A Reproducible Study on Remote Heart Rate Measurement

G. Heusch*, A. Anjos, S. Marcel

Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland

Abstract

This paper studies the problem of reproducible research in remote photo-plethysmography (rPPG). Most of the work published in this domain is assessed on privately-owned databases, making it difficult to evaluate proposed algorithms in a standard and principled manner. As a consequence, we present a new, publicly available database containing a relatively large number of subjects recorded under two different lighting conditions. Also, three state-of-the-art rPPG algorithms from the literature were selected, implemented and released as open source free software. After a thorough, unbiased experimental evaluation in various settings, it is shown that none of the selected algorithms is precise enough to be used in a real-world scenario.

Keywords: Image analysis, Remote photoplethysmography, Reproducible research, Vital signs monitoring

1. Introduction

Photoplethysmography (PPG) consists in measuring the variation in volume inside a tissue, using a light source. Since the heart pumps blood throughout the body, the volume of the arteria is changing with time. When a tissue is illuminated, the proportion of transmitted and reflected light varies accordingly, and the heart rate could thus be inferred from these variations. The aim of remote Photoplethysmography (rPPG) is to measure the same variations, but using ambient light instead of structured light and widely available sensors such as a simple webcam.

It has been empirically shown by Verkruysse et al. [14] that recorded skin

---

*Corresponding author

Email addresses: `guillaume.heusch@idiap.ch` (G. Heusch),
`andre.anjos@idiap.ch` (A. Anjos), `sebastien.marcel@idiap.ch` (S. Marcel)

colors (and especially the green channel) from a camera sensor contain subtle changes correlated to the variation in blood volumes. In their work, they considered the sequence of average color values in a manually defined region-of-interest (ROI) on the subject's forehead. After having filtered the obtained signals, they graphically showed that the green color signal main frequency corresponds to the heart rate of the subject.

Since then, there have been many attempts to infer the heart rate from video sequences containing skin pixels. Notable examples include the work by Poh et al. [9], where the authors proposed a technique where the color signals are processed by means of blind source separation (ICA), in order to isolate the component corresponding to the heart rate. In a similar trend, Lewandowska et al. [5] applied Principal Component Analysis (PCA) to the color signals and then manually selected the principal component that contains the variation due to blood flow. These two early studies empirically showed that the heart rate could be retrevied from video sequences of faces, but also highlight important limitations: the subject should be motionless, and proper lighting conditions must be ensured during the capture. According to a recent survey [7], research in remote heart rate measurement has considerably increased in the last few years, most of which focuses on robustness to subject motion and illumination conditions. Since a large number of approaches have been proposed recently, they will not be discussed here. We refer the interested reader to [7] for a comprehensive survey of existing algorithms.

Despite the vast amount of published material on the subject, there are still no standard evaluation procedures for remote heart rate estimation. One of the main reasons being the limited amount of datasets containing video sequences with accompanying physiological measurements [7]. Our work is hence motivated by the lack of standard comparison of rPPG approaches on publicly available databases. Indeed, most of the algorithms are assessed on privately-owned databases, making it difficult to compare in a fair manner or to accurately reproduce them.

As a consequence, the contribution of our work is as follows: first, we present a new, publicly available database to assess the performance of remote pulse rate measurement algorithms. This database comprises a relatively large number of subjects and various illumination conditions. Second, we experimentally test selected state-of-the-art rPPG algorithms using both the Manhob HCI-Tagging database [11] and our new database. The selected baseline algorithms are provided as an open-source, free software package that every researcher in the field can download to reproduce results presented in this paper [4].

2. Publicly available databases

2.1. Mahnob HCI-Tagging

The Mahnob HCI-Tagging dataset was not designed for rPPG benchmarks, but rather for the characterisation of multimedia content based on human emotions. For that purpose, dataset authors collected video and physiological data from 30 subjects being exposed to different audio-visual stimuli such as the display of pictures and short to medium-sized video sequences. The Manhob HCI-Tagging dataset may be used to evaluate rPPG techniques by ignoring most of its data contents and manually correlating physiological signals with the data from the professional camera located in front of the subject being tested. An example image capture for the Mahnob dataset is shown on Figure 6. Authors indicate professional lighting was used so as to maximize the video data quality. The professional camera used outputs videos of size 784 (h) by 592 (w) pixels at a rate of 60 Hz. Three electro-cardiogram (EKG) sensors placed on the participants chest measure the individual heart rate. The data extracted from those sensor readings is also made available in the dataset and are synchronized with the video recordings. Figure 1 shows the outputs of the three EKG sensors (in blue). The estimated heart rate for the subject on the video must be estimated from these by first running a QRS detector [8] which estimates the location of vertical (red) bars corresponding to true heart-beat events[1], followed-up by a consensus heuristic (voting) for selecting the estimate rate from one of the channels. Because sequences are rather short (30-60 seconds) and the subject is comfortably sitting, it is assumed the heart rate does not vary greatly during the session. Once the heart rate is estimated for every sequence in the Mahnob HCI-Tagging database (there are 3'490 videos on the dataset), it is possible to use it for rPPG benchmarking.

