B.TECH. PROJECT REPORT

On

Heart Rate Estimation using Non-contact Face Videos

BY

ASHISH GAWAI

VANDANA VARAKANTHAM

RISHIKA PATEL

DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

December 2019

Heart Rate Estimation using Non-contact Face Videos

A PROJECT REPORT

Submitted in partial fulfillment of the
requirements for the award of the degrees

of
BACHELOR OF TECHNOLOGY
in

DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING

Submitted by:
ASHISH GAWAI (160001009)

VANDANA VARAKANTHAM (160001060)

RISHIKA PATEL (160001050)

Guided by:
Dr. PUNEET GUPTA
Assistant Professor, Discipline of Computer Science and Engineering, IIT Indore

INDIAN INSTITUTE OF TECHNOLOGY INDORE
December 2019

Declaration of Authorship

We hereby declare that the project entitled "**Heart Rate Estimation using Non-contact Face Videos.**" submitted in partial fulfillment for the award of the degree of Bachelor of Technology in 'computer science and engineering' completed under the supervision of **Dr. Puneet Gupta** (Assistant professor) computer science department, IIT Indore is an authentic work.

Further, we declare that we have not submitted this work for the award of any other degrees elsewhere

Signature and name of students:

_____

Date:

Certificate

This is to certify that the project entitled "**Heart Rate Estimation using Non-contact Face Videos**." and submitted by **Ashish Gawai, Vandana Varakantham, Rishika Patel**. It is certified that the above declaration statement made by the students is correct to the best of my knowledge.

_____

Supervisor

Dr. Puneet Gupta

Assistant Professor,

Indian Institute of Technology

Indore

Date:

Preface

This report on “Heart rate estimation using non-contact face videos" is prepared under the guidance of Dr. Puneet Gupta (Assistant professor). Through this project, we have tried to give a detailed design of how we can measure heart rate without the physical contact of any devices and only through face videos, which is economically feasible. We have tried to the best of our abilities and knowledge to explain the content.

Ashish Gawai (160001009)
Vandana Varakantham (160001060)
Rishika Patel (160001050)
B.Tech 4th Year,
Indian Institute of Technology Indore.

Acknowledgments

We wish to thank Dr. Puneet Gupta (Assistant professor) for his kind support and valuable guidance. It is his help and support, due to which we became able to complete the design and technical report.

Overseeing the project meant a lot, and we learned a lot from it. We thank him for his time and efforts. Without his support, this report would not have been possible.

Abstract

In this project, we present a non-contact way of estimating heart rate. There are several methods for measuring HR. Some of the devices which are most commonly used are ECG (Electrocardiogram) devices and pulse oximetry sensors. Despite being useful, these devices can cause discomfort, irritation, or pain to the skin. That is why new solutions for non-contact measurements of HR using PPG signals have been proposed. It is possible to extract HR from the color variations and light absorption of the facial skin and even through head motions. We use face videos to extract physiological parameters using color variations and head motions of the skin. Here we try to analyze the accuracy of the proposed system, which must give similar results to the ones obtained with conventional HR monitoring systems, as mentioned above. We analyzed our results in different cases by applying different algorithms and in different scenarios. We then did the pre-processing of the data set and input them to the CNN model, after which we train and test the data-set. In this project, we give the demonstration of a low-cost, accurate video-based method for non-contact HR measurements that is automated, motion-tolerant, and which is technically and economically sound and feasible.

Contents

List of Figures:

List of Tables:

List of Abbreviations

ECG	Electro Cardio Gram
PPG	Photo Plethysmo Graphy
CNN	Convolutional Neural Network
VIS	Visible Light Source
PCA	Principal Component Analysis
FFT	Fast Fourier Transform
BPM	Beats Per Minute
BVP	Blood Volume Pulse
RGB	Red Green Blue

