Remote Heart Rate Estimation by Signal Quality Attention Network

Haoyuan Gao, Xiaopei Wu*, Jidong Geng, Yang Lv
School of Computer Science and Technology, Anhui University
Anhui University, Hefei, China
gaohaoyuan98@outlook.com, wxp2001@ahu.edu.cn,
769904039@qq.com, 1944540119@qq.com

Abstract

Heart rate estimation is very important for heart health monitoring. As a non-invasive optical technology, remote photoplethysmography (rPPG) has the advantages of non-contact, portability and low-price. However, motion and noise artifacts bring additional uncertainty to the results of heart rate estimation. Based on the signal quality assessment method, we propose a new remote heart estimation algorithm by signal quality attention mechanism and long short-term memory (LSTM) networks. The model consists of three parts: firstly, an LSTM network is used to estimate the heart rate sampling point by sampling point; secondly, a similar LSTM network predicts the signal quality; finally, an attention-based model uses the heart rates and quality scores predicted above to calculate the average heart rate of a period of time. The model allocates higher weights to the reliable heart rates estimated from high-quality signals, meanwhile, ignores unreliable results estimated from low-quality signals. Experiments show that LSTM with attention mechanism accurately estimates heart rate from corruption rPPG signal and it performs well on cross-subject tasks and cross-dataset tasks. The results also demonstrate that the scores predicted by the signal quality model is valuable to extract reliable heart rate.

1. Introduction

Heart rate (HR) is a standard vital sign measured by the number of beats per minute (bpm). Since heart rate is a critical sign for monitoring cardiovascular and chronic disease treatment, the estimation and monitoring of heart rate have become routine in the medical field. HR can be easily extracted from the electrocardiogram (ECG) or photoplethysmography (PPG) signals. However, ECG and traditional PPG methods use contact sensors such as electrodes or light receivers, which are not friendly to long-term measurements. To solve this problem, remote photoplethysmography (rPPG) [2, 32], which targets to sense the blood volume pulse (BVP) controlled by the heart beats without any contact, has been developed rapidly in recent years. The rPPG also has the superiority of low-cost, convenient, widespread and accessing multiple physiological parameters simultaneously. Extensive experiments demonstrate that rPPG can accurately estimate heart rate in a noise-controlled laboratory environment. [5, 12, 28].

The conventional rPPG technology for HR estimation follows a similar framework. Firstly, a webcam captures the video including the surface skin of a subject. Facial videos are common data for research and experiment because face detection algorithms (Viola and Jones algorithm [33, 41] or facial landmark localization algorithms [1, 24]) can easily mark face areas as regions of interest (ROIs). Foreheads [25] and cheeks [9] are also marked as ROIs that contain strong BVP signals. Then, the average values of the pixels in the ROIs are calculated and temporally concatenated to compose raw signals in red channel, green channel, and blue channel respectively. These RGB signals are further processed to remove the effect of motion and illuminate noise. HR is estimated from the noise-free and high-quality BVP signal by frequency analysis or peak detection [27, 36].

The conventional rPPG technologies focus on eliminating motion and illuminate noise to extract BVP from the RGB signals. Digital filter keeps information related to heart rate only, however, noise in similar frequency bands of HR would affect the estimation results. To solve this problem, denoising methods based on the blind source separation method (ICA [16, 23, 37] or PCA [17, 38]) are proposed to extract BVP signal. By the dichromatic reflection model, the information of color vectors can be utilized to control the demixing for component derivation. The model-based methods (CHROM [7] and POS [35]) performed well during both stationary and motion situations.

Recently, deep learning based methods have shown promise in mapping the complex physiological processes to measure remote HR, extract precise BVP signals from

*corresponding author

facial video or reconstruct high-quality signals from low-quality signals [6]. Špetlík et al. [30] proposed a novel two-step convolutional neural network to extract high-quality signal and predicted HR respectively. Chen et al. [4] presented a convolutional network based on a skin reflection model and an attention mechanism for video-based measurement of HR and breathing rate using. Yu et al. [39] applied a deep spatio-temporal networks (3D CNN based and RNN based) for reconstructing precise rPPG signals from raw facial videos. Yu et al. [40] proposed a two-stage, end-to-end method with a spatio-temporal video enhancement network (STVEN) for video enhancement and an rPPG network (rPPGNet) for rPPG signal recovery. Bousefsaf [3] designed a 3D CNN to extract features from unprocessed video streams, followed by a multilayer perceptron to regress HR. Tsou [31] proposed Siamese-rPPG, a framework based on a Siamese 3D CNN. Hu et al. [14] presented an effective time-domain attention network for remote heart rate measurement. Hu et al. [15] further designed a spatial-temporal attention network to avoid extracting redundant information from video segments and enhance long-range video temporal. Song et al. [29] designed a new framework based on generative adversarial network (PulseGAN) to generate realistic rPPG pulse signals through the signals denoised after chrominance method.

