The First Vision For Vitals (V4V) Challenge for Non-Contact Video-Based Physiological Estimation

Ambareesh Revanur
Robotics Institute
Carnegie Mellon University

Zhihua Li
Dept. of Computer Science
Binghamton University

Umur A. Ciftci
Dept. of Computer Science
Binghamton University

Lijun Yin
Dept. of Computer Science
Binghamton University

László A. Jeni
Robotics Institute
Carnegie Mellon University

Abstract

Telehealth has the potential to offset the high demand for help during public health emergencies, such as the COVID-19 pandemic. Remote Photoplethysmography (rPPG) - the problem of non-invasively estimating blood volume variations in the microvascular tissue from video - would be well suited for these situations. Over the past few years a number of research groups have made rapid advances in remote PPG methods for estimating heart rate from digital video and obtained impressive results. How these various methods compare in naturalistic conditions, where spontaneous behavior, facial expressions, and illumination changes are present, is relatively unknown. To enable comparisons among alternative methods, the 1st Vision for Vitals Challenge (V4V) presented a novel dataset containing high-resolution videos time-locked with varied physiological signals from a diverse population. In this paper, we outline the evaluation protocol, the data used, and the results. V4V is to be held in conjunction with the 2021 International Conference on Computer Vision [1].

1. Introduction

There has been a tremendous interest in the extraction of human physiological signals using just facial videos. Computer vision based physiological extraction has been gaining momentum steadily because this technology has significant benefits over traditional contact-based measurements. Firstly, these methods allow reliable estimation of Heart Rate (HR) and Respiration Rate (RR) in absence of specialized equipment such as electrocardiogram (ECG).

[1] https://vision4vitals.github.io

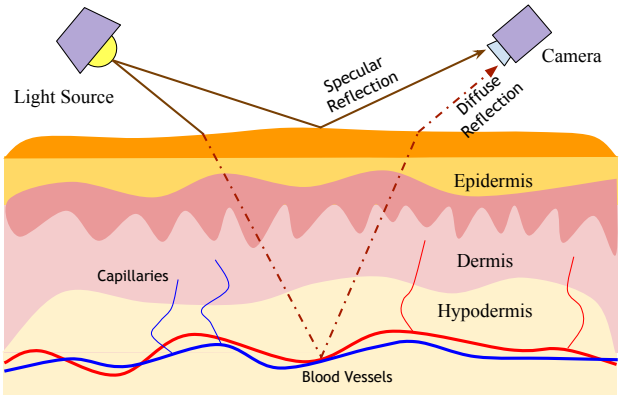

Figure 1. Skin reflection model [33]

These methods depend only on the video feed recorded from a general RGB camera readily available in a commodity smartphone. Secondly, these methods operate without any contact with the subject. Hence, video-based physiology estimation promotes social distancing and is more patient-friendly than contact-based devices. Thirdly, these methods aid in remote diagnosis of patients located in remote areas where quality healthcare facilities are limited. These methods have a wide range of applications including telehealth, deep fake detection, affective computing, human behavior understanding, and sports.

Clinicians use FDA-approved devices such as an electrocardiogram (ECG), a chest belt, and a photoplethysmography (PPG) device for extracting human physiology signals such as HR and RR. Since PPG is closely related to the V4V challenge, we describe the functioning of a PPG device. It is a contact-based device capable of extracting the subtle imperceptible color changes induced as a result of periodic changes in the volume of blood flowing in the un-

Table 1. Summary of previous challenges and recent methods in the area of video-based non-contact physiology estimation. Many of the previous evaluation protocols use non-overlapping segments. In V4V 2021, we use a continuous frame-level error measurement metrics.

