HEART RATE ESTIMATION IN INTENSE EXERCISE VIDEOS

Y. Napolean[1], A. Marwade[1]*, N. Tomen[1], P. Alkemade[2], T. Eijsvogels[3], J.C. van Gemert[1]*

TU Delft[1], VU Amsterdam[2], Radboud UMC[3]

ABSTRACT

Estimating heart rate from video allows non-contact health monitoring with applications in patient care, human interaction, and sports. Existing work can robustly measure heart rate under some degree of motion by face tracking. However, this is not always possible in unconstrained settings, as the face might be occluded or even outside the camera. Here, we present IntensePhysio: a challenging video heart rate estimation dataset with realistic face occlusions, severe subject motion, and ample heart rate variation. To ensure heart rate variation in a realistic setting we record each subject for around 1-2 hours. The subject is exercising (at a moderate to high intensity) on a cycling ergometer with an attached video camera and is given no instructions regarding positioning or movement. We have 11 subjects, and approximately 20 total hours of video. We show that the existing remote photo-plethysmography methods have difficulty in estimating heart rate in this setting. In addition, we present IBIS-CNN, a new baseline using spatio-temporal superpixels, which improves on existing models by eliminating the need for a visible face/face tracking. We will make the code and data publically available soon.[1]

***Index Terms*—** Heart rate estimation, challenging new dataset, spatio-temporal superpixels.

1. INTRODUCTION

Remote photo-plethysmography (rPPG) [1], allows non-contact heart rate estimation. This facilitates applications where contact sensors are difficult such as infant health monitoring, or athlete monitoring in large scale events like marathons, where access to individual sensors is unavailable. Initial methods of rPPG estimation [2, 3] needed the subject to sit motionless in front of a camera. Other approaches [4, 5, 6, 7] typically used face detectors and Fourier analysis based techniques. Some methods additionally use skin tracking/segmentation to estimate heart rate [8, 9]. Other approaches cast heart rate estimation from video as a blind source separation problem [10, 11]. Most existing rPPG methods including recent deep learning based approaches [12, 13] typically operate within a 'constrained' setting: minor pose and camera angle variations, subjects are cooperative, are not occluded and do not perform high-speed movements. Enforcing such constraints is not possible in real-life situations. For instance in a sports scenario, it is hard to control lighting conditions and rapid pose changes.

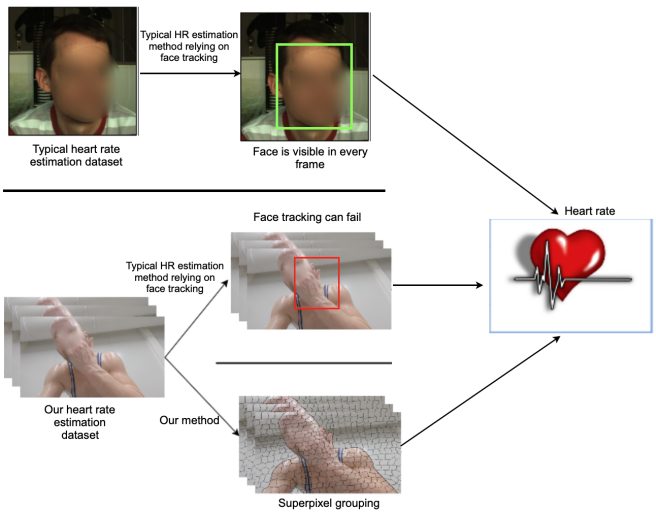

Fig. 1. Methods tracking skin pixels and their color change for heart rate estimation can perform poorly when the subject's face is not visible. Here we introduce the IntensePhysio dataset, which includes high heart rate variability and is recorded in a relatively 'unconstrained' setting, making it challenging for existing methods. We propose the IBIS-CNN method which relies on spatio-temporal grouping of pixels instead of face tracking, which improves on baseline models.

