Concurrent video-based heart-rate measurement from human faces for large groups of participants - an improved CNN approach

Thomas Pursche (Author)
University of Wuppertal, School of Electrical, Information and Media Engineering, Automatic Control
Wuppertal 42119, Germany
Email pursche@uni-wuppertal.de

Marcel Walz
University of Wuppertal, School of Electrical, Information and Media Engineering, Automatic Control
Wuppertal 42119, Germany
Email marcel.walz@icloud.com

Roland Clauß
University of Wuppertal, School of Electrical, Information and Media Engineering, Automatic Control
Wuppertal 42119, Germany
Email clauss@uni-wuppertal.de

Bernd Tibken
University of Wuppertal, School of Electrical, Information and Media Engineering, Automatic Control
Wuppertal 42119, Germany
Email tibken@uni-wuppertal.de

Roland Clauß
University of Wuppertal, School of Electrical, Information and Media Engineering, Automatic Control
Wuppertal 42119, Germany
Email clauss@uni-wuppertal.de

Bernd Tibken
University of Wuppertal, School of Electrical, Information and Media Engineering, Automatic Control
Wuppertal 42119, Germany
Email tibken@uni-wuppertal.de

Reinhard Möller
University of Wuppertal, School of Electrical, Information and Media Engineering, Automation/Computer Science
Wuppertal 42119, Germany
Email r.moeller@uni-wuppertal.de

Karol Niewiadomski
University of Wuppertal, School of Electrical, Information and Media Engineering, Automation/Computer Science
Wuppertal 42119, Germany
Email karol.niewiadomski@uni-wuppertal.de

Abstract—Measuring the heart rate is the most essential foundation for arbitrary applications in medicine and sports, as well as in many other cases of application, like psychology. The conventional ways of measuring the heart rate like pulseoxymetry, electrocardiogram, wrist belts and so on are afflicted with some very own disadvantages. The most prominent point is the need of direct contact between the measurement unit and the participant, which often leads to discomfort and irritation of the skin, to mention only a few. To avoid this crucial disadvantage, several complete contact-free measurement techniques were developed in recent years. The most promising technique is the analysis of recorded video data of a human face. While these techniques were developed further and further, they stay limited to only one participant at the same time. As a solution for this circumstance, we proposed in a multi-person measurement approach using convolutional neural networks, in a previous paper. In this new paper, the first attempt is enhanced in terms of run time and accuracy by using an adapted algorithm for HR estimation. It will be presented that the evolutional approach is able to estimate the heart rate not only for one, but for a huge number of participants with a very small runtime and high accuracy.

Index Terms—CNN, heart-rate measurement, ICA, BSS, photoplethysmography

I. INTRODUCTION

It is indisputable that the heart rate (HR) is the foundation for various medical and psychological applications. The application cases reach from stress and sleepiness monitoring, towards physical application for top athletes to optimize the training effort towards all basic medical application, where vital parameters are required. The actual conventional HR measurement techniques - like electrocardiography (ECG), sphygmomanometry, pulse oxymetry and photoplethysmography (PPG) - lead to reliable results but have some disadvantages and limitations. The main disadvantage is that they are all contact-based, which sometimes leads to skin irritations and discomfort among the patients. Another disadvantage is the often very high price and the limitation in usability. Here the video based contact-free measurement techniques, which were presented first by Pavlidis et. al [1] in 2007, are able to avoid some of these disadvantages. The work of Pavlidis et al. was the foundation for several applications like [2] and [3], who used thermal video data of a participant to estimate its HR. A subsequent development was presented by Verkruysse et al. [4] who proved that only ambient light is required to estimate the HR with sufficient accuracy of the results. This method is also the foundation for the here proposed method, that was first presented in 2012 in [5]. The innovation was the use of the independent component analysis to increase accuracy of the HR measurement enormously. This method was developed in case of robustness as well as runtime and accuracy in [6] and [7] by the use of the Hilbert Huang transform and neural networks. For the sake of completeness, different measurement techniques are mentioned in the following. The research teams around Balakrishnan [8] respectively Rubinstein [9] are using algorithms to detect minimal motion.
in the human face to estimate the participant’s HR. Machine learning approaches were presented by Monkaresi et al. [10] and Tarssenko et al. in [11] by using ARX methods. Of course there are many more, but all of them are evolutionary steps of the presented techniques.

