# Visual Heart Rate Estimation with Convolutional Neural Network

Radim Špetlík
http://cmp.felk.cvut.cz/~spetlrad/
Vojtěch Franc
http://cmp.felk.cvut.cz/~xfrancv/
Jan Čech
http://cmp.felk.cvut.cz/~cechj/
Jiří Matas
http://cmp.felk.cvut.cz/~matas/

Department of Cybernetics
Center for Machine Perception (CMP)
Czech Technical University in Prague
Karlovo nám. 13
Prague 121 35
Czech Republic

**Abstract**

We propose a novel two-step convolutional neural network to estimate a heart rate from a sequence of facial images. The network is trained end-to-end by alternating optimization and validated on three publicly available datasets yielding state-of-the-art results against three baseline methods. The network performs better by a 40% margin to the state-of-the-art method on a newly collected dataset.

A challenging dataset of 204 fitness-themed videos is introduced. The dataset is designed to test the robustness of heart rate estimation methods to illumination changes and subject's motion. 17 subjects perform 4 activities (talking, rowing, exercising on a stationary bike and an elliptical trainer) in 3 lighting setups. Each activity is captured by two RGB web-cameras, one is placed on a tripod, the other is attached to the fitness machine which vibrates significantly. Subject's age ranges from 20 to 53 years, the mean heart rate is ≈ 110, the standard deviation ≈ 25.

## 1 Introduction

Heart rate (HR) is a basic parameter of cardiovascular activity [1]. HR value is used broadly – from monitoring of exercise activities to prediction of acute coronary events. Contact HR measurement is performed as simply as by palpating the pulse or by sophisticated built-on-purpose devices, *e.g.* pulse oximeters or electrocardiographs. The more expensive the device, the more precise and reliable the measurement.

Visual HR estimation, *i.e.* HR estimation from a video sequence or direct feed from a camera, recently received a lot of attention [5, 16]. Compared to the contact methods, visual HR methods deliver precise readings using cheap measuring devices (such as web-cameras). By not requiring a physical contact, the subject's comfort is improved. This is particularly important for patients with acute skin conditions. Also, the recorded material need not to be primarily designed for HR estimation enabling for *ex post* analysis.

Accuracy of visual HR estimation depends on recording conditions. Published visual HR methods are highly sensitive to motion and light interference, thus requiring subject's

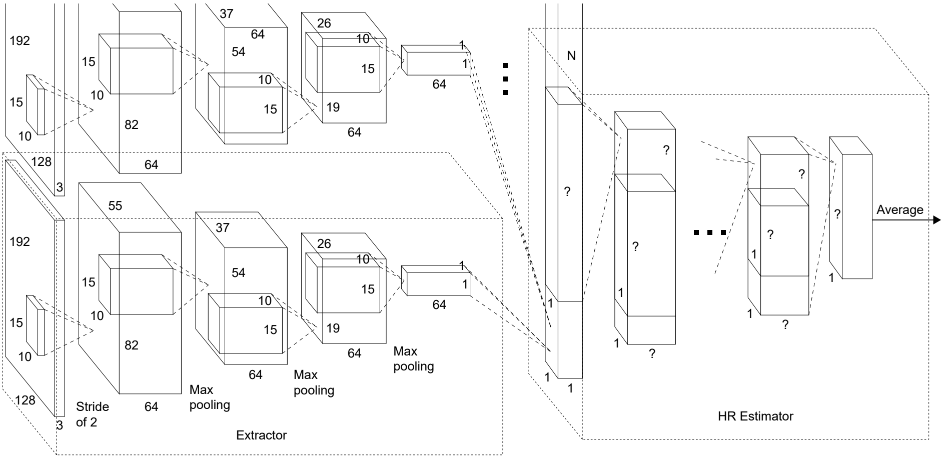

Figure 1: An illustration of the architecture of the "heart rate convolutional neural network". The *Extractor* component of the network is run over a temporal image sequence of faces. The signal is fed to *HR estimator* component and a heart rate is predicted. The architecture of the estimator differs between databases.

cooperation. Also, an engineered, complicated approach-specific signal processing pipeline is used that consists of several consecutive steps (*e.g.* [8] or [22]).

In this paper, we propose to estimate a heart rate (HR) remotely with a two-step convolutional neural network (CNN). The network is trained end-to-end by alternating optimization and is robust to illumination changes and subject's motion. The input of the network is only a roughly aligned image of the face.

