Measuring Heart Rate With an RGB Camera For Real-Time General Health Monitoring

Vladimir Pechi[1]
Supervisor(s): Jorge Martinez Castaneda[1]**, Kianoush Rassels**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 26, 2025

Name of the student: Vladimir Pechi
Final project course: CSE3000 Research Project
Thesis committee: Jorge Martinez Castaneda, Kianoush Rassels, Christoph Lofi

***Abstract*—Heart rate (HR) is a critical indicator of an individual's health, serving as a key metric for detecting potential cardiac issues. This paper explores a method for real-time heart rate measurement using RGB camera footage, aimed at general health monitoring. The proposed method utilizes a convolutional neural network (CNN) to generate a 3D mesh of the subjects' facial features. The movement over time of the points in this mesh is used to compute a signal that captures the small pulsatile movements corresponding to the mechanical motion of blood being pumped through the veins. This signal is filtered, and motion sources are separated using principal component analysis (PCA). The most periodic component, the one with the highest frequency, is considered to correspond to the heart rate, and it's frequency is used to estimate the heart rate. The proposed method is tested using the ECG-Fitness dataset, characterized by challenging environmental conditions such as significant subject motion and dim lighting conditions. Experimental results demonstrate the method's capability for real-time applications, though further enhancements are needed to improve robustness under difficult environmental conditions.**

***Index Terms*—biotelemetry, signal, real-time systems, image processing, image models, neural networks, filter, eigen analysis, principal component analysis, DCT, FFT, Fourier transform, harmonic analysis**

I. Introduction

THE heart's role in circulating oxygen and nutrients throughout the body is critical. If any issues are limiting the heart's ability to function properly, then the impact on an individual's well-being will be significant. As such, measuring the heart rate is paramount, since it is a good indicator of a person's cardiac health. Bradycardia, heart rate below 60 beats per minute, and tachycardia, heart rate above 100 beats per minute, are considered to be abnormal. They are often indicators of underlying conditions and diseases.

Traditionally, the heart rate is measured using contact methods like an electrocardiogram (EKG), or a finger pulse oximeter. These methods can cause discomfort and skin irritation to patients [1], which is an impediment to consistent measurements as a way to generally monitor the health of individuals. Therefore, developing reliable and accurate non-invasive, non-contact heart rate measurement methods could greatly improve the adoption of health monitoring.

RGB cameras, like the Logitech C920 used to collect the videos for the employed dataset, are a mature and inexpensive technology that is widely available. Given this high availability and inherent potential reach, research into health monitoring based on RGB cameras has broadly already been conducted. There are two main principles when it comes to camera based heart rate measurements: remote Photoplethysmography (rPPG) and imaging Ballistocardiography (iBCG). RPPG relies on extracting heart related bio-signals based on small changes in the color of the skin. By contrast, iBCG based methods take into account the microscopic movements of the skin generated by the mechanical motion of the blood pumping through the veins. [2]

Many proposed methods rely substantially on a limited set of experimental conditions to work reliably, since signal degradation due to improper lighting conditions or subject motion can cause erroneous results. Individuals are often asked to stand as still as possible, but this limits the usability of the technology in real-life scenarios. This paper expands on the method presented by Haque et al., with the intent of further improving the robustness of heart rate measurements to natural human motion, while minimizing the incurred computational complexity.

II. Background

There are two main approaches, color-based Photoplethysmography [3]–[5] and motion-based Ballistocardiography [6]–[8]. Within each category, there are many different kinds of techniques that can be used to isolate and extract the heart rate bio-signal. Four influential approaches with good results for each category that were considered are succinctly presented below. The first two are rPPG based methods, while the next two are based on the principles behind Ballistocardiography.

In their 2014 paper, "Remote Heart Rate Measurement From Face Videos Under Realistic Situations" [3], Li et al. address the challenges that realistic subject motion and lighting variations can have on traditional rPPG methods. They make use of face tracking using facial landmarks to minimize the impact of head motions and additionally discard noisy portions of the signal. To minimize the impact of lighting related noise, they apply a normalized least mean squares (NLMS) filter, using background light values as a reference.

Traditional rPPG methods which rely on blind source separation techniques, can face challenges in detecting heart rate during motion. G. de Haan and V. Jeanne introduce a chrominance-based approach that significantly enhances motion robustness [4]. They analyze the chromatic variations caused by blood volume variations across different color channels, separating the pulse-related signal from motion-induced noise. An in-depth analysis on how motion enters the pulse-signal as noise is presented, as well as how the chrominance approach suggested minimizes it.

