VIPL-HR: A Multi-modal Database for Pulse Estimation from Less-constrained Face Video

Xuesong Niu[1,2], Hu Han[1,3,⋆], Shiguang Shan[1,2,4], and Xilin Chen[1,2]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Peng Cheng Laboratory, Shenzhen, China
[4] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
xuesong.niu@vipl.ict.ac.cn; {hanhu, sgshan, xlchen}@ict.ac.cn

Abstract. Heart rate (HR) is an important physiological signal that reflects the physical and emotional activities of humans. Traditional HR measurements are mainly based on contact monitors, which are inconvenient and may cause discomfort for the subjects. Recently, methods have been proposed for remote HR estimation from face videos. However, most of the existing methods focus on well-controlled scenarios, their generalization ability into less-constrained scenarios are not known. At the same time, lacking large-scale databases has limited the use of deep representation learning methods in remote HR estimation. In this paper, we introduce a large-scale multi-modal HR database (named as VIPL-HR), which contains 2,378 visible light videos (VIS) and 752 near-infrared (NIR) videos of 107 subjects. Our VIPL-HR database also contains various variations such as head movements, illumination variations, and acquisition device changes. We also learn a deep HR estimator (named as RhythmNet) with the proposed spatial-temporal representation, which achieves promising results on both the public-domain and our VIPL-HR HR estimation databases. We would like to put the VIPL-HR database into the public domain[1].

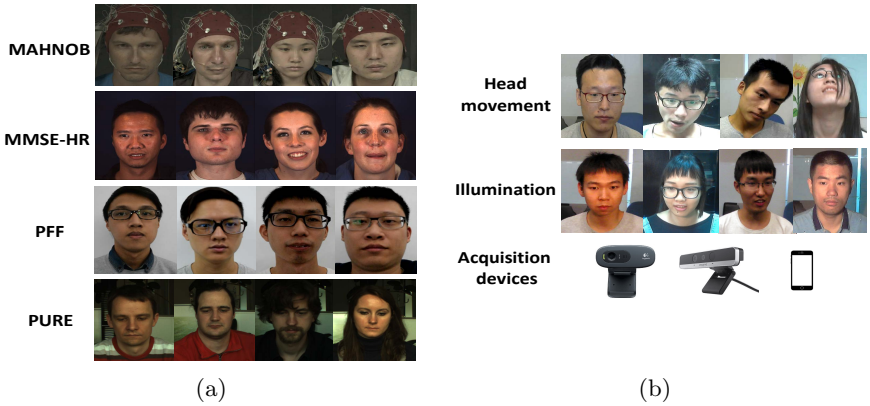
1 Introduction

Heart rate (HR) is an important physiological signal that reflects the physical and emotional activities, and HR measurement can be useful for many applications, such as training aid, health monitoring, and nursing care. Traditional HR measurement usually relies on contact monitors, such as electrocardiograph (ECG) and contact photoplethysmography (cPPG), which are inconvenient for the users and limit the application scenarios. In recent years, a growing number of studies have been reported on remote HR estimation from face videos [15,16,1,3,10,21,23].

⋆ H. Han is the corresponding author.
[1] http://vipl.ict.ac.cn/database.php

The existing video-based HR estimation methods mainly depend on two kinds of signals: the remote photoplethysmography (rPPG) signals [15,16,3,10,21,23] and the ballistocardiographic (BCG) signals [1]. There is no doubt that the idea of estimating HR from face videos could be a convenient physiological platform for clinical monitoring and health caring in the future. However, most of the existing approaches only provide evaluations on private databases, leading to difficulties in comparing different methods. Although a few public-domain HR databases are available [18,21,19,2,9], the sizes of these databases are very limited (usually smaller than 50 subjects). Moreover, these databases are usually captured in a well-controlled scenario, with minor illumination and motion variations (see Fig. 1(a)). These limitations will lead to two issues for the remote HR estimation: i) it is hard to analyze the robustness of individual HR estimation algorithms against different variations and acquisition devices; ii) it is difficult to leverage the deep representation learning approaches in remote HR estimation, which are believed to have the ability to overcome the limitation of hand-crafted methods designed on specific assumptions [13].

To overcome these limitations, we introduce the VIPL-HR database for remote HR estimation, which is a large-scale multi-modal database recorded with various head movement, illumination variations, and acquisition device changes (see Fig. 1(b)).



(a) (b)

Fig. 1: A comparison between (a) the public domain databases (MAHNOB, MMSE-HR, PFF, PURE) and (b) our VIPL-HR database in terms of the illumination condition, head movement, and acquisition device.