2.2. The COHFACE dataset

The COHFACE dataset[2] is composed of 160 videos and physiological signals collected from 40 healthy individuals. The data collection campaign spawned several days. The data acquired in this new corpus includes more realistic conditions as compared to the well-controlled setup for recording the Manhob HCI-Tagging data.

---

[1]the MNE package was used for this purpose. http://mne-tools.github.io/stable/index.html
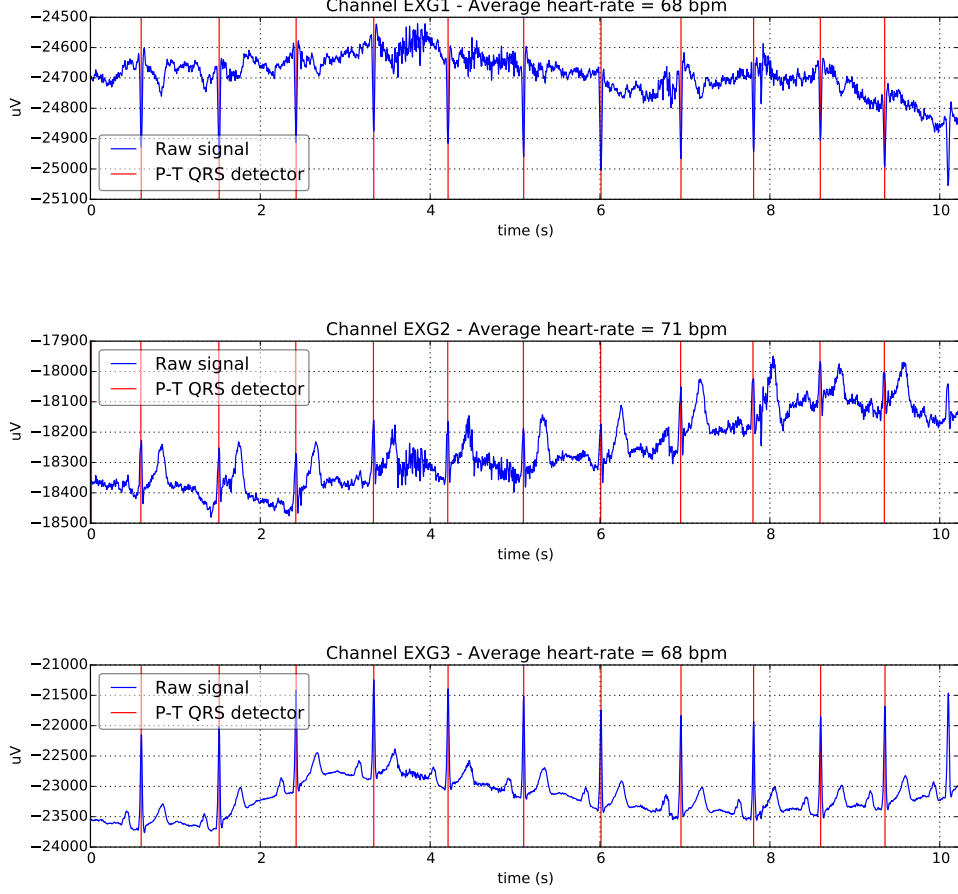
[2]https://idiap.ch/dataset/cohface

Figure 1: Example of collected electro-cardiogram signals. Signals provided with the database are depicted in blue. Red stripes indicate the result of QRS detection. The estimated heart rate is shown on the top of each graph in terms of bpm.

The age distribution from the 40 individuals is shown in Figure 2. The average subject age is 35.6 years old, with a standard deviation of 11.47 years. Gender-wise, there were 12 women (30%) and 28 men (70%).