Existing deep learning methods use the convolution kernel, RNN structure, attention mechanism, or combined models to reconstruct noise-free BVP signals or measure HR. To increase the interpretability and generalization of the deep learning model, inspired by [19, 34], we proposed a new signal-quality attention LSTM based model for remote HR estimation (SQA-rPPG). Different from the existing attention-based networks, a supervised signal-quality assessment network (SQN) [10] was used in SQA-rPPG to measure HR. The SQA-rPPG pays more attention to the segment with less noise, and ignores the segment which is corrupted by noise and difficult to extract the heart rate.

The main contributions of this work include: 1) A real-time heart rate estimation algorithm based on LSTM (LSTM-rPPG) is designed to predict heart rate at the current moment when a new sampling point is input. 2) With the output of LSTM-rPPG as the key value, an attention mechanism weighted by signal quality assessment model is proposed to measure average heart rate. 3) Both the quality score assessment model and the heart rate estimation model are based on the LSTM structure, which allows the model to estimate the average heart rates of the rPPG signals in arbitrary length without retraining the model.

The rest of the article is structured as follows: In section 2, we first introduce the datasets used in our paper, and then introduce the detail of LSTM-rPPG model and the detail of SQA-rPPG model. In section 3, we present the performance of SQA-rPPG on cross-subject tasks, cross-webcam tasks and cross-dataset tasks. We also compare SQA-rPPG and LSTM-rPPG with other denoising algorithms. In section 4, we present the conclusions.

2. Materials and Methods

2.1. Datasets

Three facial video datasets with annotated PPG signals are used in this study. The first database IIPHCI is created by the lab of intelligent information and human computer interaction [11]. It consists the facial videos collected by microsoft lifecam studio (M-cam) and Aoni A36 webcam (A-cam) simultaneously. The M-cam captured 30 frames per second (fps) and A-cam runs 25 fps. At the same time of the video recording, a pulse oximeter (CMS50E) with a sampling rate of 60 Hz is worn on the subject's finger to record the PPG signals. Overall, the dataset contains 312 minutes of facial videos of 16 males and 10 females performing resting tasks, talking tasks and facial rotation tasks, respectively.

The second dataset UBFC-Phys [26] is a public dataset for psychophysiological studies. It contains 56 participants following resting task T1, talking task T2 and arithmetic tasks T3. During the experience, the participants were filmed and were wearing a wristband that records their PPG signals with a sampling rate of 64Hz. The frame rate of the videos is about 35 fps.

The third dataset LGI-PPGI [21] records the facial videos of 20 males and 5 females in the range of 25-42 years. It consists of resting session, facial motions session, gym session, urban conversation session. The sampling rate of PPG signal is 60Hz and the frame rate of the video is 25 fps.

2.2. Data Preprocessing

2.2.1 Heart rate label generation

The HR label of facial video is generated from the PPG signal collected at the same time of the video recording. Each sampling point needs a label as the hidden feature, so we set a sliding window with the length of eight seconds and the slide step of one sampling point to measure the HR. Frequency domain analysis using Fast Fourier Transform (FFT) preprocess the segments in the sliding window. The maximum value in the frequency domain is chosen and multiplied by 60 to get the number of beats per minute. In order to achieve accuracy of 0.1 bpm in HR estimation, in the FFT calculation, we added additional zeros to the segments. Figure 1 shows the distribution of heart rates we calculated on different databases. The HR between 70 bpm and 80 bpm has the largest proportion (39.36%) on IIPHCI, and the HR between 80 bpm and 90 bpm has the largest proportion (48.53%) on UBFC. On LGI, heart rate between 50 bpm and 60 bpm (23.28%), 60 bpm and 70 bpm (25.37%),

