Journal of Information Hiding and Privacy Protection
DOI:10.32604/jihpp.2021.026380
Article

Heart Rate Detection Based on Facial Video

Yudan Zhao[*] and Chaoyu Wang

Wuhan Polytechnic University, Wuhan, 430023, China
[*]Corresponding Author: Yudan Zhao. Email: 13994647589@163.com

Received: 23 December 2021; Accepted: 03 January 2022

Abstract: Heart rate is an important data reflecting human vital characteristics and an important reference index to describe human physical and mental state. Currently, widely used heart rate measurement devices require direct contact with a person's skin, which is not suitable for people with burns, delicate skin, newborns and the elderly. Therefore, the research of non-contact heart rate measurement method is of great significance. Based on the basic principle of Photoplethysmography, we use the camera of computer equipment to capture the face image, detect the face region accurately, and detect multiple faces in the image based on multi-target tracking algorithm. Then the region segmentation of the face image is carried out to further realize the signal acquisition of the region of interest. Finally, peak detection, Fourier analysis and wavelet analysis were used to detect the frequency of PPG and ECG signals. The experimental results show that the heart rate information can be quickly and accurately detected even in the case of monitoring multiple face targets.

Keywords: Face recognition; heart rate detection; PPG signal

1 Introduction

Contact heart rate tools, which require direct contact with the skin of the person being measured, are expensive, inefficient and unsuitable for certain situations, such as monitoring the heart rate of skin burn patients and infants in neonatal intensive care [1], and may increase the risk of COVID-19 transmission. Therefore, the research of noncontact heart rate measurement method is of great significance.

Heart rate detection generally has two schemes, one by obtaining the body surface bioelectric measurement, biological potential changes are collected to form ECG signal map; second, photoplethysmography [2] technology uses a conventional camera to collect human face images and measure the tiny periodic color changes of facial skin caused by heart rate activities to obtain the basic physiological parameters of human body [3], it is more convenient and comfortable than the traditional contact heart rate measurement technology.

When the light irradiates the skin, it will be reflected and transmitted, and the hemoglobin concentration in the blood changes with the pulse. At this time, the changes of blood component concentration corresponding to the changes of light absorption amount are collected to form photoplethysmography (PPG) signal. To put it simply, the essence of measuring heart rate with optics and optical sensors is the conversion between photoelectric signals [4]. Fig. 1 illustrates the mechanism of this signal conversion.

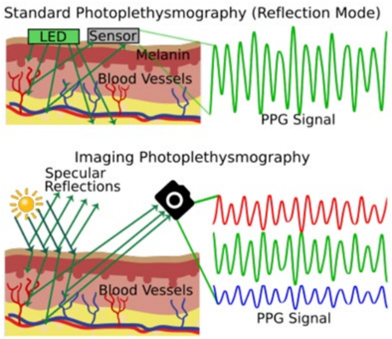


Figure 1: Photoelectric volume pulse wave measurement of heart rate principle

In standard PPG, the light source and photodetector are in direct contact with the skin, and most of the light that reaches the blood vessels is returned to the detector. In imaging PPG, the camera is a spatial sensor that records signals from a distance and uses ambient light.

In this paper, we detect heart rate through facial video. Next, this paper will introduce the test stages of the following three experiments: target detection task, image segmentation task and the acquisition and analysis of heart rate signal. The aim of the experiment is to detect the heart rate information of the subjects quickly and accurately under the condition of monitoring multiple face targets simultaneously.

2 Experiments

2.1 TensorFlow Provides Retinaface Face Detection Platform

Target detection is an important problem in image processing. In addition to object classification, the detection task also needs to find out the position and category of objects in the input image. Face detection also belongs to target detection, here we only need to detect the target position.