Given the large VIPL-HR database, we further propose a deep HR estimator, named as RhythmNet, for robust heart rate estimation from face. RhythmNet takes an informative spatial-temporal map as input and adopts an effective training strategy to learn the HR estimator. The results of within-database and cross-database experiments have shown the effectiveness of the proposed approach.

The rest of this paper is organized as follow: Section 2 discusses the related works of remote HR estimation and existing public-domain HR databases. Section 3 introduces the large-scale VIPL-HR database we have collected. Section 4 provides the details of the proposed RhythmNet. After that, Section 5 evaluates the proposed method under both within-database and cross-database protocols. Finally, the conclusions and future work are summarized in Section 6.

2 Related Work

2.1 Remote HR Estimation

The possibility of using PPG signals captured by custom color cameras was firstly introduced by Verkruysse et.al [22]. Then many algorithms have been proposed, which can be generally divided into blind signal separation (BSS) methods, model-based methods and data-driven methods.

Poh et.al first applied the blind signal separation (BSS) to remote HR estimation [15,16]. They applied independent component analysis (ICA) to temporal filtered red, green, and blue (RGB) color channel signals to seek the heartbeat-related signal, which they assumed is one of the separated independent components. A patch-level HR signal calculation with ICA was performed later in [8] and achieved the state-of-the-art on the public-available database MAHNOB-HCI [18].

Another kind of PPG-based HR estimation methods focus on leveraging the prior knowledge of the skin model to remote HR estimation. Haan and Jeanne firstly proposed a skin optical model of different color channels under the motion condition and computed a chrominance feature using the combination of RGB signals to reduce the motion noise [3]. In a later work of [24], pixel-wise chrominance features are computed and used for HR estimation. A detailed discussion of different skin optical model used for rPPG-based HR estimation is presented in [23], and the authors further proposed a new projection method for the original RGB signals to extract pulse signals. In [12], Niu et al. further applied the chrominance feature [3] to continuous estimation situations.

Besides hand-crafted methods, there are also some data-driven methods designed for remote HR estimation. Tulyakov et.al [21] divided the face into multiple regions of interest (ROI) to get a matrix temporal representation and used a matrix completion approach to purify rPPG signals. In [2], Hsu et al. generated the time-frequency maps from different color signals and used them to learn an HR estimator. Although the existing data-driven approaches attempted to build learning based HR estimator, they failed to build an end-to-end estimator. Besides, the features they used remain hand-crafted, which may not be optimum for the HR estimation task. In [13], Niu et al. proposed a general-to-specific learning strategy to solve the problem of representing HR signals and lacking data. However, they didn't investigate the choice of color spaces and missing data situation.

Instead of the PPG-based HR measurement methods, the ballistocardiographic (BCG) signals, which is the subtle head motions caused by cardiovascular

circulation, can also be used for remote HR estimation. Inspired by the Eulerian magnification method [25], Balakrishnan et al. tracked the key points on face and used PCA to get the pulse signal from the trajectories of feature points. [1]. Since these methods are based on subtle motion, the subjects' voluntary movements will introduce significant influence to the HR signals, leading to limited use in real-life applications.

Although the published methods have made a lot of progress in remote HR measurement, they still have limitations. First, the existing approaches are usually tested on well-controlled small-scale databases, which could not represent the real-life situations. Second, most of the existing approaches are designed in a step-by-step way using hand-crafted features, which are based on some specific assumptions and may fail in some complex conditions. Data-driven methods based on large-scale database are needed. However it is not as simple as visual attribute learning via deep neural network[5], in which the dominant information in the face image is related to the visual attribute learning task. The information in the face image related to the HR estimation task is quite weak.

2.2 Public-domain Databases for Remote HR Estimation

Many of the published methods reported their performance on private databases, leading to difficulties in performance comparison by other approaches. The first public domain database was introduced by Li et.al [10]. They evaluated their method on the MAHNOB-HCI database [18], which was designed for emotion detection and the subjects performed slight head movement and facial expressions. Later in 2016, Tulyakov et.al introduced a new database MMSE-HR [21], which was part of the MMSE database [26] and the subjects' facial expressions were more various. However, these two databases were originally designed for emotion analysis, and the subjects' motions were mainly limited to facial expression changes, which was far from enough for real-world remote HR estimation.

There are also a few public-available databases specially designed for the task of remote HR estimation. Stricker et al. firstly released the PURE database collected by the camera of a mobile sever robot [19]. Hus et al. released the PFF database containing 10 subjects under 8 different situations [2]. These two databases are limited by the number of subjects and recording situations. In 2018, Xiaobai et al. proposed a database designed for HR and heart rate variability (HRV) measurement [9]. Since this database aims at HRV analysis and all the situations in this database are well-controlled, making it very easy for remote HR estimation.

