applied sciences

MDPI

Article

Performance Evaluation of rPPG Approaches with and without the Region-of-Interest Localization Step

Žan Pirnar, Miha Finžgar * and Primož Podržaj

Faculty of Mechanical Engineering, University of Ljubljana, Aškerčeva cesta 6, 1000 Ljubljana, Slovenia; zan.pirnar@fs.uni-lj.si (Ž.P.); primoz.podrzaj@fs.uni-lj.si (P.P.)
* Correspondence: miha.finzgar@fs.uni-lj.si

Abstract: Traditionally, the first step in physiological measurements based on remote photoplethysmography (rPPG) is localizing the region of interest (ROI) that contains a desired pulsatile information. Recently, approaches that do not require this step have been proposed. The purpose of this study was to evaluate the performance of selected approaches with and without ROI localization step in rPPG signal extraction. The Viola-Jones face detector and Kanade–Lucas–Tomasi tracker (VK) in combination with (a) a region-of-interest (ROI) cropping, (b) facial landmarks, (c) skin-color segmentation, and (d) skin detection based on maximization of mutual information and an approach without ROI localization step (Full Video Pulse (FVP)) were studied. Final rPPG signals were extracted using selected model-based and data-driven rPPG algorithms. The performance of the approaches was tested on three publicly available data sets offering compressed and uncompressed video recordings covering various scenarios. The success rates of pulse waveform signal extraction range from 88.37% (VK with skin-color segmentation) to 100% (FVP). In challenging scenarios (skin tone, lighting conditions, exercise), there were no statistically significant differences between the studied approaches in terms of SNR. The best overall performance in terms of RMSE was achieved by a combination of VK with ROI cropping and the model-based rPPG algorithm. Results indicate that the selection of the ROI localization approach does not significantly affect rPPG measurements if combined with a robust algorithm for rPPG signal extraction.

Keywords: remote photoplethysmography; vital signs monitoring; remote sensing

Citation: Pirnar, Ž.; Finžgar, M.; Podržaj, P. Performance Evaluation of rPPG Approaches with and without the Region-of-Interest Localization Step. *Appl. Sci.* **2021**, *11*, 3467. https://doi.org/10.3390/app11083467

Academic Editor: Yannick Benezeth

Received: 16 March 2021
Accepted: 9 April 2021
Published: 13 April 2021

1. Introduction

Remote photoplethysmography (rPPG) is an optical measurement technique for detecting minute blood volume variations in cutaneous microcirculation using a digital camera [1]. It allows non-contact measurements of various physiological parameters: pulse rate (PR) and its variability, blood pressure, pulse transit time, etc.

Traditionally, rPPG approaches first require localization of the region of interest (ROI) [2,3]. The most basic solution is a manual definition of ROI [1,4]; however, the most commonly used are automated ROI localization approaches, such as (1) combination of face detection, tracking and skin mask refinement (the latter being optional) and (2) living-skin detection [5,6]. In the first approach, Viola-Jones frontal face detector has been suggested [7] for detecting faces in video recordings and has been widely used ever since in rPPG studies [8]. Detected ROI is then tracked throughout the entire recording using a dedicated algorithm, most commonly Kanade–Lucas–Tomasi (KLT) tracker. An optional step of skin mask refinement (i.e., ROI refinement) can be carried out in several ways, e.g., by resizing bounding box containing a detected face [7,9], by skin-color segmentation in various color spaces (RGB [10], RGB-H-CbCr [11], YCbCr [12]), by extracting selected facial landmarks [13,14], or by applying the skin detector based on maximization of mutual information [15]. The described automated approach is, however, only applicable in the case of facial video recordings, and its performance depends on the presence of motion artifacts

and the appearance of face in the recordings. In the second approach, video recordings are first segmented into several regions, from which rPPG pulse waveform signals are extracted. Next, based on the characteristics of rPPG signals, pulse and noise are differentiated from the extracted signals, and finally, the regions containing pulse information are labeled as skin regions [5]. Drawbacks of this approach are its high computational complexity, the presence of the causality dilemma caused by the interdependence of the skin detection and pulse waveform signal extraction, and the requirement for spatio-temporally coherent local segmentation [16].

Recently, new approaches have been proposed that do not require a ROI localization step [16,17]—these are the approaches with no image processing front-end [17]. The idea behind them is that, since we are interested only in the rPPG pulse waveform signal and not in the ROIs and their location specifics, we can treat a camera as a single rPPG sensor [16]. In the Full Video Pulse extraction (FVP) approach, a temporal mean of a color signal is used as a feature to differentiate between the image background and foreground (i.e., skin pixels) by first producing a set of weighting masks, then extracting rPPG signals from each mask, and finally combining the signals into a single pulse waveform signal [16]. Another approach is single-element rPPG, in which the average pixel intensity values of the full frames are extracted from the entire video recording [17]. In contrast to FVP, the existing skin reflection model-based rPPG algorithms showing the best general performance exhibit a significant performance drop in extracting target pulse waveform signal when applied in single-element rPPG. Therefore, this approach should be used with a dedicated rPPG algorithm, e.g., SoftSig [17]. SoftSig consists of two steps, a projection step and a selection step, in which pulse waveform signals candidates are extracted and the best candidate rPPG signal (in terms of predefined quality metrics) is selected, respectively [17].

The selection of the image processing front-end approach in rPPG strongly affects the signal-to-noise ratio (SNR) of the extracted rPPG pulse waveform signal [8,18–20]. Li et al. [18] evaluated the performance of seven different image processing front-end approaches (in combination with model-based CHROM rPPG algorithm [21]) in PR extraction from the publicly available Univ. Bourgogne Franche-Comté Remote Photoplethysmography (UBFC-RPPG) data set [6]. The studied approaches included Viola-Jones frontal face detector with KLT tracker and different ROI refinement algorithms: (1) no refinement, (2) ROI cropping [7], (3) RGB-based skin-color detection [10], (4) skin detection based on the maximization of mutual information [15], (5) adaptive RGB-based skin-color detection, (6) graph-cut based skin segmentation [22] and (7) facial landmarks segmentation [14]. The worst result was obtained for Viola-Jones face detection algorithm with KLT tracking with no ROI refinement, whereas the overall best performance was reached when skin detection based on the maximization of mutual information was applied for refining ROIs [15]. It is to be noted that the studied video recordings covered resting and mathematical game playing scenarios only. Fouad et al. [19] reported that Viola-Jones outperforms Normalized-Pixel-Difference-based face detector proposed by Liao et al. [23], KLT outperforms Camshift [24], and skin detector based on the maximization of mutual information [15] outperforms RGB-H-CbCr [11] in terms of root mean square error (RMSE) of average PRs. The authors used the UBFC-RPPG data set [6] and extracted the pulse waveform signals using blind-source-separation-based rPPG algorithm FastICA [25]. Later, Li et al. [20] additionally assessed the performance of selected approaches with different ROI localization steps in rPPG-based PR measurement: Viola-Jones face detector [26] and KLT [27], in combination with (1) no additional algorithms, (2) skin detector based on the maximization of mutual information [15], and (3) facial landmarks detector [14]. The best overall performance, in terms of correlation coefficient, percentage of PR values below 2.5 and 5 BPM errors, mean absolute error (MAE), RMSE, and SNR, was achieved by the combination of Viola-Jones, KLT and skin segmentation, and the worst by the combination of Viola-Jones and KLT. The studied data set was UBFC-RPPG [6], and CHROM [21] was used for the final rPPG signal extraction. Zhao et al. [8] studied performance of four different face detection and tracking algorithms applied in rPPG: (1) landmarks detection [14]

on all frames and (2) landmarks detection on the first frame only in combination with (a) KLT tracker, (b) Circulant Structure with Kernels (CSK) tracker [28], and (c) Sum of Template and Pixel-Wise Learners (STAPLE) tracker [29]. Two different publicly available data sets (PURE [30], UBFC-RPPG [6]) and one private data set (Self-RPPG) were used and Plane-Orthogonal-To-Skin (POS) [31] algorithm was applied for final rPPG signal extraction. The best overall result (in terms of SNR, MAE, and RMSE) for PURE was achieved by landmarks detection on each frame, for UBFC-RPPG by an approach relying on KLT tracking, and for Self-RPPG by the approach relying on CSK tracking. Recently, Woyczyk et al. [32] indirectly studied the performance of the following ROI localization approaches: (1) detecting a face on the first frame using Viola-Jones frontal face detector and keeping the position of the bounding box fixed throughout the entire recording, (2) detecting a face on the first frame using Viola-Jones frontal face detector and then tracking it using KLT tracker, and (3) the same approach as the previous one with an addition of statistical skin classifier proposed by Jones and Rehg [33]. Their performance in terms of accuracy and MAE was tested on two data sets containing a total of 1000 ten-seconds-long, lossy compressed video recordings recorded under challenging illumination conditions. The authors reported that, in the case of using the green channel signal as the final pulse waveform signal, the best overall result was achieved by the first approach (accuracy – defined as the ratio between the number of true positives [i.e., PRs within the range of $\pm$5 BPM around the reference PR] – of 0.1709 and MAE of 23.53 BPM), in the case of applying CHROM [21] rPPG algorithm, by the third approach (accuracy of 0.3411 and MAE of 16.77 BPM) and in the case of applying POS [31] algorithm, by the second approach (accuracy of 0.3455 and 15.95 BPM).