The datasets used for visual HR evaluation reflect the hidden assumptions of the methods – subjects don't move and are lighten by a daylight or a professional studio light source. We therefore collected a new ECG-Fitness dataset with interfering lighting setup and subjects performing rapid movements in all three axes.

The contributions of this work are the following: (i) we propose a novel method of estimating heart rate (HR) from a video sequence which replaces a common engineered signal processing pipeline with a single convolutional neural network (HR-CNN) trained end-to-end, (ii) new challenging publicly available database is presented and used to evaluate robustness of visual HR methods to severe motion blur and light interference, (iii) the method is evaluated on three publicly available datasets against three state-of-the-art methods with standardized protocols achieving superior prediction accuracies.

## 2 Related work

We are interested in HR estimation performed remotely by monitoring peripheral circulation of blood, *i.e.* in non-contact reflective photoplethysmographic (NrPPG) HR estimation.

To the best of our knowledge, there are three recent works reviewing field of NrPPG HR estimation. Work of Hassan et al. [5] from 2017 provides the most recent comparison. The research based on both illumination variance and ballistocardiographic motions is reviewed. Paper by Sun and Thakor [16], from 2015, provides a survey of a large body of literature focused on contact and NrPPG methods, there referred to as imaging PPG. The differences between the discussed methods are shown on different choices taken during the procedure of obtaining the NrPPG measurements. The earliest work of Liu et al. [9] tracks the rapid development of the NrPPG approaches between years 2007 and 2012. The introduction of cheap and relatively precise measuring devices, i.e. web cameras and alike, is identified as the main reason behind the expansion of the NrPPG field.

There are over 60 studies on heart rate estimation using NrPPG. Most of them are performed on private datasets with *ad hoc* evaluation procedures. Only one of them [8] is validated on a publicly available dataset.

Recently, Heusch et al. [6] reimplemented two baseline HR estimation approaches [4, 21] and a method of Li et al. [8]. Also, several experimental protocols were introduced enabling a comparison of the NrPPG methods. Authors tested the three methods on the MAHNOB HCI-Tagging database [14]. Heusch et al. provided all reimplemented codes and also collected a publicly available database COHFACE[1]. Since the three reported studies are the only ones tested on public datasets, we will discuss only these three.

An approach of Haan et al. [4] (referred to as CHROM) is based on combining color difference, *i.e*. chrominance, signals. First, skin-color pixels are found in each frame of input sequence. Then, an average color of skin pixels is computed in each frame and projected on a proposed chrominance subspace. The projected signals are bandpass filtered separately in the XY chrominance colorspace and projected into a one-dimensional signal. The algorithm is shown to outperform blind source separation methods on a private dataset of 117 static subjects.

Li et al. [8] (referred to as LiCVPR) finds bottom part of a face in the first frame of a sequence and tracks it with the Lucas-Kanade tracker [10]. An average green value of the region of interest is computed in each frame and corrected for illumination changes. Background is segmented and its average green value is used to mitigate illumination variations with a Normalized Least Mean Squares filter. Then, subject's non-rigid motions are eliminated by discarding the motion-contaminated segments of the signal. Finally, temporal filters are applied and Welch's power spectral density estimation method is used to estimate the HR frequency. Experiments are performed on two datasets, a private one and the public MAHNOB HCI-Tagging database.

The last considered approach is Spatial Subspace Rotation (referred to as 2SR) by Wang et al. [23]. First, skin pixels are found in each frame. Then, subspace of skin pixels in the RGB space is built for each frame in the spatial domain. The rotation angle between the spatial subspaces is computed and analyzed between consecutive frames. Authors claim that no bandpass filtering is required to obtain the NrPPG signal. The method is validated on a private dataset consisting of 54 videos. Performance of algorithm under various conditions such as skin tone, subject's motion and recovery after a physical exercise is examined resulting in Pearson's correlation coefficient of 0.94.

# 3 Method

We propose a convolutional neural network to estimate a heart rate (HR) of a subject in a video sequence (denoted as HR-CNN). The input of the network is sequence of images of a subject's face in time. The output is a single scalar – predicted HR.

