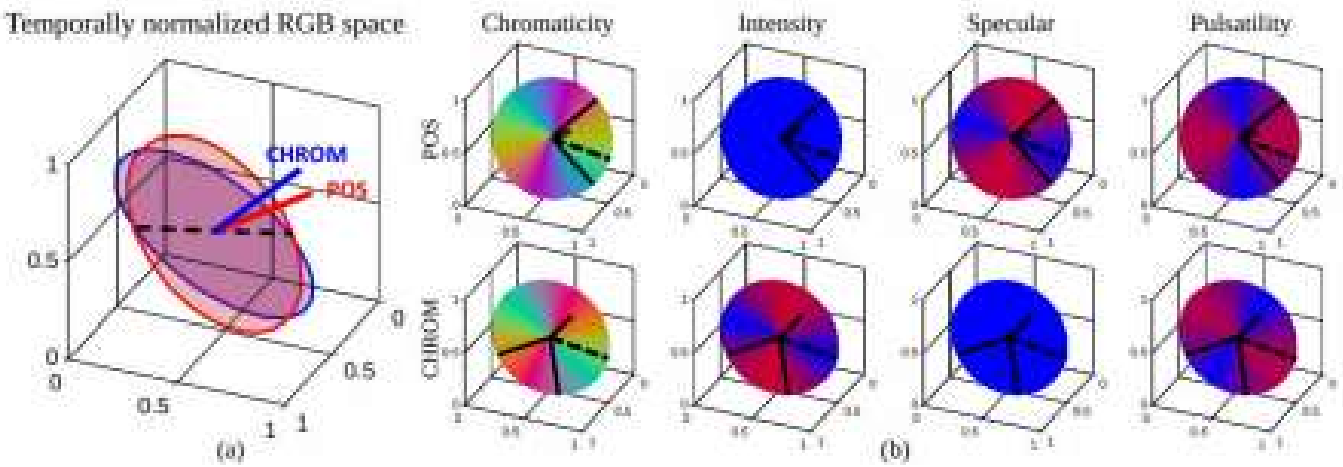

Figure 8: (a) The projection planes of POS and CHROM in the temporally normalized RGB space. (b) The projection planes of POS and CHROM have different chromaticity distributions

3. SYSTEM DESIGN AND SOFTWARE ARCHITECTURE

3.1 System Design

3.1.1 System Model

Traditional Methods: We used Viola-Jones [12] face detection technology to automatically detect the face of the subject. This step provides bounding box coordinates defining the subject face. In the Viola-Jones algorithm, handmade simple Haar features were first created. Then the image was converted into an integral image. The integral image was the calculated version of the source image. Each point in the integral image was the sum of pixels above and to the left of the corresponding pixel in the source image. However, instead of making additions for each pixel value for all features - the integrated image was used to take advantage of several subtractions to achieve the same result.

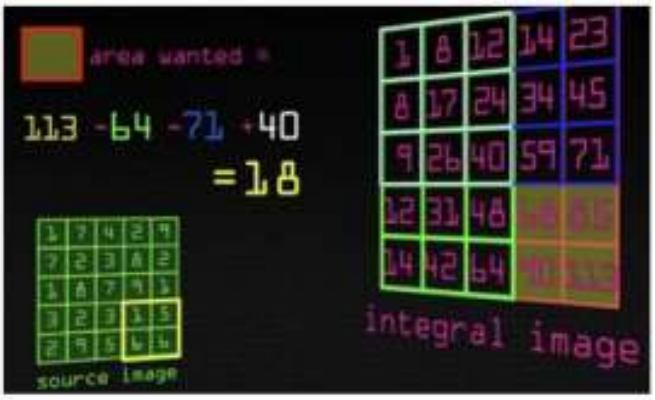

Figure 9: Using the integral image for the area wanted calculation [16]

After that, the delta values for each feature on an image region were calculated and a machine learning meta-algorithm, Adaboost, was trained for each feature. A classifier was created for each Haar feature. And these classifiers were considered "weak" classifiers. A weak classifier was trained for each feature using AdaBoost. When the training was complete, models were sorted by error rate and selected the best weak classifiers based on a threshold value and useful classifiers were added to the attentional cascade. Attentional cascade was a set of weak classifiers that were trained when used together to make a powerful classifier. After that, the cascade was loaded and the image was gradually passed through each classifier and the result was obtained. After

detecting the face, we needed to make skin detection and remove non-skin pixels. The skin detection was performed on every frame to filter out non-skin pixels. The area of interest, the skin part, is our ROI [10] piece. The pixels in the ROI were spatially averaged, the process repeated for each video frame. The result of this process was then used to obtain the rPPG signal.

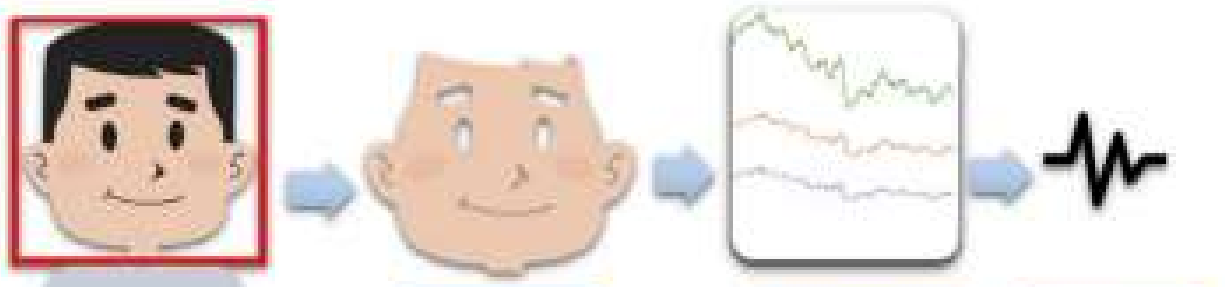

Figure 10: Obtained rPPG signal from ROI [10]

For HR extracting, FFT was applied to the selected signals and their power spectrum was obtained. The frequency corresponding to the highest power of the spectrum in an operational frequency band is determined as the pulse frequency.