Balakrisnan et al. introduce a novel method for heart rate extraction by measuring the movement caused by the influx of blood pumped by each heartbeat. They consider the other sources of involuntary head movement, such as "the oscillatory motion that keeps the head in dynamic equilibrium" [6], or movement caused by respiration. A 1D signal is obtained by tracking feature points. The signal is then filtered, eliminating low frequencies (less than 0.75 Hz), containing the respiration related movement, and very high frequencies (above 5 Hz). This upper limit would correspond to a heart rate of 300 beats per minute, which far exceeds the range of possible values. However, Balakrishnan et al. discovered that "harmonics and other frequencies higher than 2Hz provide useful precision needed for peak detection". The movement is decomposed into its sources by applying principal component analysis, and the most periodic component is considered to represent the heart rate.

In their 2016 paper, "Heart Rate Measurement from Facial Video" [7], Haque et al. expanded on the work of Balakrishnan et al. by developing a Ballistocardiographic approach that more directly considers the impact of internal (facial expressions) and external head movement, as well as other potential sources of signal degradation such as motion blur or improper lighting. They also present approaches to mitigate these effects. They found that a lot of noise is induced due to particularly low quality frames and developed a metric upon which they would be eliminated, resulting in more stable landmark trajectories. Additionally, they found that different facial landmark tracking algorithms could be combined to better mitigate each other's limitations, resulting in more reliable heart rate measurements.

For general health monitoring, not only should the data be processed as it is being received, but the subject should be able to act as they wish, unburdened by arbitrary constraints that may be needed for the system to function optimally. This makes computational complexity and especially reliability to environmental conditions essential. Photoplethysmography-based methods are inherently more dependent on lighting conditions, while movement based methods can potentially even operate in darkness, by adapting the approach to use depth cameras. To that end, the approach presented by Haque et al. will be primarily adopted for this paper. As established by Hassan et al. in their review of non-invasive heart rate monitoring [9], the approach not only performs well in a simplistic scenario, but suffers some of the least degradation in a more dynamic scenario that features more changeable lighting conditions and increased subject motion.

More recent learnings from Cheng et al. [8] will be integrated to improve upon the method. Cheng et al. discovered that by using anterior-posterior movement rather than the standard vertical-traces captured directly, they were able to better capture the heart-rate related pulsatile movement. Therefore, they were able to extract a more accurate bio-signal, leading to improved results over their previous method.

III. Proposed Method

For general-purpose health monitoring, computational complexity is also an important factor to consider, as a real-time system could provide significant assistance within a clinical setting. Therefore, the method described below will attempt to iterate on the method provided by Haque et al. [7] with the intention of reducing computational costs while negating or at least minimizing any losses in accuracy.

A. Face Quality Assessment and Face Landmark Detection

The method presented in "Heart Rate Measurement from Facial Video" [7], first passes the captured video frames through a face quality assessment (FQA) step, that seeks to detect and remove the lower quality frames that induce a disproportionate amount of noise within the signal. This is an important step that can potentially be implemented efficiently, given the right quality metric is chosen.

Following that, the face is detected as the region of interest (ROI) using the classifier introduced by Viola and Jones [11] and then facial landmarks are detected using two separate methods. In most non-contact heart rate measurement approaches, this ROI selection and tracking component is the most expensive computationally. This is the case for the Haque et al. approach as well, therefore changes here have the highest potential to reduce the computational load of the approach overall.

To that end, the ROI selection and face landmark detection stages from the paper were replaced by a fast and lightweight convolutional neural network based model made publicly available by Google [12]. This model is capable of detecting the presence of faces within an image and then generating a mesh of three-dimensional landmarks mapping out the subject's face as seen in Fig. 2. This model was chosen not only due its accurate, real-time performance, but also for its consistency across individuals from different geographic areas; as well as between male and female subjects. This careful analysis of potential bias alleviates some of the ethical concerns over data bias. Adopting this will result not only in an improvement in runtime performance, but will also enable the usage of anterior-posterior traces, allowing for the improvements identified by Cheng et al [8] to be incorporated. The Z-axis coordinates provided by the model are synthetic measurements that are "relative to the face center of mass and are scaled proportionally to the face width". [12]

In addition to the mesh of landmarks visible in Fig. 2, the model also produces, for each frame, a set of confidence values for face presence, face detection, and landmark tracking. Experimentally, it has been found that these are highly correlated with face resolution, brightness, and out-of-plane face rotation, which are the sources of noise identified by Haque et al. to disproportionately contribute to "the most erroneous segments coming from low quality face frames" [7]. As such, face quality assessment will be implemented by setting a minimum of 0.75 for these confidence values. Frames that fall below this confidence value will be discarded. This approach is beneficial, since it allows for similar noise elimination while not introducing another processing step that would add to the runtime of the algorithm.