2.2.1. Data collection

Individuals were asked to look into a conventional webcam connected to a laptop during a period of approximately 60 seconds. We registered both video and physiological signals (contact photoplethysmography and respiration) using a biomedical toolkit that synchronized the 3 input channels. The
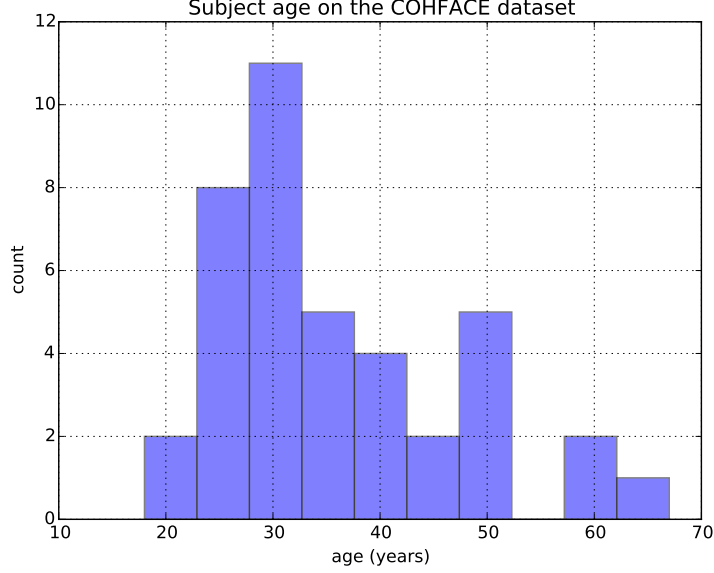
Figure 2: Age distribution for subjects at the COHFACE dataset.

physiological data shall be used as ground-truth for benchmarking rPPG algorithms.

For each of the 160 videos, the subject's face is recorded by a commodity webcam (Logitech HD Webcam C525) during a full minute, while physiological readings are taken by a Blood-Volume Pulse (BVP) sensor and a respiration belt, both from Thought Technologies (BVP model SA9308M, belt model SA9311M). The BVP sensor measures changes in skin reflectance to near-infrared lighting caused by the varying oxigen level in the blood due to heart beating. The respiration belt is composed of a mechanical coiling system that simply measures thoracic stretch. Both sensors are connected, following advice from Thought Technologies, to a computer running Microsoft Windows via their 2-channel USB-based acquisition system (ProComp2). Data from the physiological sensors together with the video stream from the webcam is synchronized and recorded using Thought Technologies' BioGraph Infiniti Software suite, version 5. With the standard settings, BioGraph Infiniti v5 was able to output a video stream with a resolution of 640 x 480 pixels at 20 frames per second, together with BVP measurements 256 times per second and, respiration sensor readouts 32 times per second. Figure 3 shows two images as captured by the webcam. Figure 4 shows typical signals acquired in sync with the video data, corresponding to the subject's pulse (top, in blue) and the respiration readings (bottom, in blue).

(a)   (b)

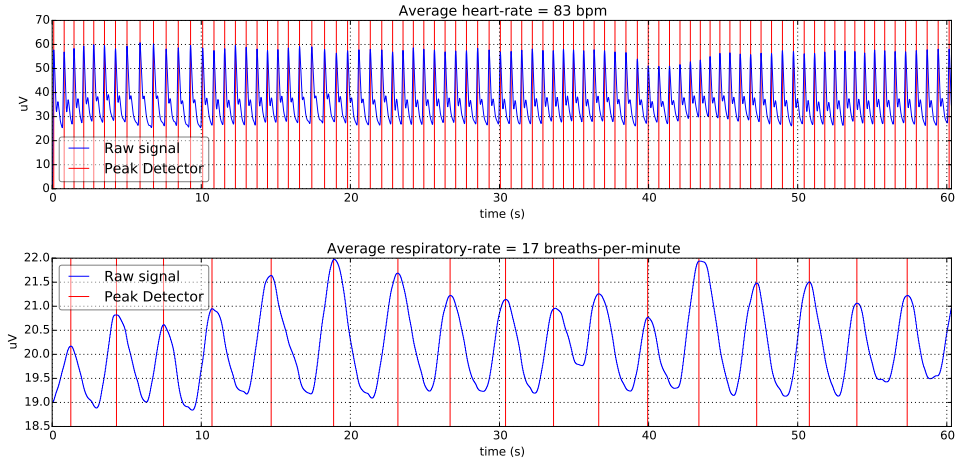Figure 3: Example images of the COHFACE dataset. (a) Studio (b) Natural

Figure 4: Example signals acquired for each session in the COHFACE dataset.

To estimate the heart and respiratory rates from the signals read-out from our physiological sensors, a simple peak detector was deployed. While the Mahnob HCI-Tagging dataset contains EKG signals from which a QRS complex can be detected, pulse signals readout by the contact PPG sensor used at the COHFACE dataset acquisition look sensibly simpler (see Figure 4) and just applying a Pan and Tompkins detector [8] would not work as expected.