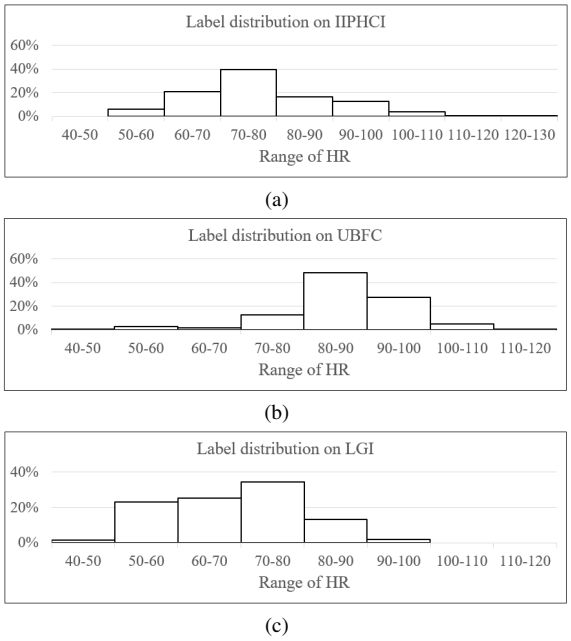


Figure 1. The distribution of heart rates on different datasets.

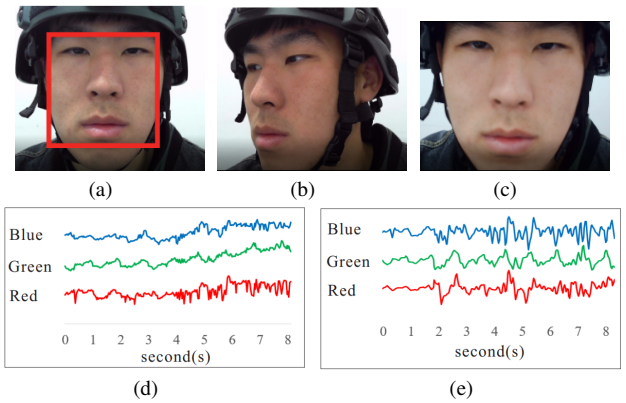


Figure 2. Examples of data used in our experiment. (a) A frame in the video recording by M-cam on IIPHCI. In the red box is the artificially selected ROI. (b) A frame in the video recording by M-cam on IIPHCI when the subject performs the rotation task. (c) A frame in the video recording by A-cam on IIPHCI. (d) An example of raw RGB signals. (e) An example of preprocessed RGB signals, also the input of our models.

70 bpm and 80 bpm (34.46%) have larger proportions. The different distributions of labels on different datasets can verify the generalization of the model on cross-dataset tasks.

2.2.2 RGB signals extraction

The input of LSTM-rPPG and SQA-rPPG is the RGB signals (r_t, g_t, b_t) in three channels. We manually selected the facial region in the first frame as the ROI of the entire video, and the average values of pixels in the ROIs of each frame construct the raw RGB signals. In order to increase the generalization of the model, the RGB signals have been preprocessing by band-pass filter and normalizing. We selected third-order Butterworth digital filter in Scipy Library of Python. We used a two-second sliding window to calculate the standard deviation and normalize the data to verify LSTM-rPPG and SQA-rPPG have the potential for real-time system. Fig 3 shows the examples of frames on IIPHCI dataset, the ROI we artificially selected, the raw RGB signals and the preprocessed RGB signals.

2.3. LSTM-rPPG

Recurrent neural network (RNN) is a typical artificial neural network which is commonly used for ordinal or temporal problems, such as language translation, natural language processing and image captioning. Long short-term memory (LSTM) is a popular RNN architecture, which was introduced by Sepp Hochreiter and Juergen Schmidhuber as a solution to the problems of vanishing gradient and long-term dependencies [13]. Formulas (1-6) describe the internal details of LSTM.

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{\boldsymbol{f}}\boldsymbol{x}_t + \boldsymbol{R}_{\boldsymbol{f}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\boldsymbol{f}}) \tag{1}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{\boldsymbol{i}}\boldsymbol{x}_t + \boldsymbol{R}_{\boldsymbol{i}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\boldsymbol{i}}) \tag{2}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{\boldsymbol{o}}\boldsymbol{x}_t + \boldsymbol{R}_{\boldsymbol{o}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\boldsymbol{o}}) \tag{3}$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_{\boldsymbol{c}}\boldsymbol{x}_t + \boldsymbol{R}_{\boldsymbol{c}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\boldsymbol{c}}) \tag{4}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \tag{6}$$

Here, $\boldsymbol{c}_t$ is the cell in the hidden layers of the neural network, $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, $\boldsymbol{o}_t$, $\boldsymbol{h}_t$ are the input gate, forget gate, output gate and hidden state of the LSTM, respectively. These gates control the flow of information which is needed to predict the output in the network. σ and $\odot$ are the logistic sigmoid activation and element-wise multiplication respectively.