	HR	RR	Datasets used	Evaluation Protocol
Related Challenges				
The 1st Remote Physiological Signal Sensing (CVPR'20) [15]	✓	✗	VIPL-HR-v2, OBF	10s non-overlapping segment
The 2nd Remote Physiological Signal Sensing (ICCV'21)	✓	✓	VIPL-HR-v2, OBF	Continuous metric: Inter-beat interval
Recent Methods				
MTTS-CAN (NeurIPS'20) [17]	✓	✓	AFRL, MMSE-HR	30s non-overlapping segment
Feature Disentanglement (ECCV'20) [23]	✓	✓	OBF, VIPL-HR, MMSE-HR	30s non-overlapping segment
1st Vision-For-Vitals Challenge				
Ours (ICCV'21)	✓	✓	*V4V dataset*	Continuous metric: Frame-level HR/RR

derlying skin tissues. In a simplified setup, it consists of a light emitter and receiver. While the emitter is used to focus the light beam on the skin tissue, the receiver is used to record the intensity of light transmitted back to the PPG device. It is known that the absorption spectrum of (oxy-) hemoglobin lies in the color band corresponding to green [32]. Accordingly, the emitter and receiver are designed to capture the periodic color variations in the frequency range of heart rate. Studies have also shown that the respiration rate can be extracted either through motion analysis [10] or using the PPG signals [31]. However, a PPG device is a contact-based method and does not offer the attractive benefits offered by a non-contact-based method. To this end, several video-based non-contact remote physiology estimation methods have been advanced.

The video-based physiology estimation methods [27, 32, 5, 33] exploit the reflectance properties of the skin (typically facial region) with an aim to extract the human physiological signals. Often, the skin tissue is modeled under Shafer's Dichromatic Reflection Model (DRM) [33] that provides a way to model the behavior of the light energy incident on surfaces. As shown in Fig. 1, the light incident on the skin tissue reflects back to the camera as two components - specular and diffuse reflectance. A fraction of the incident light energy that is reflected right off the skin surface is the specular component. This appears as a glossy/shiny reflection on the image captured using the camera. The diffuse reflectance is the light component that passes through the blood-rich tissues under the skin and is then transmitted out. Therefore, the diffuse component contains the signature of physiological signals, while the specular component does not. Similar to PPG, a careful analysis of the variations in the diffuse component of the reflected light shows a pulsatile signal in the frequency range of heart rate. Therefore, video-based physiological measurements techniques are often referred to as Remote-PPG (rPPG) methods.

Several methods have been advanced for the extraction of rPPG signals [5, 33, 27, 32, 3, 17, 23, 12, 19]. Owing to the advancements in computer vision and deep learning, remarkable results have been achieved on the task of human physiology estimation. However, there are two drawbacks with existing methods. First, it is not clear how these various methods compare in naturalistic conditions, where spontaneous movements, facial expressions, and illumination changes are present. Second, most previous benchmarking efforts focused on posed situations. No commonly accepted evaluation protocol exists for estimating vital signs in spontaneous behavior with which to compare them. Therefore, in the 1st Vision-For-Vitals 2021 challenge, we introduce a new dataset called *Vision-For-Vitals (V4V) dataset* which includes challenging elements such as spontaneous behavior. We also contribute a new evaluation metric for a stronger benchmarking on the *V4V dataset*.

2. Related works

Table 1 summarizes the related challenges and most recent methods in the area of remote video-based human physiological estimation.

Related challenges. In conjunction with CVPR'20, the 1st Remote Physiological Signal Sensing (RePSS) challenge was organized for estimation of heart rate using RGB videos. The challenge consisted of a training dataset that was drawn from VIPL-HR-v2 [22] and a test set that was drawn from VIPL-HR-v2 [22] and the OBF [14] datasets. There were about 2500 samples, each 10 seconds long, in the training dataset. In order to test the efficacy of different methods employed in the challenge, the organizers used a segment-level evaluation protocol, i.e., one heart rate prediction for a 10s segment.