Deep Learning-Based Methods: We planned to carry out our measurements with deep learning methods, which was our main approach. We hoped that deep learning reduced error rates as a result of these measurements. We used the model of MTTS-CAN to obtain the heart rate signals. This method processes RGB values captured by cameras with functions that also contain certain calculations for various external factors. These external factors include non-physiological variations such as the flickering of the light source, head rotation, and facial expressions. In this method, there are Temporal Shift Modules that will facilitate the exchange of information between frames. These modules provide superior performance in both latency and accuracy. MTTS-CAN also calculates the respiratory rate along with the heartbeat. Since respiration and pulse frequencies cause head and chest movements of the body, calculating these two values together had a great impact on the accuracy of the values compared to independently calculated models. [5]

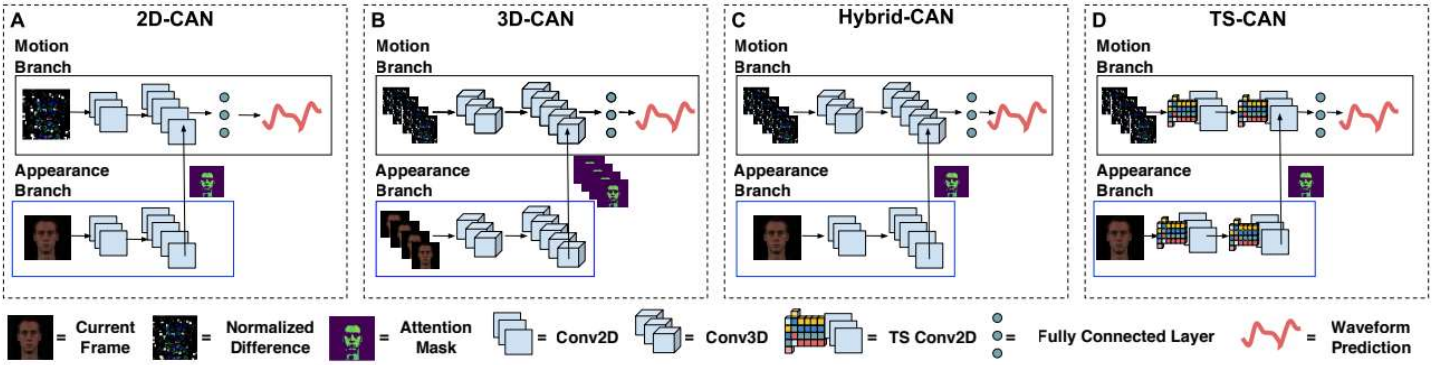

Figure 11: Comparison of TS-CAN study with several convolutional attention networks (CAN) [5]

3.1.2 Flowchart for Proposed Algorithms

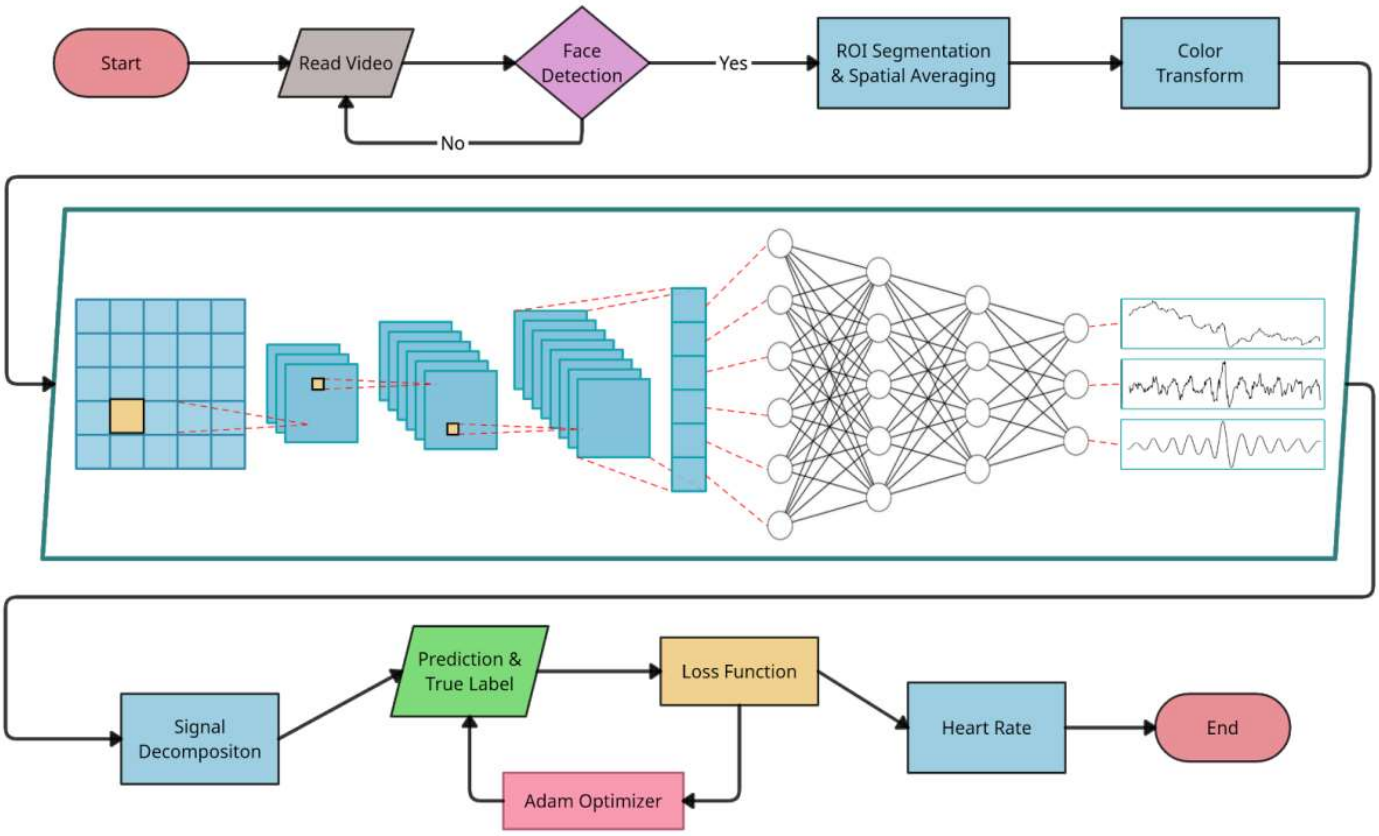

Figure 12: Flowchart of the proposed algorithm

3.1.3 Comparison Metrics

Root Mean Square Error, RMSE [9] is the standard deviation of the prediction errors. Prediction errors are called residuals. Residuals can be measured by how far data points are from the regression line. The RMSE value is a measure of how far these residues have spread.