The purpose of this work is to study the effect of approaches with and without the ROI localization step (i.e., with and without the image processing front-end), on the (1) success rate of extracting rPPG signals from video recordings, (2) SNR and (3) RMSE of average PRs estimated from video recordings. Performance evaluation is to be carried out on the publicly available data sets offering uncompressed and compressed (both lossy and lossless) videos covering various scenarios. Extraction of final rPPG signals will be extracted using two state-of-the-art rPPG algorithms relying on distinct principles and differing in sensitivity to the presence of non-skin pixels within ROI: model-based POS [31] and non-model-based (i.e., data-driven) Spatial Subspace Rotation (2SR) [34].

2. Materials and Methods

2.1. Data Sets

We used three publicly available data sets in our study: *Public Benchmark Dataset for Testing rPPG Algorithm Performance* (hereinafter PBDTrPPG) [35], *LGI-PPGI-Face-Video-Database* (LGI-PPGI-FVD) [36] and *Pulse Rate Detection Dataset* (PURE) [30].

PURE [30] consists of 59 lossless compressed one-minute-long facial video recordings (recorded at 30 fps and 640 $\times$ 480 pixels per frame) of ten subjects (eight males and two females) in sitting position during six different controlled scenarios (resting with no intended head motion, talking, slow head translation of 7% of the face height per second, fast head translation of 14% of the face height per second, small head rotation of approximately 20 deg, and large head rotation of approximately 35 deg) and reference pulse waveform signals recorded with pulse oximeter at the sampling rate of 60 Hz. We used all video recordings from the data set and divided them into four groups: resting (10 recordings), talking (9 recordings), head translation (20 recordings), and head rotation (20 recordings).

LGI-PPGI-FVD [36] consists of 100 uncompressed video recordings (recorded at 25 fps and 640 $\times$ 480 pixels per frame) of 25 subjects (20 males and 5 females) during four different sessions (resting with no intended head motion, head and facial motion, exercising on a cycle ergometer in a gym with no set restrictions, and talking in a real-world urban environment) together with reference pulse waveform signals recorded with a pulse oximeter at the sampling rate of 60 Hz. Gym recordings are five minutes long, whereas the length of other recordings is approximately one minute. When it comes to actual

public availability of the data, only 24 recordings of six subjects are available. We used only the recordings covering resting (six recordings) and gym (six recordings) scenarios. By selecting the gym recordings, we covered a challenging scenario of uncontrolled motion in a real-world environment (uncontrolled lighting, presence of multiple subjects), which is not present in the other two data sets.

PBDTrPPG consists of lossy compressed (H.264 codec) facial video recordings (recorded at 30 fps and 1080 × 1920 pixels per frame) of three male subjects during three challenging scenarios (lighting variation in combination with different skin tones, motion scenarios, and resting after exercise, causing significant changes in PR) and reference ECG recordings (recorded at 1024 Hz). Only videos covering five different lighting conditions and three different skin tones (total of 15 recordings; two additional lighting scenarios with only one recording per each scenario were excluded) were included in our study from this data set. We excluded motion and resting after exercise scenarios because there are only three and one video(s) available, respectively. In addition, the type of motion scenarios is equivalent to that from the PURE data set, so inclusion of these recordings would not represent a major contribution to our study.

With the selected data sets we included uncompressed and compressed videos, recordings made in controlled and uncontrolled environments and covering several different scenarios (rest, talking, different types of controlled and uncontrolled motion) and challenges (lighting variation and various skin tones).

2.2. Studied Approaches with and without ROI Localization Step

We applied five different approaches for PR extraction from video recordings—four with the ROI localization step and one without it:

- Viola-Jones [26] frontal face detector + ROI width reduction to 60% of its original width [7] + KLT tracker [27,37] (hereinafter VK),
- Viola-Jones [26] frontal face detector + facial landmarks detector [13] + KLT tracker [27,37] (hereinafter VK-LMK),
- Viola-Jones [26] frontal face detector + RGB-H-CbCr skin-color segmentation method [11] + KLT tracker [27,37] (hereinafter VK-RGBHCbCr),
- Viola-Jones [26] frontal face detector + skin detector on the maximization of mutual information [15] + KLT tracker [27,37] (hereinafter VK-Conaire) and
- Full Video Pulse Extraction [16] (hereinafter FVP), which is by default combined with the POS rPPG algorithm.

The basis for all approaches with an image processing front-end is a combination of Viola-Jones frontal face detector and KLT tracker due to its predominant application in rPPG studies. Viola-Jones frontal face detector was applied on the first frame of each recording. In the case more than one face was detected, the largest bounding box was selected for further processing. From this step onwards, the methodology split in separate directions:

- In the VK approach, we reduced the width of the detected ROI to 60% of its original width [7].
- In VK-LMK we identified facial landmarks within the original (unresized) ROI containing a detected face using the Discriminative Response Map Fitting approach [13]. From the identified 66 landmarks, we used nine of them to define new ROI, which covered cheeks, nose, and mouth area [38].
- In VK-RGBHCbCr and VK-Conaire, we kept the original size of the detected ROI.

The KLT tracker was then initialized by identifying feature points within the original bounding box containing the detected face using Good Features to Track method [37] and then propagated throughout the entire recording. For a current frame, the tracker attempted to find the points from the previous frame and then estimate a transformation matrix consisting of the affinity parameters between the old points from the previous frame and new points from the current frame. Estimation of the transformation matrix was done using MSAC algorithm [39], a type of RANSAC algorithm, which is, in general,

used for robust estimation of parameters of a selected mathematical model. Once the transformation matrix was defined, it was used to transform the edges of the ROIs defined in the list above from a previous frame to a current one. In the case of VK and VK-LMK, the newly transformed ROI represented the final ROI that was used for further processing, whereas in the case of VK-RGBHCbCr and VK-Conaire, the ROI was refined using RGB-H-CbCr skin color segmentation model [11], and skin detector based on the maximization of mutual information [15] were applied, respectively. The presented steps were applied on each of the following frames. The only studied approach without the ROI localization step was FVP [16]. Its implementation followed that of its authors [16], with an exception that we did not remove non-pulsatile components from rPPG signals in order to ensure easier comparison with other studied approaches. The selection of the studied approaches with image processing front-end was based on the frequency of their use in research—Zhao et al. reported that the most cited detection and tracking algorithm in rPPG research are Viola-Jones face detector, facial landmarks, KLT, and skin detection [8].

2.3. Applied rPPG Algorithms for Pulse Waveform Signal Extraction

We selected two state-of-the-art rPPG algorithms, POS [31] and 2SR [34], to extract final pulse waveform signals from the information extracted from video recordings using the approaches presented in the previous subsection. In POS, raw RGB signals extracted from ROIs on each frame are projected to a plane that is orthogonal to temporally normalized skin and, therefore, exhibits large pulsatile and low specular strength [31]. Since POS is built on a physiological reasoning, it works well also when skin mask is noisy. In 2SR, spatial subspace of the pixels within the ROIs is defined, and the measurement of temporal rotation of the defined space is used for extracting a pulse waveform signal. In contrast to POS, the quality of SNR drops significantly when the percentage of non-skin-pixels falls within 10 to 30%. The reasoning behind the selection of the presented algorithms was based on their sensitivity to skin mask noise and their performance in comparison to other state-of-the-art rPPG algorithms [31,34].

2.4. Performance Evaluation of the Studied Approaches with and without ROI Localization Step

Performance of the studied approaches was assessed by three quality metrics: (1) success rate, (2) modified SNR metric proposed by de Haan and Jeanne [21], and (3) RMSE of the difference between reference- and rPPG-derived PRs.

