A feasibility study of a video-based heart rate estimation method with convolutional neural networks

1st Senle Zhang
Department of Biomedical Engineering
Hefei University of Technology
Hefei, China
zhangsenle@mail.hfut.edu.cn

2nd Rencheng Song
Department of Biomedical Engineering
Hefei University of Technology
Hefei, China
rcsong@hfut.edu.cn

3rd Juan Cheng
Department of Biomedical Engineering
Hefei University of Technology
Hefei, China
chengjuan@hfut.edu.dn

4th Yunfei Zhang
Senturing Technologies Ltd
Vancouver, BC, Canada
yunfeizhang0616@gmail.com

5th Xun Chen
Department of Electronic Science and Technology
University of Science and Technology of China
Hefei, China
xunchen@ustc.edu.cn

***Abstract*—Remote photoplethysmography (rPPG) is a kind of video-based heart rate (HR) estimation technique which has widely potential applications in health monitoring and human-computer interaction. However, the accuracy of conventional rPPG methods is easily disturbed by motion and illumination artifacts. Recently, some deep learning based rPPG methods have attracted many attentions due to its good performance and robustness to noise. This paper proposes a new rPPG scheme using a convolutional neural network (CNN) to map the pulse accumulated image to corresponding true heart rate value, where the spatial-temporal input images are constructed with raw pulses from conventional rPPG methods. In order to check the feasibility and ideal performance of this method, synthetic rPPG pulses are built using real electrocardiograph (ECG) or blood volume pulse (BVP) signals through a modified Akima cubic Hermite interpolation. We test the proposed method in three cases, subject dependent, subject independent, and also a cross-dataset one. The experimental results show that our method performs well in heart rate value estimation with synthetic rPPG pulses even for the cross-dataset case (mean absolute error HR_{mae} = 4.36 BPM, root mean square error HR_{rmse} = 6.26 BPM, mean error rate percentage HR_{mer} = 5.46%). This pilot study verifies the feasibility of the proposed method and provides a solid foundation for the follow-up research with real rPPG pulses.**

Keywords—heart rate estimation, synthetic rPPG pulses, convolutional neural network, feasibility studies

I. INTRODUCTION

Heart rate is an important parameter that reflects the physiological and psychological conditions of the human body. Remote photoplethysmography (rPPG) is a kind of video-based heart rate (HR) measurement technique which is non-invasive and convenient for users to achieve a contactless and long-term monitoring. Therefore, it has received increasing attentions and interests from researchers.

Although rPPG is an attractive technique for remote heart rate estimation, it is easily contaminated with motion and illumination artifacts. The noise due to movements can be divided into two types, the rigid and non-rigid motions. The former one is caused by head rotations or posture changes while the latter one is usually owing to facial expressions or blinks. Illumination variations, such as the variations of indoor light or reflections of computer screens also bring noise into rPPG signals. Those noise can easily bury the weak rPPG signal which makes the rPPG extraction a difficult problem.

After a decade of research, various methods have been proposed to overcome the above difficulties. In general, conventional rPPG methods include the blind source separation methods, model-based methods and motion compensated methods etc. The latest review in this regard can be referred to [1]. Particularly, Xu *et al.* introduced an independent vector analysis in [2] to eliminate illumination interference by extracting common changes in the background and face. In [3], de Hann proposed a chrominance signal method to overcome the motion artifacts. Wu *et al.* introduced a Euler video amplification technology in [4] to improve the SNR of rPPG signals. Although these methods have made good progresses, they are generally designed under controlled environments based on hand-craft features and cumbersome signal processing techniques, which are not robust in realistic situations.