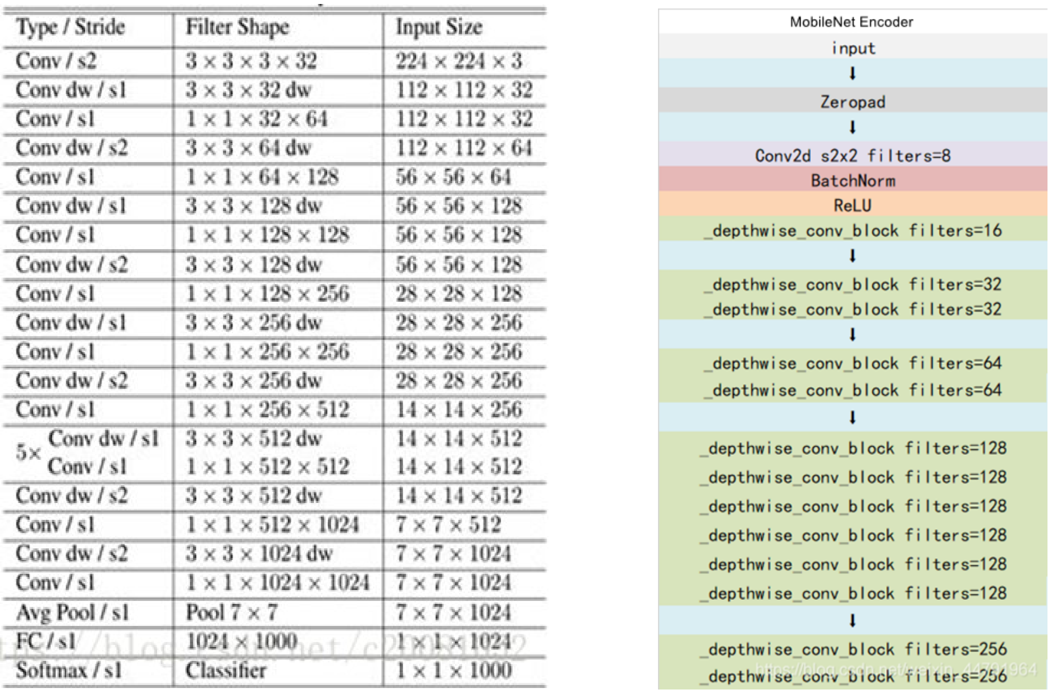
In this paper, we build a Retinaface face detection platform based on TensorFlow. Retinaface [5] is a robust single stage face detector that takes advantage of combined extra supervision and self-supervised multitasking learning to locate faces at the pixel level at different scales. Specifically, the Retinaface face detector contributes in the following five areas:

- Five human FACE landmarks were manually marked in the WILDER FACE dataset, and with the help of this additional monitoring signal, significant improvement was observed in hard FACE detection.
- The branch of self-supervised mesh decoder is further added to predict pixel-level 3D shape face information in parallel with the existing supervised branch.
- On the WILDER FACE Hard test set, the average RetinaFace accuracy (AP) was 1.1% higher than the most advanced average accuracy (AP = 91.4%).
- In iJB-C test set, RetinaFace enables the existing method (ArcFace) to improve face validation results FAR = 1E-6, TAR = 89.59%).
- Lightweight backbone network with RetinaFace can run VGA resolution images in real time on a single CPU core.

In the actual training here, Retinaface uses two types of networks as the backbone feature extraction network during actual training. They are MobilenetV1-0.25 and Resnet. High accuracy can be achieved with Resnet, and real-time detection on the CPU can be achieved with MobilenetVl-0. 25.

MobileNet [6] model is a lightweight deep neural network. The core idea of MobileNet model is deeply separable convolution. Deep separable convolution divides the complex general convolution process into two steps, deep convolution and point convolution, which greatly reduces the number of parameters of the model, reduces the difficulty of training, and thus improves the calculation speed. Figs.

2(a) and 2(b) below introduces the structure of MobileNet.



Type / Stride	Filter Shape	Input Size
Conv / s2	3 × 3 × 3 × 32	224 × 224 × 3
Conv dw / s1	3 × 3 × 32 dw	112 × 112 × 32
Conv / s1	1 × 1 × 32 × 64	112 × 112 × 32
Conv dw / s2	3 × 3 × 64 dw	112 × 112 × 64
Conv / s1	1 × 1 × 64 × 128	56 × 56 × 64
Conv dw / s1	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 128	56 × 56 × 128
Conv dw / s2	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 256	28 × 28 × 128
Conv dw / s1	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 256	28 × 28 × 256
Conv dw / s2	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 512	14 × 14 × 256
5× Conv dw / s1 Conv / s1	3 × 3 × 512 dw 1 × 1 × 512 × 512	14 × 14 × 512 14 × 14 × 512
Conv dw / s2	3 × 3 × 512 dw	14 × 14 × 512
Conv / s1	1 × 1 × 512 × 1024	7 × 7 × 512
Conv dw / s2	3 × 3 × 1024 dw	7 × 7 × 1024
Conv / s1	1 × 1 × 1024 × 1024	7 × 7 × 1024
Avg Pool / s1	Pool 7 × 7	7 × 7 × 1024
FC / s1	1024 × 1000	1 × 1 × 1024
Softmax / s1	Classifier	1 × 1 × 1000

(a)