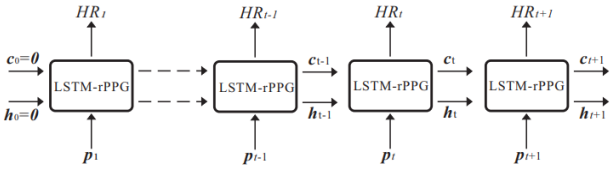


Figure 3. Details of LSTM-rPPG predicting heart rate.

Type	Activation Fun	Neurons	Params	Ops
FC	selu	4	12	24
LSTM	sigmoid/tanh	16	1.3k	2.6k
LSTM	sigmoid/tanh	16	2k	4k
FC	selu	4	64	128
FC	-	1	4	8

Table 1. Architecture of LSTM-rPPG. The FC is a fully connected layer. The Activation Fun is the activation function used by this layer. The Neurons is the number of neurons in this layer. The Params is the number of parameters in this layer and the Ops is the number of float operations requied for this layer (ignore bias parameters).

We designed an LSTM based model to predict heart rate directly from RGB signals (LSTM-rPPG). Table 1 and Fig 3 show the configuration of LTSM-rPPG. It is designed as a many to many RNN architecture and every input sampling point has an output as HR at the current moment. The model contains three fully connected layers. The vector of new sampling points $\boldsymbol{p}_t = (r_t, g_t, b_t)$ in RGB channels at time t is the input of LTSM-rPPG. New feature representation is calculated by encoding through a fully connected layer with selu activation function. The hidden state features are put into the two-layer-LSTM module and the module predict new output $\boldsymbol{h}_t$ through the vectors $\boldsymbol{h}_{t-1}$ and $\boldsymbol{c}_{t-1}$ that retain the information at the previous moment. Two fully connected layers estimate the heart rate. We used a clip function at last to limit the network output to a range of [40, 220]. Table 1 also shows the number of operations (ops) required for each forward propagation and the amount of weight parameters (params) required by the network. This is a lightweight network that can operate in real time.

2.4. SQA-rPPG

The attention mechanism in deep learning is based on this concept of directing your focus, and it pays greater attention to certain factors when processing the data. We assume that it is easier to obtain reliable heart rates based on high-quality signals, while the heart rate information is corrupted by noise and difficult to extract in the low-quality signals. Based on this assumption, SQA-rPPG focuses on high-quality signals and ignores low-quality signals.

Signal quality assessment methods based on morphological features [18] or statistical features [8] have been validated that can indirectly improve the accuracy of heart rate estimation. In this paper, we chose the LSTM-based quality assessment network (SQN) in paper [10] because its output size is consistent with LSTM-rPPG and gradients for the weights within the network are easily calculated by back-propagation. SQN is also designed as a many to many RNN architecture and the rPPG signals in the G channel

Type	Activation Fun	Neurons	Params	Ops
FC	selu	8	8	16
LSTM	sigmoid/tanh	8	0.5k	1k
FC	selu	8	64	1
FC	sigmoid	1	8	16

Table 2. Architecture of SQN.

are the input features of SQN. The peaks of heartbeat are more obvious in high-quality segments, which can speed up network training and increase the generalization of the network. Low-quality signals are corrupted by various noises, and it is difficult to design a general denoising model, which affects the generalization of the model.

Table 2 shows the detail of SQN and Fig 4 illuminates the detail of SQA-rPPG. SQA-rPPG contains two LSTM networks for HR prediction and signal quality assessment respectively. For each input sampling point, both LSTM-rPPG and SQN have the output as HR HR_t and quality score s_t as formula (7)(8). A soft attention weighted is computed by formula (9). New average HR between time t and time $t+n$ is calculated by formula (10).

$$HR_t = LSTM{-}rPPG(\boldsymbol{h}_{t-1}, \boldsymbol{c}_{t-1}, \boldsymbol{p}_{t-1}) \tag{7}$$

$$s_t = SQN(\boldsymbol{h}'_{t-1}, \boldsymbol{c}'_{t-1}, \boldsymbol{p}_{t-1}) \tag{8}$$

$$w_{t+t'} = \frac{e^{s_{t+t'}}}{\sum_{k=t}^{t+n} e^{s_k}}, t' = 0, 1, ...n \tag{9}$$

$$\overline{HR} = \sum_{k=t}^{t+n} w_k HR_k \tag{10}$$

3. Experiments

The platform used to implement and validate our models was Pytorch 1.10.1, the GPU used to train the networks was Nvidia 1080ti, the learning rate was set to 0.01. During the training process, the maximum gradient norm was clipped to 5. With different recording webcams, the IIPHCI dataset is divided into M-cam dataset and A-cam dataset. The models in all experiments were trained using the video recording by M-cam on IIPHCI. The parameters in LSTM-rPPG were optimized and the parameters of SQN trained in paper [10] were shared in our model. The parameters in LSTM-rPPG and SQN were used as the initial parameters of SQA-rPPG, then SQA-rPPG was further fine-tuned using the data from M-cam. The mean absolute error (MAE) of heart rate estimation was the criterion and the optimized target of the models.