The 2nd Remote Physiological Signal Sensing (RePSS) challenge was held recently[2], and it included samples drawn from VIPL-HR-V2 and OBF. In this challenge, the orga-

[2]in conjunction with ICCV'21

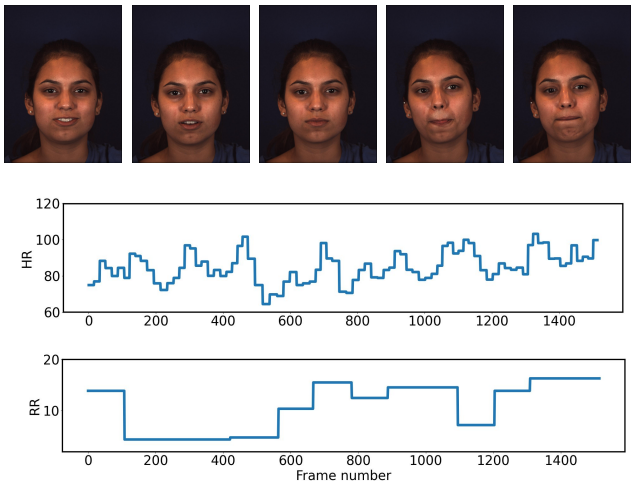

Figure 2. Subject participating in a cold presser (T8) task

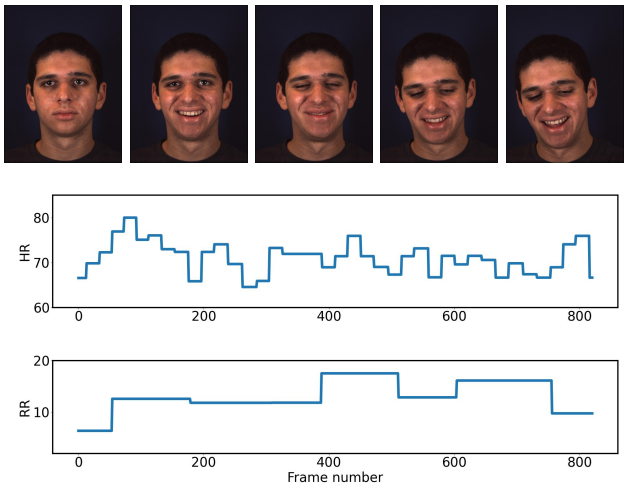

Figure 3. Subject participating in a silly song (T6) task

nizers introduced an inter-beat-interval (IBI) based metric for measuring the performance of the participants' methods. Unlike the segment-level metric used in the 1st RePSS, the IBI-based metric is more granular as it penalizes missed/extra heartbeat predictions. In the 2nd RePSS challenge, the organizers also included respiration rate as a separate challenge track. Our V4V challenge is similar in spirit to the 2nd RePSS as we used granular evaluation metrics and have both tracks (HR and RR). Further, we introduce a newly curated dataset that contains challenging elements such as spontaneous behavior and varied physiological signals as part of our V4V challenge.

Methods. Traditional rPPG extraction methods typically involve two stages. The signal is first extracted based on the rPPG principles, and then the signal processing methods are used to compute the HR and RR. In [32], the pixels in the green channel are used to extract the physiological signals since it contains a strong signature of the pulsating rPPG signal. Some methods such as [28, 27, 4] use an ICA method to determine the underlying rPPG signal followed by 3rd order Butterworth bandpass filtering to obtain the power spectrogram whose peak corresponds to the heart rate (in a valid range of 0.7Hz - 2.5Hz). Tarassenko [31] et. al propose a pipelined approach that uses face tracking and pole selection mechanism to estimate heart rate, respiration rate, and oxygen saturation. However, these methods are susceptible to noise, motion, and lighting conditions of the environment.

Advancements in deep learning [12, 17, 13, 24, 29, 18] have made it possible to achieve remarkable performance in the task of video-based human physiological estimation. One of the early popular methods in this direction is DeepPhys [3]. This model is trained on many facial videos in a supervised fashion, where ground truth blood volume pulse was used as the label. The key idea is to use separate branches for modeling motion and appearance. The latter branch aids the former by providing attention over facial pixels. Similarly, a more recent method MTTS-CAN [17] used the attention mechanism in conjunction with Temporal Shift Modules [16] to achieve real-time performance on the task. In [23], the authors demonstrated an effective method to disentangle spatio-temporal representation of the video called MSTmaps into noisy signals and physiological signals. In summary, deep learning methods have demonstrated reliable performance on datasets containing relatively stable physiological signals. In the V4V challenge, we introduce a dataset in which the physiological signals vary significantly due to the elicitation of spontaneous emotions.