MobileNet Encoder
input
↓
Zeropad
↓
Conv2d s2x2 filters=8
BatchNorm
ReLU
_depthwise_conv_block filters=16
↓
_depthwise_conv_block filters=32
_depthwise_conv_block filters=32
↓
_depthwise_conv_block filters=64
_depthwise_conv_block filters=64
↓
_depthwise_conv_block filters=128
_depthwise_conv_block filters=128
_depthwise_conv_block filters=128
_depthwise_conv_block filters=128
_depthwise_conv_block filters=128
_depthwise_conv_block filters=128
↓
_depthwise_conv_block filters=256
_depthwise_conv_block filters=256

(b)

Figure 2: (a): MobilenetV1-1 structure; (b): Mobilenetv1-0.25 structure

In Fig. 2(a), Conv DW is hierarchical convolution, followed by a 1 × 1 convolution for channel processing. MobilenetVl-0. 25 built in our code is a network with the number of mobilenetV1.1 channels compressed to l/4 of the original, as shown in Fig. 2(b).

Retinaface uses the FPN [7] structure to build up the valid feature layers of the last three Mobilenet shapes. The construction method is very simple. Firstly, 1 × 1 convolution is used to adjust the number of channels for the three effective feature layers. After adjustment, Upsample and Add were used for feature fusion of upsampling.

Through this part of the operation, we get P3, P4, P5 three effective feature layers. Retinaface to further enhance feature extraction, the SSH [8] module is used to enhance the receptive field. The idea of SSH is very simple. Three parallel structures are used to replace the effect of 5 × 5 and 7 × 7 convolution by stacking 3 × 3 convolution: the one on the left is 3 × 3 convolution, two 3 × 3 convolution is used to replace 5 × 5 convolution, and three 3 × 3 convolution is used to replace 7 × 7 convolution on the right.

As shown in Fig. 3, through feature enhancement extraction, three effective feature layers, SSH1, SSH2 and SHH3, have been obtained. After obtaining these three effective feature layers, we need to obtain the prediction results through these three effective feature layers. Then, the prediction results of Retinaface are divided into three categories, the regression prediction results of box and the regression prediction results of face key points.

- Classification prediction results are used to judge whether the prior box contains objects.
- The regression prediction result of the box is used to adjust the prior box to obtain the predietion box.
- Regression prediction results of face key points are used to adjust the prior box to obtain face key points, each face key point needs two adjustment parameters, a total of five face key points.

After the adjustment and judgment, non-maximum suppression is also needed. The function of non-maximum suppression is to screen out the box with the largest score of the same category in a certain area.

Then the prediction results are decoded, the prediction results of Retinaface are used to judge whether the prior frame contains human faces, and the prediction frame and human face key points are obtained by adjusting the prior frame containing human faces.

We calculate loss bv using the prediction results of processed real frames and corresponding pictures.

- Box Smooth Loss: regression Loss to obtain the predicted results of all positive labeled boxes.
- MultiBox Loss: cross entropy Loss for obtaining all kinds of prediction results.
- Lamdmark Smooth Loss: regression Loss of prediction results of all positive labeled key points.

When calculating losses, it should be noted that Box Smooth Loss calculates all the prior frames that are identified as containing faces internally, while Lamdmark Smooth Loss calculates all the prior frames that are identified as containing faces internally and containing key points of faces (In the annotation of some face frame because of the Angle and clarity of the problem is no key face).