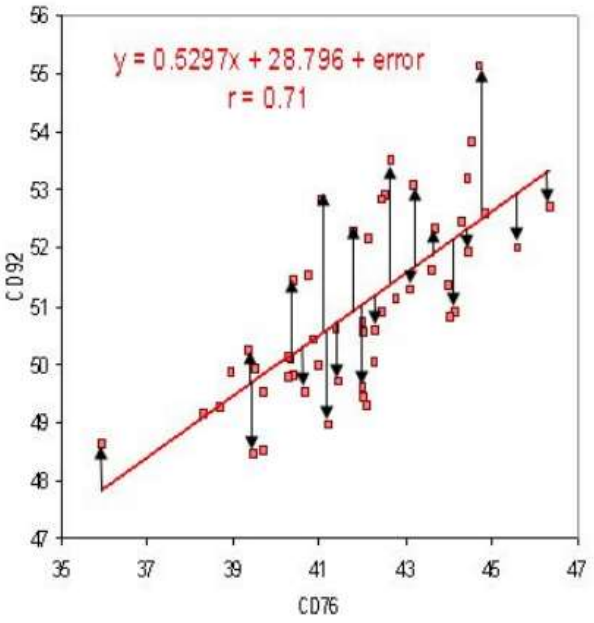

Figure 13: Residuals on a scatter plot [17]

RMSE value can be calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \quad (2)$$

Mean absolute error, MAE [11] is a measure of errors between observations expressing the same phenomenon. MAE is calculated as follows:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n} \quad (3)$$

Signal-to-Ratio, SNR [8], is defined as the ratio of signal power to the noise power and calculated as follows:

$$SNR = \frac{P_{signal}}{P_{noise}} \quad (4)$$

The Mean Signal-to-Noise Ratio, shortly MSNR, is a kind of matrix eigenvalue parsing method. Constructs the SNR function, predicts the separation matrix by eigenvalue decomposition or generalized eigenvalue decomposition. With this algorithm, the closed-form solution can be found without the iterative optimization process. MSNR can be calculated as follows:

$$MSNR = \frac{1}{N}\sum_{k=1}^{N}\{10log_{10}(\frac{S_k(f=f^*)}{\Sigma_{f\in F^*}S_k(f)})\} \quad (5)$$

The correlation coefficient (r) is a measure of how close the points on a scatter plot are to the linear regression line. The correlation coefficient can be calculated as follows:

$$r = \frac{Cov(X, Y)}{\sqrt{s_x^2\, s_y^2}} \quad (6)$$

where Cov(X, Y) is the covariance and can be calculated as follows:

$$Cov(X, Y) = \frac{\Sigma(X - X)(Y - Y)}{n - 1} \quad (7)$$

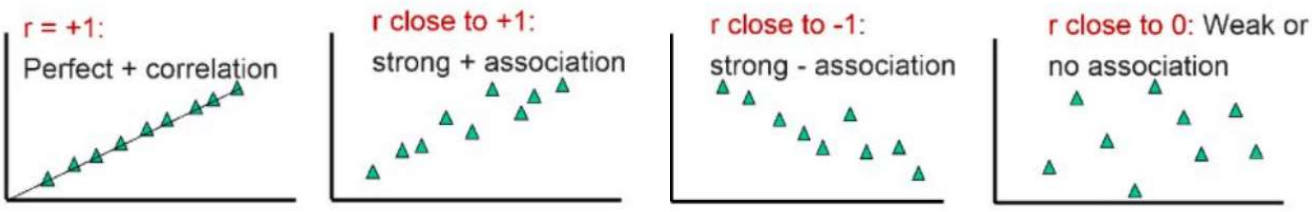

Figure 14: Example scatter plots for correlation coefficient [18]

3.1.4 Data Set or Benchmarks

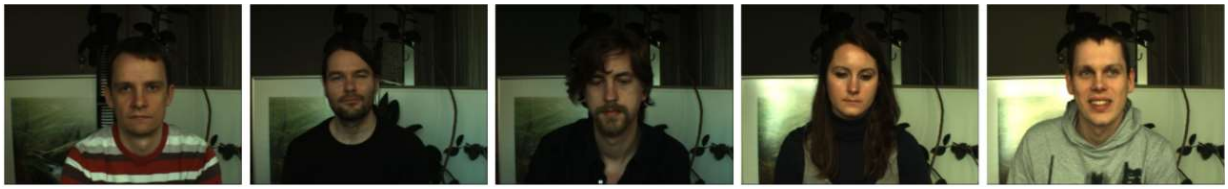

Figure 15: A few examples from the PURE dataset [7]

Pulse Rate Detection Dataset - PURE [7] data set consists of 10 persons (8 male, 2 female) that were recorded in 6 different setups:

- Head is steady
- Talking without head movements
- Slow translation
- Fast translation - twice slow translation speed - average speed was 7% of the face height per second
- Small head rotation up to 20°
- Medium head rotation up to 35°

So there is a total number of 60 sequences of 1 minute each. We can see a few example frames in Figure 16. The image sequences of the head and the reference pulse measurements were recorded. The videos were captured at a frame rate of 30 Hz with a cropped resolution of 640x480 pixels and a 4.8mm lens. Reference data were captured using a finger clip pulse oximeter that provides pulse rate wave and SpO2 readings with a 60 Hz sampling rate. [13]

Figure 16: An example set from the UBFC dataset [6]

In the UBFC [6] database, there are 42 records created with a simple low-cost webcam at 30fps with a resolution of 640x480 in uncompressed 8-bit RGB format. A transmissive pulse oximeter was used to obtain the ground truth PPG data. The subjects sit in front of the camera at a distance of about 1m with their faces visible. The environment was well-lit. The subjects were required to play a time-sensitive mathematical game. This increased their heart rate. All experiments were conducted indoors with a varying amount of sunlight and indoor illumination. We saw some examples in Figure 16.