Chapter 1

Introduction

1.1 Background

Human health is a vital factor in society's growth and progress. The HR has been estimated using conventional electronic sensors such as ECG devices, oximeter sensors, thermal imaging sensors, and few other devices. Besides, commercially available wearable devices such as fitness watches, chest bands, and ankle belts have also been used to estimate the HR of a person. However, such devices are not commonly available or not always accessible or are expensive in daily life. Further, these devices can cause harm to the body or can cause irritation, discomfort, or pain when used for long periods, and it is difficult to use for the people who have fragile skin. Besides, these devices can damage the fragile skin of premature newborn babies or older people. So, a non-contact means of detecting HR is preferable in such cases. Non-contact HR measurement through a webcam would also aid telemedicine and allow an ordinary person to track HR without purchasing special devices. It is possible to extract HR from the color variations and light absorption of the facial skin and even through head motions. For this, we use face videos to extract physiological parameters using color variations and head motions of the skin. So, to do this, we generally use video cameras, as they are often low in cost solutions for sensing and are readily available. The non-contact physiological parameters of measuring HR have been derived from the cardiovascular system of the human body. The cardiovascular system helps the blood to flow in the body through continuous blood pumping by the heart. The heart pumps the blood through the blood vessels of this cardiovascular system, and for each heartbeat, blood flow produces color variations in the facial skin. So, it is possible to extract HR from the color variations and light absorption of the skin.

Recent developments in this field have led to automated non-contact ways for the estimation of HR. These methods are also inexpensive. So, we can estimate the HR using ambient light and a camera. Due to the moderate price of the cameras and the fact that the method is comparatively easy to execute in, for example, a smartphone app gives this method the possibility to become famous. To our knowledge, for non-contact measurement of HR, there are no such simple methods. However, conventional methods for continuous monitoring of HR are in practice.

1.2 Problem Statement

In this project, we analyzed the methods of measuring HR with low-cost cameras and ambient light developed by BCG Balakrishnan [4], Chrom Dehan [3], Green Channel [1], and ICA Poh [2] in the hope of achieving similar results. There are several methods for measuring HR, and the most common ones are Electrocardiogram Devices and Pulse oximetry sensors. Despite being useful, these devices can cause discomfort, irritation, or pain. That is why new solutions for non-contact measurements using photoplethysmography have been proposed. It is possible to obtain the HR from the color variations and light absorption of the facial skin. We use face videos for this purpose. From face videos, we extract the physiological parameters using color variations of the skin. Cameras are often low-cost solutions for sensing. Here we try to analyze the accuracy of the system, which must give similar results to the ones obtained with conventional HR monitors like Pulse oximetry sensors. We analyzed the result in different cases and scenarios, and we analyze the HR using CNN and train and test the data set.

Chapter 2

Algorithm Analysis and Objectives

We have used four different algorithms in the analysis of HR, which are Green Channel [1], ICA Poh [2], Chrom Dehan [3], and BCG Balakrishnan [4]. The detailed proposed methods of the mentioned algorithm are as follows:

2.1 BCG BALAKRISHNAN

In this algorithm, we use the small head motions caused by the cardiac cycle for the extraction of information about the cardiovascular activities from the face videos. The periodic movement of blood starting from the heart to the head through the carotid arteries and the abdominal aorta causes the periodic movement of the head. Our algorithm detects the pulse from this movement.

The process involved in the algorithm:

1. Region Selection and Tracking
2. Temporal filtering
3. PCA Decomposition.
4. Signal Selection
5. Peak Detection

2.2 CHROM DEHAN

Chrom robust pulse rate from chrominance based rppg:

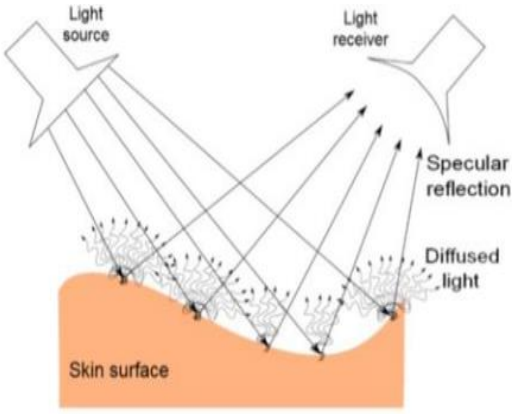


Figure 1: Illustration of specular and diffuse reflection [1]

Figure 1 shows the interaction between the skin surface and the light source. The light is incident

on the skin surface, and there are two types of reflection from the skin surface, specular and diffused reflection. The diffuse reflection changes color with the blood volume of the skin by reflecting the light inside the body, while the specular reflection displays the color of the light source and is not affected by changes in blood volume.

