2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

SynRhythm: Learning a Deep Heart Rate Estimator from General to Specific

Xuesong Niu[1,2], Hu Han *,[1], Shiguang Shan[1,2,3], 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]CAS Center for Excellence in Brain Science and Intelligence Technology
xuesong.niu@vipl.ict.ac.cn; {hanhu, sgshan, xlchen}@ict.ac.cn

***Abstract*—Remote photoplethysmography (rPPG) based non-contact heart rate (HR) measurement from a face video has drawn increasing attention recently because of its potential applications in many scenarios such as training aid, health monitoring, and nursing care. Although a number of methods have been proposed, most of them are designed under certain assumptions and could fail when such assumptions do not hold. At the same time, while deep learning based methods have been reported to achieve promising results in many computer vision tasks, their use in rPPG-based heart rate estimation has been limited due to the very limited data available in public domain. To overcome this limitation and leverage the strong modeling ability of deep neural networks, in this paper, we propose a novel spatial-temporal representation for the HR signal and design a general-to-specific transfer learning strategy to train a deep heart rate estimator from a large volume of synthetic rhythm signals and a limited number of available face video data. Experiment results on the public-domain databases show the effectiveness of the proposed approach.**

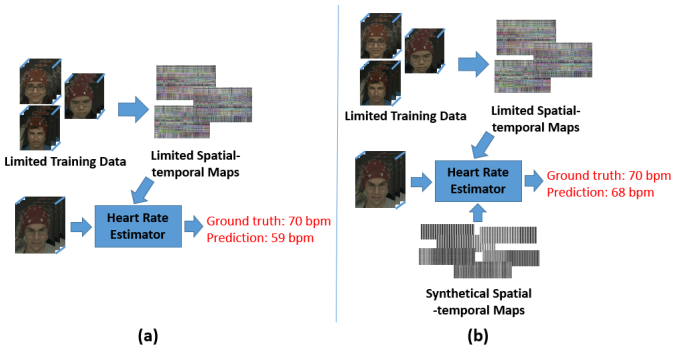

Fig. 1. (a) Due to the small number of available data for training, directly learning a deep HR estimator could get overfitting. (b) With prior knowledge from related domains, such as image classification and synthetic rhythm signals, a deep learning based HR estimator could obtain improved generalization ability.

I. 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 methods usually rely on contact monitors, such as electrocardiograph (ECG) and contact photoplethysmography (cPPG), which are inconvenient for the users in many application scenarios. Recently, remote HR measurement based on remote photoplethysmography (rPPG) from face videos has drawn increasing attention, and many effective methods have been proposed [1], [2], [3], [4], [5], [6], [7].

Existing rPPG based remote HR estimation methods are mainly designed using hand-crafted features and signal processing methods, such as chrominance feature [4] or signal filters [5], which are based on some certain assumptions with respect to the skin reflection and face movement. At the same time, data-driven methods, especially deep learning methods, are believed to have the ability to handle complicated variations and have achieved great success in many tasks like image classification [8] and object detection [9]. However, when it comes to the task of rPPG-based HR estimation, the effectiveness of deep learning could be limited, because most of the existing remote heart rate estimation methods are tested on some small-scale self-collected databases which are usually not available to the public domain. Meanwhile, a good representation of face sequences is also important for learning an accurate HR estimator, but there are only a few attempts have been made on this aspect [10], [11].

In this paper, we present a deep HR estimator learned in a general-to-specific fashion using synthetic spatial-temporal representation. Specifically, we firstly propose a spatial-temporal map representation which can effectively model the periodic signal from face sequences. Then we generate a large scale of rhythmical spatial-temporal maps, which will be further used for pre-training. Finally, the pre-trained model with the knowledge of mapping periodic signals to the number of periods is used to adapt to the HR spatial-temporal maps from face sequences. Experiments on the public-available databases show the effectiveness of our methods.

The main contributions of this work include: (i) we propose a general-to-specific transfer learning method for building a deep rPPG-based HR estimator from synthetic rhythm spatial-temporal maps and (ii) we propose the first known work that uses deep learning to directly estimate HR from spatial-

*H. Han is the corresponding author.

temporal representation of the HR signals from face video sequences.

II. Related Work

In this section, we briefly review previous methods on remote heart rate estimation and transfer learning.