The limitation to only one participant in the front of the camera is the one crucial disadvantage of all of these presented methods. A first attempt to measure more than one participant at the same time was presented by Poh et al. [12], who used conventional algorithms. This approach was not further pursued so Pursche et al. presented in 2019 [13] an approach using convolutional neural networks, which are specialized neural networks, to estimate the HR of up to 9 persons simultaneously with recorded video data of the participant’s faces. The reasons for using CNN are given by the high efficiency in identifying and detecting human faces in video data proven such as in [14], [15] and [16]. As presented in [13] the proposed algorithm is able to derive the HR of up to 9 persons with a good accuracy, but relatively high runtime. The main aim of the here proposed paper is to optimize runtime as well as accuracy of the algorithm for a high amount of participants.

The paper is organized as follows. Because there are no new mathematical foundations in this work, a mathematical part is not presented. Section II deals with the experimental setup and gives an overview of the general approach. Hereafter the status quo is presented and discussed in detail in Section III. Section IV deals with the optimization and presents the improvements in runtime and accuracy. The paper is concluded in Section V by a summary of the achievements and gives a short prospect to future work.

II. DESIGN CONFIGURATION
A. Experimental setup

![Fig. 1. Snapshot of a test video sequence with 3 participants. Without identified faces.](image1)

The employed experimental setup is very similar to the one used in [5]. Also, nearly every other video-based HR measurement approach uses a similar structure. A participant is placed in front of a camera - often a conventional low budget webcam - in a distance of about 0.5m (see Fig 3 left). The video, recorded by the webcam owns the following specifications: a color video with a resolution of 800x448 pixels, 24 Bit RGB, 8 bit per channel. The video has a length of about 120 seconds and is recorded with 30 fps, so 3600 frames in total. When multiple persons are measured, the test setup slightly changes. Depending on the number of participants, the distance between camera and peoples faces rises up to 3m (see Fig 3 right) to obtain every participants face completely in the recorded frames. An exemplary obtained frame can be seen in Fig. 1 without and in 2 with identified faces.

![Fig. 2. Snapshot of a test video sequence with 3 participants. With identified faces.](image2)

![Fig. 3. Experimental Setup](image3)

It was taken care that the participants moved as little as possible and that the background was relatively plain monotonic and there were no fast and rapid changes in the lighting conditions. Each combination in the amount of up to 9 persons was recorded several times with different participants. As reference measurement an pc based ECG was chosen. The applied ECG is a Biocare ECG-2000, with 12 channels, a sampling rate of 1000Hz and a frequency band from 0.05 to 250MHz to cover all possible HR cases and beyond. Based on the fact, that the ECG is very expensive, only one participant was provided with the ECG and all other subjects were equipped with common chest belts. In contrast to previous studies, there were no changes in the activation level of the participants, so every person was nearly at his or her resting level. The resting level of all participants was about 60Bpm. The employed hardware was Intel Core i/-7500U 2.7GHz pc with Windows 10 and 8GB of Ram.
B. Heart-rate measurement

The heart beat is the measurable result of the heart’s action. The alternating rhythmic contraction and relaxation of the heart muscle is controlled by electrical impulses, which are formed in the sinus node located in the right atrium. These electrical impulses pass through multiple stations into the left and right ventricle. In consequence of the alternating electric impulses the heart contracts and relaxes, which leads to a movement of blood through the arteries and veins of a human. This blood movement can be interpreted as a pulse wave traveling throughout the vessel system with alternating maxima and minima. The heart beat, or heart rate, is defined as the distance between two main maxima. To measure the heartbeat there are a couple of different techniques. The foundation of the here used algorithm is the photoplethysmography. This technique works with the detection of the different fillings in the vessel system for the contracting respectively the relaxation phases. By this amplitude, differences in the HR of a person can be measured. In [1] Pavlidis et al. proved that this technique is capable for a complete contact free HR measurement. That this technology is able to estimate a persons HR only by video data of his or her face, was proven in many publications.