A recent analysis [14] shows that large, high speed motion, tracking based methods can fail. There are methods that have been proposed to handle partial facial occlusion [15][16], but these methods rely on extracting information from other parts of the face that are visible and do not address facial occlusion (as depicted in Fig. 1). To develop a method for heart rate estimation from video in a practical setting (like sports, for example the last image from the left in the second row of Fig. 2), this must be addressed.

Early datasets, MAHNOB [17] and COHFACE [18] feature almost no subject motion. The PURE dataset features minimal controlled head motion [19]. An important step towards more realistic scenarios is the seminal ECG-Fitness dataset [20] where subjects perform fitness-related motions

* Equal contribution

[1] https://github.com/ynapolean/IBIS-CNN

Fig. 2. Samples from our IntensePhysio rPPG dataset. Note the major pose changes, specular reflections on the background and on the subject's skin and instances where the face region is occluded or not fully in frame.

under varying lighting conditions, demonstrating that previous visual heart rate estimation methods performed poorly in this scenario. This ECG-Fitness dataset, however, assumes that faces are still strictly visible, allowing for solutions limited to face tracking.

In this work, we introduce IntensePhysio: a new, publicly available, challenging dataset with ample variance in the heart rate, fast motions, severe appearance changes. It presents a completely unconstrained setting wherein subjects are allowed exacerbated motion, causing their face to be sometimes occluded, or they can go out of frame and exhibit greatly varying face angles. In addition to the new dataset, we analyze existing rPPG approaches and provide an intentionally simple baseline method called IBIS-CNN that replaces face tracking with spatio-temporal superpixels. Bobbia[21] introduced an algorithm called Iterative Boundaries implicit Identification for superpixel Segmentation (IBIS) [21] for real-time computation of temporal superpixels for rPPG. We propose IBIS-CNN as a novel application of IBIS for heart rate estimation in a deep learning setting. IBIS-CNN performs on-par to existing methods on existing datasets, but significantly outperforms existing work on more realistic scenarios as exemplified in our new IntensePhysio dataset. The contributions of this paper are: (i) The new, public, IntensePhysio dataset with facial occlusions, and high speed motion (ii) A simple baseline method using superpixels which does not rely on the face to extract heart rate. (iii) We demonstrate the difficulty of the dataset for existing rPPG work and (iv) show that our simple baseline performs equally well on existing datasets, yet significantly better on our more realistic dataset.

2. METHOD

2.1. The IntensePhysio dataset

We collected video recordings and heart rate as the ground truth. The videos are recorded with a Go-Pro Hero 7 black camera at a resolution of $1,920 \times 1,080$ pixels at ~ 60 frames per second. The ground truth heart rate (in beats per minute) is measured every second, using the Polar Vantage-M heart rate sensor (chest strap).

For illumination we use LED lights (47W), natural light is also allowed to enter through a small window. Recordings occur during different times of the day, resulting in variation of lighting conditions. There are a total of 11 subjects (3 female and 8 male) and all participants are working out on an ergometer and are given no instructions except to do a workout. This dataset features the largest heart rate range and variation: from 51 bpm up to 186 bpm with a standard deviation of 25 bpm. IntensePhysio dataset statistics are presented in Table 1. Subjects exercised at a moderate to high intensity. They did not receive any camera-related instructions, such as to face the camera or be in the frame. This means that the subject's motion is 'unconstrained' and consequently the subject's face could be occluded or might not be visible at all, as illustrated in Fig. 2.

Dataset Statistics	
Mean heart rate (ground truth)	129
Std. deviation heart rate (ground truth)	25
Max. heart rate (ground truth)	186
Min. heart rate (ground truth)	51
No. of subjects	11
No. of videos	15
Avg. video runtime	1hr. 12min.
Frames per second	59.94

Table 1. Statistics of IntensePhysio. The dataset features a wide range of heart rates, from resting to intense workout. Heart rate values are in beats per minute.

2.2. Simple rPPG baseline: IBIS-CNN

To accompany the IntensePhysio dataset, we implement a simple baseline method for rPPG estimation using temporal superpixels and a convolutional neural network. The approach is shown in Fig. 3.