A. Remote Heart Rate Estimation

The possibility of using face videos to estimate HR remotely was first reported by Verkruysse et.al in 2008 [12]. Since then, a number of methods have been proposed for remote HR estimation.

Independent component analysis (ICA) was used in [1] to decompose a multivariate temporal signal into independent non-Gaussian signals, of which one is expected to be the heart rhythm signal. In a later work of [3], temporal filters, such as the moving average filter and bandpass filter, were applied to reduce the noise in the temporal signal sequence.

Under a certain face motion assumption, Haan and Jeanne proposed a chrominance difference feature for remote HR estimation [4]. They computed the chrominance feature based on two orthogonal projections of Red-Green-Blue (RGB) space to reduce the influence of face motion. In the work of [13], instead of a region-wise calculation, a pixel-wise chrominance feature computation method was used.

Recent studies on HR measurement focus on how to select region of interest (ROI) from the face. In [15], Kumar et al. proposed a method to combine the green channel signals of different ROIs using the frequency characteristics. Lam et al. used multiple randomly selected patches from the face as ROI, and used a majority vote rule to decide the final HR estimation [16]. Tulyakov et al. divided the face into multiple ROI regions, and used a matrix completion approach to purify the temporal signals [6]. Niu et al. provided a multi-patch ROI strategy for HR estimation and introduced the problem of continuous HR estimation [7].

Besides of the color-based HR measurement methods, a motion-based method was proposed in [2]. Inspired by the Eulerian magnification method [17], they tracked subtle head motions caused by cardiovascular circulation, and get the pulse signal from the trajectories of multiple tracked feature points. Since the method is based on subtle motion, no additional voluntary head movements are allowed, leading to very limited use in real applications.

In addition to the methods using handcrafted features for remote HR estimation, there are a few attempts to build a learning-based HR estimator. In [11], Hsu et al. combined the frequency domain features of RGB channels as well as ICA components and used support vector regression (SVR) to estimate the HR. Hsu et al. [10] generated the time-frequency maps from pre-processed green channel signals and used them as input of a VGG-16 model to estimate the HR. Although these methods attempts to build learning based HR estimator (as opposed to signal analysis based estimator), they failed to build an end-to-end estimator. In addition, the features they used remain handcrafted, which may not be optimum for the HR estimation task.

Many of the existing methods reported their performance on their private databases, leading to difficulties in performance comparisons with each other. Li et al. [5] firstly introduced the MAHNOB-HCI database [14] and proposed a complete framework for remote HR estimation, which achieved the state-of-the-art HR estimation accuracy. In the latter work of Tulyakov et al. [6], a more challenging database MMSE-HR has been proposed.

In summary, the published methods for HR measurement still have limitations. First, the existing approaches usually made some particular assumptions, which may limit the application scenarios. Second, most of the published approaches are designed in a step-by-step way requiring sophisticated knowledge and experiences.

B. Transfer Learning for Domain Adaption

Transfer learning refers to the methods that transfer knowledge between different but related task domains. It is commonly believed that training and test data are sampled from the same distribution, and most learning methods require rebuilt when the distribution of test data changes. However, it is very expensive, and often impossible for many real-world applications to collect and label a new large-scale training set. In such a case, knowledge transfer from related task domains is desirable in order to build a robust model based on a limited amount of training data in the target domain.

Although deep learning methods have achieved great success in many tasks, it is believed that a large scale of training data is needed in order to learn a robust neural network. When the available training data is limited, a good initialization of the network with prior knowledge has been found to be very helpful [18]. There are many successful applications using transfer learning for domain adaption. In [19], although the number of images used for age estimation is very small(a few thousands), fine-tuning the network pre-trained on large-scale face recognition database still leads to a promising age estimation. In the task of face anti-spoofing, given a few small-scale public databases, Keyurkumar et al. fine-tuned the neural network pre-trained on large-scale face databases to extract texture features that are discriminative between genuine and spoof faces [20].

While the propose approach also uses a deep transfer learning method similar to [19], [20] to build our deep HR estimator in an end-to-end fashion, the proposed approach has its novelty. We design the general-to-specific training strategy specifically for the HR estimation task by generating synthetic rhythm signals.