After testing, we decided to adopt a simple peak-detector available as free software[3], which deployment results in the detected peaks (bars in red) also shown in Figure 4. The estimated heart and respiratory rates can then be

---

[3]https://github.com/demotu/BMC/blob/master/functions/detect_peaks.py

easily calculated from the peak positions as shown on the title bars of the same figure.

2.2.2. Acquisition Protocol and Illumination Changes

Each of the 40 individuals was asked to sit still in front of the webcam so that the face area is captured in full, for four sessions that lasted about 1 minute. Illumination on the scene was changed once as to create two types of lighting conditions:

1. **studio**, for which we closed the blinds, avoiding natural light, and used extra light from a spot to well illuminate the subject's face, and
2. **natural**, in which all the lights were turned off and the blinds were opened.

The four video sequences (2 with studio lighting and 2 with natural lighting) can be used to evaluate the performance of rPPG algorithms in either matched or unmatched settings. Figure 3 shows the differences in illumination for the two conditions of acquisition for the COHFACE dataset.

2.2.3. File Formats and Metadata

The choice of formats for data distribution aimed on improving readability and access using free software tools. Data for the video stream in each session and for each individual is shipped in MP4 format inside a commonly available AVI movie container. Data from the physiological sensors is distributed using standard HDF5 containers, with information such as the date of birth of the subject, his gender, the illumination condition and of course the recorded pulse and respiration signals.

2.2.4. Comparison to Mahnob HCI-Tagging

Table 1 summarizes the differences between the only two public datasets for rPPG benchmarking. While the Mahnob HCI-Tagging dataset provides more video samples, the COHFACE dataset allows more realistic testing for the use of rPPG due to its increased data variability (more subjects and more realistic data acquisition conditions).

3. Selected Baseline Algorithms

In this section, algorithms which have been selected for establishing baselines on our new database are presented. Three different approaches to re-

Table 1: Key figures of the COHFACE dataset when compared to the Mahnob HCI-Tagging dataset for the purpose of remote non-contact PPG benchmarking.

	HCI-Tagging	**COHFACE**
Subjects	30	40
Samples	3490	160
Camera	784x592@60Hz	640x480@20Hz
Physiological Signals	$3 \times$ EKG	BVP + Respiratory Belt
Illumination	Studio	Studio + Natural

trieve the pulse signal have been chosen and are detailed in the following subsections.

3.1. CHROM

The so-called CHROM approach [3] is relatively simple, and has been shown to outperfom previous baselines such as ICA [9] and PCA [5]. The algorithm first applies a color filter based on a one-class SVM to find skin-colored pixels in each frame of the input sequence. Then, the mean skin color value is projected onto the proposed chrominance subspace, which aims to reveal subtle color variations due to blood flow. The final pulse trace is obtained by first bandpass filtering the temporal signals in the XY chrominance colorspace, and then combining the two dimensions of this colorspace into a one-dimensional signal (see Figure 5). Note that in our implementation, we did not use a one-class SVM as the skin color filter, but a recent approach proposed by Taylor [13]. An example of the obtained skin mask with this algorithm is shown in Figure 5.
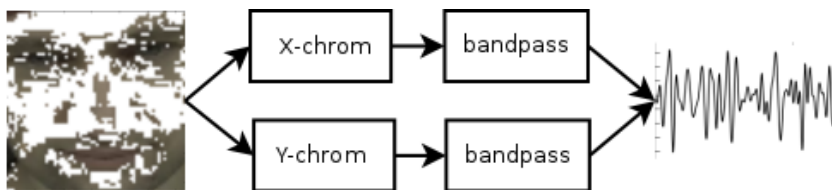

Figure 5: Flowchart of the CHROM algorithm to retrieve the pulse signal.

This approach was applied to no less than 117 static subjects, where different skin tones were represented. The recordings have been made in a controlled environment and using professional studio illumination. Obtained performances for the CHROM algorithm, as well as for the baselines used for comparison are almost perfect (with a Pearson's correlation coefficients between 0.97 and 1). Also, another experiment was presented on exercising subjects, either on a bike or on a stepping device. As expected, performance

degrades, but the CHROM algorithm was shown to better mitigate the effect of subject's motion on the rPPG signal, as compared to previous baselines.