For each frame of the video not removed by the FQA, the model will be run, and a set of 478 facial landmarks will be recorded. A subset of these landmarks will be selected corresponding to well vascularized areas such as the forehead and cheeks, as well as highly stable points around the eyebrows, eye sockets, mouth, and nose bridge. This subset can be seen in Fig. 3. The movement of these facial landmarks over time will constitute the signal traces containing the heart rate bio-signal. An example signal trace, before filtering, can be seen in Fig. 4.

B. Motion Signal Filtering

To extract the heart rate from the traces, the first step that must be done is to process the signal, attempting to eliminate noise. To that end, following the findings of Balakrishan et al. [6] and Haque et al. [7], the signal is first filtered using an 8-th order band-pass Butterworth filter. Heart rate should be within the 45 to 300 beats per minute range. Therefore, the cut-off frequencies that will be used will be 0.75 Hz and 5 Hz. Subsequently, to further smooth the traces and mitigate the impact of noise, a moving average filter will be applied. [7]. The effect of applying these filters is visualized in Fig. 5 and Fig. 6.

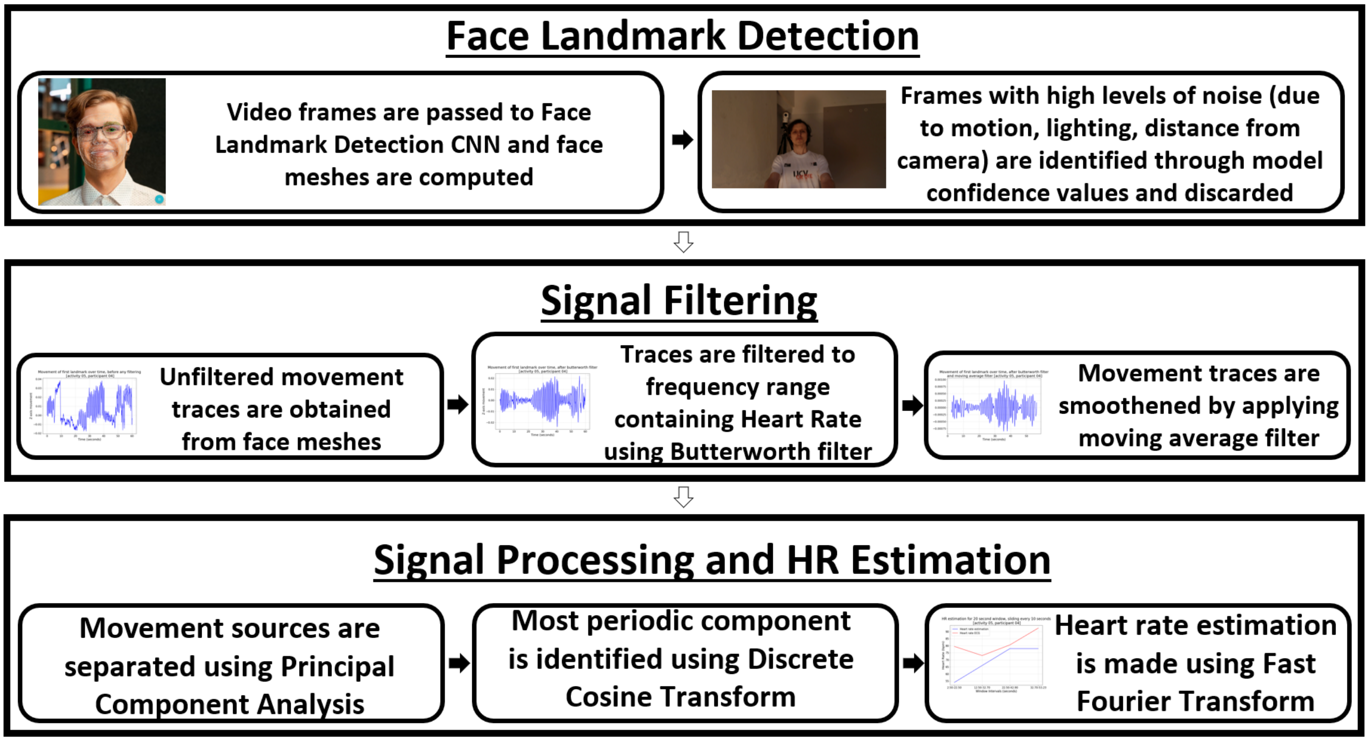
Face Landmark Detection
Video frames are passed to Face Landmark Detection CNN and face meshes are computed
Frames with high levels of noise (due to motion, lighting, distance from camera) are identified through model confidence values and discarded
Signal Filtering
Unfiltered movement traces are obtained from face meshes
Traces are filtered to frequency range containing Heart Rate using Butterworth filter
Movement traces are smoothed by applying moving average filter
Signal Processing and HR Estimation
Movement sources are separated using Principal Component Analysis
Most periodic component is identified using Discrete Cosine Transform
Heart rate estimation is made using Fast Fourier Transform