3. *V4V dataset* curation

As part of the V4V challenge, we curated a database called the *V4V dataset* by carefully selecting subjects from the Multimodal Spontaneous Emotion database (BP4D+) [37] along with a number of new subjects that are collected as BP4D+ extension. In this section, we describe the data collection process, distribution of subjects, and data annotation process used for creating the *V4V dataset*. In this section, we also describe the significance of the dataset for V4V challenge since the dataset contains challenging variations in the physiological signals induced through emotion elicitation.

3.1. Data collection and annotation

The *V4V dataset* was curated with the goal of obtaining a large-scale emotional corpus for human behavioral and physiological analysis. The dataset is collected at Binghamton University and includes subjects of age groups ranging from 18 to 66. It has subjects from diverse ethnicities/racial ancestries - African American, White, Asian

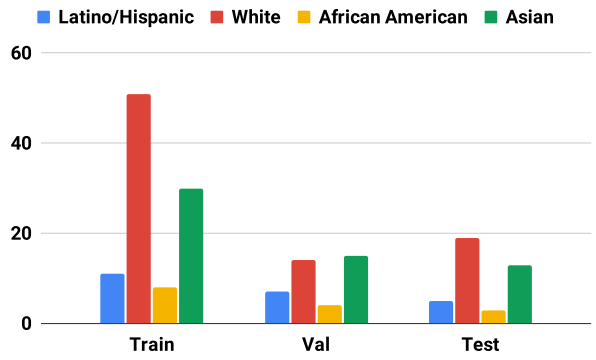

Figure 4. Distribution of videos according to the ethnicity of subject for train, validation, and test subjects.

(East and Middle-east), Hispanic/Latino, Native American. There are 179 subjects in total with a maximum of 10 experimental tasks per subject. Each task was specifically designed to induce specific emotions among participants.

For recording the facial videos of each subject, a 3D Dynamic imaging system (Di3D) was used. All the videos were recorded with a resolution of 1040×1392 pixels and a fixed frame rate of 25 FPS in portrait mode. The Di3D system also has a symmetric lighting system that is used as the light source for capturing the videos. A board has been placed in the background while recording the video to limit any background motion and noise.

For collecting human physiological data, the Biopac MP150 system was used. The specification of the device is as follows:

- Blood pressure: For monitoring the blood pressure, Biopac NIBP100D system with a measurement range of (-25mmHg, 300mmHg) was used. It contains a finger unit and an inflatable cuff that can be placed on the arm to measure blood pressure. The device recorded high-quality measurements of systolic and diastolic blood pressure and also recorded the continuous blood pressure waveform at 1000Hz.

- Heart Rate (HR) measurement: We used off-the-shelf software called Biopac AcqKnowledge to derive HR measurement from the continuous blood pressure signal. This was achieved by performing noise removal, followed by peak-to-peak time calculation. The software used an HR range preset to (40, 180) beats-per-min. The HR signal is then downsampled and synchronized with each frame.

- Respiration Rate (RR) measurement: The respiratory signal was captured by using the Biopac Respiration Belt. Similar to the extraction of HR, we used the Biopac Acqknowledge software to extract the respiration rate of the subject by computing peak-to-peak time with RR range preset to (4, 20) breaths-per-min.

Table 2. Ten tasks and the target emotion

Task	Activity performed	Emotion induced
T1	Funny joke	Happiness
T2	Watch 3D avatar of self	Surprise
T3	911 emergency call	Sadness
T4	Sound	Surprise
T5	True / False question	Skepticism
T6	Silly song	Embarrassment
T7	Dart threat	Fear
T8	Cold presser	Pain
T9	Complaining against participant	Angry
T10	Odor experience	Disgust

Table 3. *V4V Dataset*: Data split

Data Fold	Number of subjects	Number of Videos	Average video length (in sec)
Train	100	724	44.2
Validation	39	276	42.9
Test	40	358	45.8

In order to synchronize the camera and physiological devices, a controller was used to trigger the start of video capture and physiology measurement simultaneously.