[1] https://idiap.ch/dataset/cohface

The network consists of two components – *Extractor* and *HR estimator* (the architecture is shown in Fig. 1). The Extractor takes an image and produces a single number. By running the extractor over a sequence of images, a sequence of scalar outputs, a NrPPG signal, is produced. The NrPPG signal is fed to the HR estimator which outputs the HR. The two components are trained separately. First, given the true heart rate, the extractor is trained to maximize the signal-to-noise ratio (SNR). Then, the estimator is trained to minimize the mean absolute error (MAE) of the estimated and the true HR.

Let $\mathcal{T} = \{(\mathbf{x}_1^j, \ldots, \mathbf{x}_N^j, f^j) \in \mathcal{X}^N \times \mathcal{F} \mid j = 1, \ldots, l\}$ be the training set that contains $l$ sequences of $N$ facial RGB image frames $\mathbf{x} \in \mathcal{X}$ and their corresponding HR labels $f \in \mathcal{F}$. Symbol $\mathcal{X}$ denotes a set of all input images and $\mathcal{F}$ is a set of all sequence labels, *i.e.* the true HR frequencies measured in hertz.

**Extractor** Let $h(\mathbf{x}_n; \mathbf{\Phi})$ be the output of the *Extractor* CNN for the $n$-th image and $\mathbf{\Phi}$ a concatenation of all convolutional filter parameters. The quality of the extracted signal is measured by SNR using power spectral density (PSD). Given frequency $\hat{f}$

$$\mathrm{PSD}(\hat{f}, \mathbf{X}; \mathbf{\Phi}) = \left( \sum_{n=0}^{N-1} h(\mathbf{x}_n; \mathbf{\Phi}) \cdot \cos\left(2\pi \hat{f} \frac{n}{f_s}\right) \right)^2 + \left( \sum_{n=0}^{N-1} h(\mathbf{x}_n; \mathbf{\Phi}) \cdot \sin\left(2\pi \hat{f} \frac{n}{f_s}\right) \right)^2 \tag{1}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is a sequence of $N$ facial images, and $f_s$ is the sampling frequency.

Intuitively, given the true HR, the amplitude of its frequency should be high while amplitudes of background frequencies low. To measure the quality of the extractor, a SNR introduced in [4]

$$\mathrm{SNR}(f, \mathbf{X}; \mathbf{\Phi}) = 10 \cdot \log_{10} \left( \sum_{\hat{f} \in \mathcal{F}^+} \mathrm{PSD}(\hat{f}, \mathbf{X}; \mathbf{\Phi}) \Big/ \sum_{\hat{f} \in \mathcal{F} \setminus \mathcal{F}^+} \mathrm{PSD}(\hat{f}, \mathbf{X}; \mathbf{\Phi}) \right) \tag{2}$$

is used where $f$ is the true HR, $\mathcal{F}^+ = (f - \Delta, f + \Delta)$, and the tolerance interval $\Delta$ accounts for the true HR uncertainty, *e.g.* due to HR non-stationarity within the sequence. The nominator captures the strength of the true HR signal frequency. The denominator represents the energy of the background noise, the tolerance interval excluding.

The parameter $\mathbf{\Phi}$ is estimated by minimizing the loss function

$$\ell(\mathcal{T}; \mathbf{\Phi}) = -\frac{1}{l} \sum_{j=1}^{l} \mathrm{SNR}(f^j, \mathbf{X}^j; \mathbf{\Phi}). \tag{3}$$

**HR Estimator** The HR estimator is another CNN taking 1D signal – output of the Extractor – and producing the HR. The training minimizes average $\mathrm{L}_1$ loss between the predicted and true HR $f^j$

$$\ell(\mathcal{T}; \boldsymbol{\theta}) = \frac{1}{l} \sum_{j=1}^{l} \left| g\left( \left[ h(\mathbf{x}_1; \mathbf{\Phi}), \cdots, h(\mathbf{x}_N; \mathbf{\Phi}) \right]; \boldsymbol{\theta} \right) - f^j \right| \tag{4}$$

where $\boldsymbol{\theta}$ is a concatenation of all convolutional filter parameters of the HR estimator CNN, and $g\left( \left[ h(\mathbf{x}_1; \mathbf{\Phi}), \cdots, h(\mathbf{x}_N; \mathbf{\Phi}) \right]; \boldsymbol{\theta} \right)$ is the output of the CNN for a sequence of $N$ outputs of the Extractor.