3.2. LiCVPR

The algorithm pusblished by Li et al. [6] has been included in selected baselines since it is the only rPPG approach which reports results on a publicly available database [11]. Hence, it provides a reference and allows comparison between published material and our implementation of the different algorithms. This approach relies on the tracking of the bottom part of the face, defined by keypoints as shown in Figure 6. The keypoints are detected thanks to the DMRF algorithm [1] in the first frame of the sequence and are used to build the mask. The position of this mask is updated at each subsequent frame by applying an affine transformation found by a KLT tracker on features detected by a Shi-Tomasi detector [10] inside the face region. Once the temporal signal corresponding to the mean green value inside the tracked area has been built, it is corrected for illumination using the Non-Linear Mean Square algorithm with the mean green value of the background as reference. Large variations in the corrected signals, probably due to subject motion, are removed according to statistics computed on the entire database. Finally, the pulse signal is obtained by applying a detrending filter [12] to reduce slow trends in the signal, a moving average filter to remove noise and a bandpass filter to retain the range of interest corresponding to plausible heart rate values. A flowchart of the whole procedure is given in Figure 6.
Experiments are performed on two different datasets. The first one has been built with 10 subjects recorded under controlled conditions, and individuals were explicitly asked to avoid any movement. Note that this dataset has not been made public by the authors. The second one is a subset of the Manhob HCI-Tagging dataset, containing only sequences lasting more than 30 seconds. The performance of the proposed algorithm on the first, controlled video sequences is almost perfect, reaching a Pearson's correlation coefficient of 0.99. However, on the less controlled data from the Manhob HCI-Tagging database, there is a slight drop in performance with a reported correlation coefficient of 0.81.
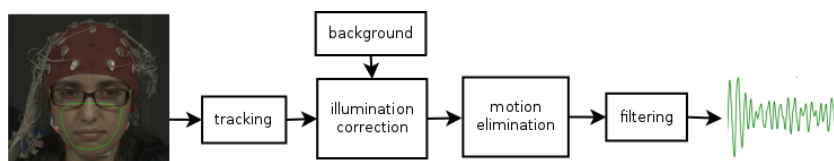

Figure 6: Flowchart of the LiCVPR algorithm to retrieve the pulse signal.

3.3. 2SR

Wang et al. recently proposed a novel algorithm termed as Spatial Subspace Rotation (2SR) [15]. In their approach, the authors consider the subspace of skin pixels in the RGB space and derive the pulse signal by analyzing the rotation angle of the skin color subspace in consecutive frames. To do so, the eigenvectors of the skin pixels correlation matrix are considered. More precisely, the angle between the principal eigenvector and the hyperplane defined by the two others is analyzed across a temporal window. As claimed by the authors, this algorithm is able to directly retrieve a reliable pulse signal, and hence no post-processing step (i.e. bandpass filtering) is required. However, this algorithm needs an accurate estimate of the skin color, since it is implicitly assumed that the skin colored pixels form a single cluster in the RGB space. In our work, the same skin color filter [13] as in the CHROM method is used, allowing a fair comparison between these two approaches. Experimental results obtained on a private dataset containing 54 video sequences show better overall performances than all the approaches considered for comparison, including CHROM. A correlation coefficient of 0.94 is reported across a wide variety of conditions, including skin tone, motion of the subject and recovery after exercise.

4. Experimental assessment of selected baselines

In this section, results obtained with the algorithms described in Section 3 on both the Manhob HCI-Tagging and the COHFACE databases are reported. We chose to use the same performances metrics as in [6], namely the root mean square error (RMSE) and the Pearson's correlation coefficient ρ. As a first step, we tried to reproduce results already published in the literature. We then go in more details in the assessment of the different approaches. A thorough evaluation is made regarding the generalization ability to unseen data, including the robustness to different acquisition conditions. For this purpose, several experimental protocols are defined in respective subsections.

For all of the investigated approaches, the face is first located in each frame using a detector based on a boosted cascade of LBP features [2]. Furthermore, the average heart rate over a video sequence is inferred from the pulse signal by detecting the maximum value in the frequency spectrum of the estimated pulse signal.

Note finally that all the data, algorithms and scripts used to produce results presented in this paper are freely available for download to ensure maximum reproducibility [4].

4.1. Baselines

As a starting point, obtained results on the same subset of the Manhob HCI-Tagging database as used by Li et al. [6] are presented. This data subset consists in 527 sequences of 30 seconds. These sequences are obtained by considering selected video files from frame 306 to frame 2135. Table 2 report the RMSE and the correlation coefficient with the ground truth on these 527 sequences. We also report, for each algorithm, the number of free parameters that have been optimized directly on this dataset.

Since testing all possible combinations of the different parameters is virtually impossible, the following strategy was adopted: because investigated approaches generally involve a serie of sequential steps, we first optimize the parameters of the first step (the threshold on the skin color probability, for instance), fix them, and keep on with the optimization of the parameters of the second step, and so on.