3.2. Emotion elicitation protocol of the *V4V dataset*

Each subject participated in 10 different tasks that were carefully designed to evoke a specific emotion. It is known that an increased emotional activity often alters the human physiological signals [34]. For example, fear arousal spontaneously increases the heart rate and respiration rate of a subject. To achieve this goal, a skilled interviewer was hired to conduct the tasks.

As shown in Table 2, the tasks included interpersonal communication, film watching, cold pressure, and physical activities. There is also a smooth transition in the emotions induced across the 10 tasks. There was a brief pause between any two tasks for self-reporting purposes.

First, the interviewer ensured that the participant felt comfortable and relaxed at the start of data collection by advancing a joke (T1). Then the subject was shown their own 3D avatar to invoke the feeling of surprise (T2). Next, the T3 task required participant to watch a short film of a 911 emergency call to elicit a feeling of sadness. In the T4 task, a loud noise was played to startle and surprise the participant. The interviewer then induced skepticism by advancing a question (T5), followed by arousal of embarrassment when the subject was required to conduct a silly task (T6). In the T7 task, the interviewer invoked fear in the participant by threatening to throw a dart at the subject. In T8, the participant was required to submerge their hand into the

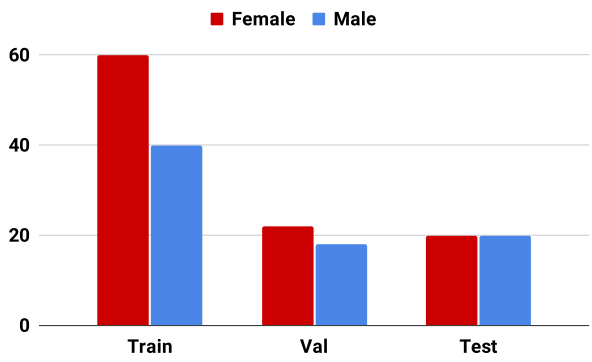

Figure 5. Distribution of videos according to subject gender for train, validation, and test splits.

ice water which invoked physical pain. The interviewer pretended that the participant demonstrated poor performance in task T8 and complained to the participant to evoke anger (T9). In the final task, T10, the subject experienced a smelly odor to evoke the feeling of disgust. At the end of each task, the subject was asked to report the emotions experienced from a list of choices and also rate the emotional intensity in a 5-point rating style.

Owing to the carefully designed emotion elicitation, the *V4V dataset* contains challenging intra-video physiological variations. Further, due to the nature of the tasks T6-T10, they often are associated with large head movements (> 10 deg) adding an additional element of challenge for physiological estimation. Therefore, *V4V dataset* offers desirable elements for benchmarking approaches effectively.

3.3. Post-processing

After creating the ground truth, we eyeballed each of the heart rate and respiration rate sequences to discard any video that had noisy readings, e.g. Shaking the contact-based device during emotion elicitation tasks. We ensured that every video had exactly 1 HR and 1 RR reading per frame of the video after aligning the physiological signals. After processing the data as described, we obtained 179 subjects and 1358 videos with heart rate and respiration rate readings.

3.4. V4V Challenge phases and dataset split

The 1st V4V challenge was organized in two phases. In the first phase, the participants used the validation set to improve performance of their methods, and in the second phase, the participants evaluated their method on the test set. A public leaderboard was maintained in both the phases indicating scores obtained by the methods on the evaluation metrics described in Sec. 4.

As shown in Table 3, we used 1000 videos in phase 1 of the challenge where the training set included 724 videos and validation set included 276 videos. At the start of phase-1, we released the training dataset and validation set videos. The training set contains the compressed videos, heart rate signal, respiration rate signal, and raw 1000 Hz blood pressure waveform. At the start of phase-2, we released the test set videos (a total of 358 videos) and validation-set ground truth as well.