III. Proposed Approach

Fig. 2 gives a diagram for the proposed approach of learning a deep HR estimator from general to specific via transfer learning. Generally speaking, we firstly use ImageNet [21] and a large amount of synthetic rhythm spatial-temporal maps to pre-train our deep HR regression model. Then the pre-trained

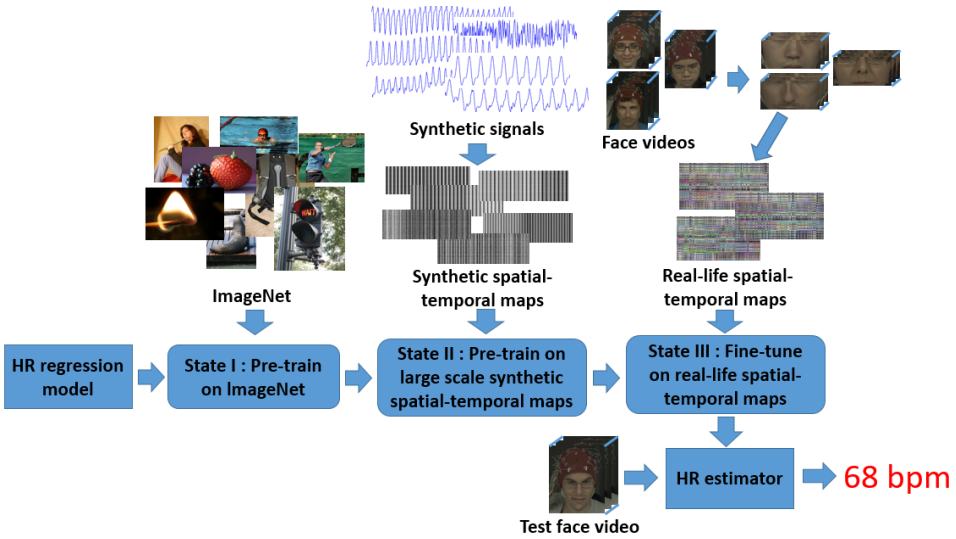

Fig. 2. A diagram of the proposed deep heart rate estimator using a spatial-temporal representation and general-to-specific transfer learning.

model was transferred to the real HR estimation task where only a small portion of operational face video data is available in this target domain.

A. Spatial-temporal Map for Representing HR Signals

According to [22], the most informative facial part containing the color changes due to heart rhythms is the cheek area. This area contains much less non-rigid motions such as eye blink. Therefore, we choose the cheek area as the ROI to get the raw RGB signals. Specifically, we use an open source face detector[1] to localize 81 facial landmarks (see Fig. 3), and calculate the bounding box of the cheek area of the face based on the landmarks. Since the facial landmarks detection is able to run at a frame rate of more than 30 fps, we perform landmarks detection on every frame in order to get stable ROI localizations in a face video sequence.

As shown in Fig. 3, after getting the cheek area, we resize this area into a $M \times N$ rectangle for the convenience of computing. Then we divided the cheek rectangle into n block ROIs. Since each facial landmark has a fixed semantic meaning, we can assume that different blocks are aligned. Based on the n blocks, we can generate a spatial-temporal representation map for each face video sequence. As stated in [12], average pooling is helpful to reduce the sensor noises of heart rhythm signals. Let $C(x, y, t)$ denotes the value at location (x, y) of the t^{th} frame from one of the R, G and B channels, and the average pooling of the i^{th} block ROI for each channel at time t can be presented as

$$\overline{C}_i(t) = \frac{\sum_{x,y \in ROI_i} C(x, y, t)}{|ROI_i|} \quad (1)$$

where $|ROI_i|$ denotes the area of a block (the number of pixels). So, for each face video we obtain $3 \times n$ temporal sequences with the length of T for R, G, and B channels, e.g., $\mathbf{C}_i = \{\overline{C}_i(1), \overline{C}_i(2), \cdots, \overline{C}_i(T)\}$, where C donates one of the R, G and B channels and i donates the index of the ROI. Then, max-min normalization is applied for each temporal signal, and the values of the temporal series are scaled into [0, 255]. Finally, we directly place the n temporal sequences into rows, and get a spatial-temporal map representation for each channel. Eventually, we get a spatial-temporal representation from the raw RGB video sequence with the size of $n \times T \times 3$ as the input of our deep HR estimation network.