The model is trained to predict the scalar valued instantaneous heart rate in beats per minute (bpm) for an input video. We break an input video sequence into non-overlapping clips of one minute each. For each clip we generate K temporal superpixels using IBIS [21] where K refers to a user-defined value for the number of superpixels, which we set to 300.

The input images are converted to the $CIE\ l^*a^*b^*$ color space. Then pixels are grouped iteratively based on their closest 'seeds', which are the initial values for superpixel centers. Pixel grouping is done according to both the chromatic similarity D_{lab} and spatial proximity $D_{spatial}$ of the pixel with the associated seed. The metric used to quantify this distance, D_{total} is given as

$$
\begin{aligned}
D_{total} &= D_{lab} + \theta * D_{spatial}, \text{ where} \\
D_{lab} &= \left\| (l,a,b)_{i^{th}pixel} - (l,a,b)_{k^{th}seed} \right\| \\
D_{spatial} &= \left\| (x,y)_{i^{th}pixel} - (x,y)_{k^{th}seed} \right\|
\end{aligned}
$$

for the i^{th} pixel and the k^{th} seed. $\|.\|$ is a Euclidean distance and θ is defined by $\theta = 1/c^2$ where c is a user-defined *compacity* parameter. The seeds value is propagated temporally,

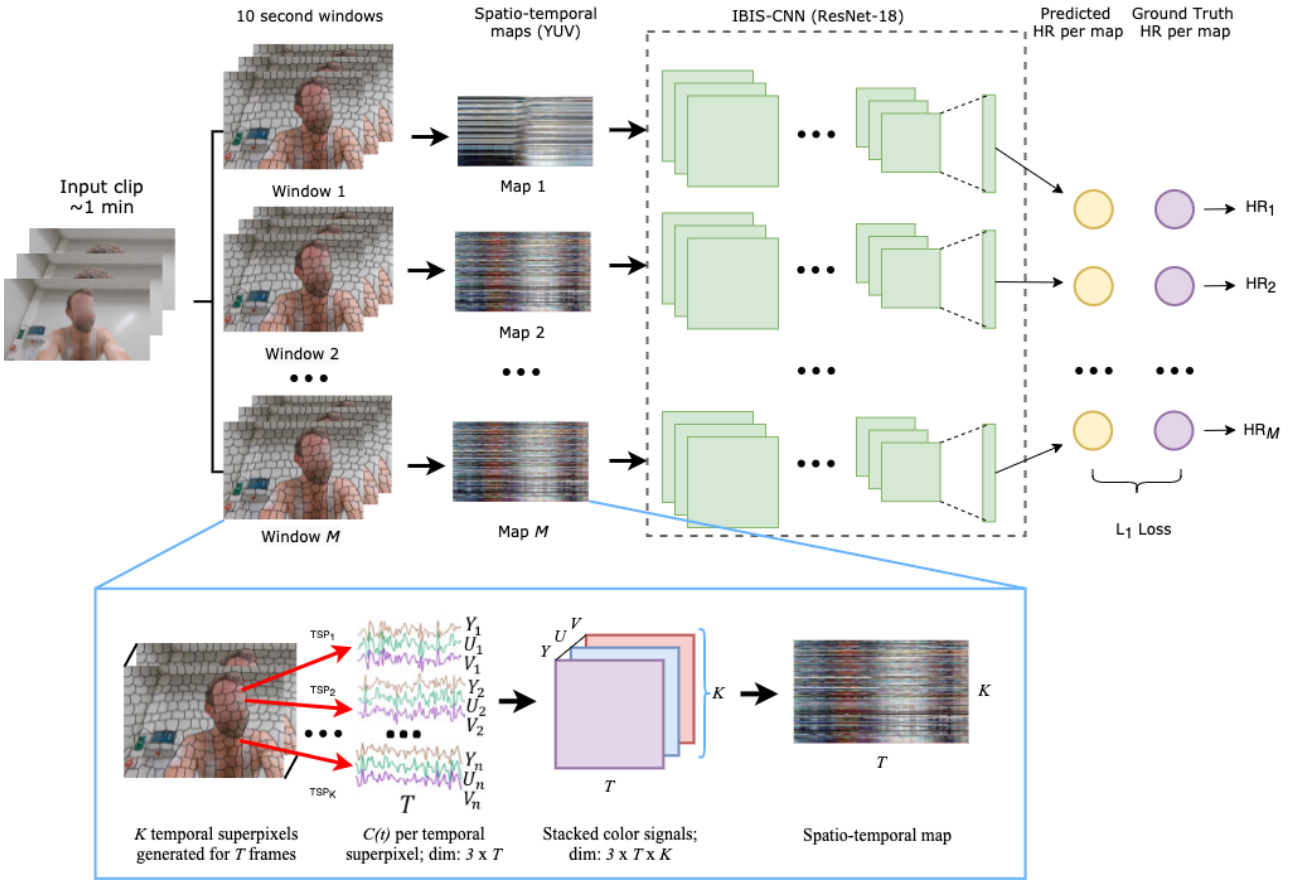

Fig. 3. An overview of the IBIS-CNN model. A clip is split into M windows of 10 seconds each. For each frame, K superpixels are extracted and per superpixel, the average YUV color is stacked over time, where T is the number of frames and generate a spatio-temporal map. The predicted heart rate per map from the CNN are averaged to estimate the HR per video.

to generate an output of average RGB color value per temporal superpixel per frame.