**Discussion** Our first experiments were conducted on a non-challenging database. A simple argument maximum in the PSD of extractor's output $\hat{f} = \arg\max_f \mathrm{PSD}(f, \mathbf{X}, \mathbf{\Phi})$ gave MAE less than 3. However, this simple HR estimation was not robust to corruption of the extracted signal in other datasets, *e.g.* video compression, non-stationarity of a subject's HR and subject's motion. Therefore, we introduced the HR estimator CNN.

**Training** In all experiments, the Extractor was trained on the training set of the PURE database (see Sec. 4.1) and fixed. Data augmentation including random translation, cropping and rotation was applied to each frame of the training sequence. Brightness was randomly adjusted for a whole sequence. The HR estimator was trained for each database separately. In case of the ECG-Fitness dataset, the sequences were split to 10 seconds clips to account for rapid HR changes.

**Implementation details** Both networks use a standard chain of convolution, MaxPool and activation functions. Before the first convolution layer and after every MaxPool layer, a batch normalization was inserted. Exponential Linear Units [2] were used as the activation functions. Dropout was used. Batch normalization was initialized with weights randomly sampled from a Gaussian distribution with $\mu = 0$ and $\sigma = 0.1$, convolution layers were initialized according to the method described in [3]. Both network components were trained using PyTorch library, Adam optimizer was used with learning rate set to 0.0001 in case of the Extractor and to 0.1 in case of the HR estimator. For both training setups, a set of all input facial RGB images $\mathcal{X} = \mathbb{R}^{192\times128}$. Faces were found by a face detector, the bounding boxes were adjusted to the aspect ratio 3 : 2 to cover the whole face, cropped out and resized to 192 × 128 pixels. The set of true HR $\mathcal{F} = \{\frac{40}{60}, \frac{41}{60}, \ldots, \frac{240}{60}\}$ in case of extractor and $\mathcal{F} = \mathbb{R}^{0+}$ in case of estimator.

# 4 Experiments

An open source Python package `bob.rppg.base`[2] provided by Heusch et al. [6] was used for the computations. The same error metrics reflecting discrepancy between the true and predicted HR were used. In particular, the root mean square error (RMSE) and the Pearson's correlation coefficient, were used as in [6]. In addition, mean absolute error (MAE) was computed. We test the proposed HR-CNN method on four datasets (standard: COHFACE, MAHNOB, PURE, newly collected: ECG-Fitness) against three baseline methods (LiCVPR [8], CHROM [4], 2SR [23]).

## 4.1 Datasets

To the best of our knowledge, there are three publicly available datasets for evaluation of HR estimation methods. In MAHNOB dataset [14], the ground truth is derived from an electrocardiogram. The PURE dataset [15] and COHFACE dataset [6] contain the ground truth from pulse oximeters. Devices performing pulse oximetry differ in both software and hardware implementation. Also, they are prone to inaccuracy due to various conditions (subject's health status, motion, external lighting) [7, 18, 19] and produces errors in the

[2] https://gitlab.idiap.ch/bob/bob.rppg.base

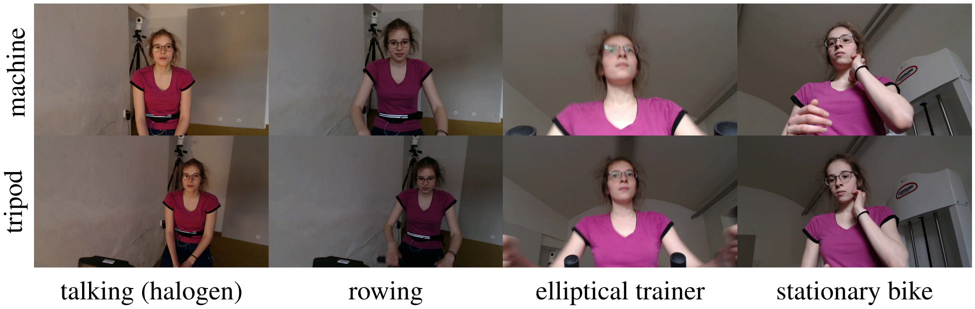

Figure 2: The ECG-Fitness dataset. Images captured by a camera attached to: the currently used fitness machine (top), a tripod placed as close as possible to the first camera (bottom).

ground truth. We collected an ECG-Fitness dataset where the ground-truth HR is given by an electrocardiogram. Bellow, the datasets are described in more detail.