Fig. 1. Block diagram of the proposed system, image of person exercising sourced from ECG-Fitness dataset [10]

Fig. 2. MediaPipe Face Landmark Detection model [12] on image of author

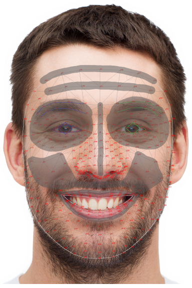

Fig. 3. Selected subset of landmarks shown with a gray outline over the sample picture freely provided by Google as reference for facial landmarks [12]

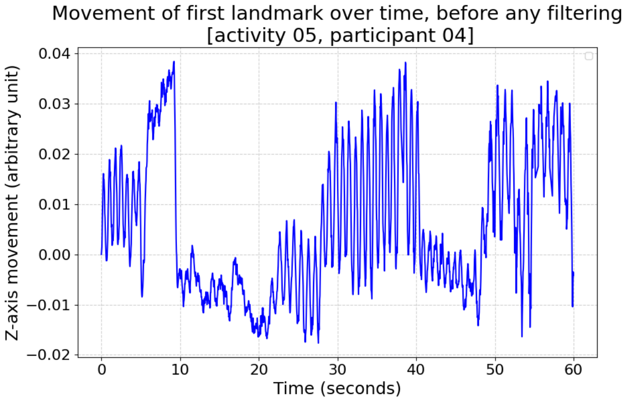

Fig. 4. Signal trace representing the movement of specific face landmark over time, before any filtering is performed

C. Motion Signal Processing

In order to isolate the pulsatile movement corresponding to the heart rate, from other sources of head motion, principal component analysis (PCA) will be applied to the filtered trajectories. The highest frequency component isolated by PCA should be the one corresponding to the heart rate. It will be identified by performing a discrete cosine transform (DCT) on all components, selecting the one with the highest frequency magnitude. Finally, an inverse DCT will be performed to retrieve the original signal and then a Fast Fourier Transform (FFT) will be employed to determine the frequency corresponding to the heart rate. [7]

IV. Experiments

Due to concerns over sensitive personal data further detailed in the responsible research section, no testing data was collected for this project. Instead, appropriate pre-existing datasets were accessed and used to evaluate the performance of the method described above. For a general health monitoring

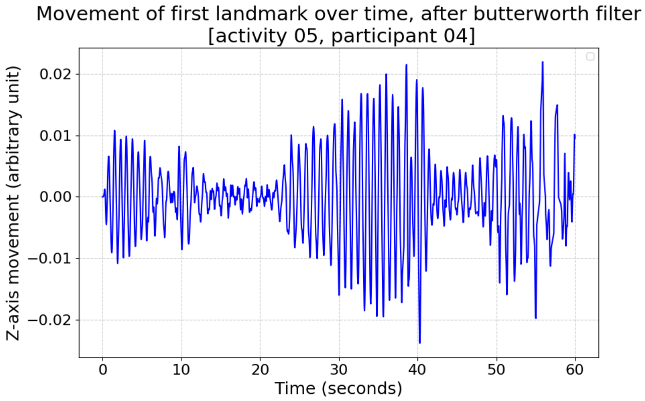

Fig. 5. Signal trace representing the movement of specific face landmark over time, after applying the Butterworth bandpass filter

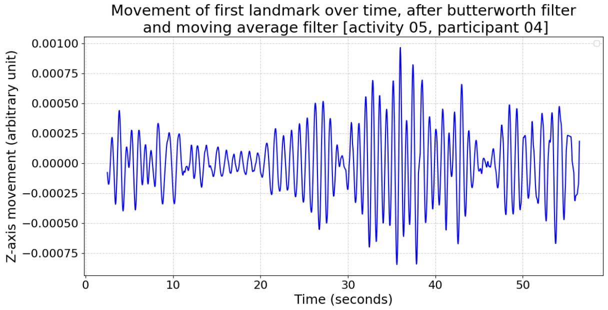

Fig. 6. Signal trace representing the movement of specific face landmark over time, after applying the butterworth bandpass and moving average filters

context, a system should be able to operate well no matter the environmental conditions, and without limiting a patient's activities. Therefore, to suitably represent the various conditions that would occur when such a system is deployed in practice, a dataset should contain at least a moderate amount of subject motion, as well as a variety of different lighting conditions.