The process involved in the algorithm:

1. In image number I, the intensity of a given pixel registered by the camera in color channel $C \in R, G, B$, can be modeled as $C_i = I_{Ci}(\rho C_{dc} + \rho C_i + S_i)$. Here I_{Ci} indicates the intensity of the light source for the color channel C, ρC_{dc} shows the stationary part of the color channel C skin reflection coefficient, ρC_i displays the zero-mean time fraction of the blood volume pulse, and S_i is the additive specular reflection contribution.
2. Under white light, the normalized skin tone, $[R, G, B]/\sqrt{(R^2 + G^2 + B^2)}$ is the equal for everyone: $[R_s, G_s, B_s] = [0.7682, 0.5121, 0.3841]$ after assuming a fixed skin-tone.
3. $R_s = 0.7682R_n$, $G_s = 0.5121G_n$, $B_s = 0.3841B_n$, where R_s, G_s, and B_s are the standardized RGB channels.
4. Assuming white light, to the diffuse reflection component of all channels, the specular reflection adds equal component of specular fraction of white light. By this we can understand that specular reflection component can be removed by using color difference and this is chrominance signals. We can create two orthogonal chrominance signals called X and Y from three color channels RGB.

$$X_s = \frac{R_s - G_s}{0.7682 - 0.5121} = 3R_n - 2G_n.$$

$$Y_s = \frac{R_s + G_s - 2B_s}{0.7682 + 0.5121 - 0.7682} = 1.5R_n + G_n - 1.5B_n.$$

5. $S = X_f - \alpha Y_f$, $\alpha = \sigma(X_f)/\sigma(Y_f)$ where $\sigma(X_f)$ and $\sigma(Y_f)$ are the standard deviations of X_f and Y_f, respectively, and we used X_f and Y_f, the band passed filtered versions of X_s, Y_s, respectively for best results.

2.3 GREEN CHANNEL

The process involved in the algorithm:

1. PPG's theory is that blood absorbs more light than the surrounding tissue, and changes in blood volume have a reciprocal effect on transformation or reflection.
2. Pixel values (PV, 8bit, 0-255) for RGB channels are read for each of the movie frames, which provides the PV(x, y, t) values, where x is the horizontal position, y is the vertical position, and t denotes the time of the frame rate.
3. Filter and normalize.
4. On PV(t) signals, perform fast fourier transforms in Matlab and determine the pulse rate.

2.4 ICA POH

The process involved in the algorithm is ROI Detection, ROI separation into RGB channel and raw trace, independent component analysis of RGB traces, component selection (Always select second component), FFT on the selected component, and peak detection.

2.5 Improvements in BCG Balakrishnan

We made the following changes in the BCG Balakrishnan algorithm to get more accurate HR.

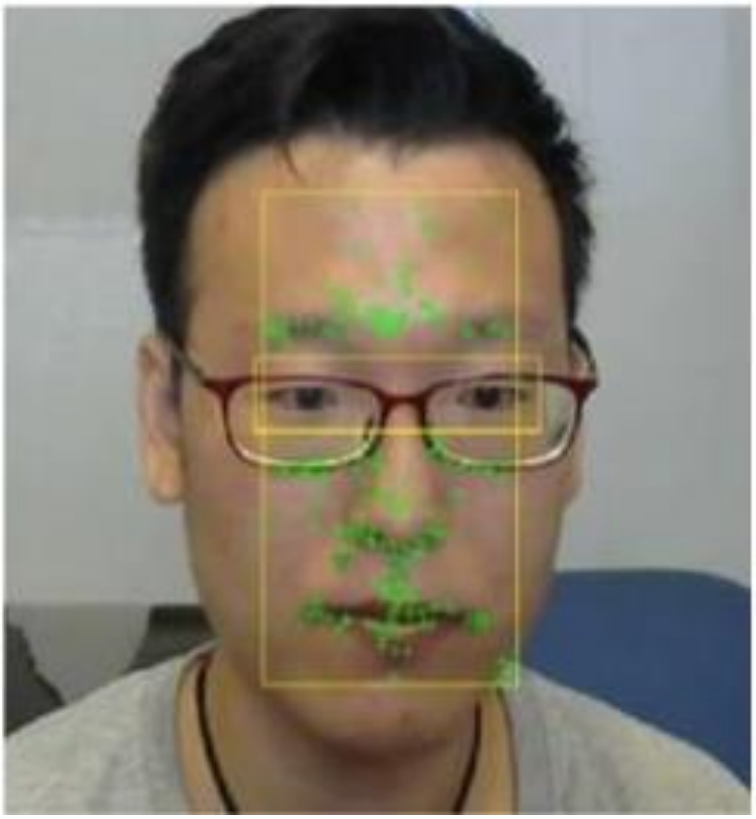


Figure 2: Illustration of ROI detection [4]

In the original algorithm, they used OpenCV Matlab inbuilt tool for face detection. Moreover, sometimes blinking affect our results. Thus, we made changes in the code. We detect face without using Matlab open CV toolbox to detect eyes. As can see in Figure 2, it tracks the feature points on the face and removes features point from the eyes area. From the rectangle, we choose to use the middle 50 percent width-wise and the top 90 percent height-wise for ensuring that the whole rectangle is within the facial area. Due to the above changes, in some scenarios, we are getting good results.