**COHFACE** Heusch et al. [6] collected 160 videos from 40 healthy individuals, each 60 seconds long. Two experimental setups with studio and natural lighting were introduced. Logitech HD Webcam C525 was used to record a video signal and a contact rPPG sensor to capture a blood volume pulse. rPPG sensors are generally less reliable than electrocardiographs. The videos are compressed in MPEG-4 Visual, *i.e.* MPEG-4 Part 2, bit rate $\approx 250$ kb/s, resolution $640 \times 480$ pixels, 20 fps, which gets $\approx 5 \times 10^{-5}$ bits per pixel. In other words, the videos were heavily compressed and in the light of recent findings of McDuff et al. [11] the NrPPG signal is almost certainly corrupted.

**MAHNOB** 3739 videos of 30 young healthy adult participants are available. The corpus contains one color and five monochrome videos for each recording session. The videos were recorded in a controlled studio setup. For full details, please see the database manual[3] [14]. The lengths of the videos vary from 1 to 259 seconds. Subjects in the videos watch emotion-eliciting clips. Every session is accompanied by rich physiological data that include readings from electroencephalograph, electrocardiograph, temperature sensor and respiration belt. The videos are compressed in H.264/MPEG-4 AVC compression, bit rate $\approx 4200$ kb/s, 61 fps, $780 \times 580$ pixels, which gets $\approx 1.5 \times 10^{-4}$ bits per pixel. Again, the videos are heavily compressed.

**PURE** 10 subjects performing 6 different activities (sitting still, talking, four variations of rotating and moving head) were recorded by an industry grade RGB camera by Stricker et al. [15]. In total 60 videos, 1 minute each, were captured with a 30 Hz frame rate along with a contact tPPG signal from a finger clip pulse oximeter. Unlike the two previously described datasets, here the signal from the camera was stored in lossless PNG files. A frame $640 \times 480$ pixels is $\approx$ 390kB, which gets $\approx 10$ bits per pixel. That is $2 \times 10^5$ times more than COHFACE and $\approx 6 \times 10^4$ times more than MAHNOB.

**ECG-Fitness** We collected realistic corpus of subjects performing physical activities on fitness machines. The corpus is available on http://cmp.felk.cvut.cz/∼spetlrad/ecg-fitness/.

[3] https://mahnob-db.eu/hci-tagging/media/uploads/manual.pdf

17 subjects (14 male, 3 female) performing 4 different activities (speaking, rowing, exercising on a stationary bike and on an elliptical trainer, see Fig. 2) were captured by two Logitech C920 web cameras and FLIR thermal camera. The FLIR camera was not used in the current study. One Logitech camera was attached to the currently used fitness machine, the other was positioned as close as possible to the first camera on a tripod. Three lighting setups were used: (i) natural light coming from a nearby window, (ii) 400W halogen light and (iii) 30W led light. The artificial light sources were positioned to bounce off the walls and illuminate the subject indirectly. Two activities (speaking and rowing) were performed twice – once with the halogen lighting resulting in a strong 50 Hz temporal interference, once without. In case of 4 subjects, led light was used during the recording of all activities. In total 204 videos from web cameras, 1 minute each, were recorded with 30 fps, 1920 × 1080 pixels and stored in an uncompressed YUV planar pixel format. The age range of subjects is 20 to 53 years. During the video capture, an electrocardiogram was recorded with two-lead Viatom CheckMe$^{TM}$Pro device with CC$_5$ lead. The ground-truth HR was computed with Python implementation of Pan-Tomkins algorithm [17]. The lowest measured HR is 56, the highest 159. The mean HR is 108.96, standard deviation 23.33 bpm.

The dataset covers the following challenges: (i) large subject's motion (possibly periodic) in all three axis, (ii) rapid motions inducing motion blur, (iii) strong facial expressions, (iv) wearing glasses, (v) non-uniform lighting, (vi) light interference, (vii) atypical non-frontal camera angles.

## 4.2 Evaluation

Face bounding boxes in case of PURE and ECG-Fitness datasets were detected by a commercial implementation of *WaldBoost* [13] based detector[4]. Bounding boxes for the MAHNOB and COHFACE datasets were provided by `bob.rppg.base` package.