The operating principles of the proposed method are illustrated in Fig. 4. At the beginning a video with an arbitrary amount of peoples faces in it is loaded into the algorithm. In the first step the Face Tracker analyzes the whole video, frame per frame, if and how many faces are inclosed in it. The Face Tracker is based on a face identification algorithm that was presented by Zang et al. [14] and implemented by Antoshchenko et al. in [17]. The algorithm has a multiple layer structure and is highly capable of identification and localization of human faces in video data by optimization through recalibration. The algorithms output is divided into three parts. The face classification, a stochastic parameter, which describes the probability of a face detection. The facial landmarks indicates the x and y coordinates in the frame, where outstanding features like eyes, nose and mouth angles can be found. At least the parameter bounding box gives the coordinates of a rectangle, which includes the found face completely. It is obvious that, if more than one face was found, these parameters are given for every face. The region of interest (ROI) can either be put as the whole bounding box or, which has proven to be more efficient in runtime as well as in accuracy, computed in a defined way through the received facial landmarks. After the preprocessing has taken place and the algorithm found a minimum of one face, the whole video is decomposed into single RGB channels. For further contemplation only the green channel is of interest, because it proved well in multiple publications. After that detrending and normalization are performed. For noise reduction the Hilbert Huang transform, which has proven capable in elimination of changes in lighting conditions, as well as movement of the recorded person in front of the camera, is conducted. Before this step, the independent component analysis (ICA) is used, which is related to the Blind Source Separation (BSS). The complete mathematics can be found in [5] and [18]. The last step consist of several filtering techniques to obtain and finally analyze the corresponding HR.

<table>
<thead>
<tr>
<th>#</th>
<th>abs. time [s]</th>
<th>rel. time [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1149.72</td>
<td>96.72</td>
</tr>
<tr>
<td>2</td>
<td>1691.89</td>
<td>98.28</td>
</tr>
<tr>
<td>3</td>
<td>2307.02</td>
<td>91.91</td>
</tr>
<tr>
<td>4</td>
<td>2749.22</td>
<td>93.33</td>
</tr>
<tr>
<td>5</td>
<td>3609.06</td>
<td>94.53</td>
</tr>
<tr>
<td>6</td>
<td>4242.54</td>
<td>95.09</td>
</tr>
<tr>
<td>7</td>
<td>5536.81</td>
<td>97.77</td>
</tr>
</tbody>
</table>

**TABLE I** Absolute and relative runtime results

**III. CURRENT STATUS**

As stated in [13] the reference group consists of 30 different participants. They were clustered in groups of up to 9 people in permuting order. Several videos of each resulting group were recorded. This leads to the examination of 7 different settings with a diverging amount of simultaneously faces in the video. Due to high costs only one participant was monitored with an ECG, while the others were supplied with common chest belts to measure the reference HR. The results were compared towards the HR obtained by the video-based approach. The outcome of all 7 test scenarios is stated in Table I. In the first column the number of faces in the video are stated. For each test setting the absolute time is stated as well as the efficiency per face, defined as the quotient of total time to number of detected faces. In the next column relative time for each considerable step is given. These steps are CNN, which describes the face identification, Corr the correction time for the video data, ICA time taken for the independent component analysis and the last one, PD describes the amount of time needed for the peak detection and the finally HR estimation. Table II describes in the first row the deviation of HR with ΔHR and in the second row face detections accuracy with CNNp.
the need of performing a complete ICA for every face results in additional runtime. It can be seen, that there is an optimum for the efficiency of the runtime for 7 participants (see Fig. 5 (b)). The reason for this is the fact, that the runtime can be divided into two parts, a fixed and a variable part. Processing time for the CNN is mostly fixed, because every pixel of each frame has to be processed. For the process time it is irrelevant if there is one or more faces in the frame. The variable time is given by the processing of localization data, because this one correlates directly with the amount of faces in a frame. This leads to the conclusion that for a few number of faces, the fixed part is relatively high and descends with the ascending number of faces found in the picture. Consequently, the variable part rises due to the localization process for each face. This leads to the case, that the efficiency of the runtime has its optimum at seven participants in a picture. If the number of participants reaches more than seven people, the efficiency is getting worse (see Fig. 5 (b)+(c)). The results of the basic approach can be described as follows. First of all only the face identification via CNN and the ICA have noticable influence on the whole relative runtime. In particular, for each setting the time used for the CNN ranges from 92% to 98%, which directly results in the insight, that this is the most time consuming part of the algorithm and therefore the most worthwhile objective for a further optimization. The combined ICA’s and intensity evaluation share varies from 1.6% to 6.5%. All other parts - like correction, peak detection and filtering - contribute less than 0.5% and are therefore not noteworthy. This leads to the conclusion that the ICA and the CNN are the most promising candidates for an overall runtime optimization.