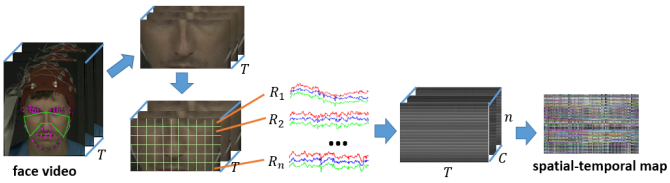

Fig. 3. An illustration of our algorithm to generate the spatial-temporal map representation for a face video. A RGB face video with T frames will finally be converted to a three-channel image with the size of $n \times T$.

B. Synthetic Rhythm Generation

There are only about 600 video sequences for training and testing, which are far from enough to learn a robust deep HR estimation model. To address this problem, we propose an algorithm to generate synthetic heart rhythm to replicate the color changes caused by real heart rhythm.

To be specific, we first use a sine function to represent the basic periodic part of the synthetic signal, and limit the frequency to [0.7, 4] Hz, corresponding to an HR range of [42, 240] bpm. Since the HR signal from a stable subjects face is caused by the entire cardiac cycle, which is a four-stage physical activity, a twice frequency of the basic HR signal will be introduced. This phenomenon can also be observed when we apply Fourier transform to the ECG signal, and a twice frequency of the HR frequency can be clearly seen as

[1] https://github.com/seetaface/SeetaFaceEngine

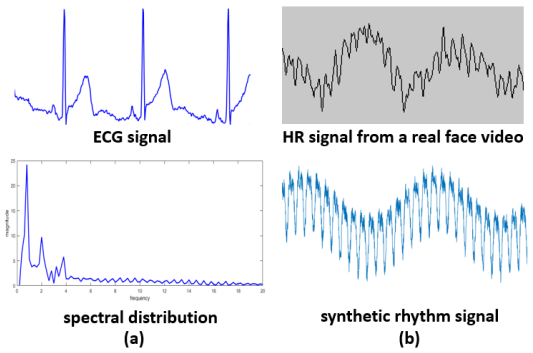

Fig. 4. (a) An ECG signal (top) recorded by an ECG machine, and its corresponding spectral distribution map in the frequency domain (bottom). (b) A temporal HR signal (top) computed from a real face video when a person keeps still, and a temporal HR signal (bottom) computed from a synthetic rhythm sequence.

a peak magnitude in the frequency domain (see Fig. 4(a)). Based on this observation, we overlay a twice-periodic signal to the original signal to simulate the cardiac cycle. At the same time, breathing rhythm will also introduce a periodic signal to the HR signal in a face video sequence. Therefore, we further add a periodic signal range from 5 bpm to 20 bpm (the typical range of the respiratory rhythm). Finally, in order to overcome the noise introduced by facial movement or illumination changes, a random step signal and a random Gaussian noise are added to the original signal. The final formulation of the generated signal S can be presented as follow,

$$\begin{aligned} S =& M_1 \sin(\omega_1 t + \phi) + \\ & 0.5 M_1 \sin(2\omega_1 t + \phi) + \\ & M_2 \sin(\omega_2 t + \theta) + \\ & P_1 Step(t - t_1) + P_2 Step(t - t_2) + N(t); \end{aligned}$$

where M_1 and M_2 are the magnitudes randomly sampled from [0, 1]; ω_1 and ω_2 are the frequencies of cardiac cycle and breath activity; ϕ and θ are the corresponding phases which are randomly sampled from $[0, 2\pi]$. $Step(t)$ is a step signal and t_1 and t_2 are randomly chosen in the range of $[0, T]$. P_1 and P_2 are the probabilities from a Bernoulli distribution. N donates the Gaussian noise function. From Fig. 4(b), we can see that the synthetic signals generated by the proposed approach are able to replicate the real signals when the subject is stable.

As shown in Fig. 2, after generating a large-scale synthetic rhythm database (each sequence has a length of T frames), the spatial-temporal representation maps are calculated from these synthetic signals following the algorithm in Section III-A.