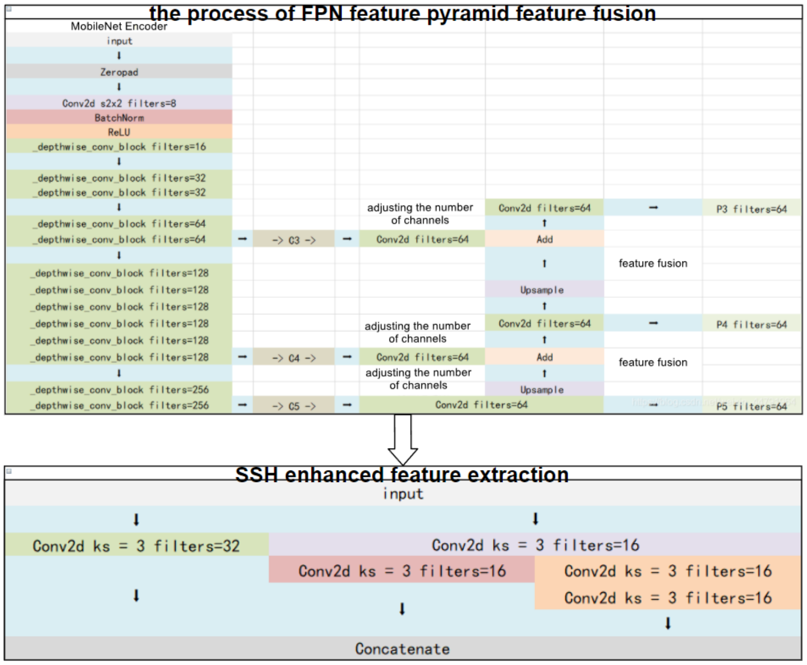

the process of FPN feature pyramid feature fusion
MobileNet Encoder
input
Zeropad
Conv2d s2x2 filters=8
BatchNorm
ReLU
_depthwise_conv_block filters=16
_depthwise_conv_block filters=32
_depthwise_conv_block filters=32
_depthwise_conv_block filters=64
_depthwise_conv_block filters=64
_depthwise_conv_block filters=128
_depthwise_conv_block filters=128
_depthwise_conv_block filters=128
_depthwise_conv_block filters=128
_depthwise_conv_block filters=128
_depthwise_conv_block filters=128
_depthwise_conv_block filters=256
_depthwise_conv_block filters=256
-> C3 ->
-> C4 ->
-> C5 ->
adjusting the number of channels
Conv2d filters=64
Add
Upsample
feature fusion
P3 filters=64
P4 filters=64
P5 filters=64
SSH enhanced feature extraction
input
Conv2d ks = 3 filters=32
Conv2d ks = 3 filters=16
Conv2d ks = 3 filters=16
Conv2d ks = 3 filters=16
Conv2d ks = 3 filters=16
Concatenate

Figure 3: FPN feature pyramid feature fusion and SSH enhance feature extraction process

Through the above steps, we can obtain the position of prediction boxes on the original picture, and these prediction boxes are filtered. These filtered boxes can be drawn directly on the image to obtain the results, see Fig. 4 below.

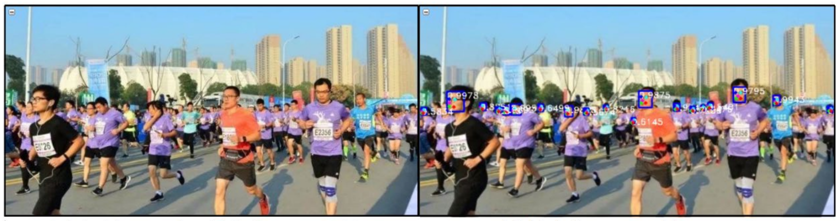

Original image of multiple face recognition
Multiple face recognition targets

Figure 4: Target recognition task results

2.2 PyTorch Builds an Unet Semantic Segmentation Face Parsing Platform

Next, we will perform region segmentation on the recognized faces, also known as face analysis, in order to find ROI and extract PPG signals of more representative ROI.

lmage segmentation needs to classify different pixels in the image. Compared with object detection, image segmentation is more detailed and difficult. In semantic segmentation, each pixel has its own label attribute.

Unet [9] is an excellent semantic segmentation model, and its main execution process is similar to other semantic segmentation models. Unet can be divided into three parts: main feature extraction, enhanced feature extraction and prediction. Fig. 5 shows the Structure of Unet network.