In order to ensure that the methods performed fairly across different population groups, we used gender and ethnicity information self-reported by the subjects to create the data folds. Fig. 4 shows the ethnic distribution of subjects of train, validation, and test set. We tried to balance the ethnic distribution keeping them similar across the data splits to avoid the biases brought by skin tone and affective attributes. Fig. 5 presents the distribution of videos according to the gender of the subject, which has been adjusted to have near-balanced distribution, with the number of female subjects slightly more than the number of male subjects in the train and validation set.

4. Evaluation metrics

In the existing literature, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson Correlation Coefficient (R) have been used commonly for evaluating the efficacy of the proposed method. The caveat is that almost all of the existing works construct 30-second non-overlapping segments and predict a single HR/RR value per segment. While this is a good measure of the performance of the method, it has some potential drawbacks.

Unlike other physiological datasets, the *V4V dataset* consists of short video clips that have significant intra-video variations in the HR and RR (as seen in Fig. 2 and Fig. 3) for the duration of the video. Therefore, an accurate method should be able to predict HR and RR at a more fine-grained frame-level rather than at a segment-level. Further, contact based devices predict the heart rate continuously, instead of one prediction for every 30s. This is especially useful in diagnosing Atrial fibrillation (Afib) [26] which is associated with high Heart Rate Variability. Based on these observations, we propose a metric that computes error by taking into account the frame-level HR/RR.

In the 1st V4V challenge, we employ the following three metrics - MAE, RMSE, and R - at a frame-level rather than at segment-level. We denote these granular evaluation metrics as cMAE, cRMSE, and cR respectively.

$$cMAE = \frac{\Sigma_i |\widehat{HR}_i - HR_i|}{N} \tag{1}$$

$$cRMSE = \sqrt{\frac{(\Sigma_i |\widehat{HR}_i - HR_i|^2)}{N}} \tag{2}$$

$$cR = \mathbf{PearsonCorrelation}(\widehat{HR}, HR) \tag{3}$$

Table 4. Results obtained on the *V4V dataset* (test set) sorted by cMAE (lower is better ↓)

Approach	cMAE (↓)	cRMSE (↓)	cR (↑)
Stent et. al [30]	9.22	14.18	0.47
Hill et. al [9]	9.37	14.59	0.44
Kossack et. al [11]	10.15	15.38	0.44
Ouzar et. al [25]	11.60	14.90	0.20
Baseline (Green [32])	15.45	20.73	0.05

where HR_i is the ground truth heart rate of the frame i in the test set, $\widehat{HR}_i$ is the predicted heart rate for the frame i in the test set and N is the total number of frames in the test set. Similarly, we define equivalent evaluation metrics for Respiration Rate using RR_i and $\widehat{RR}_i$.

5. Methods used in the V4V challenge

There were four teams that participated in the V4V challenge. In Table. 4 results obtained by different teams and a baseline (Green) method [32, 20] have been listed.

5.1. Estimating Heart Rate from Unlabelled Video

Stent et. al [30] use a self-supervised approach based on [7] to overcome challenges that are typically faced by a supervised approach such as imprecise and noisy data collection owing to complex ground truth capturing setup, proprietary hardware computations, etc. Further, by virtue of the self-supervised nature of the proposed method, it can also tackle any domain shift problem that arises during the test phase. First, an image of size 192×128 is extracted by using a face detector [36] and padded with an additional 25% buffer. Then a PPG estimator consisting of 3DCNN [35] is used to extract the rPPG signal.

The core idea of this approach is to formulate a self-supervised framework by utilizing a video resampling module to augment the dataset with new heart rate labels. E.g., the heart rate of the subject can be increased by squeezing video. Similarly, stretching the video decreases the heart rate. The module is not only used to increase/decrease the heart rate by controlling the speed for video, but it is also used to resample the output rPPG signal with corresponding inverse frequency. Based on this idea, positive and negative pairs are constructed for an effective contrastive learning [8, 2].