Success rate was defined as a percentage of successfully processed video recordings among the total number of recordings for each studied approach. Whenever an approach failed (regardless of the processing stage at which a failure occurred), further processing was stopped and the processing of a given video recording was marked as unsuccessful.

SNR was calculated as the ratio between the energy within 5 bins around the first harmonic (or fundamental frequency) plus 11 bins around the second harmonic and the remaining energy within the frequency band of expected PRs in humans (i.e., [30, 240] BPM) (de Haan and Jeanne [21] took 10 bins around the second harmonic, but it is unclear, how the authors positioned this window relative to the second harmonic):

$$SNR = 10\log_{10}\left(\frac{\sum_{f=30}^{240}(w_t(f)S(f))^2)}{\sum_{f=30}^{240}(1-w_t(f)S(f))^2}\right) \quad (1)$$

where f denotes the PR frequency (expressed in BPM), w_t the applied binary window and $S(f)$ the power spectrum of rPPG pulse waveform signal. Fundamental frequency was defined as the peak frequency within the defined frequency band in the power spectrum of reference pulse waveform signals (obtained by Fast Fourier Transform). In the case a window around the second harmonic fell outside the set PR band, we extended the spectrum accordingly (this is the second modification of the SNR metric proposed by de Haan and Jeanne [21] that we applied).

Before calculating SNR, the reference signals were pre-processed by (1) applying first order Butterworth bandpass filter with lower and upper cut-off frequencies corresponding to 40 and 180 BPM, respectively, and (2) resampling to the sampling rate that matched the frame rate of the accompanying video recordings. The reason for applying bandpass filter with cut-off frequencies different from the upper- and lower frequencies in the frequency band corresponding to expected PRs in humans is the fact that we adopted the values of the actual data in order to achieve cleaner reference signals. SNR values were calculated for windowed signals (512-frame-long sliding window was applied). The results are visually presented in the form of boxplots if the size of each group within the studied scenario is larger than or equal to five, whereas strip plots are used otherwise.

RMSE was defined as the square root of the average of squared differences between the estimated ($PR_{\mathrm{rPPG_i}}$) and actual, i.e., reference, PR values ($PR_{\mathrm{ref_i}}$):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(PR_{\mathrm{rPPG_i}} - PR_{\mathrm{ref_i}})^2}, \quad (2)$$

where n denotes the total number of estimated PRs. PR values were measured using a sliding window of 512 frames; we took into the account that, in general, a PR is not constant throughout the entire recording, not even in a resting condition.

Statistical evaluation of the results was carried out using one-way analysis of variance (ANOVA). ANOVA was used in the cases when number of samples in each studied group was larger than or equal to five.

We set the length of the sliding processing window to 128 frames for POS, 60 frames for 2SR and 128 frames for the FVP approach. The selection of the window lengths are partially adjusted to each algorithm to ensure their reliable performance: FVP tends to perform worse at shorter window lengths and is more sensitive to window size if compared to the approaches with image processing front-end [16]; in 2SR, optimal window length should include at least a half of the cardiac cycle, so its selection depends on the camera frame-rate and PR of the subject [34]; in POS, similarly as in 2SR, window length should capture at least one cardiac cycle (shorter window lengths are preferred so that instantaneous distortions can be suppressed as quickly as possible [31]).

3. Results

Studied approaches failed several times due to the following reasons: (1) no face was detected on the first frame of the recording, (2) KLT tracker lost all the feature points, (3) no skin pixels were detected using the RGB-H-CbCr skin-color detector, and (4) no skin pixels were detected using skin detector based on maximization of mutual information. Table 1 lists the recordings in which one of the presented issues occurred; it can be seen that Viola-Jones frontal face detector failed in some gym recordings, KLT tracker failed (i.e., lost all tracked points) in one gym recording (in this particular recording, the subject was moving his head very fast and at some point his face was occluded), skin-color-based segmentation failed in at least one of the videos from each studied data set, and the skin detector based on maximization of mutual information failed in one gym recording only.

Table 1. The identified issues during ROI localization step using the studied approaches and the IDs of the recordings in which the issues occurred. VK-Conaire: Viola-Jones combined with skin detector based on the maximization ofmutual information and Kanade–Lucas–Tomasi tracker; VK-RGBHCbCr: Viola-Jones combined with RGB-H–CbCr skin-color segmentation and Kanade–Lucas–Tomasi tracker; KLT: Kanade–Lucas–Tomasi.

Issue/Data Set	PURE	PBDTrPPG	LGI-PPGI-FVD
no detected face on the first frame	/	/	*alex_gym, angelo_gym, david_gym*
KLT tracker failure	/	/	*felix_gym*
no skin pixels detected (using VK-RGBHCbCr)	03-01, 03-04, 04-02, 05-04	P3LC1	*harun_gym*
no skin pixels detected (using VK-Conaire)	/	/	*harun_gym*

The reasoning behind unsuccessful skin detection is due to the limited training set applied for determining the threshold values in both skin detection algorithms, whereas in the case of Viola-Jones frontal face detector failure, the appearance of faces on the first frames was the reason for unsuccessful detection. For each occurrence of the presented issues, we discarded the obtained results and marked the processing of the particular video recording as unsuccessful. Based on the number of unsuccessfully processed recordings, we defined the success rates of the studied approaches: 88.37% of VK-RGBHCbCr, 94.19% of VK-Conaire, 95.35% of VK and VK-LMK, and 100% of FVP.

Figure 1 shows the examples of successfully localized ROIs for a selected recording from each of the studied data sets using four different approaches with an image-processing front-end.

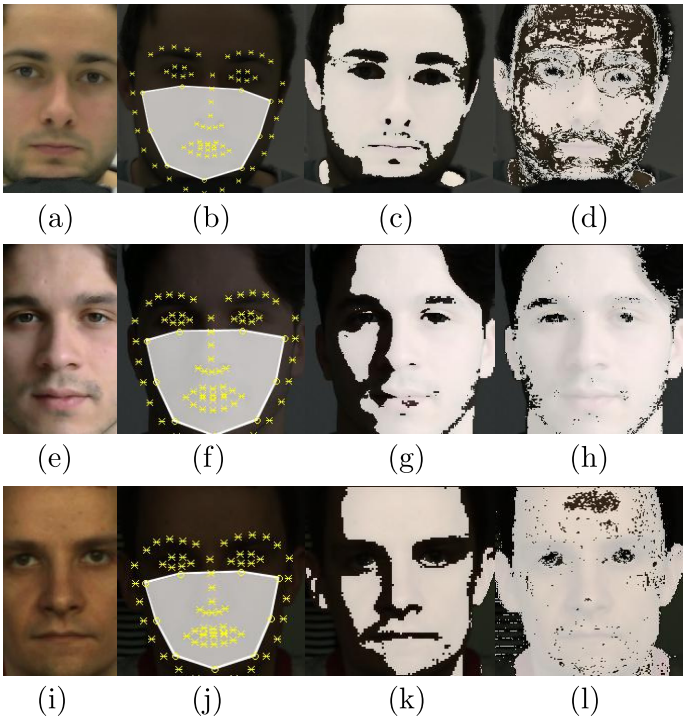

Figure 1. Results of the region of interest (ROI) localization on the selected recordings from studied data sets: recording P1LC3 from PBDTrPPG using (**a**) VK, (**b**) VK-LMK, (**c**) VK-Conaire, (**d**) VK-RGBHCbCr; recording *angelo_resting* from LGI–PPGI–FVD using (**e**) VK, (**f**) VK-LMK, (**g**) VK-Conaire, (**h**) VK-RGBHCbCr; recording 01-01 from PURE (**i**) VK, (**j**) VK-LMK, (**k**) VK-Conaire, (**l**) VK-RGBHCbCr. VK: Viola-Jones combined with ROI width reduction and Kanade–Lucas–Tomasi tracker; VK-Conaire: Viola-Jones combined with skin detector based on the maximization of mutual information and Kanade–Lucas–Tomasi tracker; VK-LMK: Viola-Jones combined with landmarks detection and Kanade–Lucas–Tomasi tracker; VK-RGBHCbCr: Viola-Jones combined with RGB–H–CbCr skin-color segmentation and Kanade–Lucas–Tomasi tracker; FVP: Full Video Pulse Extraction rPPG algorithm.