C. General to Specific Deep Transfer Learning

As shown in Fig. 2, in order to avoid the risk of over-fitting, we calculated from the large-scale synthetic spatial-temporal maps to learn a prior knowledge of transforming a RGB video sequence into a HR value in an end-to-end fashion. Specifically, our training strategy could be divided into three stages. Firstly, we train our model using the large-scale image database ImageNet [21] for image classification task in order to obtain network parameter initialization. Then we use the synthetic spatial-temporal maps to further guide the network to learn the prior knowledge of mapping a RGB video sequence into a HR value. With this prior knowledge, we can further fine-tune the neural network for the final HR estimation task using only a limited number of face videos.

IV. Experimental Evaluation

In this section, we provide evaluations of proposed approach from the following perspectives: (i) average HR estimation accuracy per video on the public-domain MAHNOB-HCI database [14] and MMSE-HR database [6] and (ii) the effectiveness of the key components in our method.

A. Experimental Settings

Different measures have been proposed for evaluating the performance of HR estimation methods, such as the HR error (HR_e) between the estimated HR (HR_{est}) and ground-truth HR (HE_{gt}), the mean and standard deviation of the HR error (HR_{me} and HR_{sd}), the root mean squared HR error (HR_{rmse}), and the mean of error rate percentage (HR_{mer}) [5]. In this paper, we use HR_{me}, HR_{sd}, HR_{rmse} and HR_{mer} to report all the results.

In this paper, we conduct experiments on the public-domain MAHNOB-HCI database [14] and MMSE-HR database [6], which has been widely used for remote HR estimation [5], [6]. The MAHNOB-HCI database is a multi-modal database with 20 videos per subject and 27 subjects in total, and the MMSE-HR database includes 102 videos and heart rate information from 40 participants. All the subjects from these two databases participated in the experiment of emotion elicitation and implicit tagging, during which the HR may float because of the change of subjects emotions. The ground-truth HRs are calculated based on the ECG signal provided in the databases.

In order to test the effectiveness of the proposed method, we randomly divide all the databases into three folds and perform cross-validation for training and test. For each 30-second video, we randomly sample 100 sequences of short video clips, each containing 300 frames. We totally get 52,700 spatial-temporal maps for MAHNOB-HCI database and 10,000 maps for MMSE-HR database. The HR estimation of each video is calculated as the average of the prediction results of the 100 sequences of video clips.

For the proposed approach, we use a rectangle ROI of 100×200, and divide it into 200 blocks (10×20 grids). We choose ResNet-18 [23] for feature learning in our regression model. For each fold of the experiment, our network is trained with a learning rate of 0.001 and a maximum of 30 epochs. L1 loss is used as the regression loss function in our experiments.

B. HR Estimation Results

1) Average HR Estimation per Video: In these experiments, following the test protocol in [5], [6], we compare the

TABLE I

THE RESULTS OF ESTIMATING AVERAGE HR PER VIDEO USING DIFFERENT METHODS ON THE MAHNOB-HCI DATABASE (BEST PERFORMANCE IN BOLD).

Method	HR_{me} (bpm)	HR_{sd} (bpm)	HR_{rmse} (bpm)	HR_{mer}
Poh2010 [1]	-8.95	24.3	25.9	25.0%
Poh2011 [3]	2.04	13.5	13.6	13.2%
Balakrishnan2013 [2]	-14.4	15.2	21.0	20.7%
Li2014 [5]	-3.30	6.88	7.62	6.87%
Haan2013 [4]	-2.89	13.67	10.7	12.9%
Niu2017 [7]	-0.38	10.81	8.72	11.5%
Tulyakov2016 [6]	3.19	5.81	6.23	5.93%
Hus2014 [11]	**-0.20**	11.32	11.31	12.8%
Proposed method	0.30	**4.48**	**4.49**	**4.37%**

TABLE II

THE RESULTS OF ESTIMATING AVERAGE HR PER VIDEO USING DIFFERENT METHODS ON THE MMSE-HR DATABASE (BEST PERFORMANCE IN BOLD).

Method	HR_{me} (bpm)	HR_{sd} (bpm)	HR_{rmse} (bpm)	HR_{mer}
Li2014 [5]	11.56	20.02	19.95	14.64%
Haan2013 [4]	9.41	14.08	13.97	12.22%
Tulyakov2016 [6]	7.61	12.24	11.37	10.84%
Proposed method	**-0.01**	**6.86**	**6.83**	**6.21%**

proposed HR estimation method with several state-of-the-art methods for estimating the average HR given a video clip with 30 seconds. The baseline methods we use for comparisons are describe in [1], [3], [2], [4], [5], [6], [7]. The performances of [1], [3], [2], [5] are taken from [5], and the results of [4], [6], [7] are from [7]. At the same time, we re-implement the method in [11], which is a state-of-the-art learning-based HR estimation method, and report its performance under the same protocol. The results of MAHNOB-HCI and MMSE-HR are listed in Table I and II.