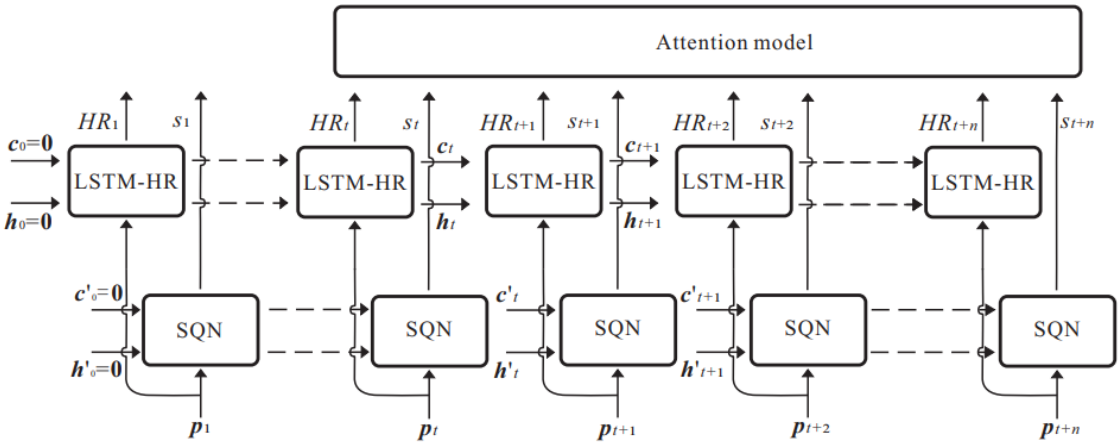


Figure 4. Architecture of SQA-rPPG. The $p_t = (r_t, g_t, b_t)$ is the RGB signals and the input of LSTM-HR. The g_t in p_t is the input of SQN. HR HR_t and quality score s_t predict by LSTM-HR and SQN respectively. The attention model combines $(HR_t, HR_{t+1}, ..., HR_{t+n})$ and $(s_t, s_{t+1}, ..., s_{t+n})$ and calculates the average HR between time t and time $t+n$.

The trained model was applied to the other facial video datasets to further test its generalization to different environments and different webcams. We used a linear interpolation algorithm to keep the sampling rate of different RGB signals from different webcams consistent. We compared SQA-rPPG with different denoise algorithms on different datasets. We also discussed the relationship between MAE of heart rate estimation and quality score from SQN.

3.1. Cross-subject experiments

The videos recording by M-cam on the IIPHCI dataset were used to test the generalization of SQA-rPPG to different subjects in the same dataset, i.e., the same lighting and artificial noise environment. We randomly selected the RGB signals of 17 subjects as the training set. Multiple rounds of experiments were performed using different subjects among the remaining subjects as the test set and the verification set. The model with the smallest MAE in the verification set is used to measure the HR and calculate the MAE of HR estimation on the test set.

Table 3 shows that SQA-rPPG performs well on cross-subject task. On the resting task, the RGB signals are least affected by noise, which makes HR estimation easy and the average MAE is 1.54 bpm. On the talking task and the rotation task, RGB signals were corrupted and the heart rate is difficult to estimate. The average errors are 3.32 bpm and 4.65 bpm respectively on talking task and the rotation task. On talking task, although SQA-rPPG has a large error of the subject XB (9.28 bpm), it has good performance of the subject WR (1.49 bpm), the subject XD (1.82 bpm), the subject XMX (1.8 bpm), the subject ZXS (1.1 bpm). On rotation task, artificial motion noise bring a larger MAE of HR estimation and the MAEs of most subjects were around 3. The difficulty level of rest tasks, talking tasks, and rotation tasks increase in order. The average error of SQA-rPPG on the three tasks is 3.17. which shows SQA-rPPG can accurately estimate the subject's heart rate.

Subject	Resting	Talking	Rotation	Average
WR	0.66	1.49	3.19	1.78
XB	2.82	9.28	3.22	5.11
XD	1.01	1.82	2.85	1.89
XJY	2.01	4.59	3.41	3.34
XK	1.32	3.15	11.25	5.24
XMX	1.33	1.8	6.78	3.30
ZXS	2.44	1.1	5.81	3.12
ZYT	0.76	2.57	3.75	2.36
ZZF	1.51	4.08	1.62	2.40
Average	1.54	3.32	4.65	3.17

Table 3. MAE of heart rate estimation in cross-subject experiment on M-cam dataset (unit: bpm).