The sizes of ROIs are shown in Table 2. Mean absolute ROI sizes are, in general, the largest for videos from LGI–PPGI–FVD and the smallest for videos from PURE (differences in sizes are more than two-fold); mean relative ROI sizes range from 4.9 to 13.2% of the total frame size for LGI–PPGI–FVD recordings, from 3.0 to 6.5% for PURE and from 0.6 to 1.4% for PBDT–rPPG. The differences arise from different distances between the subjects and the camera, as well as different video frame sizes between the three studied data sets. Note that FVP does not rely on a common ROI localization step; therefore it is not included in Table 2.

Table 2. ROI sizes (in pixels) expressed as a mean ± standard deviation for different image-processing front-ends and all studied data sets. VK: Viola-Jones combined with ROI width reduction and Kanade–Lucas–Tomasi tracker; VK-Conaire: Viola-Jones combined with skin detector based on the maximization of mutual information and Kanade–Lucas–Tomasi tracker; VK-LMK: Viola-Jones combined with landmarks detection and Kanade–Lucas–Tomasi tracker; VK-RGBHCbCr: Viola-Jones combined with RGB–H–CbCr skin-color segmentation and Kanade–Lucas–Tomasi tracker; FVP: Full Video Pulse Extraction rPPG algorithm.

	VK	**VK-LMK**	**VK-RGBHCbCr**	**VK-Conaire**
PURE	19,063 ± 3581	9114 ± 1591	13,407 ± 5339	19,984 ± 5252
PBDTrPPG	28,849 ± 14,196	12,731 ± 4498	19,514 ± 16,015	19,128 ± 9699
LGI-PPGI-FVD	38,060 ± 13,863	14,962 ± 7106	32,916 ± 12,141	40,550 ± 15,970

Figure 2 shows the performance of studied approaches in combination with two rPPG algorithms (POS and 2SR) in extracting rPPG signals from video recordings from PURE in terms of SNR.

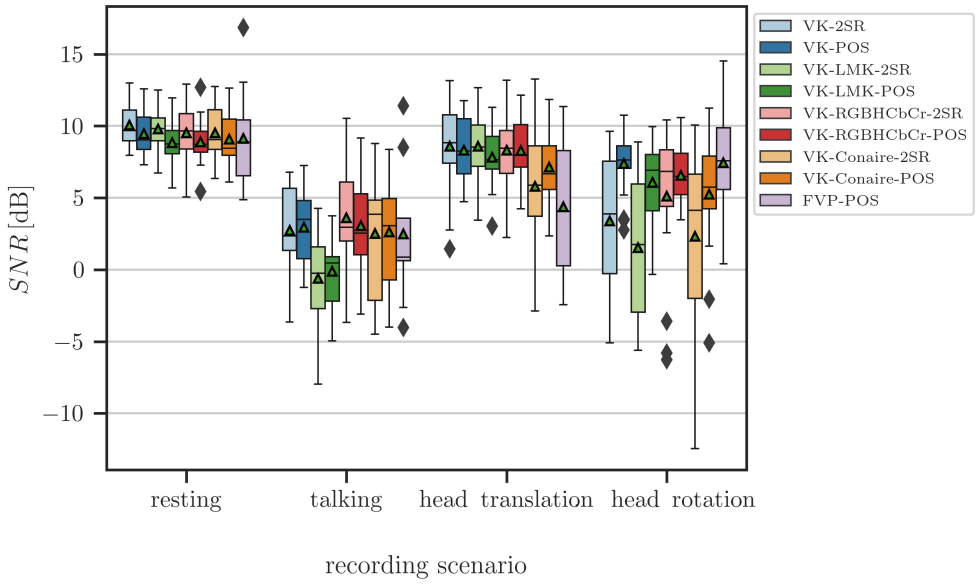

Figure 2. Performance of the studied approaches with and without ROI localization step in combination with two rPPG algorithms (POS and 2SR) in terms of SNR for PURE. VK: Viola-Jones combined with ROI width reduction and Kanade-Lucas-Tomasi tracker; VK-Conaire: Viola-Jones combined with skin detector based on the maximization of mutual information and Kanade-Lucas-Tomasi tracker; VK-LMK: Viola-Jones combined with landmarks detection and Kanade-Lucas-Tomasi tracker; VK-RGBHCbCr: Viola-Jones combined with RGB-H-CbCr skin-color segmentation and Kanade-Lucas-Tomasi tracker; FVP: Full Video Pulse Extraction rPPG algorithm.

Black horizontal lines denote median SNR values, upper and bottom sides of the boxes the 25th and 75th percentiles, and whiskers extend to the values not considered as outliers, which are marked with black diamonds; green triangles denote mean SNR values.

The largest SNR values are achieved for the resting scenario, followed by the head translation, the head rotation and the talking scenarios. For the resting and talking scenario, one-way ANOVA showed no statistically significant differences (at $\alpha = 0.05$) of mean SNR values between all pairs of studied approaches ($p = 0.954$ and $p = 0.260$). The boxplot of FVP and VK-RGBhCbCr SNR values exhibits the widest interquartile range for the resting and talking scenarios, respectively. For the head translation scenario, one-way ANOVA showed statistically significant differences in mean SNR values between FVP and (1) VK-POS ($p = 0.0013$), (2) VK-2SR ($p = 0.0004$), (3) VK-LMK-POS ($p = 0.0098$), (4)

VK-LMK-2SR ($p = 0.0004$), (5) VK-RGBHCbCR-POS ($p = 0.0023$), (6) VK-RGBHCbCR-2SR ($p < 0.0020$). The boxplot of VK-Conaire-2SR SNR values exhibits the widest interquartile range for this recording scenario. For the head rotation scenario, one-way ANOVA showed statistically significant differences in mean SNR values between VK-POS and VK-LMK-2SR ($p = 0.0002$), VK-POS and VK-Conaire-2SR ($p = 0.0029$), VK-2SR and FVP ($p = 0.0427$), VK-LMK-POS and VK-LMK-2SR ($p = 0.0123$), VK-LMK-2SR and VK-RGBHCbCr-POS ($p = 0.0029$), VK-LMK-2SR and FVP ($p = 0.0001$), VK-RGBHCbCr-POS and VK-Conaire-2SR ($p = 0.0267$), and VK-Conaire-2SR and FVP ($p = 0.0022$). Similarly as in the previous recording scenario, the boxplot of VK-Conaire-2SR SNR values exhibits the widest interquartile range.

Figure 3 shows the SNR results of the studied approaches tested on PBDTrPPG.