Owing to the self-supervised nature of the proposed method, the method is able to train on the test samples directly. The paper also implements several hand-crafted tricks to further improve performance of the their method. The authors employ specialist models to account for multi-mode lighting conditions in the test set and also propose a confidence model to mitigate errors caused by faulty PPG prediction.

5.2. Beat-to-Beat Cardiac Pulse Rate Measurement From Video

Hill et al. [9] proposed a hybrid-CAN-RNN framework by incorporating a bi-directional GRU on the top of CAN [17]. The original Hybrid-CAN has two branches: (1) the appearance branch, and (2) the motion branch. The appearance branch deploys 2D convolutions to extract the spatial skin features. The temporal module leverages 3D convolutions to capture the temporal relationships. An attention module is used to bridge the connection between motion and appearance branch, and the generated attention masks make the network focus on the useful signals.

In addition, two layers of GRU (the first is bi-directional) are used for learning the longer-term rPPG waveform transitions, and the reason is that the convolutional layers can only capture local spatial and temporal features. Besides, a synthesized dataset is used to improve the model generalization ability, and other datasets such as AFRL [6] and UBFC [1] are used for training. The variety of the training data improves the model generalization ability, which helps to deal with more complicated environmental biases.

5.3. LCOMS Lab's approach to the Vision For Vitals (V4V) Challenge

To model the spatial and temporal features concurrently, Ouzar et al. [25] make use of the residual connected 3D Depthwise Separable Convolution layers. The depth-wise separable convolution significantly reduces the parameters and saves the computational time, and achieves satisfactory results. Moreover, their method can run in real-time both in CPUs and GPUs.

Different from the others that directly feed the images with the background to the deep networks, this paper uses a precise face swapping-based segmentation method [21] to exclude the background and only keep the facial regions. The paper shows that most of the heart rate values are inside the range of 70 BPM to 90 BPM. And to tackle this data imbalance issue, they performed an offline data augmentation to the sequences that have minority heart rate values.

Although the framework is end-to-end trainable and is superior in speed and simplicity, the basic idea is still treating each task as a one-stage regression problem. The model predicts one average heart rate value in two seconds video segment. Proposing a solution to predict frame-wise heart rate (e.g. combining frame-specific features with segment features) can be a future work of this paper.

5.4. Automatic region-based heart rate measurement using remote photoplethysmography

Kossack et. al [11] utilize a popular classical method called "plane orthogonal to skin" (POS) [33] to measure the heart rate. The core idea of this V4V submission is to extract a signal from the region of the face that has the

strongest signature of rPPG in it. In order to achieve this goal, the face region is divided into five subregions - forehead, right cheek, left cheek, nose, and the entire face is considered as the fifth subregion. For each of these ROIs, a score is computed and the ROI with the best score is used to determine the heart rate of the subject.

First, POS projection is applied to each subregion to extract the rPPG signal followed by FFT and bandpass filtering to obtain a power spectrum density for each subregion. Next, a scoring function is designed for the determination of the best region of interest. This function is dependant on two parameters, (1) Maximum magnitude ($Mmax$) and (2) Row wise sum of the correlation matrix of 5 rPPG signals ($Csum$). The sum of $Mmax$ and $Csum$ yields the final score which is used to determine the best region of interest. Since this method does not take a deep learning approach, the method can be used directly on the test set.

5.5. Results

Various approaches are evaluated for cMAE, cRMSE, cR as shown in the Table 4. It is interesting to note that the self-supervised formulation used by [30] is on top of the leaderboard demonstrating good performance on continuous evaluation metrics. However, their approach also involves carefully handcrafted tricks. The method presented in [9] used additional datasets for training and obtained cMAE of 9.37. All participants chose to participate only in the HR subchallenge. In summary, the results obtained by different methods indicate that continuous prediction is still a challenging task and there is scope for further improvements.

6. Conclusion

In this paper, we have presented the first Vision-for-Vitals challenge for benchmarking the performance of different rPPG methods on a newly curated large-scale dataset called *V4V dataset*. The dataset contains desirable attributes necessary for benchmarking various approaches including challenging elements such as spontaneous behavior and varied HR/RR signals. Further, in order to benchmark different methods effectively, we evaluate various approaches using granular frame-level error metrics rather than segment-level error metrics employed by previous methods. The results show that there is room for further improvement in methods and evaluation protocols used for non-contact video-based human physiological estimation.