Let $C_p(t)$ denote the 3-dimensional per-channel average YUV color signal of a superpixel p in frame t. The 3 average YUV values and grouping all K superpixels gives a 2d matrix of size $K \times 3$. Stacking these over a temporal window of T frames, we obtain a $3 \times K \times T$ tensor, which can be seen as a $K \times T$ color image which we call a *spatial-temporal map*. This map is 10 seconds long, while each clip lasts around 1 minute. We slide a temporal window over the clip to obtain M maps per clip using a window of 10 seconds (stride = 0.5s).

3. EXPERIMENTS

We compare against two state of the art rPPG methods in HR-CNN [20] and RhythmNet [12]. Both these methods involve pre-processing the input video frames using face-detection (with an additional alignment step in RhythmNet), cropping and re-sizing using specific toolboxes. We use train-test splits as specified by the authors. For our dataset, we make use of a train-test split featuring 8 subjects in the train set and 3 in the test set. The number of superpixels is $K = 300$.

Due to lack of publicly available code for RhythmNet [12], we re-implemented it as per the authors description in their paper. The HR-CNN [20] model and code is publicly available. We evaluate with mean absolute error (MAE) and root mean square error (RMSE) calculated over the number of spatio-temporal maps. Results in Table 2 show that our method achieves similar performance to others on PURE and ECG-Fitness, validating our baseline. Yet, our baseline significantly outperforms others on our new IntensePhysio dataset.

All models show poor performance on our dataset. To es-

	Model	PURE	ECG-Fitness	IntensePhysio
RMSE	HR-CNN	**11.00**	19.15	25.27
	RhythmNet	19.67	20.47	32.67
	IBIS-CNN (ours)	11.99	**17.03**	**22.01**
MAE	HR-CNN	**8.72**	14.48	22.19
	RhythmNet	17.46	16.82	28.36
	IBIS-CNN (ours)	9.39	**13.75**	**16.53**

Table 2. RMSE and MAE results for the proposed IBIS-CNN and two baseline methods in HR-CNN and RhythmNet on the test sets of PURE, ECG-fitness and IntensePhysio (ours).