It can be seen from the results that our method achieves very promising results with an HR_{rmse} of 4.49 bpm in MAHNOB-HCI and 6.83 bpm in MMSE-HR, which are much smaller than previous methods. At the same time, in order to evaluate the consistency between the ground truth HR and the estimated

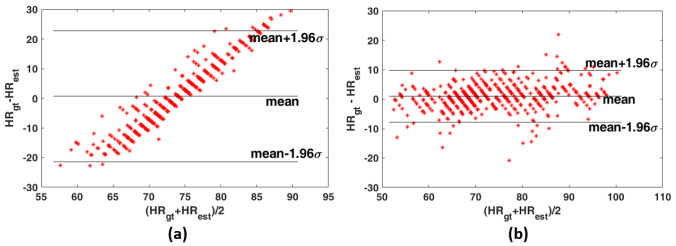

Fig. 5. The Bland-Altman plots demonstrating the agreement of the HR_{est} and HR_{gt} for (a) Hus2014 [11] and (b) the proposed method. The lines represent the mean and 95% limits of agreement.

TABLE III

EFFECTIVENESS OF THE INDIVIDUAL STAGES IN THE PROPOSED TRAINING VIA DEEP TRANSFER LEARNING (BEST PERFORMANCE IN BOLD).

Training Stage	HR_{me} (bpm)	HR_{sd} (bpm)	HR_{rmse} (bpm)	HR_{mer}
Stage-I	**0.14**	4.72	4.72	4.51 %
Stage-II	0.27	4.53	4.53	4.45 %
Stage-III	0.30	**4.48**	**4.49**	**4.37%**

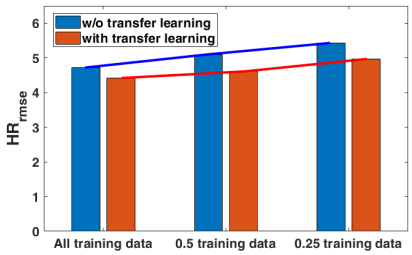

Fig. 6. The changes of the HR estimation accuracy (in terms of HR_{rmse}) with reduced video clips in the training set.

HR, we draw a Bland-Altman plot [24] for the result of MAHNOB-HCI database in Fig. 5. The Bland-Altman plot for another state-of-the-art learning-based method Hus2014 [11] is also given for comparison in Fig. 5. It can be seen that our method has a much smaller standard deviation and a better consistency.

2) Effectiveness of Individual Parts in Our Method: We further analyze the effectiveness of our approach from different aspects on the MAHNOB-HCI database. We first test the performance of our three-stage training strategy step-by step and report the results of each stage in Table III. Specifically, Stage-I denotes training the HR estimator directly using the face videos; Stage-II denotes training the HR estimator with a model pre-trained on ImageNet [21], and Stage-III denotes training the HR estimator using the proposed general-to-specific training strategy. From the results, we can see that each stage in our training strategy could improve the performance, and after all the three training stages, our method achieves the best performance.

In order to further validate the effectiveness of the transfer learning in our method, we reduce the number of video clips used for training by one-half and three quarters, and report the changes in HR estimation accuracy using and not using our deep transfer learning strategy. The results are shown in Fig. 6. It could be seen that with reduced training data, the HR_{rmse} rises for both methods, but using the proposed deep transfer learning leads consistently lower HR estimation error than not using transfer learning.

We then perform three-fold cross-validation using the protocol by dividing the database both randomly and subject-exclusively, and the results are given in Table IV. From the results, we can see that different partition of the database does lead to different HR estimation accuracy. This is because that

TABLE IV
THE RESULTS OF THE PROPOSED APPROACH USING A SUBJECT-EXCLUSIVE PROTOCOL AND A RANDOM-SPLIT PROTOCOL.