The existing public-domain databases for remote HR estimation can be found in Table 1. As we can see from Table 1, all these databases are limited in either the number of subjects or the recording situations. A large-scale database recorded under real-life variations is required to push the studies on remote HR estimation.

Table 1: A summary of the public-domain databases and VIPL-HR database.

	Subject	**Illumination**	**Head movement**	**Recording devices**	**Video**
MAHNOB [18]	27	*L*	*E*	*C*	527
MMSE-HR [21]	40	*L*	*E*	*C*	102
PURE [19]	10	*L*	*S*	*C*	60
PFF [2]	13	*L/D*	*S/SM*	*C*	104
OBF [9]	106	*L*	*S*	*C/N*	2,120
VIPL-HR	**107**	**L/D/B**	**S/LM/T**	***C/N/P***	**3,130**

L = Lab Environment, D = Dark Environment, B = Bright Environment, E = Expression,
S = Stable, SM = Slight Motion, LM = Large Motion, T = Talking,
C = Color Camera, N = NIR Camera, P = Smart Phone Frontal Camera

3 The VIPL-HR Database

In order to evaluate methods designed for real-world HR estimation, a database containing various face variations, such as head movements and illumination change, and acquisition diversity, is needed. To fill in this gap, we collected the VIPL-HR database, which contains more than one hundred subjects under various illumination conditions and different head movements. All the face videos are recorded by three different cameras and the relative physical measurements, such as HR, SpO2, and blood volume pulse (BVP) signal, are also simultaneously recorded. In this section, we introduce our VIPL-HR database from three aspects: i) setup and data collection, ii) video compression, and iii) database statistics.

3.1 Setup and Data Collection

We design our data collection procedure with two objectives in mind: i) videos should be recorded under natural conditions (i.e., head movement and illumination change) instead of well-controlled situations; and ii) videos should be captured using various recording devices to replicate the common case in daily life, i.e., smartphones, RGB-D cameras, and web cameras. The recording setup is arranged based on these two targets, which includes a computer, an RGB web-camera, an RGB-D camera, a smartphone, a finger pulse oximeter, and a filament lamp. The details of the device specifications can be found in Table 2.

Videos recorded from different devices are the core component of VIPL-HR database. In order to test the influence of cameras with different recording quality, we choose the widely used web-camera Logitech C310 and the color camera of RealSense F200 to record the RGB videos. At the same time, while smartphones have become an indispensable part of our daily lives, remote HR estimation from the videos recorded by smart phone cameras has not been studied yet.

Table 2: Specifications and settings of individual recording devices used in our VIPL-HR database.

Device	Specification	Setting	Output
Computer	Lenovo ThinkCentre	Windows 10 OS	N/A
Color camera	Logitech C310	25fps 960×720 color camera	Color videos
RGB-D camera	RealSense F200	30fps, 640×480 NIR camera 1920×1080 color camera,	Color videos NIR videos
Smart phone	HUAWEI P9 frontal camera	30fps, 1920×1080 color camera	Color videos
BVP recoder	CONTEC CMS60C	N/A	HR, SpO2, and BVP signals
Filament lamp	N/A	50Hz	N/A

Thus, we use a HUAWEI P9 smart phone (with its frontal camera) to record the RGB face videos for the potential applications of remote HR estimation on mobile devices. Besides recording the RGB color videos, we also record the NIR face videos using a RealSense F200 to investigate the possibility of remote HR estimation under dim lighting conditions. Related physiological signals, including HR, SpO2, and BVP signals, are synchronously recorded with a CONTEC CMS60C BVP sensor.

The recording environmental setup is illustrated in Figure 2. The subjects are asked to sit in front of the cameras at two different distances: one meter and 1.5 meters. A filament lamp is placed aside the cameras to change the light conditions. Each subject is asked to sit naturally in front of the cameras, and daily activities such as talking and looking around are encouraged during the video recording. HR changes of the subject after exercises are also taken into consideration. The smartphone is first fixed in front of the subject for video

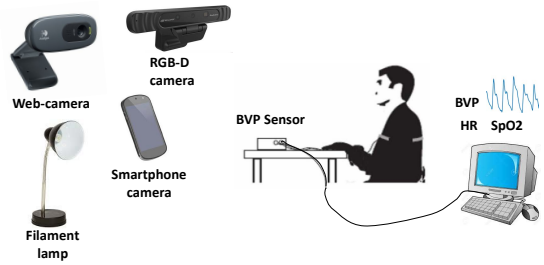


Fig. 2: An illumination of the device setup used in collecting the VIPL-HR database.