The approach presented by Haque et al. [7] is evaluated using a common dataset for non-contact heart rate measurement, the MAHNOB-HCI dataset [13]. This is a highly appropriate dataset for a general purpose health monitoring approach, since it features a normal amount of internal and external face motion due to the human-computer interaction, affect measuring scenario. Unfortunately, due to difficulties related to the timely acceptance of a data management plan, obtaining access to this dataset during the course of the project was not possible. Therefore, a direct comparison between the approach presented above, and the original one by Haque et al. will not be possible.

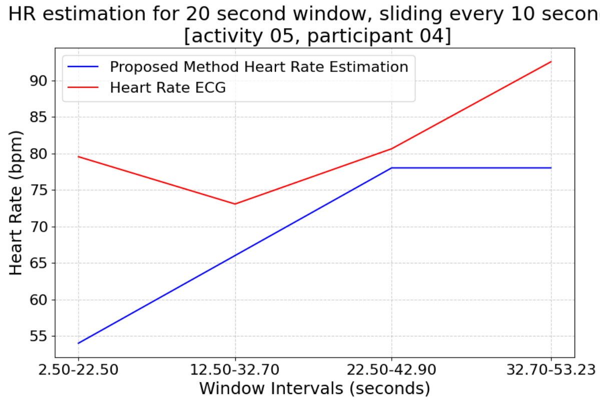

Fig. 7. Result of heart rate extraction on signal showed in Fig. 4, Fig. 5 and Fig. 6

As a substitute, access was possible to the ECG-Fitness dataset collected by Spetlik et al. [10]. This dataset features videos and electrocardiograph measurements from 17 subjects as they speak or perform physical exercise using different fitness machines. There are also three different lighting scenarios: natural light from a window, a bright 400W light, and a dim 30W light. Due to the presence of more variable environmental conditions, as well as the presence of intense, periodic, exercise motion, this constitutes a scenario that covers even more of the range of possible conditions a system in the field could encounter. This makes the ECG-Fitness dataset a suitable substitute for MAHNOB-HCI, but as it represents a much more challenging context, a direct comparison of the results would be misleading.

The method described in the section above was applied to all of the videos included in the ECG-Fitness dataset. The approach provides an average heart rate estimation over a 20-second window, using a sliding window. These estimations were compared against the average heart rate measured by the EKG over the same window to determine the following statistical measures: mean absolute error (MAE), standard deviation of error (SD), root mean squared error (RMSE) and Pearson correlation coefficient (r). These can be found in the Results section below.

In addition, to evaluate the feasibility of real-time measurements, details about the runtime of the algorithm were also collected. The approach was implemented in Python 3.8 making extensive use of optimized scientific libraries such as NumPy, OpenCV and SciPy. The experiments were run using a Windows 10 laptop equipped with an AMD Ryzen 9 5900HS processor. The experiments made use of a single processor thread, as efforts to improve the performance by parallelizing the process were not feasible during the duration of the project. However, the video can be segmented into several segments of contiguous frames, and then processed by separate instances of the face landmark detection model [12], As further presented in the Results and Discussion section, this is a promising approach for further optimization.

V. Results & Discussion

When creating the ECG-Fitness dataset to analyze the performance of their convolutional neural network (HR-CNN) approach [10], Spetlik et al. also evaluated other state-of-the-art methods, including the chrominance based approach (CHROM) [4] and that introduced by Li et al. (LiCVPR) [3] described in the Background section. Alongside those two, there is also data on an rPPG approach based on a detailed skin reflection model (2SR) introduced by Wang et al. [5]. Table I presents a comparison of results of these different methods on the ECG-Fitness dataset. From these results it can seen

TABLE I
AVERAGE RESULTS ON ECG-FITNESS DATABASE

	Proposed method*	HR-CNN	CHROM	LiCVPR	2SR
MAE	37.66	14.48	21.37	63.25	43.66
SDE	4.97	-	-	-	-
RMSE	38.23	19.15	33.47	67.67	52.86
Pearson Correlation Coefficient	0.0545	0.50	0.33	-0.02	0.06

$*$ on videos where the FQA did not discard more data than would allow for the creation of at least one window

that the proposed method performs better than those described by Li et al. and Wang et al., but significantly worse than those of De Haan et al. and Spetlik et al.. As detailed below, many of the shortcomings of the proposed method are related to the difficulty of extracting accurate face landmarks given conditions as difficult and variable as those introduced by the ECG-Fitness dataset. A model trained with such scenarios in mind, might enable the proposed method to derive much more accurate heart rate estimations, and should be investigated in the future.