Acknowledgement

This research was supported in part by the Bill & Melinda Gates Foundation (BMGF) under grant INV-002154, the NSF under grant CNS-1629898, and the Center of Imaging, Acoustics, and Perception Science (CIAPS) of the Research Foundation of Binghamton University.

References

[1] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124, 2019.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 2020.

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

[4] Ananyananda Dasari, Sakthi Kumar Arul Prakash, László A. Jeni, and Conrad Tucker. Evaluation of biases in remote photoplethysmography methods. *NPJ Digital Medicene*, 2021.

[5] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10), 2013.

[6] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2014.

[7] John Gideon and Simon Stent. Remote photoplethysmography from un- labelled video. In *Proceedings of the IEEE international conference on computer vision*, 2021.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[9] Brian Hill, Xin Liu, and Daniel McDuff. Beat-to-beat cardiac pulse rate measurement from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

[10] Rik Janssen, Wenjin Wang, Andreia Moço, and Gerard De Haan. Video-based respiration monitoring with automatic region of interest detection. *Physiological measurement*, 2015.

[11] Benjamin Kossack. Automatic region-based heart rate measurement using remote photoplethysmography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

[12] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *European Conference on Computer Vision*, 2020.

[13] Eugene Lee, Annie Ho, Yi-Ting Wang, Cheng-Han Huang, and Chen-Yi Lee. Cross-domain adaptation for biometric identification using photoplethysmogram. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[14] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *2018 13th IEEE International*

Conference on Automatic Face & Gesture Recognition (FG 2018), 2018.

[15] Xiaobai Li, Hu Han, Hao Lu, Xuesong Niu, Zitong Yu, Antitza Dantcheva, Guoying Zhao, and Shiguang Shan. The 1st challenge on remote physiological signal sensing (repss). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[16] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[17] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020.

[18] Xin Liu, Ziheng Jiang, Josh Fromm, Xuhai Xu, Shwetak Patel, and Daniel McDuff. *MetaPhys: Few-Shot Adaptation for Non-Contact Physiological Measurement*. Association for Computing Machinery, 2021.

[19] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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

[21] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018.

[22] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision*. Springer, 2018.

[23] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *European Conference on Computer Vision*, 2020.

[24] Ewa Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising remote vitals measurements using inverse attention, 2020.

[25] Yassine Ouzar, Djamaleddine Djeldjli, and Frédéric Bousefsaf. Lcoms lab's approach to the vision for vitals (v4v) challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

[26] Tania Pereira, Nate Tran, Kais Gadhoumi, Michele M. Pelter, Duc H. Do, Randall J. Lee, Rene Colorado, Karl Meisel, and Xiao Hu. Photoplethysmography based atrial fibrillation detection: a review. *NPJ Digital Medicene*, 2020.

[27] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10), 2010.

[28] Sakthi Kumar Arul Prakash and Conrad S Tucker. Bounded kalman filter method for motion-robust, non-contact heart rate estimation. *Biomedical optics express*, 9(2), 2018.

[29] Mohammad Sabokrou, Masoud Pourreza, Xiaobai Li, Mahmood Fathy, and Guoying Zhao. Deep-hr: Fast heart rate estimation from face video under realistic conditions. *Expert Systems with Applications*, 2021.

[30] Simon Stent and John Gideon. Estimating heart rate from unlabelled video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

[31] Lionel Tarassenko, Mauricio Villarroel, Alessandro Guazzi, João Jorge, DA Clifton, and Chris Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35(5), 2014.

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

[33] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7), 2016.

[34] Yan Wu, Ruolei Gu, Qiwei Yang, and Yue-jia Luo. How do amusement, anger and fear influence heart rate and heart rate variability? *Frontiers in Neuroscience*, 13, 2019.

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

[36] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, 2017.

[37] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.