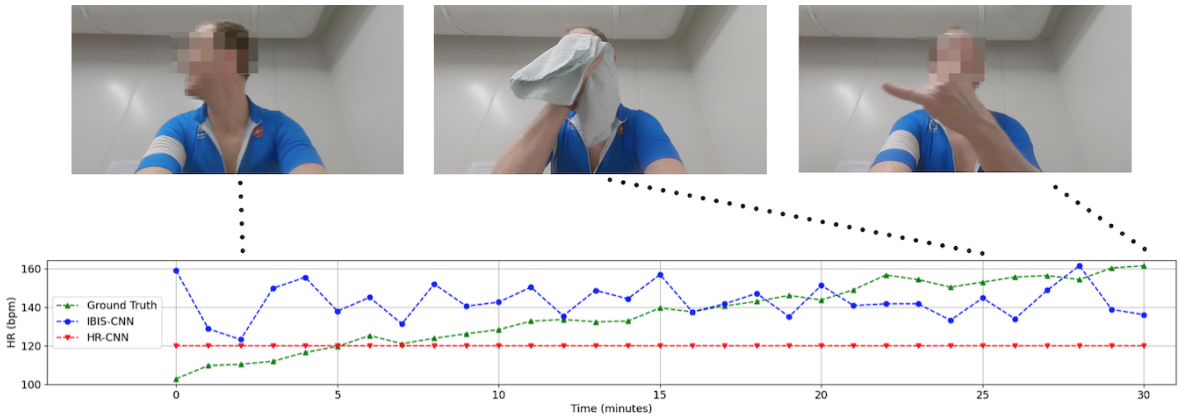

Fig. 4. Predictions of the IBIS-CNN model compared with he ground truth and HR-CNN on subjects from the test set. The HR-CNN model outputs a constant value, while our predictions exhibit learning behavior, following the ground truth.

tablish the challenging nature of the IntensePhysio dataset for methods that rely on face tracking, we run an off the shelf face detector (Dlib [22]) and find that *on average we are able to detect faces in ∼28% of frames per video in our dataset.On the other hand we observed a successful face detection in ∼74% frames per video in ECG-Fitness*. Since the face region (particularly the cheeks and forehead) contain the most information on heart rate [13][20], this leads to a poor performance when models rely on face detection and tracking.

The predictions of our IBIS-CNN model and HR-CNN along with the ground truth are shown in Fig. 4. A few sample frames from the video are also presented with the dotted lines connecting them to the minute of extraction. In these frames, there is a lot of specular reflection on his skin (due to sweating), the subject's face is not visible. This happens particularly around the 30^{th} minute, where the error is larger. The HR-CNN model outputs a constant value, indicating that it might not be learning relevant rPPG features. The IBIS-CNN model is able to predict heart rate reasonably well despite occlusions, specular reflections etc. However, we also note that the predictions from IBIS-CNN are not always correlated with the ground truth, especially in the second plot of 4. This could be because of rapid motion and would be interesting future direction of research. IBIS-CNN does not solve all the challenges posed by the IntensePhysio dataset but is an alternative baseline with a new approach.

Comparison of input representations. We compare two existing methods for temporal superpixel generation - IBIS [21] and TS-PPM [23]. We find that the performance of TS-PPM is lower than that of IBIS-CNN (as seen in Table 3 possibly because, in the IBIS method pixel membership is constrained to not vary beyond a certain threshold to maintain temporally coherent RGB traces.

As a pre-processing step, the IBIS superpixel generation is more computationally efficient than the TS-PPM. To generate the superpixel results, IBIS processes 5.33 frames/sec on average whereas TS-PPM was averages 0.56 frames/sec. Thus, the IBIS method is more suited to the rPPG estimation task. These results were obtained using PURE dataset.

	Model	Validation error for PURE
RMSE	TS-PPM [23]	13.97
	IBIS-CNN (ours)	**11.99**
MAE	TS-PPM [23]	10.81
	IBIS-CNN (ours)	**9.39**

Table 3. Comparison between temporal superpixel methods for heart rate estimation.

4. CONCLUSIONS

We present IntensePhysio, a challenging new dataset for heart rate estimation. The dataset features large subject motion with frequent face occlusions and cases of facial region absent from the frame entirely. Through a comparative study, we observe a considerable degradation in the performance of the existing state of the art methods on this new dataset, especially methods relying on face detection and tracking. This highlights IntensePhysio as a challenging dataset for heart rate estimation, indicating that occlusion and non-visible facial regions are key factors for this performance degradation. Hence, we propose IBIS-CNN as a new baseline method for heart rate estimation (using temporal superpixels) which significantly outperforms state of the art methods on our challenging new dataset in addition to the existing ones. However, the IBIS-CNN baseline predictions are not strongly correlated with the actual ground truth always.This shows that there are further problems posed by our dataset that require addressing. Also, there is a need for a thorough hyperparameter tuning while generating these temporal superpixels using IBIS, specific to our task. So, it would be of significance to investigate the possibility of developing IBIS-CNN as an end to end learnable pipeline.