3.2. Cross-webcam experiments

We used the SQA-rPPG trained in section 3.1 to predict the RGB signals extracted from the videos captured by another webcam A-cam. M-cam and A-cam recorded the same subject's facial video at the same time, so the heart rate label and illumination environment are the same. The difference is the angle of shooting and the hardware parameters of webcam. Because of the different sampling rates, an interpolation algorithm is required to keep the sampling rates of the RGB signals consistent. The A-cam is fixed on one end of the wearing connecting rod, and the other end is fixed on the cap on the subject's head, which is mainly used to cancel the noise from the rotation movement, and the sit and talk environments are similar to the M-cam.

Subject	Resting	Talking	Rotation	Average
WR	1.31	3.07	2.82	2.4
XB	1.23	1.86	4.2	2.43
XD	0.95	2.23	1.48	1.55
XJY	1.16	2.54	3.08	2.26
XK	1.26	5.46	3.95	3.56
XMX	1.00	3.47	7.04	3.84
ZXS	1.41	2.82	1.06	1.76
ZYT	2.47	2.73	2.88	2.69
ZZF	1.53	2.84	1.47	1.95
Average	1.37	3.00	3.11	2.49

Table 4. MAE of heart rate estimation in cross-webcam experiment with SQA-rPPG trained on M-web dataset and tested on A-web dataset (unit: bpm).

Table 4 shows the results of the SQA-rPPG on cross-webcam task. The average MAE of all tasks is 2.49 bpm, and the MAEs on resting task, talking task and rotation task are 1.37 bpm, 3.00 bpm, and 3.11 bpm respectively. The subjects in the test set are the same as that in section 3.1, and there are no subjects in the training set, so it is a simultaneous cross-subject and cross-device task. Under similar conditions, resting task and talking task achieve similar results, illustrating that SQA-rPPG has good robustness to across-webcam task. The overall average rate of SQA-rPPG in A-cam data is better than that of M-cam, which also indicates that corrupted rPPG signals are difficult to extract heart rate accurately and effectively.

3.3. Compared methods

Since SQA-rPPG are applied to extract HR from RGB signals, we compared SQA-rPPG and LSTM-rPPG with Green [32], ICA [22], CHROM [7], POS [35], 1D-CNN [30] on multiple databases. Among them, Green, ICA, CHROM and POS are applied to extract BVP signals from RGB signals. Then the BVP signal is used in frequency domain analysis to measure the heart rate. The methods are implemented by iPhys Toolbox [20]. While 1D-CNN is used as the HR estimator in paper [30], it can directly estimate the heart rate from the BVP signal.

Table 5 shows the average error of SQA-rPPG in all data sets is 4.68 bpm, which is similar to the best result of CHROM (4.62 bpm). For SQA-rPPG on M-cam dataset, the training and test sets have the same environment, it is easy to optimize model parameters and get the best performance on the cross-subject task. On A-cam dataset, LGI dataset and UBFC dataset, it is cross-dataset for the model. In supervised deep learning models (1D-CNN, LSTM-rPPG, SQA-rPPG), SQA-rPPG achieves better results on A-cam dataset and LGI dataset. LSTM based models can estimate the heart rate of signals of different lengths without retraining, which is more flexible than CNN.

Methods	M-cam	A-cam	LGI	UBFC	Average
Green	5.26	4.38	5.53	14.17	7.34
ICA	6.64	3.47	5.81	6.71	5.66
CHROM	7.14	1.9	5.05	4.39	4.62
POS	11.21	2.51	7.87	5.98	6.89
1D-CNN	3.61	2.84	8.44	5.41	5.08
LSTM-rPPG	3.89	3.01	7.12	6.48	5.13
SQA-rPPG	3.17	2.49	7.05	6.01	4.68

Table 5. MAE of heart rate estimation by each method per dataset. 1D-CNN, LSTM-rPPG, and SQA-rPPG are supervised learning models. A-cam, LGI and UBFC are cross-dataset task for the supervised learning models (unit: bpm).