Table 3: Details of the nine recording situations in the VIPL-HR database.

Situation	Head movement	Illumination	Distance	Exercise	Phone position
1	S	L	1m	No	Fixed
2	LM	L	1m	No	Fixed
3	T	L	1m	No	Fixed
4	S	B	1m	No	Fixed
5	S	D	1m	No	Fixed
6	S	L	1.5m	No	Fixed
7	S	L	1m	No	Hand-hold
8	LM	L	1m	No	Hand-hold
9	S	L	1m	Yes	Fixed

S = Stable, LM = Large Motion, T = Talking,
L = Lab Environment, D = Dark Environment, B = Bright Environment

recording, and then we asked the subject to hold the smartphone by themselves to record videos like a video chat scenario. Videos under nine different situations are recorded in total for each subject, and the details of these situations are listed in Table 3.

3.2 Database Compression

As stated in [11], video compression plays an important role in video-based heart rate estimation. The raw data of VIPL-HR we collected is nearly 1.05TB in total, making it very inconvenient for the public access. In order to balance the convenience of data sharing and completeness of HR signals, we investigate to make a compressed and resized version of our database, which can retain the completeness of the HR signals as much as possible. The compression methods we considered include video compression and frame resizing. The video compression codecs we take into consideration are 'MJPG', 'FMP4', 'DIVX', 'PIM1' and 'X264', which are commonly-used video compression codecs. The resizing scales we consider are 1/2, 2/3, and 3/4 for each dimension of the original frame. We choose one of the widely used remote HR estimation method Haan2013 [3] as a baseline HR estimation method to verify the HR estimation accuracy changes after individual comparison approaches.

The HR estimation accuracies by the baseline HR estimator on various compressed videos are given in Fig. 3 in terms of root mean square error (RMSE). From the results, we can see that the 'MJPG' video codec is better in maintaining the HR signal in the videos while it is able to reduce the size of the database significantly. Resizing the frames to two-thirds of the image resolution leads to little damage to the HR signal. Therefore, we choose the 'MJPG' codec and two-thirds of the original resolution as our final data compression solution, and we obtained a compressed VIPL-HR dataset with about 48GB. However,

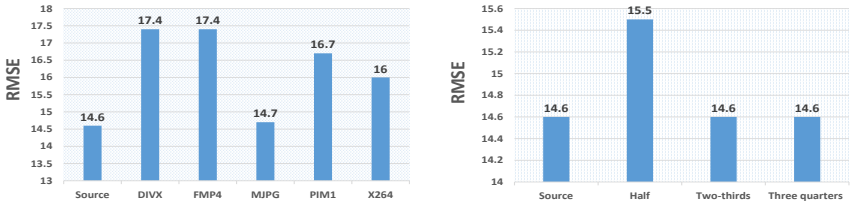


Fig. 3: Evaluations of using (a) different video compression codecs and (b) different image resolutions to compress the VIPL-HR database. All the compression methods are tested using Haan2013 [3]. Source represents the test results on uncompressed data.

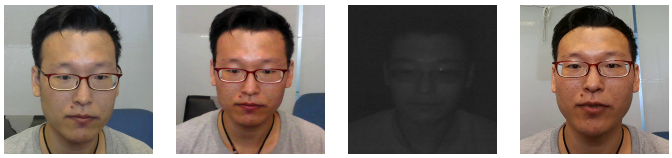


Fig. 4: Example video frames captured for one subject by different devices: (a) Logitech C310, (b) RealSense F200 color camera, (c) RealSense F200 NIR camera, (4) HUAWEI P9 frontal camera.

we would like to share both the uncompressed and compressed databases to the research community based on the researchers' preference.

3.3 Database Statistics

The VIPL-HR dataset contains a total of 2,378 color videos and 752 NIR videos from 107 participants (79 males and 28 females) aged between 22 and 41. Each video is recorded with a length of about 30s, and the frame rate is about 30fps (see Table 1). Some example video frames of one subject captured by different devices are shown in Fig. 4.

To further analyze the characteristics of our VIPL-HR database, we calculated the head pose variations using the OpenFace head pose estimator[2] for the videos with head movement (see Situation 2 in Table 3). Histograms for maximum amplitudes of the three rotation components for all the videos can be found in Fig. 5. From the histograms, we can see that the maximum rotation amplitudes of the subjects vary in a large range, i.e., the maximum rotation amplitudes are 92° in roll, 105° in pitch and 104° in yaw. This is reasonable because every subject is allowed to look around during the video recording.