5. REFERENCES

[1] Qi Zhan, Wenjin Wang, and Gerard de Haan, "Analysis of cnn-based remote-ppg to understand limitations and

sensitivities," *Biomedical optics express*, vol. 11, no. 3, pp. 1268–1283, 2020.

[2] Sungjun Kwon, Hyunseok Kim, and Kwang Suk Park, "Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone," in *2012 Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 2174–2177.

[3] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson, "Remote plethysmographic imaging using ambient light.," *Optics express*, vol. 16, no. 26, pp. 21434–21445, 2008.

[4] E Jonathan and Martin Leahy, "Investigating a smartphone imaging unit for photoplethysmography," *Physiological measurement*, vol. 31, no. 11, pp. N79, 2010.

[5] Amogh Gudi, Marian Bittner, Roelof Lochmans, and Jan Van Gemert, "Efficient real-time camera based estimation of heart rate and its variability," in *ICCV Workshops*, 2019.

[6] Amogh Gudi, Marian Bittner, and Jan van Gemert, "Real-time webcam heart-rate and variability estimation with clean ground truth for evaluation," *Applied Sciences*, vol. 10, no. 23, 2020.

[7] Huishi Zhu, Yuejin Zhao, and Liquan Dong, "Non-contact detection of cardiac rate based on visible light imaging device," in *Optics and Photonics for Information Processing VI*. International Society for Optics and Photonics, 2012, vol. 8498, p. 849806.

[8] Chuanxiang Tang, Jiwu Lu, and Jie Liu, "Non-contact heart rate monitoring by combining convolutional neural network skin detection and remote photoplethysmography via a low-cost camera," in *CVPR Workshops*, 2018.

[9] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao, "Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement," in *ICCV*, 2019.

[10] 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*, vol. 18, no. 10, pp. 10762–10774, 2010.

[11] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jedrzej Nowak, "Measuring pulse rate with a webcam — a non-contact method for evaluating cardiac activity," in *FedCSIS*, 2011, pp. 405–410.

[12] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen, "VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video," in *ACCV*, 2018.

[13] Weixuan Chen and Daniel McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *ECCV*.

[14] Changchen Zhao, Peiyi Mei, Shoushuai Xu, Yongqiang Li, and Yuanjing Feng, "Performance evaluation of visual object detection and tracking algorithms used in remote photoplethysmography," in *ICCV Workshops*, 2019.

[15] Yuichiro Maki, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi, "Remote heart rate estimation based on 3d facial landmarks," in *EMBC*, 2020.

[16] Kun Zheng, Kangyi Ci, Jinling Cui, Jiangping Kong, and Jing Zhou, "Non-contact heart rate detection when face information is missing during online learning," *Sensors*, vol. 20, no. 24, 2020.

[17] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42–55, 2011.

[18] Guillaume Heusch, André Anjos, and Sébastien Marcel, "A reproducible study on remote heart rate measurement," *arXiv preprint arXiv:1709.00962*, 2017.

[19] Ronny Stricker, Steffen Müller, and Horst-Michael Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *Symposium on Robot and Human Interactive Communication*, 2014.

[20] Radim Špetlík, Vojtech Franc, and Jirí Matas, "Visual heart rate estimation with convolutional neural network," in *BMVC*, 2018.

[21] Serge Bobbia, Duncan Luguern, Yannick Benezeth, Keisuke Nakamura, Randy Gomez, and Julien Dubois, "Real-time temporal superpixels for unsupervised remote photoplethysmography," in *CVPR Workshop*, 2018.

[22] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[23] Se-Ho Lee, Won-Dong Jang, and Chang-Su Kim, "Temporal superpixels based on proximity-weighted patch matching," in *ICCV*, 2017.