3.2 System Architecture

First, we read the videos and frame them. We had to detect a face in each frame. The first thing we needed to do predicting heart rate from the video should have been to find the face from the video and crop it. Because foreign objects in the background could cause the algorithm to work incorrectly. At this stage, we used the Viola-Jones face detection technique [12] to automatically detect the face of the subject. This step provided bounding box coordinates defining the subject face. After the face detection, we needed to make skin detection and remove non-skin pixels. The skin detection was performed on every frame to filter out non-skin pixels.

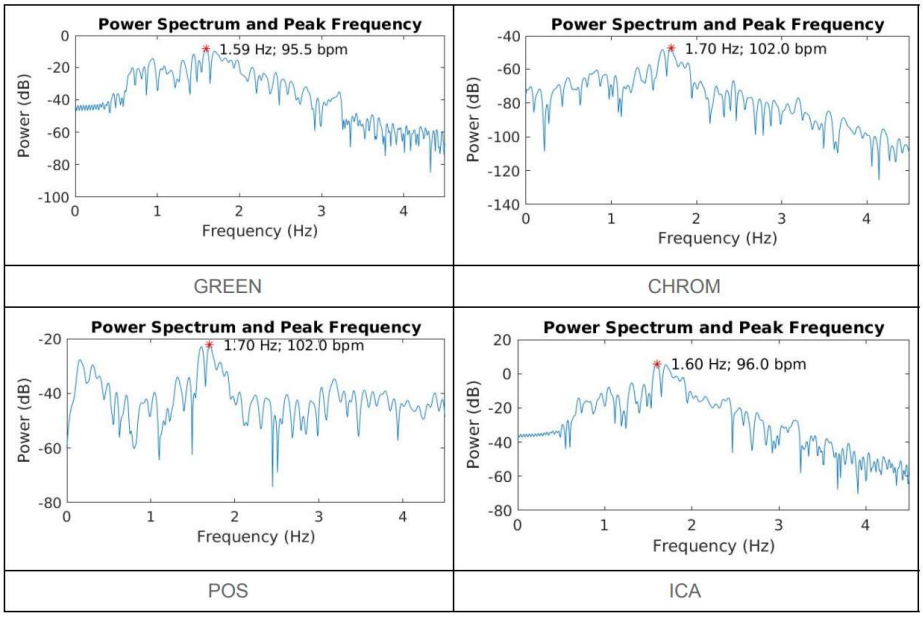

Figure 17: Results of four methods at 13.avi from the UBFC dataset

The area of interest, the skin part, was our ROI piece. After the ROI [10] region was taken, we now had data to search for the information we wanted. We constituted temporal RGB signals by making colour transformations on this data. The raw RGB signals were composed by calculating the average pixel value of the skin pixels within the ROI region over time. The RGB signals contain information about the heart rate in mixed components. Therefore, we used deep learning approaches to recover the source signals from these mixed signals. As we can see a few results in Figure 17. And then, the spectrum of the resulting components of these methods was obtained. The peaks in the components power of these methods were determined, and the index frequency of the highest peak corresponds to the heart rate frequency.

4. TECHNICAL APPROACH AND IMPLEMENTATION DETAILS

4.1 Technical Approach

4.1.1 Traditional Methods

Like traditional methods, we used iPhys library methods:

- CHROM DEHAN [1]
- ICA POH [2]
- GREEN VERCRUYSSE [3]
- POS WANG [4]

Before applying these methods, there are some operations that we all did in common. First, we need to detect the face in the video, for this, we used Viola-Jones, which is a frequently used face detection algorithm. After detecting the face, it is necessary to delete the pixels that do not have skin pixels, such as eyes, eyebrows, hair etc, from the region we worked on, called Region of interest or ROI. After that, we obtained the RGB values in these ROI pixels. We got three different signals with the RGB values we got by doing the same operations for each frame, these are red, blue and green signals. We took and processed the most suitable of these signals for processing. What separates these methods from each other is how the signals were processed. Each method gives us the heartbeat signal as output.

4.1.2 Deep Learning-Based Methods

With deep learning-based methods, we obtained the heartbeat signal from the videos by using the model of MTTS-CAN [5], which is a successful method in this regard. The schematic of MTTS-CAN is above. This model has a two-branch structure. The motion branch is used for motion modelling and the appearance branch for extracting meaningful spatial features. We worked on this by finding power spectrum density with the hamming window of the signal we obtained. First of all, by masking the frequency range we obtained between 1 and 4, we reduced the heart rate value to normal levels. Afterwards, we tried to find the maximum point in this power spectrum and multiply its frequency by 60 to get the heart rate value. While calculating the

error rate here, we found the RMSE [9], for each video. To find this, we used the estimated values and ground truth values obtained on thirty-second video segments with the starting points at one-second intervals for each video.

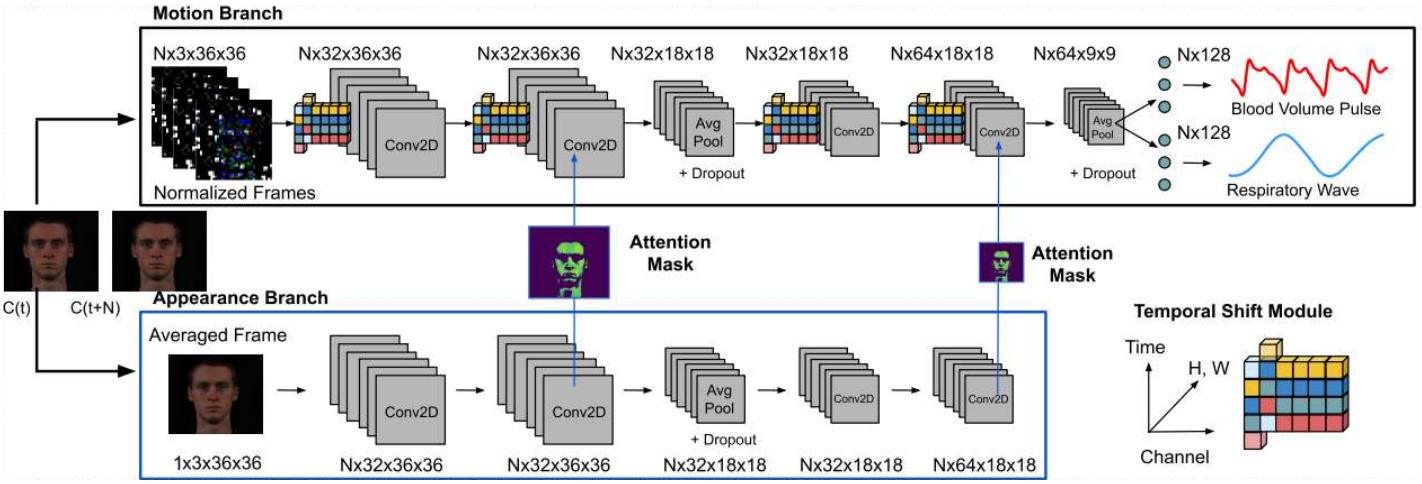

Figure 18: Multi-task temporal shift convolutional attention network for camera-based physiological measurement [5]