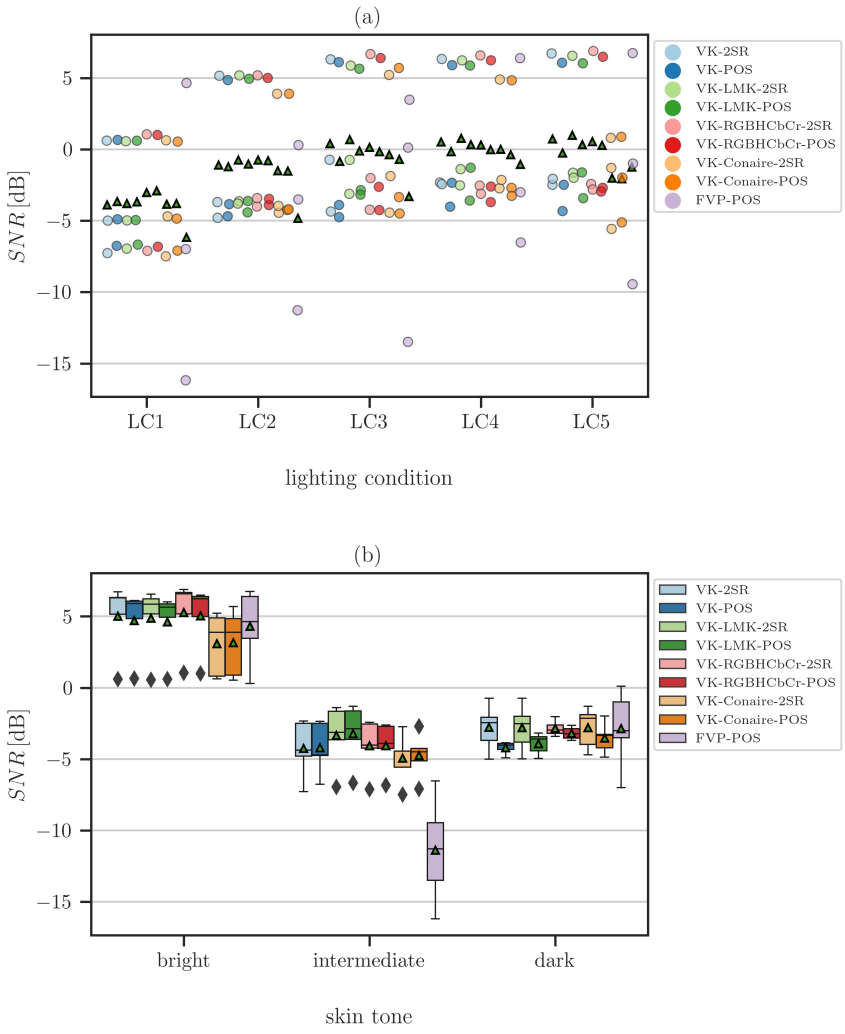

Figure 3. Performance of the studied approaches with and without image processing front-end in combination with two rPPG algorithms (POS and 2SR) in terms of SNR on the example of PBDTrPPG covering (**a**) different lighting conditions (LC1: 0.052×10^2 lux, LC2: 0.363×10^2 lux, LC3: 1.870×10^2 lux, LC4: 7.200×10^2 lux, LC5: 27.200×10^2 lux) and (**b**) skin tone challenges. VK: Viola-Jones combined with ROI width reduction and Kanade-Lucas-Tomasi tracker; VK-Conaire: Viola-Jones combined with skin detector based on the maximization of mutual information and Kanade–Lucas–Tomasi tracker; VK-LMK: Viola-Jones combined with landmarks detection and Kanade–Lucas–Tomasi tracker; VK-RGBHCbCr: Viola-Jones combined with RGB–H–CbCr skin-color segmentation and Kanade–Lucas–Tomasi tracker; FVP: Full Video Pulse Extraction rPPG algorithm.

Figure 3a shows that mean SNR values increase with an increasing lighting intensity in all studied approaches. Performances of various approaches are comparable except for FVP, which shows weaker performance at all lighting intensities, as well as VK-Conaire-2SR and VK-Conaire-POS at the largest lighting intensity. For the bright and dark skin tone recordings, the results of one-way ANOVA show no statistically significant differences in mean SNR values between all pairs of studied approaches ($p = 0.8032$ and $p = 0.6831$, respectively). For the intermediate skin tone recordings, there are statistically significant differences in mean SNR values between FVP and (1) VK-POS ($p = 0.0003$), (2) VK-2SR ($p = 0.0003$), (3) VK-LMK–POS ($p < 0.0001$), (4) VK-LMK-2SR ($p < 0.0001$), (5) VK-RGBHCbCR–POS ($p = 0.0002$), (6) VK-RGBHCbCR-2SR ($p = 0.0002$), (7) VK-Conaire-POS ($p = 0.0008$) and (8) VK-Conaire-2SR ($p = 0.0012$). In all the studied recordings from LGI–PPGI–FVD, FVP exhibits the widest interquartile range.

Figure 4 shows the performance of the studied methods for extracting rPPG signals from video recordings from LGI–PPGI–FVD in terms of SNR.

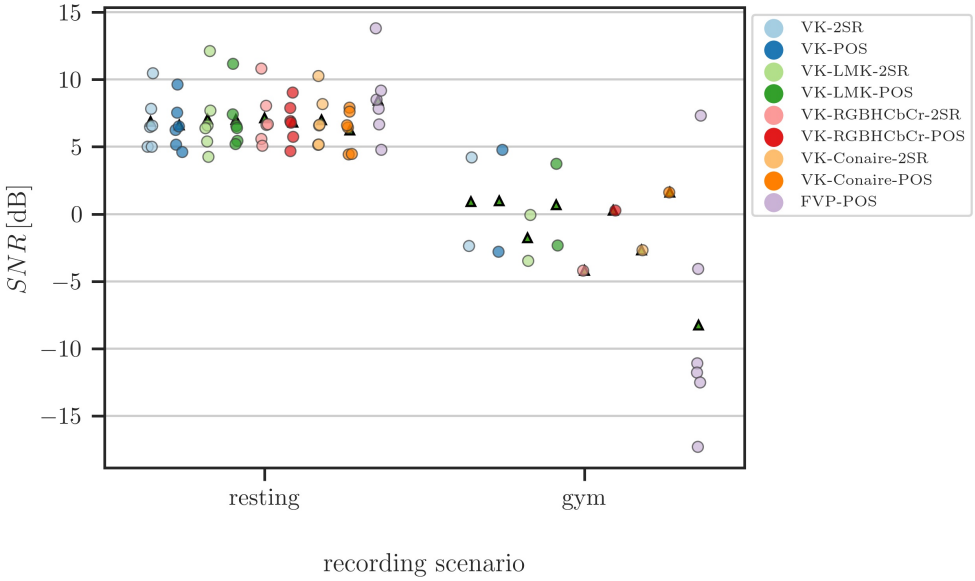

Figure 4. Performance of the studied approaches with and without image processing front-end in combination with two rPPG algorithms (POS and 2SR) in terms of SNR on the example of LGI–PPGI–FVD. VK: Viola-Jones combined with ROI width reduction and Kanade–Lucas–Tomasi tracker; VK-Conaire: Viola-Jones combined with skin detector based on the maximization of mutual information and Kanade–Lucas–Tomasi tracker; VK-LMK: Viola-Jones combined with landmarks detection and Kanade–Lucas–Tomasi tracker; VK-RGBHCbCr: Viola-Jones combined with RGB–H–CbCr skin-color segmentation and Kanade–Lucas–Tomasi tracker; FVP: Full Video Pulse Extraction rPPG algorithm.

There are no significant differences between the SNR values among the studied approaches in case of resting scenario, whereas in the case of gym recordings, FVP performed worse than the other approaches. However, it should be noted that in the case of the approaches with the image-processing front-end, four of the video recordings were not processed successfully. The SNR values of the rPPG signals extracted from these recordings are the ones that lower the mean SNR value for FVP.

Table 3 shows the results of the RMSE analysis. The results follow those of SNR, but the differences between the studied approaches seem to be more prominent. On average, the best performance is achieved by a combination of the VK approach and the SB rPPG algorithm. As expected, the RMSE is the largest in the case of low lighting conditions (LC1) and dark skin tone and the lowest in the case of resting conditions regardless of the applied approach. Similarly as in the case of SNR values, the worst performance was achieved by FVP.

Table 3. Performance of the studied approaches with and without image processing front-end in terms of RMSE (expressed as a mean ± standard deviation; measured in BPM) for PURE, PBDTrPPG and LGI–PPGI–FVD. The lowest RMSE values for each recording scenario are denoted in bold; the values in parentheses denote the number of samples in each group. VK: Viola-Jones combined with ROI width reduction and Kanade–Lucas–Tomasi tracker; VK-Conaire: Viola-Jones combined with skin detector based on the maximization of mutual information and Kanade–Lucas–Tomasi tracker; VK-LMK: Viola-Jones combined with landmarks detection and Kanade–Lucas–Tomasi tracker; VK-RGBHCbCr: Viola-Jones combined with RGB–H–CbCr skin-color segmentation and Kanade–Lucas–Tomasi tracker; FVP: Full Video Pulse Extraction rPPG algorithm.