Chapter 3

Experiments and Datasets

We analyzed our results using COHFACE and VIPL-HR datasets.

3.1 Analysis of COHFACE dataset

In the COHFACE dataset, there are four videos of about 1 minute of each Person.

- Two videos with good conditions
- Two videos with more natural (degraded) conditions.

The dataset comprises of RGB video sequences of faces, synchronized with human vital parameters such as HR and breathing rate of the recorded individuals that we consider to be subjects. The data-set contains 40 subjects with 160 one-minute-long RGB video sequences in the COHFACE data-set. In most cases, we are getting more equivalent results in the BCG algorithm. However, in the other three algorithms, the results are fluctuating due to some anomalies. After a detailed study, we came to know that IPPG signals in the COHFACE data-set are corrupted. So, there is much variation in the result of the remaining algorithms. So, for further study, we used the VIPL-HR dataset.

3.2 Analysis of VIPL-HR Dataset

We used three scenarios, and the frame rate of all three scenarios is about 25fps.

- V1 (stable scenario): The subject is asked to sit naturally in front of the camera at a distance of 1 m. The ceiling lamp of the room is turned on, and the filament lamp is turned off.
- V5 (bright scenario): The filament lamp is turned on. Other settings are the same as a stable scenario.
- V6 (long-distance scenario): The subject is asked to sit naturally in front of the camera at a distance of 1.5 m. Other settings are the same as the stable scenario.

3.2.1 Analysis of results using VIPL-HR dataset

In VIPL-HR Dataset, we analyzed our results using three scenarios, v1 (stable scenario), v5 (bright scenario), and v6 (long-distance scenario). Detailed analysis of the scenarios using sample videos follows:

For Stable Scenario (V1):

- Sample Video 1:

P18v1source1

Results:
Ground Truth = 74.66 BPM.

1. BCG BALAKRISHNAN = 72.64 BPM.
2. CHROM DEHAN = 72.5 BPM.
3. GREEN CHANNEL = 72.5 BPM.
4. ICA POH = 72.5 BPM.

Variations due to anomalies in the stable scenario:

- Sample Video 2:

P54v1source1

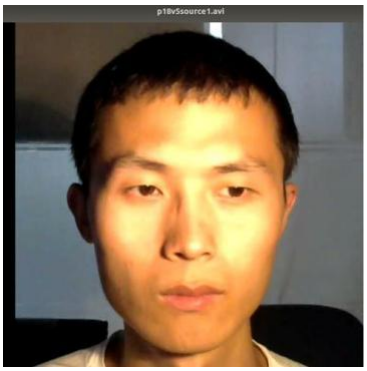
Results:
Ground Truth = 79.96 BPM.

1. BCG BALAKRISHNAN = 67.28 BPM.
2. CHROM DEHAN = 84 BPM.
3. GREEN CHANNEL = 85.5 BPM.
4. ICA POH = 57.5 BPM.

The variations in the results are due to the continuous motion of the person in the video.

For Bright Scenario (V5):

- Sample Video 3:



P18v5source1

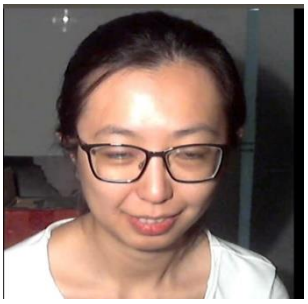
Results:

Ground Truth = 85.87 BPM.