[2] `https://github.com/TadasBaltrusaitis/OpenFace`

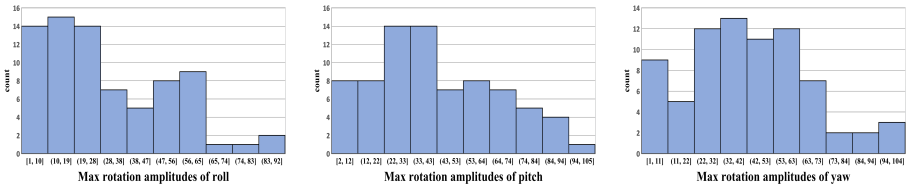


Fig. 5: Histograms for maximum amplitudes of the roll, pitch and yaw rotation components for all the videos with head movement.

At the same time, in order to quantitatively demonstrate the illumination changes in VIPL-HR database, we have calculated the mean grey-scale intensity of face area for Situation 1, Situation 4, and Situation 5 in Table 3. The results are shown in Fig. 6. We can see that the mean gray-scale intensity varies from 60 to 212, covering complicated illumination variations.

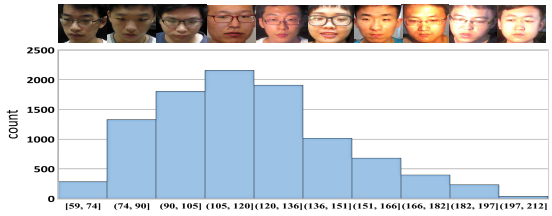


Fig. 6: A histogram of the average image intensity (gray-scale) for the videos recorded under the illumination-specific situations.

A histogram of ground-truth HRs is also shown in Fig. 7. We can see that the ground-truth HRs in VIPL-HR vary from 47 bpm to 146bpm, which covers the typical HR range[3]. The wide HR distribution in VIPL-HR fills the gap between lab-controlled databases and the HR distribution presenting in daily-life scenes. The relatively large size of VIPL-HR also makes it possible to use deep learning methods to build data-driven HR estimators.

4 Deeply Learned HR Estimator

With the less-constrained VIPL-HR database, we are able to build a data-driven HR estimator using deep learning methods. Following the idea of [13], we propose a deep HR estimation method, named as RhythmNet. An overview producer of RyhthmNet can be seen in Fig. 8.

[3] `https://en.wikipedia.org/wiki/Heart_rate`

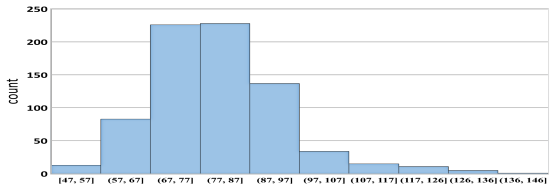


Fig. 7: The histogram showing the ground-truth HR distribution.

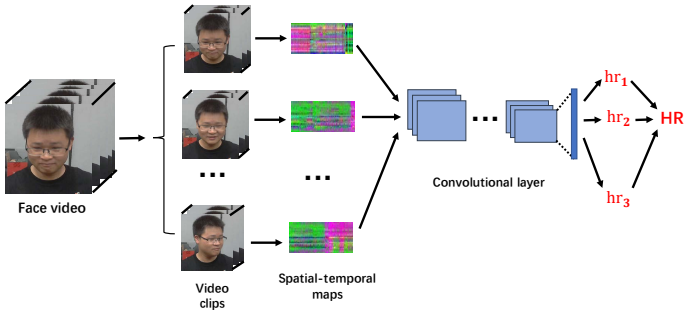


Fig. 8: Overview of RhythmNet. Given an input video sequence, we first divide it into multiple short video clips. Then spatial-temporal maps are generated from the aligned face images within each video clip to represent the HR signals, and a CNN model is trained to predict the HR per video clip. Finally, the average HR estimated for a video sequence is computed as the average of all the HRs.

4.1 Spatial-temporal Maps Generation

In order to identify the face area within the video frames, we first use the face detector provided by the open source SeetaFace[5] to get the face location and 81 facial landmarks (see Fig. 9). Since the facial landmarks detection is able to run at a frame rate of more than 30 fps, we perform face detection and landmarks detection on every frame in order to get consistent ROI localization in a face video sequence. A moving average filter is applied to the 81 landmark points to get more stable landmark localizations.