	VK		**VK-LMK**		**VK-RGBHCbCr**		**VK-Conaire**		**FVP**
	SB	**2SR**	**SB**	**2SR**	**SB**	**2SR**	**SB**	**2SR**	**SB**
					PURE				
resting	**1.08** ± **0.71** (10)	1.11 ± 0.72 (10)	1.09 ± 0.74 (10)	1.18 ± 0.72 (10)	1.31 ± 0.91 (9)	1.5 ± 1.29 (9)	1.09 ± 0.75 (10)	1.13 ± 0.78 (10)	1.48 ± 0.86 (10)
talking	**4.09** ± **3.68** (9)	8.19 ± 15.62 (9)	11.08 ± 15.72 (9)	14.54 ± 20.24 (9)	8.68 ± 16.24 (8)	9.06 ± 17.59 (8)	8.48 ± 9.5 (9)	12.7 ± 16.58 (9)	7.95 ± 9.32 (9)
head translation	**1.07** ± **0.64** (20)	1.14 ± 0.75 (20)	1.13 ± 0.83 (20)	1.20 ± 0.87 (20)	1.13 ± 0.74 (18)	1.28 ± 0.8 (18)	1.14 ± 0.71 (20)	2.98 ± 4.18 (20)	3.96 ± 4.97 (20)
head rotation	0.88 ± 0.7 (20)	2.77 ± 3.94 (20)	1.06 ± 0.97 (20)	5.53 ± 6.62 (20)	**0.84** ± **0.62** (20)	5.41 ± 12.43 (20)	3.08 ± 6.29 (20)	9.45 ± 17.76 (20)	2.47 ± 5.73 (20)
					PBDTrPPG				
LC1 ($0.052 \cdot 10^2$ lux)	25.57 ± 12.25 (3)	25.23 ± 13.09 (3)	23.27±12.02 (3)	22.92 ± 11.4 (3)	**19.66** ± **11.69** (2)	19.85 ± 12.71 (2)	24.56 ± 12.37 (3)	23.89 ± 12.15 (3)	52.28 ± 6.65 (3)
LC2 ($0.363 \cdot 10^2$ lux)	22.69 ± 13.27 (3)	25.78 ± 16.04 (3)	21.47 ± 13.42 (3)	23.81 ± 16.56 (3)	20.47 ± 12.7 (3)	21.55 ± 12.59 (3)	**18.95** ± **11.4** (3)	24.79 ± 16.89 (3)	54.84 ± 9.41 (3)
LC3 ($1.870 \cdot 10^2$ lux)	**17.37** ± **10.36** (3)	30.53 ± 28.34 (3)	21.88 ± 22.58 (3)	27.15 ± 28.57 (3)	20.23 ± 12.79 (3)	22.39 ± 15.64 (3)	18.03 ± 11.19 (3)	23.26 ± 17.84 (3)	52.17 ± 17.90 (3)
LC4 ($7.200 \cdot 10^2$ lux)	**12.84** ± **9.61** (3)	17.36 ± 15.22 (3)	15.12 ± 13.1 (3)	16.65 ± 15.11 (3)	14.36 ± 10.67 (3)	13.03 ± 8.97 (3)	14.44 ± 9.65 (3)	17.82 ± 13.57 (3)	40.76 ± 26.12 (3)
LC5 ($27.200 \cdot 10^2$ lux)	**12.44** ± **9.62** (3)	20.96 ± 19.39 (3)	15.81 ± 16.58 (3)	19.43 ± 22.08 (3)	19.76 ± 14.65 (3)	17.2 ± 15.28 (3)	24.94 ± 15.35 (3)	27.90 ± 17.85 (3)	40.62 ± 29.02 (3)
bright tone	4.37 ± 1.98 (5)	4.16 ± 1.46 (5)	4.14 ± 1.63 (5)	4.28 ± 1.6 (5)	4.27 ± 1.88 (5)	4.24 ± 1.53 (5)	4.1 ± 1.55 (5)	**4.08** ± **1.47** (5)	32.02 ± 22.30 (5)
intermediate tone	21.13 ± 10.92 (5)	21.64 ± 10.51 (5)	16.59 ± 11.79 (5)	**16.11** ± **10.76** (5)	21.41 ± 7.05 (5)	20.47 ± 9.0 (5)	25.3 ± 8.61 (5)	26.17 ± 8.79 (5)	43.74 ± 3.15 (5)
dark tone	**29.05** ± **3.53** (5)	46.12 ± 12.79 (5)	37.8 ± 9.02 (5)	45.59 ± 13.34 (5)	33.85 ± 3.51 (4)	34.66 ± 6.08 (4)	31.14 ± 3.05 (5)	40.35 ± 6.07 (5)	68.65 ± 9.62 (5)
					LGI-PPGI-FVD				
resting	2.22 ± 1.3 (6)	2.20 ± 1.31 (6)	2.23 ± 1.30 (6)	2.21 ± 1.31 (6)	**2.08** ± **1.38** (6)	2.18 ± 1.32 (6)	2.20 ± 1.31 (6)	2.20 ± 1.32 (6)	2.24 ± 1.31 (6)
gym	13.9 ± 11.16 (2)	15.6 ± 12.89 (2)	17.81 ± 9.53 (2)	33.5 ± 4.35 (2)	**2.99** ± **0.0** (1)	31.84 ± 0.0 (1)	13.77 ± 0.0 (1)	29.6 ± 0.0 (1)	39.66 ± 17.22 (6)

4. Discussion

FVP is the only approach that allowed continuous PR measurement in all studied video recordings. It would, however, fail in the case of video recordings of multiple subjects, and its performance would deteriorate in the case of significant background changes. In other approaches, face detector algorithm or KLT tracker failed in at least one of the video recordings. In general, approaches relying on Viola-Jones frontal face detector would provide false positive results in the case of mannequins and fail completely if body parts other than face were covered on video recordings.

Mean SNR and RMSE values indicate that, in most of the recording scenarios, the performance of the studied approaches is comparable. This means that (1) the obtained ROIs were large enough to cancel out the camera quantization noise and (2) the effect of non-skin pixels did not significantly affect the performance of the studied approaches. When comparing the results of the approaches combined with POS with those combined with 2SR, the POS-based solutions seem to outperform the 2SR-ones, especially in the case of more challenging scenarios. This confirms the innate sensitivity of 2SR to noisy skin mask.

In addition, there are some observable differences between the studied approaches in some recording scenarios. In the talking-scenario recordings from PURE (Figure 2), the performance of all the approaches decreases in comparison to resting scenario due to noisier ROIs and varying intensities of specular and diffuse reflection components due to the motion of the facial pixels during talking. The worst performance is achieved by VK-LMK-2SR due to the largest ratio of non-skin to skin pixels (ROIs in VK-LMK were the smallest out of all ROIs), which arises when the mouth is opened during talking. A similar result is achieved in the head rotation scenario, in which we can also observe that 2SR performs worse than POS. In the case of PBDTrPPG (see Figure 3), which contains lossy compressed video recordings, average SNRs are the lowest out of all studied data sets. The achieved SNR values are comparable, and the selection of the approach with or without an image processing front-end together with an rPPG algorithm does not affect SNR value significantly. This is most likely due to partial loss of pulsatile information or introduction of additional artifacts into the recordings, as sugggested by Wang et al. [31]. Results show that SNR values increase as light intensity increases (Figure 3a) and decreases as skin tone decreases (Figure 3b), which is expected based on the knowledge underlying the formation of rPPG signal. It should be noted that results for different skin tones are not directly comparable, because of the variable distance between subjects and a camera. In addition, intermediate and dark skin tone would, based on our perception, lie close to each other on the Fitzpatrick scale, which could explain similar SNR values achieved. Interestingly, the best performance in case of a dark skin tone was achieved by FVP, which might be due to the fact that it relies not only on a mean but also on a variance for spatial pixel combination [31]. Lastly, the results of SNR values for LGI-PPGI-FVD (Figure 4) indicate similar performance of studied approaches in both resting and gym scenarios (except for FVP at gym recordings). In the case of the gym scenario, average SNR values are greatly reduced and only up to two video recordings from the set were successfully processed using the studied approaches with image processing front-end, whereas FVP allowed processing of all videos. It is to be noted that individual SNR values that lie below the mean SNR value for FVP correspond to the videos in which other approaches failed. If we removed these results, the SNR values of FVP would be comparable to those of the other approaches. Provided that the relative ROI sizes for recordings from LGI-PPGI-FVD were the largest, unevenly distributed pulsations [40] might also affect the results.