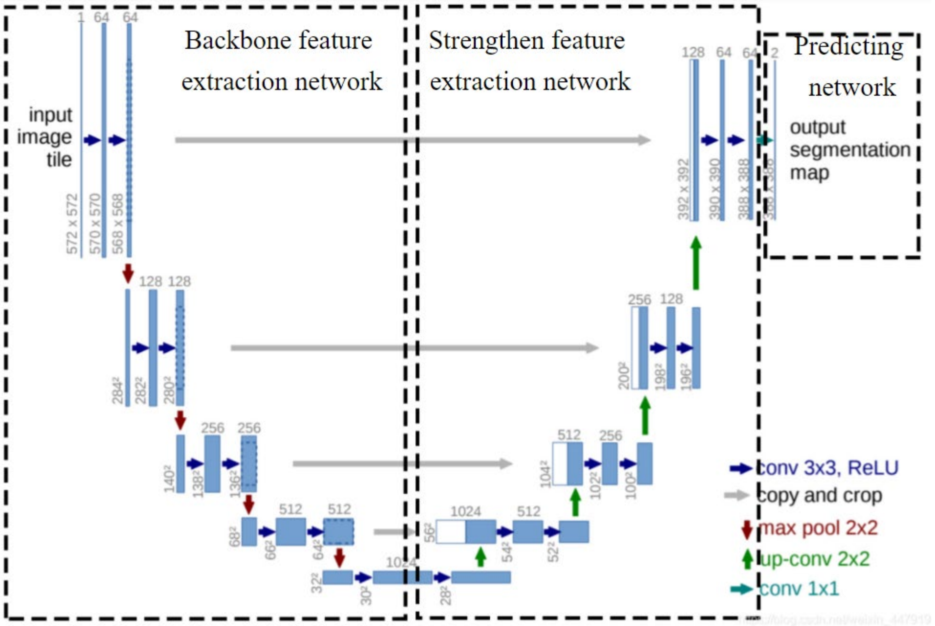


Figure 5: Unet network structure

The first part is the trunk feature extract ion part, which uses the trunk part to obtain multiple different feature layers. The trunk feature extraction part of Unet is similar to VGG, which is a stack of convolution and maximum pooling. Five initial effective feature layers can be obtained by using the main feature extraction part, and the subsequent use of these five effective feature layers can carry out feature fusion. The second part is to strengthen the feature extract ion part, using the five initial effective feature layers obtained from the main part to carry out up-sampling, feature fusion, and obtain a final effective feature layer that integrates all the features of different levels. The third part is the prediction part. The prediction part will classify each feature point by using the finally obtained effective feature layer that integrates different levels, which is equivalent to classifying every pixel point.

The trunk feature extraction part of Unet is composed of convolution layer and maximum pooling layer, and the overall structure is similar to VGG [10]. The trunk feature extraction network adopted in this paper is VGG16, and only two types of layers are used, namely, convolution layer and maximum pooling layer. The enhanced feature extraction network used by Unet is in the shape of “U”. Five initial effective feature layers can be obtained by using the backbone feature extraction network. In the enhanced feature extraction network, these five initial effective feature layers are used for feature fusion. The method of feature fusion is to up-sample and stack the feature layers.

The backbone feature extraction network completes the construction of five preliminary effective feature layers, strengthens the fusion and strengthening of the feature extraction network to five preliminary effective feature layers, and obtains an effective feature layer integrating multiple feature layers. We will use this final effective feature layer for prediction.

Helen’s face segmentation dataset used in this paper includes 2330 face images, and each face image contains masks of 11 parts. Generally, face segmentation only needs masks of face and facial features. The mask contains nine parts of the face region division plus hair and background image. Using the regional information provided by the mask file, the segmentation region of the face can be filled with different colors, so as to generate the corresponding segmentation map of the face region of each original image, as shown in Fig. 6.

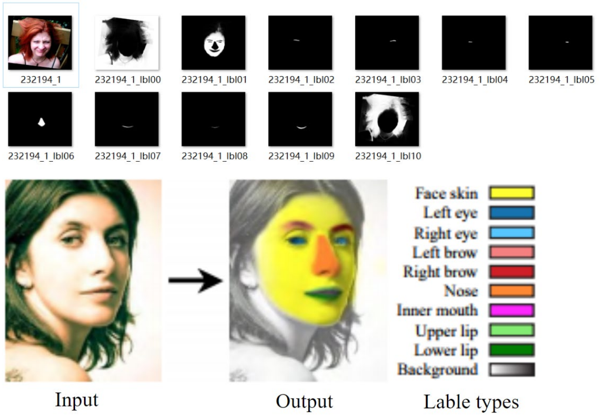


Figure 6: Above is an example of the original picture and mask below is a sample of the generated face segmentation

Since the cheeks and forehead on both sides are less affected by rigid movement, while the eyes, mouth, nose and other parts are more affected by rigid activities such as speaking, breathing and blinking, the interesting parts selected in this paper are cheeks and forehead on both sides. After the heart rate sensor obtains the independent signal intensity of the cheek and forehead, it averages the obtained signal to obtain the final effective heart rate PPG signal, and then converts it with the common ECG signal. The process enters into the next part of this paper, the analysis of heart rate signal.