Table 2: Results of our implementations of baselines on 527 sequences of the Manhob HCI-Tagging database, from frame 306 to 2135

	LiCVPR	CHROM	2SR
RMSE	**8.12**	15.40	18.4
Pearson's ρ	**0.70**	0.33	0.43
# of parameters	12	6	4

The first point to note here is that we were unable to reproduce results reported by Li et al. [6], where a RMSE of 7.62 and a correlation coefficient of 0.82 are presented. We can say with high confidence that these differences are only due to the tracking procedure and the considered background area. Indeed, the authors of [6] shared their source code, except the tracking part, hence making their algorithm only partially reproducible. However, this algorithm still significantly outperforms the two other baselines. Note also that our own implementation of the spatial subspace rotation (2SR) algorithm does reflect findings reported in [15], where the authors empirically showed that their approach outperforms CHROM [3], which is also the case here. Finally, it is important to point out that the various parameters, such as the threshold on motion elimination [6], or the window size in the overlap-add procedure [3], have been tuned *directly on the test data*, following the practice in [6]. Doing so introduces a severe bias, since the generalization ability of each algorithm to unseen data could not be established. As a consequence, we devised strict experimental protocols for subsequent experiments, specifying data for training (i.e. optimizing the free parameters) and for testing.

These various protocols are presented in the next subsections.

4.2. Unbiased performances in clean illumination conditions

In a pattern recognition framework, such as the one here, it is important to train and test proposed algorithms on different datasets. This is done to assess the generalization ability of the different approaches to unseen data. Indeed, it could be the case that some combination of free parameters would perform well on certain kind of data (say, a particular skin tone, for instance), while failing in other conditions. Using different datasets for tuning the algorithms and evaluating their performances is hence critical to have a clear insight of their generalization capabilities. In rPPG research however, there are - to the best of our knowledge - no published results following this paradigm. This could be due to the fact that databases are generally expensive to collect and hence rather small.

4.2.1. Manhob HCI-Tagging

The complete Manhob HCI-Tagging database contains 3490 video sequences recorded during 4 sessions, with constant acquisition conditions. The only difference between the sessions consists in the task assigned to the participants, and that should not affect the performance of any rPPG algorithm. However, there is a great amount of variation in sequence length, since the smallest sequence contains 556 frames and the largest 15780, the mean being 1789 frames. Different sets were built based on subjects identities, ensuring no overlaps between training and test sets. Also, since the sequence duration is usually correlated with the session, all sessions corresponding to selected subjects were included in the set. This is done to balance the mean sequence duration across the different sets. The training set has been built using data from 19 subjects and the test set with data of the remaining 10 subjects. Table 3 summarizes relevant information about this partitioning.

Table 3: Statistics for training and test sets of the Manhob HCI-Tagging database. The minimum, maximum and mean length are given in number of frames.

Set	# of sequences	min length	max length	mean length
Train	2302	556	13106	1812
Test	1188	579	15780	1745

We then applied the different baseline algorithms to the different sets of the Manhob HCI-Tagging database. Table 4 shows the Pearson's correlation coefficients obtained when parameters are optimized on the training set.

Table 4: Pearson's correlation coefficient on different sets of the Manhob HCI-Tagging database

	Training set	Test set
LiCVPR	**0.49**	**0.45**
CHROM	0.15	0.14
2SR	0.17	0.05

As can be seen on Table 4, obtained performance is generally better on the training set than on the test set, as expected. Also, it can be observed that performance is not as good as in the previous experiment (Table 2). This can be mainly explained by considering two points. First, the length of the different sequences greatly varies, and some parameters may yield to an overall better behaviour if the sequences are of the same duration (i.e. some parameters value may be best suited to long sequences). As a matter of fact, optimal parameters on the training set are different than the ones obtained in the previous experiments. Secondly, and a more obvious reason for this performance drop is that more sequences introduces more variations.

4.2.2. COHFACE

Since the COHFACE dataset contains two distinct illumination conditions, different experimental protocols have been devised. For this test however, we only consider studio conditions, i.e. when the face is well-lit by the spot (see Figure 3a). To generate training and test sets, the same approach as with the Manhob HCI-Tagging database has been applied: 24 subjects were selected to build the training set, and the 16 remaining subjects form the test set. Considering studio conditions only, this partitioning results in 48 sequences for the training set and 32 for the test set (there are 2 sequences for each subject).

Table 5: Pearson's correlation coefficient for the *studio* protocol on different sets of the COHFACE database.