Recently, some deep learning inspired rPPG methods have been proven to get better performance than the conventional ones. A common feature of these methods is the fact that they all map images containing various HR information into HR values. For example, Chen *et. al* proposed an end-to-end system in [5] to map the pulse representation images to heart rate values using convolutional attention networks. The spatial-temporal input image is defined as a difference of normalized video frames. The purpose is to remove the dominant time-invariant terms while keeping the pulse derivative information. Niu *et. al* [6] introduced another spatial-temporal representation of HR information through averaging RGB traces in multiple regions of interests. They also designed a general-to-specific transfer learning strategy to train the convolutional neural networks from a large volume of synthetic rhythm signals to achieve a faster convergence. The success of these methods lies in the powerful mapping and unified learning capabilities of deep neural networks. The training dataset can simultaneously include samples with various noise.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

978-1-5386-8344-6/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on March 19,2021 at 08:52:32 UTC from IEEE Xplore. Restrictions apply.

Therefore, the training model is adaptable to different scenarios. However, the generalization capability of deep learning model is still a crucial issue of such methods which needs to be studied carefully.

In this paper, we propose a new rPPG method based on the deep learning technology. Different from existing schemes, we introduce a novel spatial-temporal representation of HR using estimated pulses from some fast rPPG methods like CHROM [3] or POS [10] *et al.* All images are constructed with rPPG traces in a time-delayed way. This guarantees that the images have a specific and unified structure which is a nice feature to learn. Meanwhile, the input images can be generated in real time and are not disturbed too much by noise. Then a deep convolutional neural network is taken to map the spatial-temporal image to corresponding HR value.

The main goal of this paper is to take a preliminary study on the proposed method to understand the feasibility and its ideal performance such as the generalization capability. Therefore, we construct the input images using synthetic rPPG pulses instead of the real ones. The synthetic rPPG signals are generated from real electrocardiograph (ECG) or blood volume pulse (BVP) signals through a modified Akima cubic Hermite interpolation . We tested the proposed method in three cases, subject dependent, subject independent, and also a cross-dataset one. The experimental results demonstrate that the method with synthetic rPPG pulses performs well in heart rate value estimation even for the cross-dataset task. In short, the contributions of this article are threefold: first, a novel spatial-temporal representation of HR is proposed; second, the ideal performance of the method is well tested to prove the feasibility; third, the training model with synthetic rPPG signals can be further used in a transfer learning for the real one which accelerates the convergence.

The rest of this paper is organized as follows. Section II briefly describes the method, including the generation of synthetic rPPG, the construction of spatial-temporal image and the HR prediction with CNN. Then the proposed method is tested in Section III with both within-dataset and cross-dataset tasks. Finally, conclusions are drawn in Section IV.

II. METHOD

In this section, the proposed method is briefly introduced in three steps, as showed in Fig. 1. First, the synthetic rPPG signal will be interpolated with key points from the real ECG or BVP signal. Then, input images are generated from synthetic rPPG signals by a time-delayed sampling. Finally, the generated images are fed to CNN to predict heart rate values.

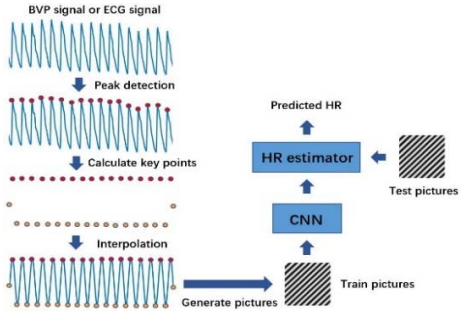

Fig. 1. An overview of the proposed method

A. Synthetic rPPG Signals Generation

The synthetic rPPG signal is interpolated from ECG (or BVP) signals as shown in Fig.1. First, a peak detection is applied on ECG signals with a predefined processing time window. Then the inter beat interval (IBI) sequence is calculated as the difference between adjacent peaks, recorded as $[A_1, A_2, \dots, A_{N-1}]$.