According to [7], the most informative facial parts containing the color changes due to heart rhythms are the cheek area and the forehead area. The ROI containing both the cheek and forehead area is determined using the cheek border and chin locations as shown in Fig. 9. Face alignment is firstly performed using the eye center points, and then a bounding box is defined with a width of w (where w is the horizontal distance between the outer cheek border points) and height $1.5 * h$ (where h is the vertical distance between chin location and eye center points). Skin segmentation is then applied to the defined ROI to remove the non-face area such as the eye region and the background area.

[5] `https://github.com/seetaface/SeetaFaceEngine`

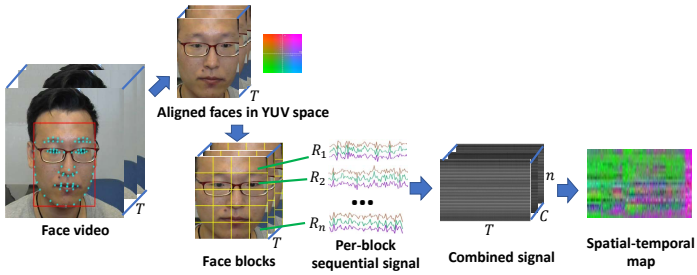


Fig. 9: An illustration of the spatial-temporal map generation producer. We first get the face area for each frame and transform the image to YUV color space, then the face area is divided into n blocks, and the average of the pixel values is concatenated into a sequence of T for C channels in each block. The n blocks is directly placed into rows, and we get the final spatial-temporal map with the size of $n \times T \times C$.

According to [13], a good representation of the HR signals is very important for training a deep HR estimator. Niu et al. directly used the average pixel values of RGB channels as the HR signals representation, which may not be the best choice to represent HR signals. As stated in [20], alternative color spaces derived from RGB video are beneficial for getting a better HR signal representation. After testing alternative color spaces, we finally choose the YUV color space for further computing. The color space transform can be formulized as

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.5 \\ 0.5 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix} \tag{1}$$

and the final spatial-temporal map generation producer can be found in Fig. 9.

4.2 Learning Strategy for HR Measurement

For each face video, we first divide it into small video clips using a fixed sliding window and then compute the spatial-temporal map, which is used to estimate the HR per video clip. The average HR of the face video is computed as the average of the HRs estimated from each clip. We choose the ResNet18 [6] as our convolutional neural network (CNN) for learning the mapping from spatial-temporal maps to HRs, which is commonly used in various computer vision tasks. The ResNet18 architecture including four blocks made up of convolutional layers and residual link, one convolutional layer and one fully connected layer for the final classification. The output of the network is one single HR value, which is normalized based on the frame rate of the face video. L_1 losses are used for measuring the distance between the predicted HR and ground truth HR.

It is important to leverage the prior knowledge of HR signals during HR estimation. To achieve this purpose, we use the synthetic rhythm data for network pre-training as stated in [13]. Specifically, our training strategy could be

divided into three stages. Firstly, we train our model using the large-scale image database ImageNet [17]. Then the synthetic spatial-temporal maps are used to guide the network to learn the prior knowledge of mapping a video sequence into an HR value. With this prior knowledge, we can further fine-tune the neural network for the final HR estimation task using the real-life face videos.

Another situation we need to consider is that the face detector may fail in a short time interval, which commonly happens when the subject's head is moving or rotating. The failing of face detection will cause the missing data of HR signals. In order to handle this issue, we randomly mask (set to zero) the spatial-temporal maps along the time dimension to simulate the missing data situation. The masked spatial-temporal maps are found to be useful to train a robust HR estimator against failing of face detection.

5 Experiments

5.1 Database, Protocol, and Experimental Settings

In this paper, we use the VIPL-HR database for within-database testing and the MMSE-HR database [21] for cross-database evaluation. Details about these two databases can be found in Table. 1. We first perform participant-dependent five-fold cross-validation for within-database testing on the VIPL-HR database. Then, we directly train the RhythmNet on the VIPL-HR database and test it on the MMSE-HR database. After that, the RhythmNet pre-trained on the VIPL-HR is fine-tuned and tested on the MMSE-HR.

Different evaluation metrics have been proposed for remote HR estimation, such as the mean and standard deviation (HR_{me} and HR_{sd}) of the HR error, the mean absolute HR error (HR_{mae}), the root mean squared HR error (HR_{rmse}), the mean error rate percentage (HR_{mer}), and Pearsons correlation coefficients r [10,21]. In this paper, we also use these evaluation metrics for all the experiments below.