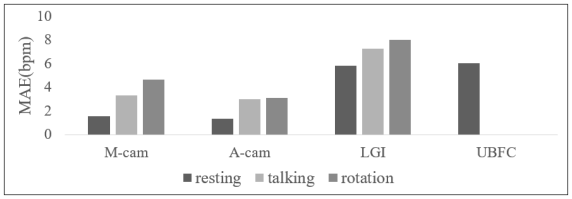
In conclusion, the supervised learning models can get better results than unsupervised learning models when the train and test environment is the same or similar, but when in cross-dataset task, the accuracy will be worse than unsupervised learning models. Compared with unsupervised learning models, the advantage of supervised learning models is that its parameters can be optimized to achieve better performance. SQA-rPPG was trained with the facial videos of 17 subjects, a small amount of data and a simple noise environment, So the cross-dataset results obtained are satisfactory.

3.4. Quality score and MAE of HR estimation

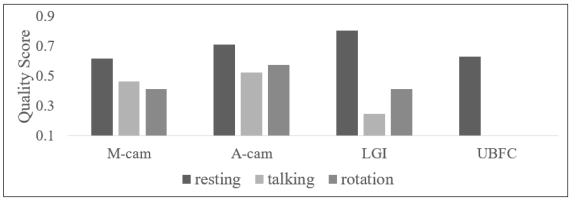
We assumed that the SQA-rPPG pays attention to good-quality signals and ignores low-quality signals. To verify our assumption, we contrasted the MAEs of SQA-rPPG and the quality scores predicted by SQN. In this experiment, the SQA-rPPG is from section 3.1 and the parameters of SQN are shared with SQA-rPPG. On the dataset of M-cam, A-cam and LGI, the data were divided into three parts (resting task, talking task and rotation task). On UBFC dataset, we used data in resting task since the annotated PPG signal in other tasks is not suitable for real-time heart rate extraction with 8 seconds sliding windows.

Fig 5a illuminates the MAEs of SQA-rPPG in different tasks of different datasets. Fig 5b shows the quality scores of SQN in different tasks of different datasets. On each dataset, resting tasks have the lowest heart rate errors and the highest quality score. On the M-cam dataset, the heart rate errors of tasks resting, talking and rotation increase sequentially,, and the corresponding quality scores decrease in order, which are 0.53, 0.34, and 0.27, respectively. Both talking motion and rotation motion bring noise and uncertainty to heart rate estimation, result in the MAE higher and the quality score is lower than the resting task.

Overall, MAE of SQA-rPPG and score quality predicted from SQN are negatively correlated, that is, the accuracy



(a)



(b)

Figure 5. Quality score and MAE of HR estimation on different tasks and datasets.

of the estimated heart rate and the quality score are positively correlated. Using attention mechanism to increase the proportion of heart rates estimated from high-quality segments can effectively improve the accuracy of estimating heart rates.

4. Conclusions

Both LSTM-rPPG and SQA-rPPG have achieved excellent performance in extracting heart rate signals from RGB signals. Heart rates estimated by the models are more accurate than traditional methods when the environment of test set is the same or similar to the training set. SQN can predict quality scores and we have proved the segment with high-quality score is easy to estimate accurate heart rates. In general, SQA-rPPG combines the signal quality score and heart rate estimation, which not only predicts a more accurate average heart rate, but also evaluates the reliability of heart rate through the quality score predicted by the internal SQN model.

References

[1] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1859–1866, 2014. 1

[2] Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3430–3437, 2013. 1

[3] Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. *Applied Sciences*, 9(20):4364, 2019. 2

[4] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 2

[5] Xun Chen, Juan Cheng, Rencheng Song, Yu Liu, Rabab Ward, and Z Jane Wang. Video-based heart rate measurement: Recent advances and future prospects. *IEEE Transactions on Instrumentation and Measurement*, 68(10):3600–3615, 2018. 1

[6] Chun-Hong Cheng, Kwan-Long Wong, Jing-Wei Chin, Tsz-Tai Chan, and Richard HY So. Deep learning methods for remote heart rate measurement: A review and future research agenda. *Sensors*, 21(18):6296, 2021. 2

[7] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1, 6

[8] Mohamed Elgendi. Optimal signal quality index for photoplethysmogram signals. *Bioengineering*, 3(4):21, 2016. 4

[9] Litong Feng, Lai-Man Po, Xuyuan Xu, Yuming Li, and Ruiyi Ma. Motion-resistant remote imaging photoplethysmography based on the optical properties of skin. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):879–891, 2014. 1

[10] Haoyuan Gao, Xiaopei Wu, Chenyun Shi, Qing Gao, and Jidong Geng. A lstm-based realtime signal quality assessment for photoplethysmogram and remote photoplethysmogram. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3831–3840, 2021. 2, 4