A. *Impact and Limitations of FQA*

Due to the high levels of subject motion and difficult lighting conditions, for certain videos within the dataset, many frames were discarded by the FQA. For a few videos in the more challenging scenarios, not enough frames could be analyzed by the face landmark detection CNN [12] to obtain even one HR measurement. Table II shows the subjects for which the heart rate could not be estimated, for each activity (participants are 0 indexed, just as in the original dataset). All of the five videos are recorded at dusk, with natural light as the primary light source, and additionally figure a large amount of internal face motion and out-of-plane face rotation, as the subjects are talking. The lighting conditions result in the skin color of the participants being highly similar to that of the background, a situation in which the face landmark detection model [12] appears to be unable to perform face detection. Table III further details the impact of the FQA by showing the average number of discarded frames, separating videos that fall within, and out of, a 20% frames dropped threshold. From the numbers presented it can be concluded that the FQA operates reliably in good-weather scenarios, but, as difficult lighting and motion conditions compound, there is a threshold beyond which the model's ability to accurately track the face landmarks degrades significantly. Robustness to these conditions might be improved by training a similar model with more data that shows greater variance across these environmental conditions.

B. *Performance Across Different Activities*

It is important to mention that due to the highly difficult circumstances imposed by some of the activities captured by the dataset, the results varied significantly between activities, and separating the data for each activity can give much more insight into the performance of the approach described by this paper. To that end, table IV presents a separate breakdown of the results for each activity. An essential acknowledgment is that activities 2 and 4 feature very high levels of rapid head motion, resulting in a very high percentage of frames discarded for most videos. This makes it so that fewer windows' worth of frames could be collected. In the case of activity 4, for all 16 videos, frames sufficient for only one HR estimation could be analyzed. This not only makes it impossible to compute a correlation coefficient, but it also means that the estimations span a much larger amount of time, i.e. the entire 60 seconds of video footage. This grants these situations a significant advantage when it comes to temporal stability.

TABLE II
VIDEOS THAT COULD NOT BE EVALUATED FOR EACH ACTIVITY

	Participants
Activity 1 Speaking while standing on rowing machine	12
Activity 2 Intense rowing	N/A
Activity 3 Moderate rowing while speaking	4, 12, 13, 16
Activity 4 Intense rowing under halogen light	N/A
Activity 5 Exercising on an elliptical trainer	N/A
Activity 6 Exercising on a stationary bike	N/A

TABLE III
AVERAGE PERCENTAGE OF FRAMES ELIMINATED BY FQA

	All videos	Videos below 20% frames dropped	Videos above 20% frames dropped
Activity 1	22.02%	2.07%	58.60%
Activity 2	51.87%	10.19%	57.43%
Activity 3	33.82%	4.33%	75.94%
Activity 4	57.78%	N/A	57.78%
Activity 5	3.90%	1.87%	36.33%
Activity 6	1.25%	1.25%	N/A
All activities	28.15%	2.46%	60.12%

As shown by the relatively good performance achieved for activities 5 and 6, a moderate level of head motion, that doesn't move the head out of plane with the camera, does not cause these issues. In fact, activities 5 and 6 feature the

TABLE IV
RESULTS ON EACH ACTIVITY

	Total number of videos	Number of unevaluated videos	MAE	SDE	RMSE	Pearson Correlation Coefficient
Activity 1	17	1	32.88	5.61	33.6	-0.0814
Activity 2	17	0	34.85	1.14	34.91	-0.0008
Activity 3	17	4	45.48	7.15	46.09	0.1706
Activity 4	16	0	36.84	0.18	36.84	N/A
Activity 5	17	0	37.35	7.42	38.10	0.0930
Activity 6	17	0	40.07	8.60	41.34	0.0907
All activities	101	5	37.66	4.97	38.23	0.0545

lowest number of frames eliminated by the FQA, as well as the consistent performance for heart rate measurement across all metrics. These two activities both have the camera much closer to the subject's face, so it can be concluded that a high resolution of the face within the input video enables much more accurate and reliable detection of the facial landmarks, and that the increase in initial signal quality has a significant impact on the final results. The difference in head position can be visualized in Fig 8.