	Training set	Test set
LiCVPR	-0.16	-0.61
CHROM	**0.23**	**0.43**
2SR	0.07	-0.31

Results in Table 5 should be compared to the ones reported in Table 2: sequences are approximatively of the same length and recorded in clean illumination conditions. In this case however, obtained results are much worse for two out of the three studied algorithms. On this database, they prove to be incapable at inferring the heart-rate reliably, since their Pearson's correlation coefficient is close to zero and even negative in the case of LiCVPR. This result, although suprising, is interesting: it emphasizes the sensitivity of these approaches to different acquisition conditions. Indeed, the COHFACE database has both lower resolution and framerate. On the other hand, the performance of the CHROM algorithm remains consistent across these different datasets. An interesting point to note is that it actually performs better on the test set than on the training set. This could mean that its performance is less sensitive to its parameters, which is a nice feature: it suggests a good generalization ability to unseen video sequences. However, it should be noted that optimal parameters for the training set of the COHFACE database are generally not the same as the one used in the very first experiment.

4.3. Evaluation on umatched conditions

An interesting and realistic scenario is to use rPPG algorithms across a wider range of aquisition conditions. Think, for instance, of a usage on a mobile device in different rooms, where the lighting conditions are not the same. It would be an additional burden to readjust different parameters each time there is a change in the environment. Hence it is important to assess the behaviour of the different approaches in various settings. Such cases have been emulated with two different experiments presented below.

4.3.1. Cross database

Although the two considered databases were partly recorded in quite similar conditions (indoor, controlled illumination), it is of interest to test the behaviour of the different approaches in a cross-database setting. For this purpose, we first tune the algorithms on the training set of the Manhob HCI-Tagging database and test them on the studio test set of the COHFACE database. For the second experiments, algorithms are optimized on the studio training set of the COHFACE database and tested on the test set of the Manhob HCI-Tagging database.

Results reported in Table 6 shows an interesting behaviour. Indeed, when algorithms are tuned on the Manhob HCI-Tagging database, their performance on the test set of the COHFACE database is better than when tuned with the same database. However, results for both LiCVPR and 2SR are still not satisfactory. While such a behaviour may seem surprising, it only means

Table 6: Pearson's correlation coefficient in cross-database testing.

	LiCVPR	CHROM	2SR
HCI → COHFACE	-0.20	**0.51**	-0.15
COHFACE → HCI	**0.25**	0.10	0.00

that optimal parameters on the training set of Manhob HCI-Tagging also suit well the data in the COHFACE studio test set. This can be due to the wider variety of sequences to tune parameters, since the training set of the Manhob HCI-Tagging database contains much more sequences than the training set of the COHFACE database. In the second experiment however, one can see that performance is generally worse than the one reported in the second column of Table 4, as expected. In this case, the limited amount of sequences in the COHFACE studio training set prevents to find optimal parameters for the more generic conditions in the test set of the Manhob HCI-Tagging database. Again, these experiments underline the extreme sensitivity of the different approaches to their respective parameters. Note finally that the performances of CHROM on both test sets are comparable, suggesting its higher stability to different configurations.

4.3.2. Illumination conditions

Since the COHFACE database contains two distinct illumination conditions (see Figure 3), one is now able to test the generalization of rPPG algorithms to different illumination conditions. The number of recordings under natural lighting is the same as in well-lit conditions, meaning that the natural test set contains 16 subjects and 32 video sequences. To test the behaviour in mismatched conditions, parameters were tuned on the studio training set, and applied to the natural test set. Performance in unmatched illumination conditions are reported in Table 7.

Table 7: Pearson's correlation coefficient on the natural test set of the COHFACE database.

	LiCVPR	CHROM	2SR
Studio → Natural	-0.24	-0.03	**0.00**

As expected, obtained results generally show a drop in performance as compared to the test set in studio conditions (Table 5). This denotes again

the need for a careful selection of the parameters, and suggest that none of the selected algorithms are suitable to achieve acceptable performance in realistic scenarios, where the acquisition conditions are unknown *a priori*.

4.4. Importance of the ROI

In the next experiments, we consider the whole COHFACE dataset, with both lighting conditions. Doing so will allow to get insights on the performances in a more generic setting in terms of illumination, corresponding to more realistic conditions. Using the full dataset, there are now 96 sequences in the training set, and 64 in the test set. Note that both sets contain sequences recorded under the two different lightings, there is no mismatch between training and testing, but rather a wider range of variability.

Since all investigated rPPG algorithms first extract skin pixels, it is interesting to see which method is best suited to perform this task. As a baseline, the complete face bounding box is used. We then compare the different algorithms when the skin pixels are either extracted using a mask on the bottom part of the face (see Figure 6), or using the skin color filter [13]. Performance using different ROI to extract skin pixels are presented in Table 8.