Our SNR values of the rPPG signals are lower in comparison with those provided by Zhao et al. [8], whereas RMSE values are lower in case of resting, talking and head translation scenarios. The authors studied the performance of several tracking algorithms (tracked ROI was a rectangular bounding box, corners of which were defined by selected facial landmark feature points). They reported the following SNR (RMSE) values for PURE video recordings: 13.26–13.42 dB (2.16–2.52 BPM) for resting scenario, 3.36–4.06 dB

(13.46–15.10 BPM) for talking scenario, 8.27–11.12 dB (2.12–3.19 BPM) for head translation and 7.05–10.46 dB (1.91–6.17 BPM) for head rotation. The results potentially indicate differences between the approaches for ROI localization, while it is important to emphasize that differences may occur due to different calculations of SNR. Although most of the studies rely on the SNR metric proposed by de Haan and Jeanne [21], there are differences in its actual calculation. Zhao et al. [8], for example, chose a frequency band of [48, 300] min^{-1} (instead of the proposed [30, 240] min^{-1}), included the energy around the third harmonic in the energy spectrum of the pulse signal (in the original implementation, only the energy around the first and the second harmonic are included), and used the same spectral window length for calculating energies around the first, second, and third harmonic (in [21], 5- and 10-bins-long windows around the first and second harmonic were used, respectively). In the case of calculating RMSE, the length of the processing window plays an important role when interpreting the results (differences of up to 10% may occur for the window lengths from 0 to 60 s [41]). The described issue results from the already exposed drawback of the rPPG research, i.e., the lack of standardized methodology [42]. There were even some attempts to define standardized report procedures for assessing rPPG PR measurements [43] with the goal to ensure direct comparison of the results of different studies, but there seems to be no general agreement on this issue. Compared to other similar studies, the worst performance in terms of SNR and RMSE is achieved by the Viola-Jones face detector with KLT tracker [18,20], which indicates that a simple original ROI width reduction applied in our study does improve the skin mask. However, on the other hand, an approach similar to ours was applied by Fouad et al. [19], but again, the obtained RMSE was more than two times lower than in other studied approaches.

The key advantages of our study are (1) the inclusion of lossy compressed, lossless compressed, and uncompressed videos from publicly available data sets covering various recording scenarios and (2) the inclusion of an approach without the image-processing front-end. These advantages also define the innovative part of our study: the proposed methodology for assessing the performance of the approaches with and without the ROI localization step. The methodology includes (1) publicly available data video recordings of various quality of subjects in different challenging conditions and (2) the inclusion of the approach without the image processing front-end, which has, to the best of our knowledge, never been used before in the evaluation of the performances of the proposed ROI localization steps.

The disadvantages are related to (1) the limited number of analysed approaches, (2) the absence of time complexity analysis of the studied approaches, and (3) the small number of studied video recordings covering motion and skin tone scenarios. These disadvantages are due to the fact that (1) we wanted to focus on the most widely used approaches (with an addition of the less known FVP approach, which relies on a completely different principle), (2) time complexity analysis would first require optimization of the programming code behind the studied approaches, and (3) the remaining 19 sets of video recordings from LGI–PPGI–FVD are not publicly available due to the limited storage space on the server that data set's creators use, whereas PBDT–rPPG offers only a small number of videos covering the skin tone challenge. It is also to be noted that the success rate and the ROI size are the only parameters that directly assessed the performance of the studied approaches. However, even ROI sizes themselves do no tell anything about the quality of the skin mask (in terms of the ratio between the skin and non-skin pixels). In SNR and RMSE metrics, potential differences between the studied approaches are masked especially in the case of POS, which has been shown to exhibit the best overall performance [31] in extracting the rPPG signal from video recordings.

The first future challenge is the optimization of the studied algorithms to allow their best performance. For example, (1) in VK-RGBHCbCr and especially in VK-Conaire, a combination of a sequence of morphological operations (erosion and dilation) could be used to further refine the ROIs (see Figure 1d,h,l for noisy ROIs identified by VK-Conaire); (2) in all approaches with image processing front-ends, identified ROIs could be

divided into multiple ROIs (multiple ROIs would also enable measurements of additional physiological parameters [44]); and (3) in all approaches, lengths of the processing window lengths could be optimized. Additionally, we could test some other skin classification approaches, such as one class support vector machine-based classifier [45] or an approach based on active contours and Gaussian mixture models [32], which do not rely on simple pixel-wise segmentation. It is to be noted that both methods require robust face detection. The second challenge is related to the analysis of the performance of the studied approaches in cases when subjects are wearing masks. Due to the fact that masks are becoming a new reality of our everyday life amid the ongoing COVID-19 pandemic, this challenge is in our opinion relevant for the rPPG research community. Recently, Speth et al. [46] created a publicly available data set containing recordings of 61 masked subjects together with an algorithm that adds a synthetic mask to a face on a selected video recording, which is, in our opinion, a valuable contribution for the rPPG community. The last challenge is related to the evaluation of various approaches for extracting rPPG signals from video recordings covering parts of the body other than face. These body parts have already been used for assessing peripheral hemodynamics [47,48].

5. Conclusions

In rPPG measurements, the selection of an approach for extracting the pulsatile information from video recordings seems not to significantly affect the extraction of rPPG signal from video recording in terms of SNR if proper rPPG algorithm that combines the information extracted from the video recording into a single pulse waveform signal is selected. On the other hand, RMSE and especially the success rate of the approaches are more affected by the selection of ROI localization approach. Therefore, when designing software for rPPG measurement system, one should adopt the software solution to the actual application to ensure as robust performance of the rPPG measurement system as possible.

Author Contributions: Conceptualization, M.F., Ž.P. and P.P.; methodology, Ž.P. and M.F.; software, Ž.P.; validation, Ž.P. and M.F.; formal analysis, Ž.P.; investigation, Ž.P. and M.F.; resources, Ž.P.; data curation, Ž.P. and M.F.; writing—original draft preparation, M.F. and Ž.P.; writing—review and editing, M.F., Ž.P. and P.P.; visualization, Ž.P.; supervision, M.F. and P.P.; project administration, P.P.; funding acquisition, P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was co-funded by Slovenian Research Agency (ARRS) (ARRS Programme code P2-0270 (C)). The APC was funded by ARRS.

Data Availability Statement: All data used in our study are made publicly available by their owners.

Acknowledgments: The authors acknowledge the financial support from ARRS.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscrip; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

2SR	Spatial Subspace Rotation rPPG algorithm
ANOVA	analysis of variance
BPM	beats per minute
CHROM	chrominance-based rPPG algorithm
CSK	Circulant Structure with Kernels
ECG	electrocardiography
FastICA	a fast algorithm for Independent Component Analysis invented by Aapo Hyvärinen at Helsinki University of Technology
FVP	Full Video Pulse Extraction rPPG algorithm

KLT	Kanade-Lucas-Tomasi tracker
LGI-PPGI-FVD	LGI-PPGI-Face-Video-Database
MAE	mean absolute error
MSAC	M-estimator Sample and Consensus
PBDTrPPG	Public Benchmark Dataset for Testing rPPG Algorithm Performance
POS	Plane-Orthogonal-to-Skin rPPG algorithm
PR	pulse rate
PURE	Pulse Rate Detection Dataset
RANSAC	Random Sample Consensus
RGB	red-green-blue color space
RGB-H-CbCr	color space consisting of red-green-blue, hue (from HSV color space), blue-difference chroma component and red-difference chroma component (both from YCbCr color space)
RMSE	root mean square error
ROI	region of interest
rPPG	remote photoplethysmography
SelfRPPG	private rPPG data set prepared by Zhao et al. [8]
SNR	signal-to-noise ratio
STAPLE	Sum of Template and Pixel-Wise Learners
UBFC-RPPG	Univ. Bourgogne Franche-Comté Remote Photoplethysmography data set
VK	Viola-Jones combined with ROI width reduction and Kanade-Lucas-Tomasi tracker
VK-Conaire	Viola-Jones combined with skin detector based on the maximization of mutual information and Kanade-Lucas-Tomasi tracker
VK-LMK	Viola-Jones combined with landmarks detection and Kanade-Lucas-Tomasi tracker
VK-RGBHCbCr	Viola-Jones combined with RGB-H-CbCr skin-color segmentation and Kanade-Lucas-Tomasi tracke
YCbCr	color space with the luma component (Y), blue-difference chroma component (Cb) and red-difference chroma component (Cr)