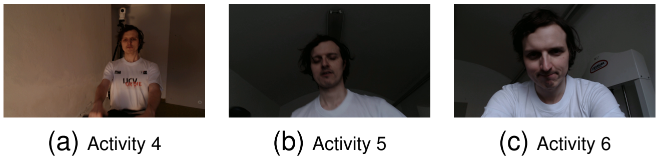
(a) Activity 4 (b) Activity 5 (c) Activity 6

Fig. 8. Difference in head position across different activities; Images available for publication from ECG-Fitness dataset [10]

C. Runtime Performance

Finally, analyzing the real-time performance aspect, table V presents the average runtime performance, in milliseconds, of all videos that could be analyzed. It can be concluded that the runtime is not highly dependent on the individual input data, as the runtime remains highly consistent across different activities. The differences in execution time can be explained in part by the different amount of operating systems operations being processed in the background. More significant differences can be observed for activities 2 and 4. For these in particular, the high levels of subject motion, result in an inability for the participant's faces to even be detected for many frames of the video. Because of that, there is significantly less data for landmarks to be extracted from, and then subsequently less signal to be filtered and processed to extract the heart rate, resulting in some greatly reduced processing times.

The average total runtime for all activities is 28.1 seconds, without landmark extraction parallelization. All the videos have a duration of exactly one minute, with 1800 frames, captured at 30 frames per second. Since it takes less than one minute to process one minute's worth of video, it is therefore possible for this method to be adapted to process live data and provide heart rate measurements in real time, with some delay.

As with other video-based approaches for measuring heart rate, ROI selection and face tracking, represented here by the "Landmark extraction" stage, remain the most time consuming. However by combining them into a single process using a CNN model based approach, a performance gain sufficient to process video faster than its duration was achieved. Moreover, this approach allows for significant parallelization, justifying its adoption over the more traditional use of Viola-Jones face detection and separate face landmark detection, as done by Haque et al. [7].

When adapting this algorithm for real-time applications, it is feasible to concurrently process the frames, and extract the face landmarks, using several parallel threads. Therefore, frames can be processed as they are captured by the camera. However, before the signal can be processed, a delay equivalent to the combined window sizes of the moving average filter, and that applied when extracting the heart rate, totaling 25 seconds, will be incurred. The signal filtering and heart rate extraction steps cannot be parallelized, so these will introduce further delay. Therefore, the total start-up delay for a real-time application utilizing the described method would be approximately 27 seconds.

TABLE V
AVERAGE RUNTIME PERFORMANCE FOR EACH ACTIVITY (IN MS)

	Landmark extraction	Signal filtering	Heart rate extraction	Total Runtime
Activity 1	28397.24	1700.48	52.78	30150.5
Activity 2	20907.97	873.41	16.88	21798.26
Activity 3	28426.58	1711.84	55.64	30194.06
Activity 4	21050.00	798.74	11.37	21860.11
Activity 5	30042.16	1973.50	64.03	32079.69
Activity 6	30650.59	2071.76	66.07	32788.42
All activities	26540.77	1519.36	44.25	28104.3836

Ultimately, estimations for the average heart rate over the past 25 seconds will be produced, with 2 seconds of added latency introduced by the signal filtering and processing steps. For a non-critical health monitoring situation, this should be sufficiently responsive to provide data that can be used to preventatively detect and address cardiac health problems before they become life-threatening.

VI. Responsible Research

The research for this paper was conducted in alignment with the principles outlined by the "Netherlands Code of Conduct for Research Integrity": Honesty, Scrupulousness, Transparency, Independence, Responsibility [14]. Care was taken to follow the standards and good practices outlined by the NWO, especially as pertains to obtaining required permissions, reporting results in a transparent, unbiased and unadulterated manner, and attributing credit to previous works through citations.

A. Ethical Considerations

Given the sensitive nature of medical data, special attention was paid to protecting participant private information and to data security. The decision to use pre-existing datasets, rather than collect new data was made in order to minimize risk to potential participants. The data management plan developed in collaboration with the supervisors, pending approval of the data stewards, ensured that the datasets used in the validation of the approach were accessed and utilized with appropriate permissions, in compliance with institutional and legal regulations. The data was maintained securely, and only images specifically labeled for publication were made directly available through this paper.

Efforts were also made to ensure that the adopted models and methods minimize biases related to sex, skin tone, or geographic origin. The facial landmark detection model employed in this study was evaluated for its consistency across a diverse pool of subjects, and no substantial explicit bias was detected, given the model's carefully considered training data [12]. In future, further work should be conducted with more diverse datasets to definitively confirm that the proposed method does not introduce, through any of the subsequent processing steps, potential disparities in the reliability of the system for individuals from groups not represented by the ECG-Fitness database.

B. Limitations and Responsible Communication

While some of the findings of this paper are promising, the limitations were communicated clearly and transparently to avoid overstatement of the results. Challenges, such as the method's reduced robustness to poor lighting and high levels of subject motion, were explicitly detailed, along with recommendations to address them in future work. This ensures that further research is conducted with a clear understanding of the current capabilities and limitations of the proposed method.