1. BCG BALAKRISHNAN = 64.60 BPM.
2. CHROM DEHAN = 123.5 BPM.
3. GREEN CHANNEL = 123 BPM.
4. ICA POH = 122.5 BPM.

Variations due to anomalies in the bright scenario:

- Sample Video 4:



P78v5source1

Results:

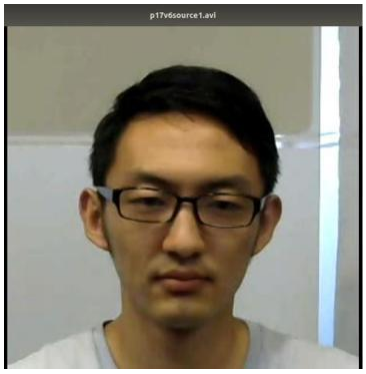
Ground Truth: 85.87 BPM.

1. BCG BALAKRISHNAN = 64.60 BPM.
2. CHROM DEHAN = 123.5 BPM.
3. GREEN CHANNEL = 123 BPM.
4. ICA POH = 122.5 BPM.

In this video, the person is moving continuously, which is causing the variations in the results.

For Long distance scenario (V6):

- Sample Video 5:



P17v6source1

Results:

Ground Truth: 74.93 BPM.

1. BCG BALAKRISHNAN = 70.43 BPM.
2. CHROM DEHAN = 70.5 BPM.
3. GREEN CHANNEL = 71.5 BPM.
4. ICA POH = 70.5 BPM.

Variations due to anomalies in the long distance scenario:

- Sample Video 6:

P74v6source1

Results:

Ground Truth: 79.96 BPM.

1. BCG BALAKRISHNAN = 89.73 BPM.
2. CHROM DEHAAN = 114 BPM.
3. GREEN CHANNEL = 50.5 BPM.
4. ICA POH = 83.5 BPM.

The variations in the results are due to the motion of the person in the video. The person is asked to sit naturally in front of the camera at a distance of 1.5 m.

3.3 Analysis of three different scenarios

- In the case of a stable scenario, we are getting good results.
- If there is slight or moderate motion in videos, it affects the results.
- Distance between person and camera also affects the result. (Shown in the third scenario).

Table 1: RMSE of algorithms in the stable scenario of VIPL-HR

ALGORITHM	RMSE
BCG BALAKRISHNAN [4]	0.138462
CHROM DEHAN [3]	1.313342
GREEN CHANNEL [1]	0.193144
ICA POH [2]	0.877131

We are comparing the dataset for different algorithms according to their subtle changes. We used the dataset to estimate the precision of the modified algorithm and standard algorithm to estimate the accuracy.

Chapter 4

Design Proposal and Experiments – CNN

We implemented the end-to-end system for video-based measurement of HR using a deep convolutional network. Using this mechanism, we estimated HR.

4.1 The pipeline of CNN

Table 2: Implementation of CNN architecture

INPUT SIZE	TYPE/STRIDE	FILTER SHAPE
72X72X3	Input A (RGB Frame)	
72X72X3	Input B (Normalized Frame)	
68X68X100	Conv2D	5X5X3X100
68X68X100	Conv2D	5X5X3X100
66X66X120	Conv2D	3X3X120
66X66X120	Conv2D	3X3X120
33X33X120	Average Pooling	2 X 2
33X33X120	Average Pooling	2 X 2
31 X 31 X120	Conv2D	3X3X120
31 X 31 X120	Conv2D	3X3X120
29X29X90	Conv2D	3X3X90
29X29X90	Conv2D	3X3X90
14X14X90	Average Pooling	2 X 2
14X14X90	Average Pooling	2 X 2
14X14X90	Multiply	(MultiplyA.output, B.output)
17460	Flatten	
1X1X512	Dense	512 Neurons (ANN)
1X1X512	Dropout	Ratio of 0.3
1X1X32	Dense	32 Neurons (ANN)
1X1X32	Dropout	Ratio 0.5
1X1X1	Dense	Output
	Check the output with fitted data (HR Values) and arrange accordingly	
	OUTPUT (AVG HR)	

4.2 HR Estimation using CNN

We are using a supervised learning technique, where we trained the normalized frame and face mask based on the corresponding HR. HR is estimated using the normalized frame difference and RGB videos. These feature frames are input in a network, and the corresponding HR is output. In our architecture, we used two inputs, and the input images are defined to be of the size 72 X 72 X 3 inspired by VGG-style CNN. Depth wise separable convolution is used as the main structure in this part.