For each video, a sliding window with 300 frames is used for generating the spatial-temporal maps. We divide the face area into 25 blocks (5×5 grids). The percentage of masked spatial-temporal maps is 50%, and the mask length varies from 10 frames to 30 frames. Since the face sizes in the NIR videos are small for the face detector, only 497 NIR videos with detected faces are involved in the experiments. The RhythmNet is implemented using the PyTorch[6] framework. The Adam solver with an initial learning rate of 0.001 is applied to train the model, and the number of maximum iteration epochs is 50.

5.2 Within-database Testing

Experiments on Color Face Videos. We first perform the within-database testing on the VIPL-HR database. We use the state-of-the-art methods (Haan2013 [3], Tulyakov2016 [21], Wang2017 [23] and Niu2018 [13]) for comparisons. The results of the individual methods are reported in Table. 4.

[6] `https://pytorch.org/`

Table 4: The results of HR estimation on color face videos on VIPL-HR database.

Method	HR_{me} (bpm)	HR_{sd} (bpm)	HR_{mae} (bpm)	HR_{rmse} (bpm)	HR_{mer}	r
Haan2013 [3]	7.63	15.1	11.4	16.9	17.8%	0.28
Tulyakov2016 [21]	10.8	18.0	15.9	21.0	26.7%	0.11
Wang2017 [23]	7.87	15.3	11.5	17.2	18.5%	0.30
RhythmNet	**1.02**	**8.88**	**5.79**	**8.94**	**7.38%**	**0.73**

Table 5: The results of HR estimation on NIR face videos on VIPL-HR database.

Method	HR_{me} (bpm)	HR_{sd} (bpm)	HR_{mae} (bpm)	HR_{rmse}	HR_{mer}	r
RhythmNet	2.87	11.08	8.50	11.44	11.0%	0.53

From the results, we can see that the proposed method could achieve promising results with an HR_{rmse} of 8.94bpm, which is a much lower error than other methods. At the same time, it can also be seen that our method achieves a better consistency on the VIPL-HR database with a much higher Pearsons correlation coefficient r of 0.73.

Experiment on NIR Face Videos. The experiments on NIR face videos are also conducted using the protocol proposed in Section. 5.1. Since the NIR face videos only have one channel, no color space transformation is used, and we get one-channel spatial-temporal maps for the deep HR estimator. Very few methods have been proposed and evaluated on the NIR data, thus we only report the results based on the RhythmNet in Table. 5.

Table 6: The results of average HR estimation per video using different methods on the MMSE-HR database.

Method	HR_{me} (bpm)	HR_{sd} (bpm)	HR_{rmse} (bpm)	HR_{mer}	r
Li2014 [10]	11.56	20.02	19.95	14.64%	0.38
Haan2013 [3]	9.41	14.08	13.97	12.22%	0.55
Tulyakov2016 [21]	7.61	12.24	11.37	10.84%	0.71
RhythmNet(CrossDB)	-.2.26	10.39	10.58	**5.35%**	0.69
RhythmNet(Fine-tuned)	**-1.30**	**8.16**	**8.22**	5.54%	**0.78**

5.3 Cross-database Testing

The cross-database experiments are then conducted based on the MMSE-HR database. Specifically, we first train our model on the VIPL-HR database and directly test it on the MMSE-HR database. We also fine-tune the model on MMSE-HR to see whether a finetuning could improve the HR estimation accuracy or not. All the results can be found in Table 6. The baseline methods we use for comparisons are Li2014 [10], Haan2013 [3] and Tulyakov2016 [21], and their performances are from [21].

From the results, we can see that the proposed method could achieve a promising performance with an HR_{rmse} of 10.58 bpm, even when we directly test our VIPL-HR pre-trained model on MMSE-HR. The error rate is further reduced to 8.22 bpm when we fine-tune the pre-trained model on MMSE-HR. Both results of the proposed approach are much better than previous methods. These results indicate that the variations of illumination, movement, and acquisition device in the VIPL-HR database are helpful to learn an HR estimator which has good generalization ability to unseen scenarios. In addition, the proposed RhythmNet leverages the diverse information contained in VIPL-HR to learn a robust HR estimator.

6 Conclusion and Further Work

Remote HR estimation from a face video has wide applications; however, accurate HR estimation from the face in the wild is challenging due to the various variations in less-constrained scenarios. In this paper, we introduce a multi-modality VIPL-HR database for remote heart estimation under less-constrained conditions, such as head movement, illumination change, and camera diversity. We also proposed the RhythmNet, a data-driven heart estimator based on CNN, to perform remote HR estimation. Benefited from the proposed spatial-temporal map and the effective training strategy, our approach achieves promising HR accuracies in both within-database and cross-database testing.