C. Transparency and Reproducibility

A key priority of this research was ensuring the transparency and reproducibility of the results. To achieve this, all code developed for this project will be made publicly available alongside this paper. This should allow other researchers to replicate the experiments.

VII. Conclusions & Future Research

This paper provides a framework for measuring heart rate from facial videos, seeking to address the limitations that prevent such a non-intrusive system from being deployed in practice: reliability and a lack of real-time performance. As such, the robust approach proposed by Haque et al. [7] is adapted in a manner that would better enable the future development of real-time applications. To that end, the Viola-Jones face detection, and the subsequent separate facial landmark tracking were replaced by a fast and lightweight convolutional neural network based model [12]. While this did lead to a substantial improvement in performance, there has also been an apparent decrease in the robustness of the approach to difficult environmental conditions relating to motion and light.

Future research should prioritize improving robustness through targeted model training on datasets featuring challenging conditions, including poor lighting, rapid, periodic movement across a diverse range of subjects. Integrating depth sensors to supplement RGB cameras could also significantly improve robustness to lighting conditions, as some approaches involving this technology have shown some promising results [15].

Additionally, the Face Quality Assessment step used to eliminate the frames that would introduce the most noise into the motion signal analyzed by the system should be further investigated to improve understanding of the correlation between the model confidence values and the detection of noise sources: improper lighting, motion artifacts, reduced face resolution and out-of-plane face rotation [7]. This improved understanding should lead to algorithmic adjustments that contribute to further improved robustness to difficult environmental conditions.

Moreover, real-time applications would benefit from implementing parallel processing during the landmark detection stage, which could reduce latency down to just the few seconds needed to filter and process the signal. Given a similar setup to that described in the Experiments section, the average heart rate over the past 25 seconds could be estimated, with only approximately 2 seconds of latency, the time needed to filter and process the motion signal. This adaptation would be sufficient to apply the method in a personal health monitoring scenario, allowing for the early detection of cardiac problems, while minimizing the discomfort caused to patients, potentially leading to higher adoption. Preventative health interventions facilitated by this early detection could lead to better health outcomes for patients.

VIII. Acknowledgments

This endeavor would not have been possible without the support and guidance of the project supervisors, Dr. Jorge Martinez Castaneda and Ir. Kianoush Rassels. Their timely and erudite feedback and technical suggestions greatly contributed to the creation of this paper.

I would also like to thank my friends and family for their continued moral and emotional support. Their encouragement has enabled me to push through all of the challenges and difficulties I encountered while conducting the research for this paper.

References

[1] H. Rafiei, M. Amiri, and J. Moghaddasi, "Skin irritation because of electrocardiograph lead in patients in intensive care unit," *International Wound Journal*, vol. 10, pp. 116–116, February 2013.

[2] M. Hassan, A. Malik, D. Fofi, B. Karasfi, and F. Meriaudeau, "Towards health monitoring using remote heart rate measurement using digital camera: A feasibility study," *Measurement*, vol. 149, p. 106804, January 2020.

[3] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 4264–4271.

[4] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, pp. 2878–2886, October 2013.

[5] W. Wang, S. Stuijk, and G. de Haan, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 1974–1984, September 2016.

[6] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3430–3437.

[7] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Heartbeat rate measurement from facial video," *IEEE Intelligent Systems*, vol. 31, pp. 40–48, May 2016.

[8] J. Cheng, B. Yue, R. Song, Y. Liu, C. Li, and X. Chen, "Motion-robust anterior–posterior imaging ballistocardiography for non-contact heart rate measurements," *Biomedical Signal Processing and Control*, vol. 86, September 2023.

[9] M. Hassan, A. Malik, D. Fofi, N. Saad, B. Karasfi, Y. Ali, and F. Meriaudeau, "Heart rate estimation using facial video: A review," *Biomedical Signal Processing and Control*, vol. 38, pp. 346–360, September 2017.

[10] R. Spetlik, J. Cech, V. Franc, and J. Matas, "Visual heart rate estimation with convolutional neural network," in *British Machine Vision Conference*, 2018.

[11] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, May 2004.

[12] G. A. for Developers, “Mediapipe solutions – face landmark detection,” https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker, Google, accessed: 2024-12-20.

[13] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE Transactions on Affective Computing*, vol. 3, pp. 42–55, January 2012.

[14] NWO, “Netherlands code of conduct for research integrity,” October 2018.

[15] C. Yang, G. Cheung, and V. Stankovic, “Estimating heart rate and rhythm via 3d motion tracking in depth video,” *IEEE Transactions on Multimedia*, vol. 19, pp. 1625–1636, February 2017.