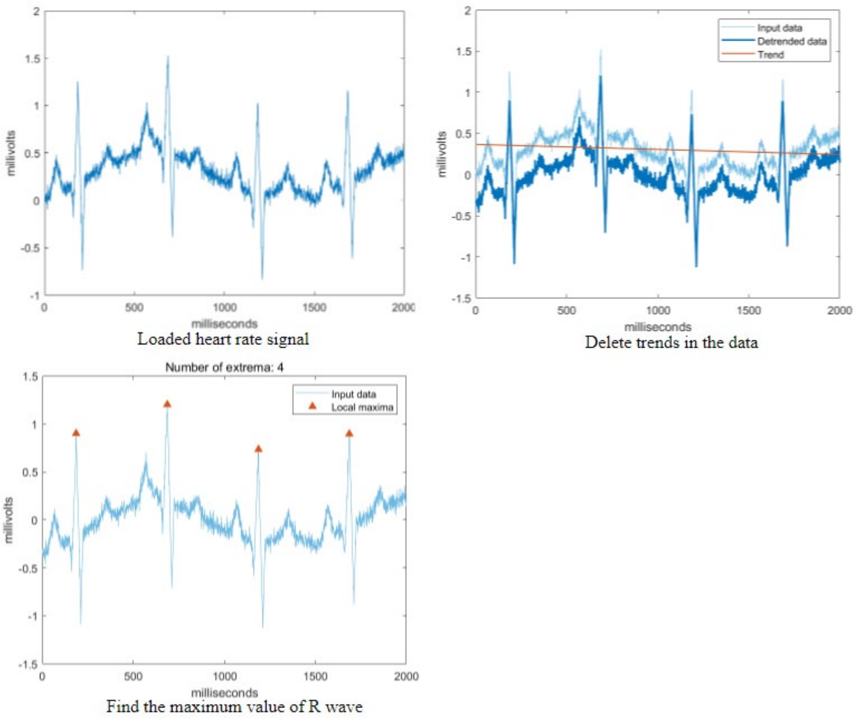
2.3 Acquisition and Analysis of Heart Rate Signal

The conversion between ECG signals and PPG signals is related in nature, and the left ventricular cardiac activity affects blood volume changes, which in turn are controlled by electrical signals from the sinoatrial node (SA). The PPG waveform sequence, amplitude and shape, contains heart and related information that is used to measure heart rate, heart rate variability, respiration rate, oxygen saturation, blood pressure, and to assess vascular function [11].

Exists in the acquisition of heart rate signal baseline drift, power frequency interference, etc., so to be from heart rate signal to derive the heart rate value, must get the heart rate signal processing, get a satisfactory heart rate signal, not only is helpful to calculate the heart rate value, also benefit from the changes of heart rate signal, judge the heart rate gatherers of physical health.

So let us take a look at how we can analyze heart rate signals to get heart rate. Due to the particularity of R wave in the heart rate signal graph, the peak detection method is used to find the peak value of R signal segment. Peak detection does not provide a detailed and complete waveform display, but captures signal key points with the highest sampling rate and only records the maximum peak value within each sampling interval. The following is the process of heart rate estimation [12].

As shown in Fig. 7, first, we load heart rate data and plot a heart rate signal, which is measured by electrodes attached to the skin and is sensitive to disturbances such as power sources and noise caused by exercise. The signal above shows a baseline offset and therefore does not represent true amplitude [13]. We can use “Delete Trend” to adjust the baseline movement to polynomial and remove it. Then find the maximum value of the R wave, which in the ECG is the large upward deflection signal indicating the expansion of the ventricular main block. R wave can be detected at the threshold peak above 0.5 mV.



Loaded heart rate signal

Delete trends in the data

Find the maximum value of R wave

Figure 7: The process of drawing heart rate signal, deleting trend of heart rate signal and searching peak of R wave of heart rate signal

Knowing the time (measured in milliseconds) between peak ECG signal s allows you to calculate the heart rate. The difference method is adopted in this paper. To be specific, the time difference between two peak values is calculated, namely the period of this period. Similarly, every two R wave peaks of this period of heart rate signal have corresponding time difference, so the frequency of the corresponding period, namely the heart rate, can be calculated. If the average heart rate during this period is required, the average difference frequency can be calculated, and the heart rate signal = 60 * (1000/peak time difference).

The purpose of preprocessing ECG and PPG signals is to obtain time-aligned and normalized signal pairs so that the critical time characteristics of both waveforms are synchronized. The steps of pre-processing are as follows, including data alignment, signal de-trending, period segmentation, time scaling, normalization, and a series of parameters related to heart health can be calculated based on the intermediate information of the processing process. Peaks and troughs can be detected on the heart rate signal, such as Reflection Index = B/A * 100. Where B is the difference between the wave peak and the wave trough Y-axis, and A is the peak value of PPG signal on the Y-axis. For example, stiffness index = H/PTT, where H is the distance from the finger tip to the heart, and PTT is the pulse transmission time obtained above. As for the images below, visualization allows us to drill down into the level of detail and improve efficiency.

Fig. 8 and Fig. 9 visually show the PPG signal alignment and peak seeking process, as well as the test results.

With regard to filtering, any sequential signal measured continuously can be represented as an infinite superposition of sinusoidal signals of different frequencies. After the decomposition of the time-domain signal by Fourier transform, it becomes the superposition of different sinusoidal signals, and then we analyze the frequency of these sinusoidal, we can transform a signal into the frequency domain. By transforming from time domain to frequency domain, we can also analyze the amplitude, power, energy and phase relations of various frequency components contained in the signal, which is to analyze the spectral characteristics of the signal.