The first layer is a 2D fully convolutional layer, referred to as 'Conv2D'. Moreover, we get the first full convolutional layer by using a kernel size of 5 X 5 X 3 X 100, where 5 X 5 denotes the filter's height and width, respectively. 3 represent RGB color channel, and 100 is used to represent the output feature map (Inspired by VGG style CNN). In the next depth wise convolutional layer, we used a kernel size 3 X 3 X 120 for filtering the input feature map. Where 3 X 3 indicates height and width, and 120 indicates the number of input channels. Then we applied average pooling instead of max-pooling of size 2 X 2. The next depth wise convolutional layer, a kernel size 3 X 3 X 90, is used for filtering the input feature map. Here 3 X 3 denotes height and width, and 90 indicates the number of the input channels. Again, we applied average pooling instead of max-pooling of size 2 X 2. A similar structure is applied to the second input. Finally, we multiply these two outputs generated by applying the convolutional layer and average pooling. After multiplying, we used 'Flatten' to transform the 2D matrix of features into a vector that can be fed into a fully connected neural network classifier. Then we applied a fully connected layer having size 1 X 1 X 512 where 1 X 1 is height and width and 512 neurons, which will be fully connected. Then a dropout layer is used to avoid problems of overfitting. The dropout ratio is set to 0.3. Again, we applied a fully connected layer of size 1 X 1 X 32, where 32 neurons will be fully connected. Moreover, again, the dropout ratio is applied, which is 0.5. The last fully connected layer has one neuron. Finally, this architecture generates a number. Check this number with a fitted HR value and arranged accordingly. Moreover, finally, it will generate an HR corresponding to each frame, whereas the average HR is taken as a final HR of the corresponding video. In a given architecture, each convolutional layer is followed by ReLU non-linearity.

4.3 Approach

CNN architecture contains three essential steps, which are pre-processing, training, and testing.

4.3.1 Pre-processing

We used the VIPL-HR dataset (Stable Scenario). There are 96 videos split into train_data and test_data. The train_data contains 1 to 70 videos, and test_data includes 71 to 96 videos. Similarly, we created two CSV files. The train_csv consists of average HR reading of training data. The test_csv consists of the average HR reading of testing data. We need two input image frames. One is the original RGB frame, and the second is a normalized frame.

1. RGB frame – To find the RGB frame, as shown in figure 3, we create a face mask with upper and lower HSV values to remove unwanted data from images (for appearance purpose). We converted RGB to HSV frames. We have done this because it is observed many times that HSV gives good output while separating the contrasting background region in the frame so that we may get better masking output.

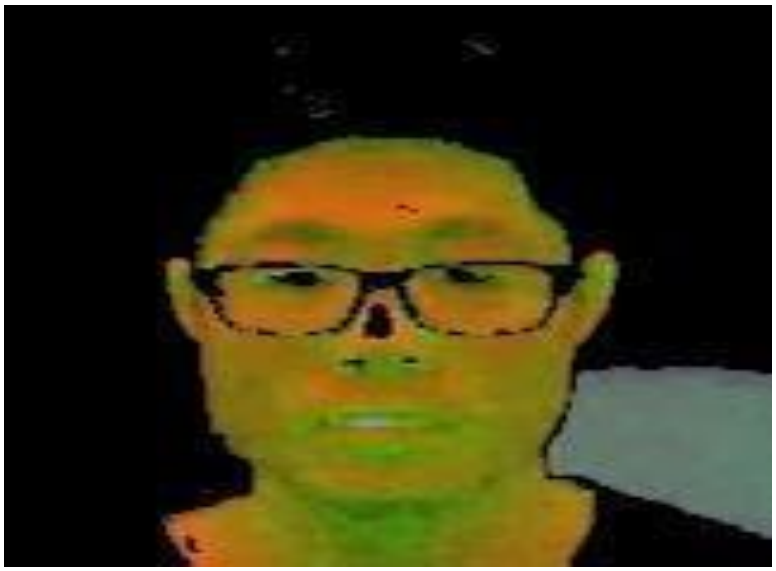


Figure 3: RGB frame

2. Normalized Frame – It shows changes between two consecutive frames. It is used to extract changes in the appearance of faces. It is used for motion representation. Figure 4 shows the normalized difference between frames at time t and t+1, which is given as input to the motion model.

Figure 4: Differenced normalized frame between t+1 and t

After pre-processing data, we convert the dataset into a suitable format to input into a neural network model for training and testing purposes. For the same, we need two input image frames and corresponding HR reading. So, we made a pair of three (RGB frame, normalized frame, and corresponding HR reading) and stored it in a CSV file using implemented code and similarly used this CSV file to create a neural network model, for training purposes. Then we reshaped all frames to size 72 X 72. 36 X 36 mentioned in the paper [5], but we found it very less. We divided the data-set into two parts x_train and x_test, to create a validation set.