We reviewed the results shortly, but the RMSE values for some videos turned out to be quite high. This is due to multiple maximum frequency points in the power spectrum density due to noise in the heartbeat signal. To prevent this, we applied signal-to-noise-ratio, SNR, while calculating power spectrum density. The formula we used for SNR calculations: [8]

$$SNR_{T,i}(f) = \frac{\sum_{j=f-\tilde{f}/2}^{f+\tilde{f}/2} P_{T,i(j)}}{\Sigma_j P_{T,i(j)} - \sum_{j=f-\tilde{f}/2}^{f+\tilde{f}/2} P_{t,i(j)}} \qquad (8)$$

We multiplied the power amplitude and SNR and took the highest value we obtained as the heart rate frequency:

$$F_{T,j} = argmax_f\{P_{T,j}(f)\, x\, SNR_{T,j}(f)\},\ j = 1,...,N \qquad (9)$$

For videos where we applied SNR, the predictions improved. We experimented with changing the ~f constant in the SNR formula to see if we could get even better results.

4.2 Implementation Details

4.2.1 Traditional Methods

In the traditional methods, we implemented CHROM DEHAN, ICA POH, GREEN VERCRUYSSE, POS WANG methods using the UBFC dataset. Before applying each method, we read the videos and split them into frames. Afterwards, we process 30-second sections of these frames at one-second intervals. For this, we first detect the face in each frame and use the skin pixels of this face. We calculate the RMSE value with the ground truth value and heart rate we obtained for each section.

To apply the CHROM DEHAN method, we convert the RGB values in the skin pixels to YCbCr values with the rgb2ycbcr method, then we process the Y, Cb, Cr signals with the help of the Hanning window. Afterwards, we estimate the heart rate with the frequency value at maximum power by taking the power spectrum of the periodogram we obtained.

In the ICA POH method, we first detrend the signal we have, then normalize it to the zero mean and choose the signal with the maximum normalized power as the BVP source. After filtering this signal with the Butterworth 3rd order filter, we normalize it and get the power spectrum density. We calculate the rate of fire covered with the frequency at the highest power obtained.

We get the green one from the RGB signals obtained in the GREEN VERCRUYSSE method. After filtering with the Butterworth 4th order filter, we calculate the heart rate based on the frequency with the highest power from the power spectrum density.

A Color-distortion filter is applied to the signals in the POS WANG method. We calculated the coverage rate by taking the frequency of the maximum power in the power spectrum densities of the signals we obtained.

4.2.2 Deep Learning-Based Methods

In the deep learning-based model, we implemented the MTTS-CAN [5] model using UBFC [6] and PURE [7] dataset with SNR [8] formulas and without SNR formulas.
First of all, we read the relevant video and text in the UBFC dataset and we read images in the PURE dataset. Then we put the ground truth values in the text into the array. After that, we put the resulting pulse prediction signal into the power spectrum density function, and in this way, we obtained the frequency and maximum peaks of the signals.

In the formula without SNR, we gave a lower limit pulse rate of 40 and an upper limit pulse rate of 240. Then, we found the frequency with power spectrum density using pulse prediction signals. So, we obtained Fmask from frequency using these pulse rates. After that, we got the FRange using FMask and in FRange with the maximum index, we got pulse rate multiplied by 60.

In the formula with SNR, we found the frequency with power spectrum density using pulse prediction signals. Then we limited this frequency from 1 to 4. After that, we applied SNR formulas using these frequencies. As a result, we find the value in the frequency with the Maximum index obtained from the SNR formula, and we multiply it by 60 to get the pulse rate.

5. EXPERIMENTAL STUDY

5.1 Experimental Setup

- Experiments must have been done indoors.
- The experiment environment must have been well-lit.
- People must have sat at a table in front of a laptop at a specified distance from the camera.
- The face of the subject should have been visible, so the subject should have not turned his/her head.

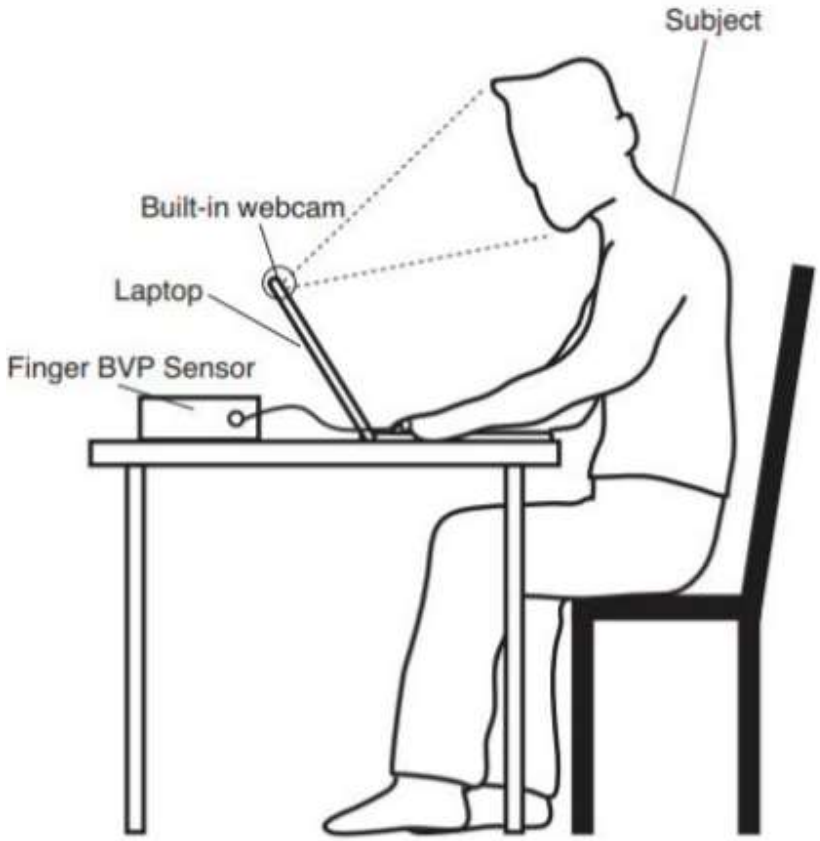

Figure 19: Experimental Setup [2]