IV. Optimization and Improvements

When analyzing the test data to evaluate the algorithm, two main potentials for improvement can be identified: the accuracy of the heartbeat calculation and the runtime. In the following sections a realization is presented for both optimization objectives and after considering the results, both are combined in a third variant of the algorithm. Only those steps of the algorithm are presented, which have been modified by the implementation of the improvement.

A. Accuracy optimization

The most promising approach to improve the accuracy, is optimizing the used ROIs. As proofed in many publications, the frontal face and cheek are the most promising areas. The ROI is calculated based on the features that the CNN locates in face identification and determines in terms of cartesian coordinates. A representation of the realized ROIs is shown in Figure 6. This leads now to three incoherent areas for each detected face. To optimize the later use of the ICA and -
as a result of this - later the runtime, the data structure is optimized. Thus, these incoherent areas are connected to one. The schematic overview of this process induced by the new data structure can be seen in Fig. 7.

![Schematic overview for optimized processing](image)

**Fig. 7.** Schematic overview for optimized processing

### B. Runtime optimization

As pointed out in Section Status quo the CNN is the most time consuming part of the complete algorithm with a minimum of 90% of the overall runtime. There exist several approaches to optimize the runtime. On option is to choose a more efficient CNN, which does not make sense in this case, because many studies have shown that this is already an optimal CNN. Another more promising way is the optimization of the CNN itself to reduce runtime remarkably without the constraint to modify the basic technical architecture. Basically the optimization approach makes use of the fact that the position of faces found in the video is nearly identical over the whole time of image acquisition. However this fact implies that the algorithm can only be used in scenarios where the subjects’ position is invariant. For example, this restriction is not limited to the initial implementation, since changes in the position of the faces can also be processed within a specific frame. In the now presented optimized algorithm, the assumption is made that the face positions are fixed and therefore is no need to redefine them on every single frame. It is sufficient to determine a mean value of the position data based on a subset of frames and to develop a template therefrom. Then this template can be applied to all frames and extracts the corresponding subsections. The amount of images that must be processed by the CNN can thus be reduced from over three thousand to a few images. Although this modification of the initial implementation is technically less complicated to implement, it has fundamental implications for the main flow of the algorithm. The first new task is now to compute the localization template. A methodical overview can be seen in Fig. 8. As a result of this optimization only a small amount of frames have to be used to gather nearly the same results when using the complete set of frames in the video.