Protocol	HR_{me} (bpm)	HR_{sd} (bpm)	HR_{rmse} (bpm)	HR_{mer}
Subject-exclusive	2.16	10.88	11.08	12.26 %
Random split	0.30	4.48	4.49	4.37 %

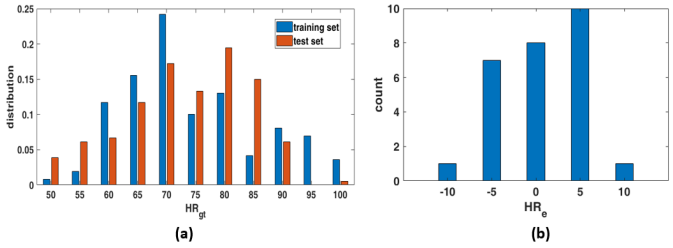

Fig. 7. (a) The HR distributions of the training and test sets in one fold of data. (b) The error distributions of the outliers testing samples which have large HR differences with respect to the ground truth HRs appeared in the training set.

different partition method will lead to different gaps between the data distributions of training and test databases (see Fig. 7(a)). We further analyze whether the network gets over-fitting with respect to individual subject (i.e., remembers individual subjects) under the random-split protocol. We calculate the HR_e for each subject and analysis the clips whose ground truth HR are the most far away from the subject's average HR in the training set. The results are shown in Fig. 7(b), and we can see that for most of these 'outlier' testing video sequences, the proposed approach still give very promising HR estimation. These studies indicate that the proposed deep HR estimator has a good generalization ability.

V. Conclusion and Further Work

Remote HR estimation from a face video sequence using a learning-based model could be challenging because of the lack of training data and various face appearance in motion, illumination, etc. In this paper, to address these limitations, we propose an end-to-end deep learning method for HR estimation, which consists of a novel spatial-temporal representation for HR estimation and a transfer learning strategy leveraging the prior knowledge from the synthetic rhythm data. Extensive evaluations of the proposed approach and comparisons with the state-of-the-art methods show the effectiveness of the proposed approach.

We notice that a large-scale training database is still very important for learning a robust deep HR estimator. In addition, new representations for the HR signals in the RGB face video sequences could benefit the training of a deep learning based HR estimator.

VI. Acknowledgement

This research was supported in part by the National Basic Research Program of China (grant 2015CB351802), Natural Science Foundation of China (61650202 and 61672496), External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843), and Strategic Priority Research Program of CAS (grant XDB02070004).

References

[1] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Opt. Express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.
[2] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proc. IEEE CVPR*, 2013, pp. 3430–3437.
[3] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in non-contact, multiparameter physiological measurements using a webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, 2011.
[4] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, 2013.
[5] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proc. IEEE CVPR*, 2014, pp. 4264–4271.
[6] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proc. IEEE CVPR*, 2016.
[7] X. Niu, H. Han, S. Shan, and X. Chen, "Continuous heart rate measurement from face: A robust rppg approach with distribution learning," in *Proc. IJCB*, 2017.
[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE ICCV*, 2015, pp. 1440–1448.
[10] M.-S. C. Gee-Sern Hsu, ArulMurugan Ambikapathi, "Deep learning with time-frequency representation for pulse estimation," in *Proc. IJCB*, 2017.
[11] Y. Hsu, Y.-L. Lin, and W. Hsu, "Learning-based heart rate detection from remote photoplethysmography features," in *Proc. IEEE ICASSP*, 2014, pp. 4433–4437.
[12] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Opt. Express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.
[13] W. Wang, S. Stuijk, and G. De Haan, "Exploiting spatial redundancy of image sensor for motion robust rppg," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 415–425, 2015.
[14] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2012.
[15] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera," *Biomed. Opt. Express*, vol. 6, no. 5, p. 1565, May 2015.
[16] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," in *Proc. IEEE ICCV*, 2015, pp. 3640–3648.
[17] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, p. 65, 2012.
[18] D. Mishkin and J. Matas, "All you need is a good init," *arXiv preprint arXiv:1511.06422*, 2015.
[19] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen, "Agenet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proc. IEEE ICCV Workshops*, December 2015.
[20] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *Proc. CCBR*, 2016, pp. 611–619.
[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.
[22] S. Kwon, J. Kim, D. Lee, and K. Park, "Roi analysis for remote photoplethysmography on facial video," in *Proc. EMBS*, 2015, pp. 851–862.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
[24] J. M. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.