Let $C^1 = [\frac{A_1}{2}, A_1, \dots, A_{N-1}, \frac{A_{N-1}}{2}]$, Then the top positions X of rPPG signal can be defined as a cumulative sum of C^1. Namely,

$$X_{i-1} = \sum_{j=1}^{i-1} C_j^1, i = 2, \dots, N+2. \quad (1)$$

Let $C^2 = \frac{1}{2}[A_1, A_1 + A_2, A_2 + A_3, \dots, A_{N-2} + A_{N-1}]$. Similarly, the bottom positions Y of rPPG signal can be defined as a cumulative sum of C^2, i.e,

$$Y_{i-1} = \frac{A_1}{2} + \sum_{j=1}^{i-1} C_j^2, i = 2, \dots, N. \quad (2)$$

Then the interpolation nodes of synthetic rPPG signal can be defined as Eq. (3),

$$\mathrm{Z} = \begin{pmatrix} 0 & X_1 & Y_1 & \cdots & X_{N-1} & Y_{N-1} & X_N & X_{N+1} \\ a_0 & 1 & -1 & \cdots & 1 & -1 & 1 & a_1 \end{pmatrix}, \quad (3)$$

where the top and bottom values of the curve are set as 1 and -1 respectively, a_0 and a_1 are random numbers between -1 and 1.

The rPPG signal is interpolated at Z with a modified Akima cubic Hermite interpolation method. The curve is finally resampled at a desired sampling rate and is further used to construct the spatial-temporal image as introduced below. It should be noted that the synthetic rPPG signal has the same IBI as real ECG or BVP. Therefore, it is more realistic compared to the synthetic signal designed in [6] which was constructed using sinusoidal curves.

B. Spatial-Temporal Image Construction

After getting synthetic rPPG signals, we construct the spatial-temporal images with the curve in a time-delayed way. Suppose there are total M points of a rPPG signal in the chosen processing window and M is an even number.

Let the first $M/2$ points be extracted and fed into the first row of a matrix. And in turn the second row is from the second point to the $\frac{M}{2}+1$ point, and so on. Therefore, we get a square matrix with size equal to $\frac{M}{2}$. This matrix can be directly converted to a grey structural image as shown in Fig. 1. A spatial-temporal image dataset can be created by repeating the above operations on different processing windows. For each image, the ground truth label is the averaged HR value in the same time window.

Since each row is obtained in a time-delayed way, the image has a specific structure which can be easily learned with CNN. In real case, the rPPG can be estimated by fast conventional methods such as CHROM or POS etc. Then the image is built in the same way using the estimated traces. If three different ROIs are selected, a color image can be synthesized using all the gray images generated within the

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on March 19,2021 at 08:52:32 UTC from IEEE Xplore. Restrictions apply.

same processing window. Compared to the way that builds image directly from RGB traces, our method ensures the generated image is not inferenced so much by noise. It also makes sure all images have a unified structure which is a good feature to learn. Finally, the way of image construction here guarantees there are overlapping parts in neighboring windows in order to achieve a continuous HR monitoring.

C. HR Prediction

The conventional rPPG methods usually take a frequency analysis to extract the HR from the estimated rPPG signal. The frequency with highest amplitude is chosen as the heart rate frequency f_{HR}, which is further converted to the heart rate value by $\mathrm{HR} = f_{HR} * 60$. However, the estimated rPPG signals generally still contaminated with noise, leading to multiple dominate frequencies in the frequency spectrum.

It is well known that the deep neural networks have a universal approximation capability to approximate continuous nonlinear mappings. The mapping from the constructed spatial-temporal image to the corresponding heart rate value is a typical nonlinear regression task, which is very suitable to be solved under the deep learning framework. Here residual neural network (ResNet) is taken to do the work. We take use of the ResNet18 model and replace the last layer of the network with a fully connected layer to predict the HR value. A L_1 loss is employed as the loss function,

$$\mathrm{Loss} = \frac{1}{T}\sum_{i=1}^{T} |\mathrm{HR}_{i,predict}(p) - \mathrm{HR}_{i,label}|, \quad (4)$$

where T is the number of total samples, p is the spatial-temporal image, $\mathrm{HR}_{i,predict}$ is the HR value predicted by CNN, and $\mathrm{HR}_{i,label}$ is the corresponding ground truth. The stochastic gradient descent (SGD) algorithm with momentum is adopted as the optimizer.