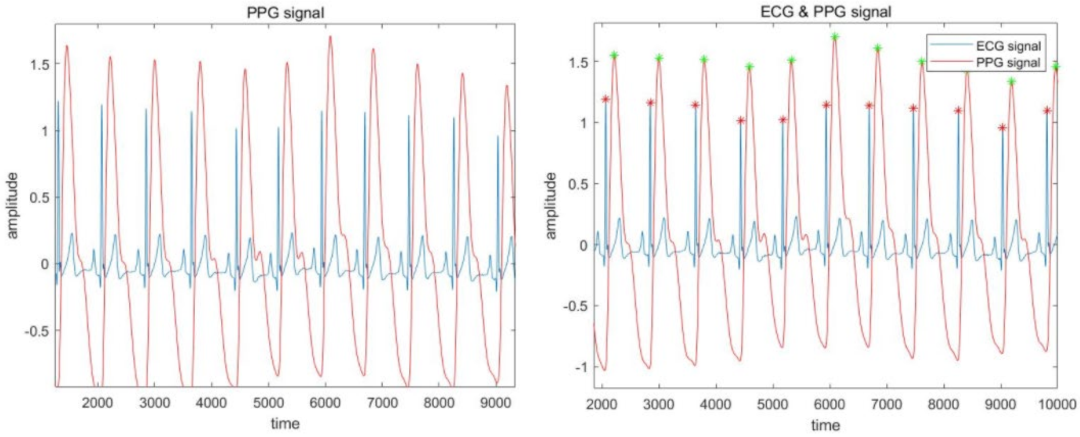


Figure 8: Parallel processing of PPG signal and ECG signal. On the left is the PPG signal aligned with the ECG signal, and on the right is the PPG signal peaking with the ECG signal

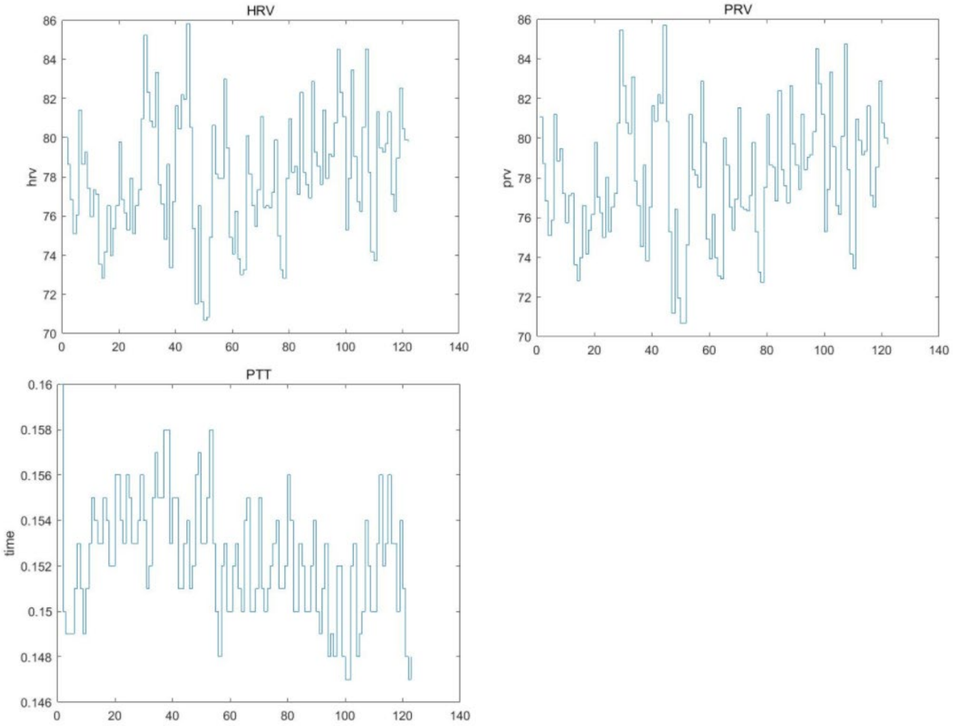


Figure 9: The figure above shows HRV index, PRV index and PTT index, respectively

A band-pass filter is used here. The band-pass frequency range is 80–120 Hz, and the corresponding heart rate range also belongs to this range. The passband is shown in Fig. 10.