In the future, besides investigating new HR signals representations, we are also going to establish models to leverage the relation between adjacent measurements from the sliding windows. Face dense alignment via 3D modeling [4] will also be studied to improve ROI alignment. In addition, detailed analysis of individual methods under various recording situations will be provided using the VIPL-HR database. We would also like to apply the proposed approach into the applications of face presentation attack detection [14], etc.

Acknowledgement

This research was supported in part by the National Key R&D Program of China (grant 2017YFA0700804), Natural Science Foundation of China (grants 61390511, 61672496, 61650202), External Cooperation Program of Chinese Academy of Sciences (grant GJHZ1843).

References

1. Balakrishnan, G., Durand, F., Guttag, J.: Detecting pulse from head motions in video. In: Proc. IEEE CVPR. pp. 3430–3437. IEEE (2013)
2. Gee-Sern Hsu, ArulMurugan Ambikapathi, M.S.C.: Deep learning with time-frequency representation for pulse estimation. In: Proc. IJCB (2017)
3. de Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rPPG. IEEE Trans. Biomed. Eng. **60**(10), 2878–2886 (2013)
4. Han, H., Jain, A.K.: 3D face texture modeling from uncalibrated frontal and profile images. In: Proc. IEEE BTAS. pp. 223–230. IEEE (2012)
5. Han, H., Jain, A.K., Shan, S., Chen, X.: Heterogeneous face attribute estimation: A deep multi-task learning approach. IEEE Trans. Pattern Anal. Mach. Intell. (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE CVPR. pp. 770–778. IEEE (2016)
7. Kwon, S., Kim, J., Lee, D., Park, K.: ROI analysis for remote photoplethysmography on facial video. In: Proc. EMBS. pp. 851–862. IEEE (2015)
8. Lam, A., Kuno, Y.: Robust heart rate measurement from video using select random patches. In: Proc. IEEE ICCV. pp. 3640–3648. IEEE (2015)
9. Li, X., Alikhani, I., Shi, J., Seppanen, T., Junttila, J., Majamaa-Voltti, K., Tulppo, M., Zhao, G.: The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In: Proc. IEEE FG. pp. 242–249. IEEE (2018)
10. Li, X., Chen, J., Zhao, G., Pietikainen, M.: Remote heart rate measurement from face videos under realistic situations. In: Proc. IEEE CVPR. pp. 4264–4271. IEEE (2014)
11. McDuff, D.J., Blackford, E.B., Estepp, J.R.: The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In: Proc. IEEE FG. pp. 63–70. IEEE (2017)
12. Niu, X., Han, H., Shan, S., Chen, X.: Continuous heart rate measurement from face: A robust rPPG approach with distribution learning. In: Proc. IJCB (2017)
13. Niu, X., Han, H., Shan, S., Chen, X.: Synrhythm: Learning a deep heart rate estimator from general to specific. In: Proc. IEEE ICPR. IEEE (2018)
14. Patel, K., Han, H., Jain, A.K.: Secure face unlock: Spoof detection on smartphones. IEEE Trans. Inf. Forensics Security **11**(10), 2268–2283 (2016)
15. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt. Express **18**(10), 10762–10774 (2010)
16. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans. Biomed. Eng. **58**(1), 7–11 (2011)
17. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
18. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affect. Comput. **3**(1), 42–55 (2012)
19. Stricker, R., Müller, S., Gross, H.M.: Non-contact video-based pulse rate measurement on a mobile service robot. In: Proc. IEEE RO-MAN. pp. 1056–1062. IEEE (2014)

20. Tsouri, G.R., Li, Z.: On the benefits of alternative color spaces for noncontact heart rate measurements using standard red-green-blue cameras. J. of Biomed. Opt **20**(4), 048002 (2015)
21. Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J.F., Sebe, N.: Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In: Proc. IEEE CVPR. IEEE (2016)
22. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Opt. Express **16**(26), 21434–21445 (2008)
23. Wang, W., den Brinker, A.C., Stuijk, S., de Haan, G.: Algorithmic principles of remote PPG. IEEE Trans. Biomed. Eng. **64**(7), 1479–1491 (2017)
24. Wang, W., Stuijk, S., De Haan, G.: Exploiting spatial redundancy of image sensor for motion robust rPPG. IEEE Trans. Biomed. Eng. **62**(2), 415–425 (2015)
25. Wu, H.Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., Freeman, W.: Eulerian video magnification for revealing subtle changes in the world. ACM Trans. Graph. **31**(4), 65 (2012)
26. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., et al.: Multimodal spontaneous emotion corpus for human behavior analysis. In: Proc. IEEE CVPR. pp. 3438–3446. IEEE (2016)