III. Experiment

A. Dataset

We evaluate our method on three public datasets: MAHNOB-HCI dataset [7], VIPL-HR dataset [8] and UBFC-RPPG dataset [9]. The MAHNOB-HCI dataset is a multi-modal dataset, which consists of 527 videos from 27 subjects. The VIPL-HR dataset is a recently published dataset for rPPG measurement under diverse situations. It contains 2378 visible light videos and 752 near-infrared videos of 107 subjects. The UBFC-RPPG dataset includes two scenarios, the simple and realistic environments, with a total of 53 videos for rPPG analysis.

In order to evaluate the proposed method, two experiments were conducted.

1) Within-dataset evaluation. We took a five-fold cross validation on the VIPL-HR dataset. We set the processing time window to 5 seconds and the sampling rate is 30 Hz. A total of 17546 pictures were generated according to the proposed method. Furthermore, subject dependent and independent cases were both considered in order to compare their performance.
2) Cross-dataset evaluation. The VIPL-HR dataset and UBFC-RPPG dataset were used for training, while the MAHNOB-HCI dataset was taken for testing. There were 2272 and 6764 pictures constructed from UBFC-RPPG dataset and MAHNOB-HCI dataset respectively. The purpose is to test the generalization capability of the proposed method.

All the generated images were up-sampled to 224 by 224 before being fed into the convolutional neural network.

B. Quality Factors

Several quality factors have been taken to evaluate the performance of the proposed method, such as the mean absolute error (HR_{mae}) ,where $\mathrm{HR}_{mae} = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{HR}_{predict}^{(i)} - \mathrm{HR}_{label}^{(i)}|$, the root mean square error (HR_{rmse}),where $\mathrm{HR}_{rmse} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{HR}_{predict}^{(i)} - \mathrm{HR}_{label}^{(i)})^2}$, and the mean error rate percentage (HR_{mer}), which is defined as $\mathrm{HR}_{mer} = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{HR}_{predict}^{(i)} - \mathrm{HR}_{label}^{(i)}| / \mathrm{HR}_{label}^{(i)}$.

C. Results

We first initialized the ResNet18 network using a pre-trained model on the ImageNet dataset. The last layer of the network was replaced with a fully connected layer to predict a single HR value. First, we fixed the ResNet18 model and only trained the last fully connected layer with 10 epochs. Then we released all parameters of the model and train another 40 epochs with a small learning rate of 10^{-4}.

The results of the proposed method are reported in Table I. It is shown that our approach has performed well in all three tests, even for the cross-dataset one. This verifies the nice generalization capability of the proposed method. The Bland-Altman plot as well as the scatter plot are also illustrated in Figs. 2 to 5 for the three tests respectively. It shows that most of the "gold standard" HR measurements lie in the range of 65 to 95 BPM. Therefore, the scatter points concentrate within this range. As observed, the prediction of ground truth beyond this typical human resting HR range may be degraded due to the lack of enough training samples. It indicates the training dataset not only needs to cover enough realistic situations to against noise but also needs to satisfy the diversity of human HR range to ensure prediction accuracy.

TABLE I: THE RESULTS OF THE PROPOSED METHOD USING SYNTHETIC RPPG SIGNALS.