5.2 Experimental Results

In Table 1, we could see the results of traditional methods which are CHROM DEHAN, ICA POH, GREEN VERCRUYSSE and POS WANG methods. In a deep learning-based method which is MTTS-CAN. For example, if we look at 17.avi for all methods, we calculated that deep learning has dropped below five. This result is quite good for us. When we look at the average RMSE values of these methods, we see that the deep learning-based method gives the best results because it has the lowest RMSE.

Video from UBFC	Traditional Methods				Deep Learning Based Method
	CHROM DEHAN	ICA POH	GREEN VERKRUYSSE	POS WANG	MTTS-CAN with SNR f~ 0,167
13.avi	10,4263	10,6027	12,358	10,6355	30,172
14.avi	16,284	22,6972	27,8859	24,4553	7,907
17.avi	18,4104	16,5873	12,84	17,787	2,207
31.avi	8,7691	8,6861	21,5627	1,1123	1,740
37.avi	6,0574	6,0336	27,3165	6,0389	3,579
40.avi	7,7885	7,821	16,7349	7,8318	0,854
42.avi	17,4668	16,6722	24,4086	45,6642	18,347
43.avi	2,824	2,8962	41,8808	3,2266	4,237
Average RMSE	14,5	16	27,2	20,6	11,421

Table 1: RMSE values of sample videos from UBFC obtained with traditional methods and deep learning-based method

In Table 2, we could see the results of deep learning-based methods on the UBFC and PURE dataset. When we look at the results we run with the UBFC dataset, we see that the heart rate calculated with the SNR formulas gives better results. For example, Video 37.avi RMSE decreased 8 to 3. However, when we look at the Pure dataset, the results are generally unchanged or worsened. For example, video 03-05 RMSE increased 0.8 to 17, this is pretty bad but there are also unchanging ones in pure, for example, video 10-6. When we look at the general RMSE, the values of UBFC have improved, while those of the pure dataset have worsened.

PURE					
Video	MTTS-CAN Without SNR	MTTS-CAN With SNR f~ 0.167	Video	MTTS-CAN Without SNR	MTTS-CAN With SNR f~ 0.167
01-01	6,215	6,224	06-02	56,048	49,427
01-02	8,843	9,148	06-03	8,733	8,484
01-03	1,908	1,908	06-04	12,230	7,422
01-04	2,665	2,676	06-05	4,817	4,838
01-05	4,744	4,656	01-06	4,410	4,410
06-06	5,346	5,075	02-01	6,262	5,989
07-01	1,814	1,848	02-02	7,462	7,212
07-02	70,445	61,444	02-03	4,717	5,108
07-03	4,046	3,763	02-04	5,512	5,508
07-04	36,659	25,839	02-05	2,548	2,490
07-05	4,352	7,329	02-06	5,059	49,283
07-06	1,694	1,778	03-01	1,771	31,311
08-01	4,110	42,089	03-02	1,725	9,292
08-02	6,807	6,935	03-03	3,615	1,772
08-03	5,508	42,883	03-04	13,747	16,465
08-04	8,154	41,708	03-05	0.83	17,894
08-05	2,920	35,057	03-06	4,942	1,553
08-06	3,008	41,041	04-01	3,138	10,411
09-01	1,792	43,087	04-02	9,352	9,661
09-02	6,261	40,799	04-03	4,035	3,325
09-03	1,519	44,229	04-04	12,682	7,193
09-04	18,135	30,395	04-05	4,736	4,621
09-05	2,351	48,319	04-06	5,616	5,048
09-06	1,508	47,319	05-01	2,991	31,759
10-01	3,422	3,157	05-02	9,795	10,939
10-02	7,227	4,867	05-03	20,956	29,435
10-03	5,343	5,426	05-04	15,198	17,085
10-04	9,330	9,331	05-05	10,126	27,253
10-05	7,794	7,759	05-06	8,009	24,073
10-06	1,173	1,165	AVG	8,449	17,322
06-01	3,150	2,834			

Table 2.a: Examples of RMSE values from processing videos in PURE with and without SNR with MTTS-CAN

UBFC					
Video	MTTS-CAN Without SNR	MTTS-CAN With SNR f~ 0.167	Video	MTTS-CAN Without SNR	MTTS-CAN With SNR f~ 0.167
1.avi	2,595	2,594	32.avi	30,173	31,737
3.avi	25,365	16,918	33.avi	2,726	2,789
4.avi	8,363	8,564	34.avi	1,782	1,356
5.avi	2,197	2,072	35.avi	7,674	4,990
8.avi	22,985	2,431	36.avi	2,980	2,373
9.avi	20,378	21,278	37.avi	8,254	3,579
10.avi	14,309	9,299	38.avi	45,581	45,446
12.avi	1,755	1,738	39.avi	10,534	10,281
13.avi	33,607	30,172	40.avi	0.845	0.854
14.avi	7,999	7,907	41.avi	32,563	14,712
15.avi	36,642	31,380	42.avi	18,307	18,347
16.avi	1,887	1,958	43.avi	4,212	4,237
17.avi	21,278	2,207	44.avi	6,277	6,081
22.avi	21,312	16,867	45.avi	23,332	17,621
23.avi	3,879	3,751	46.avi	2,179	1,982
24.avi	17,172	17,178	47.avi	9,181	9,482
26.avi	18,588	13,132	48.avi	16,655	4,764
27.avi	34,467	17,137	49.avi	9,415	6,936
30.avi	20,665	20,384	AVG	15,052	11,421
31.avi	1,946	1,740			

Table 2.b: Examples of RMSE values from processing videos in UBFC with and without SNR with MTTS-CAN

When we look at Figure 20 we can see the BVP (Blood volume pressure) signal of the 16.avi on the left-hand side. On the right-hand side, we could see a power spectrum density of 16.avi and when we look at the maximum peak in this graph, the frequency corresponds to approximately 1.8, and when we multiply this by 60, we get an average of 108 heart rate.