**Experimental protocols** We adopt the training and test split for the COHFACE and MAHNOB databases, a protocol denoted as "all", that was introduced by Heusch et al. [6] in the *bob.rppg.base* Python package. We define the splits for the PURE and ECG-Fitness database that were not covered by the protocol. In all presented experiments, the parameters of each method were trained on the training set of the particular database. The testing was done on previously unseen data – it was a common practice in the community to tune parameters of the methods directly on the test sets.

### 4.2.1 MAHNOB HCI-Tagging

The training set consists of 2302 sequences with an average length of 1812 frames. The test set contains 1188 sequences with an average length of 1745 frames.

The results are presented in Tab. 1. In case of all three measures, the HR-CNN method clearly dominates over the 2SR, CHROM and LiCVPR methods. This is not true only for the SAMC method [20]. SAMC was not known to the authors at the time of writing. Therefore, the method was not reimplemented and only the results presented in [20] are reported. Interestingly, Li et al. [8] reports Pearson's correlation coefficient of 0.81, but neither we nor Heusch et al. were able to reproduce the results. The reason is probably the unknown parameter setting of the signal extraction pipeline. In the dataset, the most informative area

[4] Eyedea Recognition Ltd. `http://www.eyedea.cz/`.

| | | COHFACE | ECG-Fitness | MAHNOB | PURE | PURE<br>MPEG-4 Visual |
|---|---|---|---|---|---|---|
| RMSE | 2SR | 25.84 | 52.86 | 21.39 | 3.06 | 12.81 |
| | CHROM | 12.45 | 33.47 | 22.36 | 2.50 | 11.36 |
| | LiCVPR | 25.59 | 67.67 | 10.21 | 30.96 | 31.10 |
| | SAMC* | — | — | 6.23 ① | — | — |
| | HR-CNN | 10.78 ① | 19.15 ① | 9.24 | 2.37 ① | 11.00 ① |
| MAE | 2SR | 20.98 | 43.66 | 13.84 | 2.44 | 5.78 ① |
| | CHROM | 7.80 ① | 21.37 | 13.49 | 2.07 | 6.29 |
| | LiCVPR | 19.98 | 63.25 | 7.41 | 28.22 | 28.39 |
| | SAMC* | — | — | — | — | — |
| | HR-CNN | 8.10 | 14.48 ① | 7.25 ① | 1.84 ① | 8.72 |
| Pearson's correlation coefficient | 2SR | −0.32 | 0.06 | 0.14 | 0.98 | 0.43 |
| | CHROM | 0.26 | 0.33 | 0.21 | 0.99 ① | 0.55 |
| | LiCVPR | −0.44 | −0.02 | 0.45 | −0.38 | −0.42 |
| | SAMC* | — | — | 0.83 ① | — | — |
| | HR-CNN | 0.29 ① | 0.50 ① | 0.51 | 0.98 | 0.70 ① |

* SAMC method [20] was not known to the authors at the time of writing. Therefore, only the results presented in [20] are reported.

Table 1: Root-mean-square error, mean average error and Pearson's correlation coefficient on test sets of the databases for three baseline methods and the proposed HR-CNN.

for HR estimation is the lower part of a face. The subjects in the dataset wear an electroencephalographic caps that either cover the forehead completely or force hair in the forehead's direction. Also, the cap's color is very similar to the skin's tone. With these limitations, the selection of a measuring area is given – LiCVPR estimates HR only from the lower part of the face. Also, subjects in the database rarely move. If a subject moves, LiCVPR removes such sub-sequence, *i.e.* sub-sequence containing "non-rigid motion", as not suitable for the estimation.

### 4.2.2 COHFACE

The COHFACE training set contains 24 subjects, the test set 32 subjects.

The database contains the most compressed videos. The results presented in Tab. 1 show that CHROM method yields the best MAE, but in case of the other measures, HR-CNN performs the best. 2SR and LiCVPR perform significantly worse. CHROM and HR-CNN methods use the whole input sequence to reconstruct the NrPPG signal and estimate the HR, while the other aggregate local estimates. The first approach seems to best account for the heavy compressed COHFACE videos.

### 4.2.3 PURE

The PURE training set contains 36 videos of 6 subjects, the test set 24 videos of 4 subjects.