Test	HR_{mae}(BPM)	HR_{rmse}(BPM)	HR_{mer}(BPM)
Subject-dependent (VIPL-HR)	4.02	5.43	5.02%
Subject-independent (VIPL-HR)	4.24	5.72	5.26%
Cross-dataset (MAHNOB-HCI)	4.36	6.26	5.46%

The above study verifies the feasibility of the proposed method using synthetic rPPG signals. Although there is no fair direct comparison, we still add some reference results from existing deep learning based rPPG methods for real data in Table 2. It is shown that for the same subject-dependent test with VIPL-HR dataset, the method in [6] gets attractive

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on March 19,2021 at 08:52:32 UTC from IEEE Xplore. Restrictions apply.

results for real data. Similarly, the 'DeepPhys' method proposed in [5] also performed well for the cross-dataset test. Those available results show that our approach has great potential for real applications.

TABLE II: SOME REFERENCE RESULTS OF DEEP LEARNING BASED RPPG METHODS FOR REAL DATA.

Method	HR_{mae}(BPM)	HR_{rmse}(BPM)	HR_{mer}(BPM)
RhythmNet [8] Subject-dependent (VIPL-HR)	5.79	8.94	7.38%
DeepPhys [5] Cross-dataset (MAHNOB-HCI)	4.57	-	-