[11] Jidong Geng, Chao Zhang, Haoyuan Gao, Yang Lv, and Xiaopei Wu. Motion resistant facial video based heart rate estimation method using head-mounted camera. In *2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, pages 229–237. IEEE, 2021. 2

[12] Mohamed Abul Hassan, Aamir Saeed Malik, David Fofi, Naufal Saad, Babak Karasfi, Yasir Salih Ali, and Fabrice Meriaudeau. Heart rate estimation using facial video: A review. *Biomedical Signal Processing and Control*, 38:346–360, 2017. 1

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[14] Min Hu, Fei Qian, Dong Guo, Xiaohua Wang, Lei He, and Fuji Ren. Eta-rppgnet: Effective time-domain attention network for remote heart rate measurement. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021. 2

[15] Min Hu, Fei Qian, Xiaohua Wang, Lei He, Dong Guo, and Fuji Ren. Robust heart rate estimation with spatial-temporal attention network from facial videos. *IEEE Transactions on Cognitive and Developmental Systems*, 2021. 2

[16] Antony Lam and Yoshinori Kuno. Robust heart rate measurement from video using select random patches. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3640–3648, 2015. 1

[17] M. Lewandowska, J. Ruminski, T. Kocejko, and J. Nowak. Measuring pulse rate with a webcam - a non-contact method

for evaluating cardiac activity. In *Federated Conference on Computer Science and Information Systems - FedCSIS 2011, Szczecin, Poland, 18-21 September 2011, Proceedings*, 2011. 1

[18] Qiao Li and Gari D Clifford. Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. *Physiological measurement*, 33(9):1491, 2012. 4

[19] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. 2

[20] Daniel McDuff and Ethan Blackford. iphys: An open non-contact imaging-based physiological measurement toolbox. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6521–6524. IEEE, 2019. 6

[21] Christian S Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. Local group invariance for heart rate estimation from face videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1254–1262, 2018. 2

[22] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 6

[23] Ming-Zher 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):10762–10774, 2010. 1

[24] Huan Qi, Zhenyu Guo, Xun Chen, Zhiqi Shen, and Z Jane Wang. Video-based human heart rate measurement using joint blind source separation. *Biomedical Signal Processing and Control*, 31:309–320, 2017. 1

[25] Jacek Rumiński. Reliability of pulse measurements in videoplethysmography. *Metrology and Measurement Systems*, 23(3):359–371, 2016. 1

[26] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 2021. 2

[27] Hang Sik Shin, Chungkeun Lee, and Myoungho Lee. Adaptive threshold method for the peak detection of photoplethysmographic waveform. *Computers in biology and medicine*, 39(12):1145–1152, 2009. 1

[28] Arindam Sikdar, Santosh Kumar Behera, and Debi Prosad Dogra. Computer-vision-guided human pulse rate estimation: a review. *IEEE reviews in biomedical engineering*, 9:91–105, 2016. 1

[29] Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1373–1384, 2021. 2

[30] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, Newcastle, UK*, pages 3–6, 2018. 2, 6

[31] Yun-Yun Tsou, Yi-An Lee, Chiou-Ting Hsu, and Shang-Hung Chang. Siamese-rppg network: Remote photoplethysmography signal estimation from face videos. In *Proceedings of the 35th annual ACM symposium on applied computing*, pages 2066–2073, 2020. 2

[32] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 1, 6

[33] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001. 1

[34] Le Wang, Jinliang Zang, Qilin Zhang, Zhenxing Niu, Gang Hua, and Nanning Zheng. Action recognition by an attention-aware temporal weighted convolutional neural network. *Sensors*, 18(7):1979, 2018. 2

[35] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 1, 6

[36] Lin Wanhua, Dingchang Zheng, Guanglin Li, Fei Chen, and Hui Zhou. Investigation on pulse wave forward peak detection and its applications in cardiovascular health. *IEEE Transactions on Biomedical Engineering*, 2021. 1

[37] Bing Wei, Xuan He, Chao Zhang, and Xiaopei Wu. Non-contact, synchronous dynamic measurement of respiratory rate and heart rate based on dual sensitive regions. *Biomedical engineering online*, 16(1):1–21, 2017. 1

[38] Yong-Poh Yu, P Raveendran, Chern-Loon Lim, and Ban-Hoe Kwan. Dynamic heart rate estimation using principal component analysis. *Biomedical optics express*, 6(11):4610–4618, 2015. 1

[39] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 2

[40] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 151–160, 2019. 2

[41] Fang Zhao, Meng Li, Yi Qian, and Joe Z Tsien. Remote measurements of heart and respiration rates for telemedicine. *PloS one*, 8(10):e71384, 2013. 1