Surprisingly, MAE (see Tab. 1) is less than 3 in case of three methods out of four. Poor results of LiCVPR are probably caused by the fact that unlike in the MAHNOB database, subjects in the PURE database were asked to perform strong head movements in two cases and to talk in one case. Also, a different video compression method was used. We believe

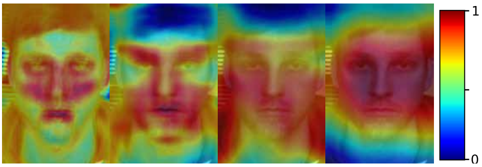

Figure 3: Sequence of Grad-CAM heatmaps of convolutional layers for the Extractor network (from the earliest 1. convolutional layer on left to the latest 4. on right). The heatmaps show that the activations in cheek and lips areas contribute to the output the most.

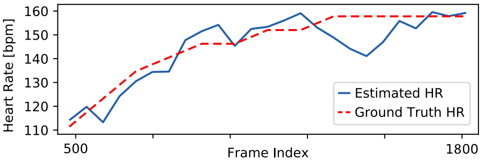

Figure 4: Output example of the HR estimator for a video with significant increase of subject's HR. The estimated (blue solid) and the true HR (red dashed) computed from a 10 second window at a 1 sec. time interval.

that the main reason behind the good prediction accuracy of the algorithms is the fact that the PURE dataset is *not compressed*. To confirm our hypothesis, we decided to perform another experiment.

The PURE dataset was compressed with the same compression method and to the same average bit rate as videos from the COHFACE database. A drop of the accuracy is visible in the results in case of all methods.

### 4.2.4 ECG-Fitness

There is 72 videos of 12 subjects in the training set and 24 videos of 4 subjects in the test set of the ECG-Fitness database. Videos from both cameras (one positioned on a tripod and the other attached to the currently used fitness machine) were used.

The results presented in Tab. 1 show that our method is the most robust one when a strong motion and heavy light interference is present in the videos. Due to the rapid movement of subjects, the face bounding boxes were not found in videos in all frames. In that cases, the last found bounding box was used. Visual inspection of the extracted faces revealed a strong clutter. The clutter and motion blur are the reason why the LiCVPR and 2SR methods do not perform well. CHROM performs better, because it averages skin-colored pixels in each frame and then performs computations on the sequence as a whole.

To provide an interpretation of what the HR-CNN actually learned, we present two insights. First, we give a "visual explanation" of the Extractor component based on Grad-CAM method [12] (Fig. 3). The method was adapted to our settings. For a given convolutional layer, its weighted activations are outputted. Sequence of layer activations provides a clue about the extractor's function. In our case, the first layer (left) "focuses" on cheeks and lips, the next one increases the importance of cheeks and reduces the importance of hair, and this trend follows in the next two layers. Next, a plot of the ground truth HR and estimated HR for a sequence with a significantly changing HR (Fig. 4) is presented. The predicted HR follows the ascending trend of true HR. The deviation of estimated HR from the true HR around frame 1500 corresponds to a moment when the subject used strong facial expressions to reflect the difficulty of the activity.

# 5 Conclusion

A two-step convolutional neural network composed from the *Extractor* and *HR Estimator* that predicts heart rate was presented. The network yields state-of-the-art results on three publicly available datasets against three baseline methods [4, 8, 23] under a standardized protocol. New challenging publicly available ECG-Fitness dataset with 60 second videos of people performing physical exercise was introduced.

The proposed method performs significantly the best on the ECG-Fitness database that contains realistic challenges. In contrast to commonly used COHFACE and MAHNOB databases, the videos are not compressed. In terms of practical impact, there is a little point in validating the heart rate estimation methods on databases where the only challenge is the compression.

# 6 Acknowledgment

The research was supported by the Czech Science Foundation project GACR P103/12/G084 and by the OP VVV project CZ.02.1.01/0.0/0.0/16_019/0000765 Research Center for Informatics and CTU student grant SGS17/185/OHK3/3T/13.

# References

[1] Yu-Hsin Chen, Hong-Hui Chen, Tung-Chien Chen, and Liang-Gee Chen. Robust heart rate measurement with phonocardiogram by on-line template extraction and matching. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2011:1957–1960, 2011. ISSN 1557-170X. doi: 10.1109/IEMBS.2011.6090552.

[2] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv:1511.07289 [cs]*, November 2015. URL http://arxiv.org/abs/1511.07289. arXiv: 1511.07289.

[3] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, March 2010. URL http://proceedings.mlr.press/v9/glorot10a.html.