4.3.2 Training

To create the neural network model, we used VGG style CNN (16 layers) as given in the paper [5] using keras with tensor flow backend. We used average pooling layers. We structured the model so that those two inputs generate a number, and in model fit, we give HR reading (y_train) as input so that it will correct output. We used very little data-set, so we used epoch = 20 and batch size = 15. The loss function of our model is the MSE between the estimated and actual HR. Further, we used the RMS prop optimizer because we do not need to adjust the learning rate. RMSprop does it automatically. For each parameter, it chooses a different learning rate. At last, we stored trained model architecture in the .json file and model weights value in the .h5 file. (Automatically it will be stored in the drive). After 20 epochs, the test score of our trained model is 7.8187.

4.3.3 Testing

Now, to test the trained model, we implemented code model_demo.py. We extracted the .json file and the .h5 file as a model and gave input frame to model_demo.py to predict HR reading corresponding to each frame. Figure 5 shows HR reading corresponding to each frame. Average HR is estimated as the final HR.

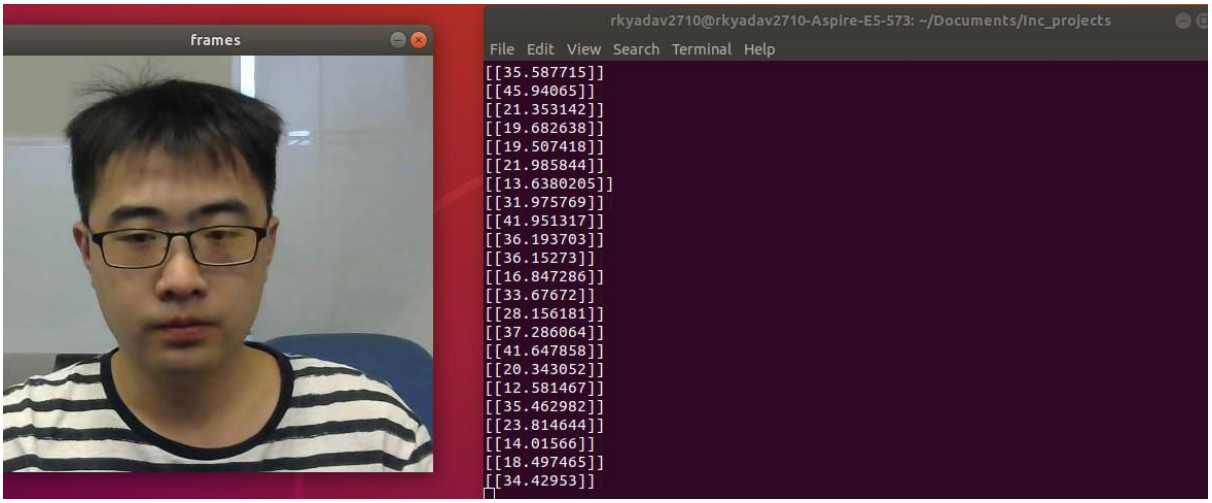


Figure 5: HR reading corresponding to each frame.

In figure 5, the image on the left side is one of the image frames from the video clip that is used as an input source for the model. The bottom-most reading on the right side gives the prediction of the model that takes into consideration the corresponding image frame on the left side, at any moment of the demonstration through a video clip.

References

1. W Verkruysse, Lars O. Svaasand, and J. Stuart Nelson. "Remote plethysmographic imaging using ambient light.", Optics express 16, no. 26, 2008, 21434-21445.
2. MZ Poh, Daniel J. McDuff, and Rosalind W. Picard. "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." Optics express 18.10, 2010, 10762-10774.
3. G De Haan, and Vincent Jeanne. "Robust pulse rate from chrominance-based rPPG.", IEEE Transactions on Biomedical Engineering 60.10, 2013, 2878-2886.
4. G. Balakrishnan, F. Durand, and J. Guttag. "Detecting pulse from head motions in video." In Computer Vision and Pattern Recognition, 2013, 3430–3437.
5. W Chen, and Daniel McDuff. "Deepphys: Video-based physiological measurement using convolutional attention networks." Proceedings of the European Conference on Computer Vision, 2018, 349-365.