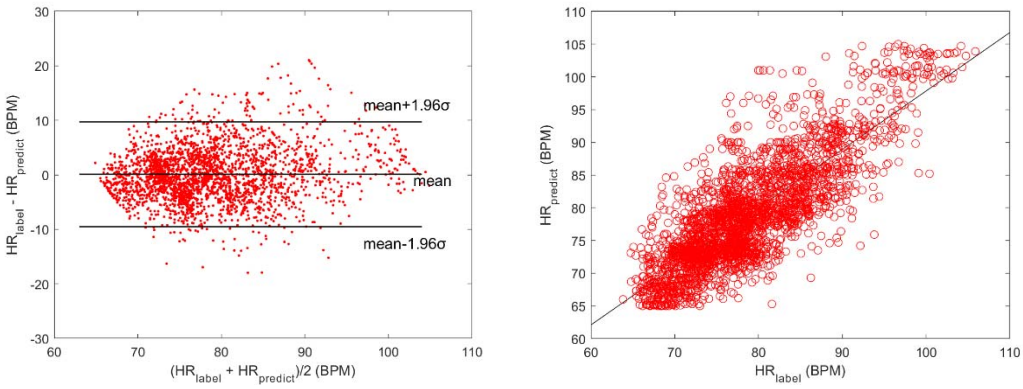

Fig. 2. Bland-Altman plot(left) and scatter plot(right): subject-dependent

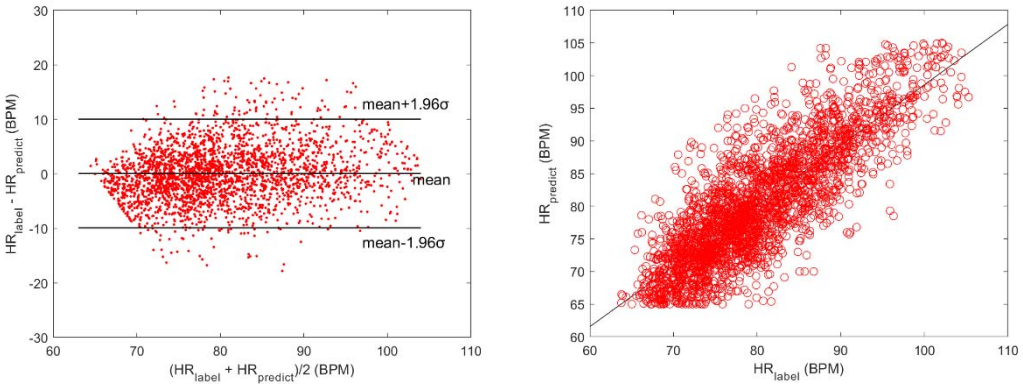

Fig. 3. Bland-Altman plot(left) and scatter plot(right): subject-independent

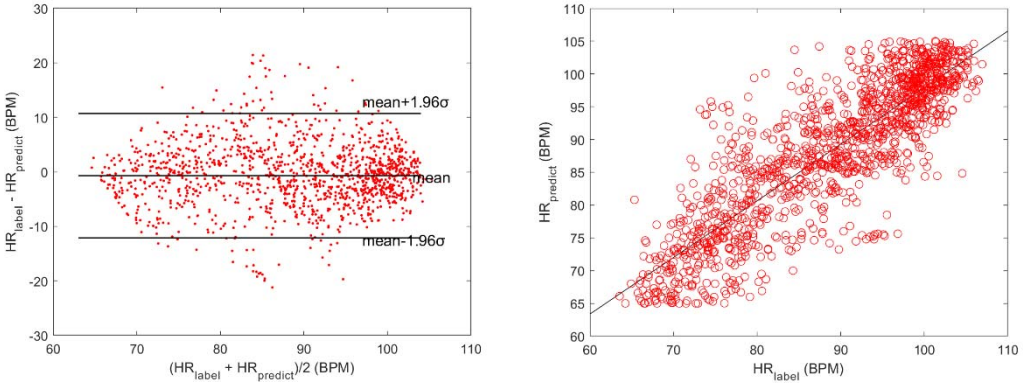

Fig. 4. Bland-Altman plot(left) and scatter plot(right): cross-dataset

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on March 19,2021 at 08:52:32 UTC from IEEE Xplore. Restrictions apply.

IV. Conclusions

This paper presented a new rPPG method based on convolutional neural network to predict heart rate. The method first extracted raw rPPG traces through conventional method like CHROM. Then the spatial-temporal image was constructed using the rPPG trace in a time-delayed way. Finally, the input image was mapped to a corresponding hear rate value by convolutional neural networks. We have created synthetic rPPG curves through interpolating the real ECG or BVP signals. Our method has been verified in three tests, including the subject-dependent, subject-independent, and the cross-dataset one. The experimental results showed that our method with synthetic rPPG pulses performed well in heart rate value estimation even for the cross-dataset task. Those preliminary results have proved the feasibility of the new method and provided a solid foundation for our follow-up research with real rPPG pulses.

Acknowledgement

This work was supported by the National Key R&D Program of China (Grant 2017YFB1002802), National Natural Science Foundation of China (Grant 81571760), and Young Elite Scientists Sponsorship Program by CAST (Grant 2017QNRC001). Rencheng Song is the person to whom correspondence should be addressed.

References

[1] Chen, X., Cheng, J., Song, R., Liu, Y., Ward, R., & Wang, Z. J. (2018). Video-based heart rate measurement: Recent advances and future prospects. IEEE Transactions on Instrumentation and Measurement, in press.

[2] Cheng, J., Chen, X., Xu, L., & Wang, Z. J. (2017). Illumination variation-resistant video-based heart rate measurement using joint blind source separation and ensemble empirical mode decomposition. IEEE Journal of Biomedical and Health Informatics, 21(5), 1422-1433.

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

[4] Wu, H. Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., & Freeman, W. (2012). Eulerian video magnification for revealing subtle changes in the world.

[5] Chen, W., & McDuff, D. (2018). Deepphys: Video-based physiological measurement using convolutional attention networks. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 349-365).

[6] Niu, X., Han, H., Shan, S., & Chen, X. (2018, August). Synrhythm: Learning a deep heart rate estimator from general to specific. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 3580-3585).

[7] Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing, 3(1), 42-55.

[8] Niu, X., Han, H., Shan, S., & Chen, X. (2018). VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. arXiv preprint arXiv:1810.04927.

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

[10] Wang, W., den Brinker, A. C., Stuijk, S., & de Haan, G. (2017). Algorithmic principles of remote PPG. IEEE Transactions on Biomedical Engineering, 64(7), 1479-1491.

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on March 19,2021 at 08:52:32 UTC from IEEE Xplore. Restrictions apply.