References

1. Verkruysse, W.; Svaasand, L.O.; Nelson, J.S. Remote plethysmographic imaging using ambient light. *Opt. Express* **2008**, *16*, 21434–21445. [CrossRef]
2. Sun, Y.; Thakor, N. Photoplethysmography revisited: From contact to noncontact, from point to imaging. *IEEE Trans. Biomed. Eng.* **2015**, *63*, 463–477. [CrossRef]
3. Unakafov, A.M. Pulse rate estimation using imaging photoplethysmography: Generic framework and comparison of methods on a publicly available dataset. *Biomed. Phys. Eng. Express* **2018**, *4*, 045001. [CrossRef]
4. Takano, C.; Ohta, Y. Heart rate measurement based on a time-lapse image. *Med. Eng. Phys.* **2007**, *29*, 853–857. [CrossRef] [PubMed]
5. Wang, W.; Stuijk, S.; de Haan, G. Living-skin classification via remote-PPG. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2781–2792. [CrossRef]
6. Bobbia, S.; Macwan, R.; Benezeth, Y.; Mansouri, A.; Dubois, J. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognit. Lett.* **2019**, *124*, 82–90. [CrossRef]
7. Poh, M.Z.; McDuff, D.J.; Picard, R.W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* **2010**, *18*, 10762–10774. [CrossRef] [PubMed]
8. Zhao, C.; Mei, P.; Xu, S.; Li, Y.; Feng, Y. Performance evaluation of visual object detection and tracking algorithms used in remote photoplethysmography. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019. [CrossRef]
9. Mestha, L.K.; Kyal, S.; Xu, B.; Lewis, L.E.; Kumar, V. Towards continuous monitoring of pulse rate in neonatal intensive care unit with a webcam. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 3817–3820. [CrossRef]
10. Kovac, J.; Peer, P.; Solina, F. *Human Skin Color Clustering for Face Detection*; IEEE: New York, NY, USA, 2003; Volume 2. [CrossRef]
11. bin Abdul Rahman, N.A.; Wei, K.C.; See, J. *RGB-H-CbCr Skin Colour Model for Human Face Detection*; Faculty of Information Technology, Multimedia University: Nairobi, Kenya, 2007; Volume 4.
12. Mahmoud, T.M. A new fast skin color detection technique. *World Acad. Sci. Eng. Technol.* **2008**, *43*, 501–505. [CrossRef]
13. Asthana, A.; Zafeiriou, S.; Cheng, S.; Pantic, M. Robust discriminative response map fitting with constrained local models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3444–3451. [CrossRef]
14. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874. [CrossRef]

15. Conaire, C.O.; O'Connor, N.E.; Smeaton, A.F. Detector adaptation by maximising agreement between independent data sources. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–6. [CrossRef]
16. Wang, W.; den Brinker, A.C.; De Haan, G. Full video pulse extraction. *Biomed. Opt. Express* **2018**, *9*, 3898–3914. [CrossRef]
17. Wang, W.; den Brinker, A.C.; De Haan, G. Single-element remote-PPG. *IEEE Trans. Biomed. Eng.* **2018**, *66*, 2032–2043. [CrossRef]
18. Li, P.; Benezeth, Y.; Nakamura, K.; Gomez, R.; Li, C.; Yang, F. Comparison of region of interest segmentation methods for video-based heart rate measurements. In Proceedings of the 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan, 29–31 October 2018; pp. 143–146. [CrossRef]
19. Fouad, R.; Omer, O.A.; Aly, M.H. Optimizing remote photoplethysmography using adaptive skin segmentation for real-time heart rate monitoring. *IEEE Access* **2019**, *7*, 76513–76528. [CrossRef]
20. Li, P.; Benezeth, Y.; Nakamura, K.; Gomez, R.; Yang, F. Model-based Region of Interest Segmentation for Remote Photoplethysmography. In Proceedings of the 14th International Conference on Computer Vision Theory and Applications, Prague, Czech Republic, 25–27 February 2019. [CrossRef]
21. De Haan, G.; Jeanne, V. Robust pulse rate from chrominance-based rPPG. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 2878–2886. [CrossRef]
22. Hu, Z.; Wang, G.; Lin, X.; Yan, H. Skin segmentation based on graph cuts. *Tsinghua Sci. Technol.* **2009**, *14*, 478–486. [CrossRef]
23. Liao, S.; Jain, A.K.; Li, S.Z. A fast and accurate unconstrained face detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 211–223. [CrossRef]
24. Bradski, G.R. Real time face and object tracking as a component of a perceptual user interface. In Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision. WACV'98 (Cat. No. 98EX201), Princeton, NJ, USA, 19–21 October 1998; pp. 214–219. [CrossRef]
25. Hyvärinen, A.; Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **2000**, *13*, 411–430. [CrossRef]
26. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I. [CrossRef]
27. Lucas, B.D.; Kanade, T. An iterative image registration technique with an application to stereo vision. In Proceedings of the IJCAI'81, 7th International Joint Conference on Artificial Intelligence—Volume 2, Vancouver, BC, Canada, 24–28 August 1981.
28. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. Exploiting the circulant structure of tracking-by-detection with kernels. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 702–715. [CrossRef]
29. Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; Torr, P.H. Staple: Complementary learners for real-time tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1401–1409. [CrossRef]
30. Stricker, R.; Müller, S.; Gross, H.M. Non-contact video-based pulse rate measurement on a mobile service robot. In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK, 25–29 August 2014; pp. 1056–1062. [CrossRef]
31. Wang, W.; den Brinker, A.C.; Stuijk, S.; de Haan, G. Algorithmic principles of remote PPG. *IEEE Trans. Biomed. Eng.* **2016**, *64*, 1479–1491. [CrossRef] [PubMed]
32. Woyczyk, A.; Fleischhauer, V.; Zaunseder, S. Skin Segmentation using Active Contours and Gaussian Mixture Models for Heart Rate Detection in Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 312–313. [CrossRef]
33. Jones, M.J.; Rehg, J.M. Statistical color models with application to skin detection. *Int. J. Comput. Vis.* **2002**, *46*, 81–96. [CrossRef]
34. Wang, W.; Stuijk, S.; De Haan, G. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE Trans. Biomed. Eng.* **2015**, *63*, 1974–1984. [CrossRef] [PubMed]
35. Hoffman, W.F.C.; Lakens, D. *Public Benchmark Dataset for Testing rPPG Algorithm Performance*; 4TU.Centre for Research Data: The Hague, The Netherlands, 2019.
36. Pilz, C.S.; Zaunseder, S.; Krajewski, J.; Blazek, V. Local group invariance for heart rate estimation from face videos in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1254–1262. [CrossRef]
37. Shi, J. Good features to track. In Proceedings of the 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 21–23 June 1994; pp. 593–600. [CrossRef]
38. Li, X.; Chen, J.; Zhao, G.; Pietikainen, M. Remote heart rate measurement from face videos under realistic situations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 4264–4271. [CrossRef]
39. Rabin, J.; Delon, J.; Gousseau, Y.; Moisan, L. MAC-RANSAC: A robust algorithm for the recognition of multiple objects. In Proceedings of the Fifth International Symposium on 3D Data Processing, Visualization and Transmission (3DPTV 2010), Paris, France, 17–20 May 2010; p. 051.
40. Kamshilin, A.A.; Nippolainen, E.; Sidorov, I.S.; Vasilev, P.V.; Erofeev, N.P.; Podolian, N.P.; Romashko, R.V. A new look at the essence of the imaging photoplethysmography. *Sci. Rep.* **2015**, *5*, 1–9. [CrossRef]

41. Mironenko, Y.; Kalinin, K.; Kopeliovich, M.; Petrushan, M. Remote Photoplethysmography: Rarely Considered Factors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 296–297. [CrossRef]
42. Finžgar, M.; Podržaj, P. Feasibility of assessing ultra-short-term pulse rate variability from video recordings. *PeerJ* **2020**, *8*, e8342. [CrossRef]
43. van der Kooij, K.M.; Naber, M. An open-source remote heart rate imaging method with practical apparatus and algorithms. *Behav. Res. Methods* **2019**, *51*, 2106–2119. [CrossRef]
44. Kumar, M.; Veeraraghavan, A.; Sabharwal, A. DistancePPG: Robust non-contact vital signs monitoring using a camera. *Biomed. Opt. Express* **2015**, *6*, 1565–1588. [CrossRef] [PubMed]
45. Wang, W.; Stuijk, S.; De Haan, G. Exploiting spatial redundancy of image sensor for motion robust rPPG. *IEEE Trans. Biomed. Eng.* **2014**, *62*, 415–425. [CrossRef]
46. Speth, J.; Vance, N.; Flynn, P.; Bowyer, K.; Czajka, A. Remote Pulse Estimation in the Presence of Face Masks. *arXiv* **2021**, arXiv:2101.04096.
47. Rubins, U.; Miscuks, A.; Lange, M. Simple and convenient remote photoplethysmography system for monitoring regional anesthesia effectiveness. In *EMBEC & NBC 2017*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 378–381. [CrossRef]
48. McDuff, D.; Nishidate, I.; Nakano, K.; Haneishi, H.; Aoki, Y.; Tanabe, C.; Niizeki, K.; Aizu, Y. Non-contact imaging of peripheral hemodynamics during cognitive and psychological stressors. *Sci. Rep.* **2020**, *10*, 1–13. [CrossRef] [PubMed]