![Runtime optimization](image)

**Fig. 8.** Schematic overview for runtime optimization

### C. Results

To evaluate the implementations performance, both optimization approaches are conducted at the scenario with 3 participants. The evaluation criterion is the same as in Table I and II. The results in Table III can be read as follows. **init** describes the results from the initial approach, **prec** and **Δ** describes the absolute respectively relative accuracy. With **time**, and **Δtime,20** is the absolute respectively relative runtime described. **i** states the number of used initial frames in this case 20 and 50. The main focus of the result evaluation for the optimization of the accuracy is, of course, the parameter **ΔHR**. As can be seen from the table, the accuracy divergence for the examined scenario could be reduced by almost a third from 4.98% to 3.54%. The fact, that the proportion of ICA computation time increases significantly due to the more complex image structures of the ROIs, is no longer significant, because in turn, the overall runtime could be slightly improved by further improvements and flattening of the algorithm. For consideration of the performance optimization results, the total running time comes into focus, although the accuracy of the calculation is still important. With a total runtime of 137.46 or 160.09 seconds, a reduction of more than 90% was achieved. However, like mentioned before, this drastic improvement comes at the cost of significantly lower accuracy results. With values of just under 9 and 10 % respectively, the deviation in calculating the heart rate is twice as high as in the initial implementation.

Looking at both optimization approaches in isolation, they fulfill their purpose perfectly. Both results in the improvement of precision, as well as in runtime optimization, are noteworthy. This insight suggests the idea of combining both approaches and so combining advantages in speed from one with the high precision of the other. Since both approaches have already been described in detail, we will refrain from further detailed documentation at this point. In terms of content, the combination of both approaches means that a template is again determined based on a defined subset of images. However,
the template now does not consist of a rectangle around the face, but from the ROIs defined in Figure 6. When evaluating the results, two scenarios with a threshold of 20 and 50 initial frames are evaluated for performance optimization. The results are recorded in Table IV. Although the results of time20 and time50 in Table III are still relatively similar, the data from combi20 and combi50 show more significant differences. Especially when setting the considered subset to 50 frames, the results are achieved, which are combined with the realization of this optimization approach to be followed. With a runtime reduction of almost 90%, the accuracy of the HR calculation, compared to the initial implementation, could be improved by more than 18%. The results obtained accurately reflect the intended goals of the optimizations. By modifying the algorithm at the correct positions both accuracy and runtime can be improved significantly.

<table>
<thead>
<tr>
<th>#</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>init</td>
<td>2307.02</td>
<td>769.01</td>
<td>91.91</td>
<td>2.01</td>
<td>4.98</td>
<td>100.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prec</td>
<td>1997.65</td>
<td>665.88</td>
<td>87.20</td>
<td>13.73</td>
<td>3.54</td>
<td>99.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δt [%]</td>
<td>▼ 13.41</td>
<td>▼ 5.12</td>
<td>▼ 33.33</td>
<td>▼ 33.33</td>
<td>▼ 28.92</td>
<td>▼ 0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time20</td>
<td>137.46</td>
<td>45.82</td>
<td>23.83</td>
<td>75.01</td>
<td>0.38</td>
<td>9.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δt [%]</td>
<td>▼ 94.04</td>
<td>▼ 74.07</td>
<td>▼ 3631.84</td>
<td>▼ 1166.67</td>
<td>▼ 98.59</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time50</td>
<td>160.09</td>
<td>33.36</td>
<td>29.40</td>
<td>69.59</td>
<td>0.32</td>
<td>8.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δt [%]</td>
<td>▼ 93.00</td>
<td>▼ 68.01</td>
<td>▼ 3622.19</td>
<td>▼ 966.67</td>
<td>▼ 80.32</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE III**

**Optimization of runtime and accuracy**

**V. Conclusion**

The presented approach demonstrated an extension for estimating the HR for a large group of people concurrently. It was proven that it is not only capable to estimate the HR for each participant in a decent accuracy, but also in a very short time in comparison to the initial implementation. By the combination of accuracy and runtime optimization, it was possible to reduce the average runtime for over 87.25% and as well increase the accuracy by over 18.27%, which is a huge step towards a "realtime" application. As a conclusion of this, it is now possible to measure the HR of a large group with up to 9 people with decent accuracy, in a sustainable overall runtime. By further optimization of the CNN, and a more efficient and skillfully dealing with the literal needed data, an efficient hardware implementation is possible. And therefore a "realtime" HR measurement for large groups of people.

**References**