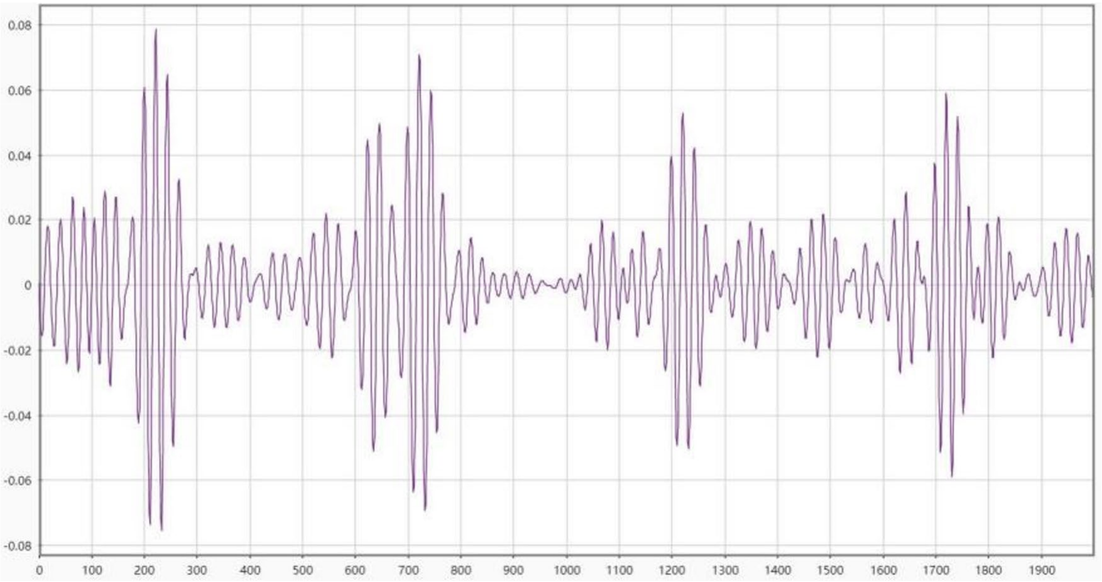


Figure 10: Bandpass filter processing signal

Time series signals such as heart rate can also be processed by using wavelet analysis, which will not be described in this paper.

3 Conclusions

The results showed that heart rate information could be detected even when multiple faces were being monitored. However, according to the author's own experiments and replication, the actual measurement effect is still far from ideal, and there is still a lot of research space.

RPPG heart rate detection based on video analysis has been attracting the attention of researchers since it was proposed. After 10 years of research, the development of research objects has changed from static to motion, and the environment has gradually changed from a single light source to a light source. The application of RPPG scenes has gradually approached the real scenes, which has a wide application prospect. This is a promising research direction.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] E. M. Nowara, D. McDuff and A. Veeraraghavan, "Combining magnification and measurement for non-contact cardiac monitoring," in *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 3805–3814, 2021.

[2] D. J. McDuff, J. R. Estepp, A. M. Piasecki and E. B Blackford, "A survey of remote optical photoplethysmographic imaging methods," in *2015 37th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6398–6404, 2015.

[3] X. Y. Li, P. Wu, Y. Liu, H. Y. Si and Z. L. Wang, "Heart rate parameter extraction based on face video," *Optics and Precision Engineering*, vol. 3, pp. 548–557, 2020.

[4] N. Martinez, M. Bertran, G. Sapiro and H. T. Wu, "Non-contact photoplethysmogram and instantaneous heart rate estimation from infrared face video," in *2019 IEEE Int. Conf. on Image Processing (ICIP)*, IEEE, 2019.

[5] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," in *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[6] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. J. Wang *et al.,* "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv Preprint arXiv:1704.04861*, 2017.

[7] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan *et al.,* "Feature pyramid networks for object detection," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017.

[8] M. Najibi, P. Samangouei, R. Chellappa and L. S. Davis, "SSH: Single stage headless face detector," in *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 4885–4894, 2017.

[9] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," Springer International Publishing, 2015.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[11] X. R. Wang, W. J. Wang, J. S. Zhao, L. Q. Kong and Y. J. Zhao, "Realtime video-based non-contact multiplayer heart rate detection during exercise," *Optical Technique*, vol. 45, pp. 6, 2019.

[12] L. Lu and J. Cheng, "Non-contact thermal infrared video heart rate detection method based on multi-region analysis," *Journal of Biomedical Engineering Research*, vol. 40, pp. 21–27, 2021.

[13] J. Y. Zhang and Z. Wu, "Multimodal face recognition based on color and depth information," *Engineering Journal of Wuhan University*, vol. 4, pp. 353–363, 2020.