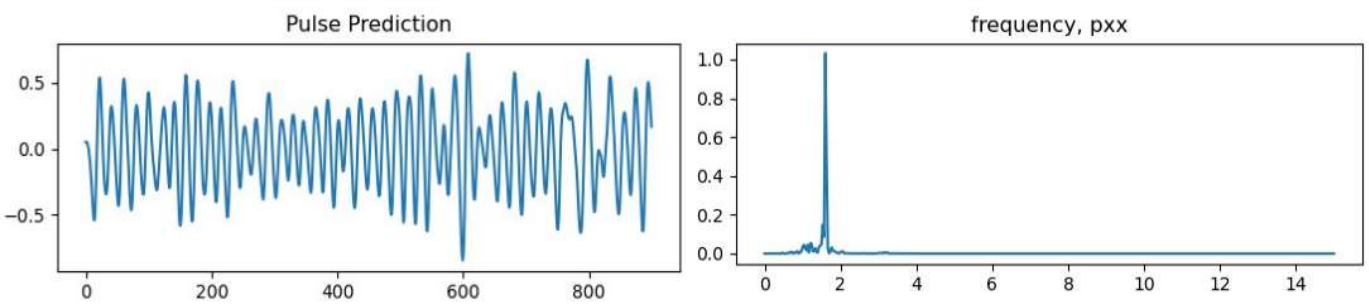

Figure 20: BVP signal and Power Spectrum Density (PSD) of 16.avi from UBFC

In Figure 21, we can see a bad example of pulse prediction. The person may have moved their head in this video. When we look at the power spectrum density of this video, we see more than one peak. In such cases, it becomes difficult to calculate the heart rate and we get wrong results.

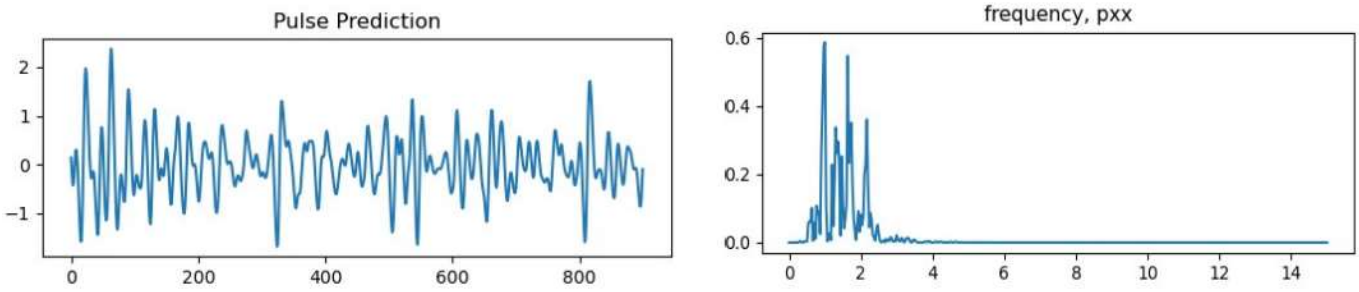

Figure 21: BVP signal and Power Spectrum Density (PSD) of 06-02 from PURE

In Figure 22, We gave different f~ values in the formula we calculated with SNR in the deep learning-based method and observed the changes in the RMSE results. In the Pure dataset, the RMSE results generally remained constant below f~ 0.167. In the UBFC dataset, the RMSE results increased above f~ 0.25. As a result, in the Pure and UBFC dataset, We see that f~ 0.167 gives the best result.

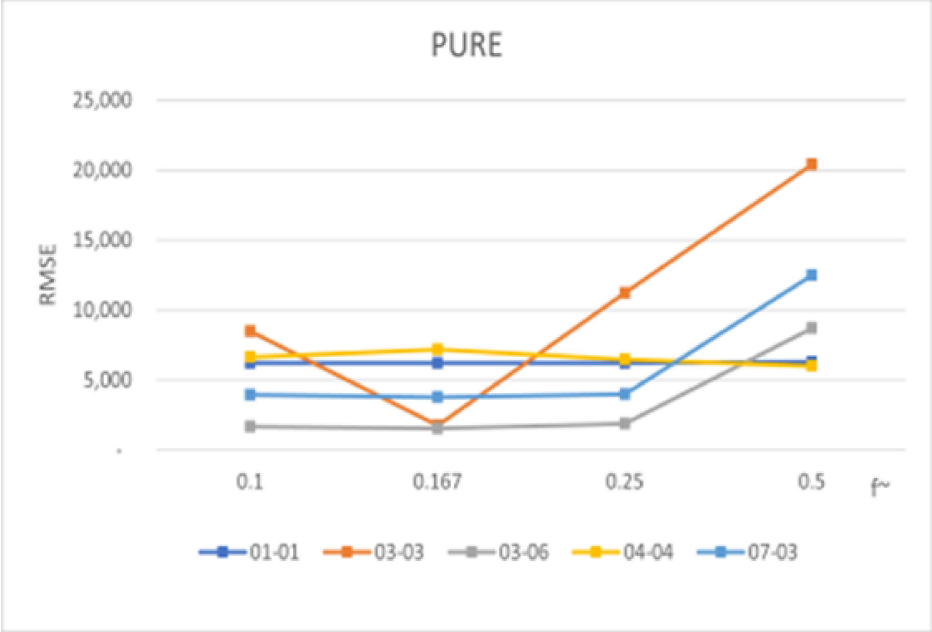

Figure 22.a: RMSE values obtained by calculating some videos from PURE with different f~ values of BVP signals from MTTS-CAN