[4] G. de Haan and V. Jeanne. Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, October 2013. ISSN 0018-9294. doi: 10.1109/TBME.2013.2266196.

[5] M. A. Hassan, A. S. Malik, D. Fofi, N. Saad, B. Karasfi, Y. S. Ali, and F. Meriaudeau. Heart rate estimation using facial video: A review. *Biomedical Signal Processing and Control*, 38:346–360, 2017. ISSN 1746-8094. doi: 10.1016/j.bspc.2017.07.004. URL http://www.sciencedirect.com/science/article/pii/S1746809417301362.

[6] Guillaume Heusch, André Anjos, and Sébastien Marcel. A Reproducible Study on Remote Heart Rate Measurement. *arXiv:1709.00962 [cs]*, September 2017. URL http://arxiv.org/abs/1709.00962. arXiv: 1709.00962.

[7] A. Huch, R. Huch, V. König, MR Neuman, D. Parker, J. Yount, and D. Lübbers. Limitations of pulse oximetry. *The Lancet*, 1:357–358, 1988.

[8] X. Li, J. Chen, G. Zhao, and M. Pietikäinen. Remote Heart Rate Measurement from Face Videos under Realistic Situations. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4264–4271, June 2014. doi: 10.1109/CVPR.2014.543.

[9] He Liu, Yadong Wang, and Lei Wang. A review of non-contact, low-cost physiological information measurement based on photoplethysmographic imaging. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2012:2088–2091, 2012. ISSN 1557-170X. doi: 10.1109/EMBC.2012.6346371.

[10] Bruce D. Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc. URL http://dl.acm.org/citation.cfm?id=1623264.1623280.

[11] D. J. McDuff, E. B. Blackford, and J. R. Estepp. The Impact of Video Compression on Remote Cardiac Pulse Measurement Using Imaging Photoplethysmography. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 63–70, May 2017. doi: 10.1109/FG.2017.17.

[12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv:1610.02391 [cs]*, October 2016. URL http://arxiv.org/abs/1610.02391. arXiv: 1610.02391.

[13] J. Sochman and J. Matas. WaldBoost - learning for time constrained sequential detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 150–156 vol. 2, June 2005. doi: 10.1109/CVPR.2005.373.

[14] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, January 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.25.

[15] R. Stricker, S. Müller, and H. M. Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062, August 2014. doi: 10.1109/ROMAN.2014.6926392.

[16] Y. Sun and N. Thakor. Photoplethysmography Revisited: From Contact to Noncontact, From Point to Imaging. *IEEE Transactions on Biomedical Engineering*, 63(3):463–477, March 2016. ISSN 0018-9294. doi: 10.1109/TBME.2015.2476337.

[17] Michał Sznajder and Marta Łukowska. Python Online and Offline ECG QRS Detector based on the Pan-Tomkins algorithm, July 2017. URL https://zenodo.org/record/826614#.WuBhhohuZPY.

[18] X. F. Teng and Y. T. Zhang. The effect of contacting force on photoplethysmographic signals. *Physiological Measurement*, 25(5):1323–1335, October 2004. ISSN 0967-3334.

[19] Narendra S. Trivedi, Ahmed F. Ghouri, Nitin K. Shah, Eugene Lai, and Steven J. Barker. Effects of motion, ambient light, and hypoperfusion on pulse oximeter function. *Journal of Clinical Anesthesia*, 9(3):179–183, May 1997. ISSN 0952-8180. doi: 10.1016/S0952-8180(97)00039-1. URL http://www.sciencedirect.com/science/article/pii/S0952818097000391.

[20] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2396–2404, June 2016. doi: 10.1109/CVPR.2016.263.

[21] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, December 2008. ISSN 1094-4087. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2717852/.

[22] W. Wang, S. Stuijk, and G. de Haan. Exploiting Spatial Redundancy of Image Sensor for Motion Robust rPPG. *IEEE Transactions on Biomedical Engineering*, 62(2):415–425, February 2015. ISSN 0018-9294. doi: 10.1109/TBME.2014.2356291.

[23] Wenjin Wang, Sander Stuijk, and Gerard de Haan. A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation. *IEEE transactions on biomedical engineering*, 63(9):1974–1984, September 2016. ISSN 1558-2531. doi: 10.1109/TBME.2015.2508602.