Table 8: Pearson's correlation coefficient for different skin color ROI on the test set of the COHFACE database.

	Face bounding box	Skin pixels	Mask
LiCVPR	-0.29	-0.16	-0.44
CHROM	**0.30**	**0.27**	0.30
2SR	0.26	0.09	**0.65**

In this case, the performance of the CHROM algorithm remains stable and comparable to the one obtained on the studio dataset (see Table 5), no matter which ROI is considered. This could be explained by the fact that the mean color is used, which should not vary much across different ROIs. This also shows that when parameters are tuned using conditions known *a priori*, this algorithm is able to cope with variability in illumination. It is interesting to see that the 2SR algorithm performs better when the mask is considered. Indeed, one could expect that it would work best when the color subspace of the considered skin pixels are compact in the RGB space, since it has been designed this way. This result suggests that with this algorithm the

ROI has a tremendous influence on the results, as evidenced by the variation in performance across the different regions. Finally, one should notice that LiCVPR's algorithm does not achieve acceptable performance no matter the ROI. This shows that this algorithm is very sensitive to variations in terms of both image resolution and illumination conditions.

5. Conclusion

In this contribution, three state-of-the-art rPPG algorithms were selected and evaluated. For this purpose, a new, publicly available database containing 40 subjects captured under two different illumination conditions has been introduced. A thorough experimental evaluation of the selected approaches has been conducted using different datasets and their associated protocols. Our reproducible research framework allows assessing performance in a principled and unbiased way. Obtained results show that only one rPPG algorithm has a stable behaviour, but overall it has been noticed that performance is highly dependent on a careful optimization of parameters. Conducted experiments also shows that generalization across conditions (i.e. resolution, illumination) should be of high concern when assessing rPPG approaches. The data, the experimental protocols and the implementation of the algorithms used in this study have been made available as open source free software[4]. We hope that this will help standardizing the comparison of remote heart rate measurement algorithms and advance the progress in this field.

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3451, 2013.

---

[4]https://pypi.python.org/pypi/bob.rppg.base

[2] C. Atanasoaei. *Multivariate Boosting with Look-up Tables for Face Processing*. PhD thesis, Ecole Polytechnique Federale de Lausanne (EPFL), 2012.

[3] G. de Haan and V. Jeanne. Robust Pulse Rate From Chrominance Based rPPG. *IEEE Trans. On Biomedical Engineering*, 60(10):2878–2886, 2013.

[4] G. Heusch, A. Anjos, and S Marcel. A Python Package for Remote Photoplesthymography. *SoftwareX, in press*, 2017.

[5] M. Lewandowska, J. Ruminski, T. Kocejko, and J. Nowak. Measuring Pulse Rate with a Webcam - a Non-contact method for Evaluating Cardiac Activity. In *Proceedings Federated Conference on Computer Science and Information Systems*, pages 405–410, 2011.

[6] X. Li, J. Chen, G. Zhao, and M. Pietikainen. Remote Heart Rate Measurement From Face Videos Under Realistic Situations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] D.J. McDuff, J.R. Estepp, A.M. Piasecki, and E.B. Blackford. A survey of remote optical photoplethysmographic imaging methods. In *IEEE Intl Conf. of the Engineering in Medicine and Biology Society (EMBC)*, pages 6398–6404, 2015.

[8] J. Pan and W. J. Tompkins. A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985.

[9] M. Poh, D. McDuff, and R. Picard. Non-contact, Automated Cardiac Pulse Measurements using Video Imaging and Blind Source Separation. *Optics Express*, 18(10), 2010.

[10] J. Shi and C. Tomasi. Good features to track. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[11] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Trans. On Affective Computing*, 3(1):42–55, 2012.

[12] M.P. Tarvainen, P.O. Ranta-aho, and P.A. Karjalainen. An Advanced Detrending Method with Application to HRV Analysis. *IEEE Transactions on Biomedical Engineering*, 49(2):172–175, 2002.

[13] M.J. Taylor and T Morris. Adaptive skin segmentation via feature-based face detection. In *SPIE Proceedings, Real-Time Image and Video Processing*, volume 9139, 2014.

[14] W. Verkruysse, L. Svaasand, and J. Nelson. Remote Plethysmographic Imaging Using Ambient Light. *Optics Express*, 16(26):21434–21445, 2008.

[15] W. Wang, S. Stuijk, and G. de Haan. A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation. *IEEE Transactions on Biomedical Engineering*, 2015.