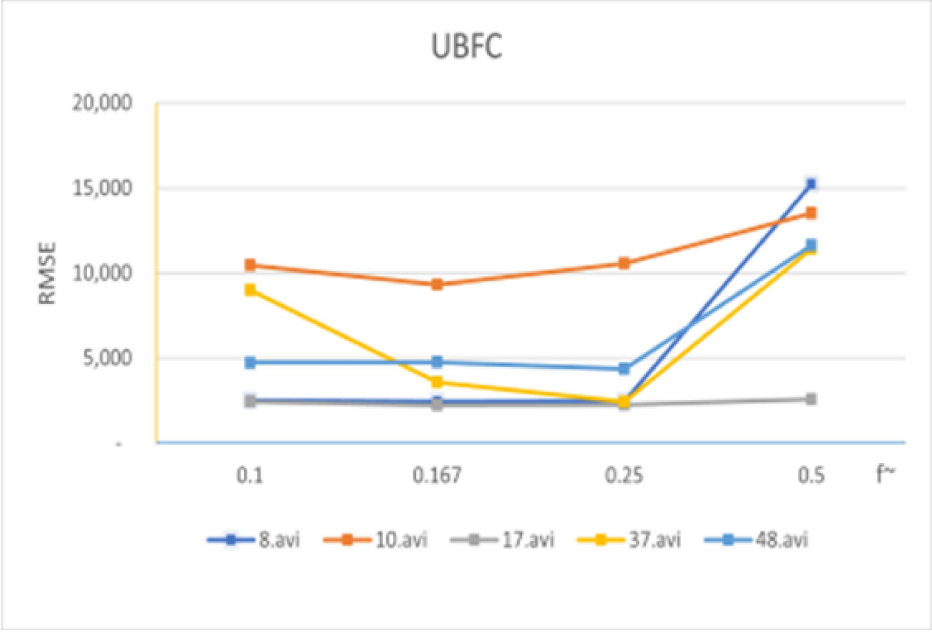

Figure 22.b: RMSE values obtained by calculating some videos from UBFC with different f~ values of BVP signals from MTTS-CAN

We can see a demo of our videos in Figure 23. We run each video with the MTTS-CAN model and get the signals that appear under the videos. We processed these signals using the SNR formula and obtained the values in the upper left corners as the estimated value. In the upper right corners, there are the actual heart rates measured with a smartwatch. We can see that the values are very close to each other. So we can say that the error rates are very low.

Figure 23: Our demo results

6. CONCLUSION AND FUTURE WORK

We worked with the traditional methods in the first term and worked with deep-based methods in the second term. According to the information from literature studies and our studies throughout the year, we can say that deep learning-based methods generally give more correct and faster results than traditional methods. In addition, when we used SNR to calculate heart rate based on the Blood Volume Pulse (BVP) signal resulting from deep learning-based methods, we observed a significant improvement in some results. As a result, we can say that deep learning-based methods play an important role in the development of rPPG technologies and their introduction into our daily lives.

In the pandemic period, telehealth and remote health monitoring have become increasingly important and people widely expect that this will have a permanent effect on healthcare systems. These tools can help reduce the risk of discovering patients and medical staff to infection, make healthcare services more reachable, and allow doctors to see more patients. In this context, we believe that it will find a place both in health centres and in all kinds of electronic devices. As we can see from the technology news that comes out every day, leading universities of education and leading companies in technology have also concentrated on rPPG studies and both contribute to the literature with research to solve the problems in rPPG or develop new methods. In the next few years, it seems quite possible to open the front camera of our mobile phone and measure our heart rate while sitting at home. Of course, there is no limit to the number of applications to which this technology will be integrated.

REFERENCES

[1] De Haan, G., & Jeanne, V. (2013). Robust pulse rate from chrominance-based rPPG. IEEE Transactions on Biomedical Engineering, 60(10), 2878-2886

[2] Poh, M. Z., McDuff, D. J., & Picard, R. W. (2010) Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics express, 18(10), 10762-10774

[3] Vercruysse, W., Svasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. Optics express, 16(26), 21434-21445.

[4] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," IEEE Transactions on Biomedical Engineering, vol. 64, no. 7, pp. 1479–1491, 2016

[5] Xin Liu, Josh Fromm, Shwetak Patel, Daniel McDuff, "Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement", NeurIPS 2020, Oral Presentation (105 out of 9454 submissions)

[6] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, (2017), Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognition Letters

[7] Stricker, R., Müller, S., Gross, H.-M. "Non-contact Video-based Pulse Rate Measurement on a Mobile Service Robot" in Proc. 23st IEEE Int. Symposium on Robot and Human Interactive Communication (Ro-Man 2014), Edinburgh, Scotland, UK, pp. 1056 - 1062, IEEE 2014

[8] Remote Photoplethysmography Using Nonlinear Mode Decomposition, Halil Demirezen, Cigdem Eroglu Erdem Marmara University Department of Computer Engineering, Goztepe, Istanbul, Turkey, pp. 1060– 1064, 2018.

[9] https://en.wikipedia.org/wiki/Root-mean-square_deviation (access time: 28.06.2021 14:05)

[10] W. Wang, S. Stuijk, and G. De Haan, "Unsupervised subject detection via remote ppg," IEEE Transactions on Biomedical Engineering, vol. 62, no. 11, pp. 2629–2637, 2015

[11] https://en.wikipedia.org/wiki/Mean_absolute_error (access time: 28.06.2021 14:10)

[12] Paul Viola and Michael Jones (2001), Robust Real-time Object Detection, International Journal of Computer Vision

[13] Stricker, R., Müller, S., Gross, H.-M. "Non-contact Video-based Pulse Rate Measurement on a Mobile Service Robot" in Proc. 23st IEEE

[14] www.joom.com (access time: 06.02.2021 12:10)

[15] New insights on the super-high resolution for video-based heart rate estimation with a

semi-blind source separation method Rencheng Song, Senle Zhang, Juan Cheng, Chang Li, Xun Chen

[16] medium.com/@aaronward6210 (access time: 05.02.2021 22:26)

[17] www.statisticshowto.com (access time: 23.01.2021 17:28)

[18] sphweb.bumc.bu.edu (access time: 28.01.2021 19:12)

APPENDICES

Inside a CD:

- Presentation Video
- Original Poster
- Thesis
- Source Codes