Multimodal Heartbeat Rate Estimation from the Fusion of Facial RGB and Thermal Videos

Johansen, Anders Skaarup ; Henriksen, Jesper Wædeled; Haque, Mohammad Ahsanul; Jahromi, Mohammad Naser Sabet; Nasrollahi, Kamal; Moeslund, Thomas B.

Published in:
Eleventh International Conference on Machine Vision, ICMV 2018

DOI (link to publication from Publisher):
10.1117/12.2523385

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Multimodal Heartbeat Rate Estimation from the Fusion of Facial RGB and Thermal Videos

Anders S. Johansen, Jesper W. Henriksen, Mohammad A. Haque,
Mohammad Naser Sabet Jahromi, Kamal Nasrollahi and Thomas B. Moeslund

Visual Analysis of People Lab, Aalborg University, Denmark

ABSTRACT

Measuring Heartbeat Rate (HR) is an important tool for monitoring the health of a person. When the heart beats the influx of blood to the head causes slight involuntary movement and subtle skin color changes, which cannot be seen by the naked eye but can be tracked from facial videos using computer vision techniques and can be analyzed to estimate the HR. However, the current state of the art solutions encounter an increasing amount of complications when the subject has voluntary motion on the face or when the lighting conditions change in the video. Thus the accuracy of the HR estimation using computer vision is still inferior to that of a physical Electrocardiography (ECG) based system. The aim of this work is to improve the current non-invasive HR measurement by fusing the motion-based and color-based HR estimation methods and using them on multiple input modalities, e.g., RGB and thermal imaging. Our experiments indicate that late-fusion of the results of these methods (motion and color-based) applied to these different modalities, produces more accurate results compared to the existing solutions.

Keywords: Heartbeat Rate, Facial Video, RGB, Thermal, Multimodal

1. INTRODUCTION

Heartbeat Rate (HR) is a physiological parameter that present the condition of cardiovascular system of our body. For many years HR has been an important tool for many fields such as medical science,[1] forensics,[2] adapting interactive experiences, and psycophysiological studies.[3] The classical method of HR measurement would be to employ a physical sensor on the body, as electrodes used in Electrocardiogram (ECG).[1] However, for situations where physical sensors would be attached for a prolonged duration, it has been shown that there is a direct correlation between the duration of electrode application and persistence of erythema.[4]

In the recent years more and more studies have shown that computer vision-based unobtrusive methods are getting closer to achieving satisfactory level of accuracy in HR measurement.[5] For extracting HR from facial video, the two main methods are motion tracking of specific points within a Region of Interest (ROI),[6,7] and measuring minor changes in skin color.[2,8,9] From the technical point of view, while the former is using Ballistocardiography (BCG)[6] the later is using Photoplethesmography (PPG).[10]

While considering different visual modalities being used in HR measurement from videos, two major modalities are thermal and RGB. Garbey et al.[11] used thermal imaging to capture HR from large superficial blood vessels. Instead of choosing ROI in the facial region, they chose ROI on the neck and wrists such that they could get clear view of these blood vessels. The method produced decent results but states that there are issues when there is voluntary head movement coming from other sources than heartbeat. The improvements in thermal technology also lead to the study of Ref. 12 which showed that it was possible to estimate the HR by using the movements of superficial vasculature.

Takano et al. first utilized the facial skin color changes in video to estimate HR.[13] They recorded the variations in the average brightness of the ROI a rectangular area on the subjects cheek to estimate HR. Few years later, Poh et al. proposed a method that used ROI mean color values from R, G and B channels as color traces from facial video, and employed Independent Component Analysis (ICA) to separate the periodic signal sources and a frequency domain analysis of an ICA component to measure HR.[10] Kwon et al. improved Pohs method by using merely green color channel instead of all three Red-Green-Blue (RGB) color channels.[8] Tulyakov et al.[9] investigated a potential solution by proposing a method

Corresponding author's email: mah@create.aau.dk

called self-adaptive matrix completion, which would automatically select the most reliable areas in the video for heartbeat signal so that areas that would provide erroneous signals could safely be discarded.

Balakrishnan et al. proposed a method for HR estimation which was based on invisible motion in the head (instead of skin color change) due to pulsation of the heart muscles, which can be obtained by a BCG.[6] In this approach, some feature points were automatically selected and tracked on the ROI of the subjects RGB facial video. Then Principle Component Analysis (PCA) was applied to measure HR. Haque et al.[14] showed improved results compared to Ref. 6, 10, 15 by fusing a trajectory based method which tracks the motion of the head, and skin color changes on the face to extract heartbeat signal. One of the notable point from Ref. 14 is that unlike other available methods it can show visible heartbeat peaks in time domain from the heartbeat signal from facial video. However, the accuracy of this method is also inferior to the physical sensor-based ECG.

To replace physical sensors and enable remote HR estimation, computer-vision methods need to reliably provide measurements of similar accuracy as invasive physiological measurements. With current state of the art methods the HR estimation is approaching a satisfying accuracy with stable conditions in laboratory environment.[9, 14, 16] However, these methods encounter difficulties when they encounter changes in illumination or head motion of the subject. There has been many proposed solution to this issue, some trying to remove the erroneous part of the signal, others have tried to design methods that use novel input modalities, such as thermal or depth imaging as opposed to regular RGB images. However, instead of investigating unimodal methods, possible solutions to this problem could be fusing different input modalities such as RGB- and thermal-images which does not suffer from the other modalities weaknesses. In this approach, the traditional RGB imaging could be fused with thermal[11, 12] that isn't as susceptible to changes in light. This could have potential for increasing accuracy and it could be interesting to see how a fusion of existing solutions affects the accuracy of the estimated HR. Thus, in this paper, we investigated the performance of HR estimation by employing different fusion approaches between RGB and thermal facial videos. A semi-supervised weighted fusion approach has also been proposed to obtain better accuracy. To the best of our knowledge, this is the first attempt to combine RGB and thermal modalities in order to achieve a better estimation of HR.

The rest of the paper is organized as follows. Section 2 describes the proposed system for estimating HR using RGB and thermal fusion. Section 3 describes experimental environment and evaluation procedure. Section 4 presents the obtained results. Finally, Section 5 concludes the paper.

2. THE PROPOSED METHODOLOGY

Similar to Ref. 14, we extract the color and motion traces from facial video as the raw heartbeat signal. The actual HR is then extracted after a series of signal processing steps. This section first describes the steps of HR estimation method from color or motion traces from each modalities (RGB and thermal) and then describes the fusion methodology.

2.1 Video Acquisition and Preprocessing

The first step of the proposed system is video acquisition in different modalities. Color RGB of frontal facial images are captured by a Microsoft Kinect Version2. The thermal data was captured by an Axis Q1922 thermal camera. Figure 1 shows examples of the original RGB and thermal video frames. After collecting the raw data, we synchronized the facial image frames in both modalities by following the capturing time stamps. Figure 1 shows that the original video frames present a large portion of the subject body in the space of the acquisition room. For the raw heartbeat signal extraction we just focus on the face. Thus, on the synchronized data modalities, we applied haar-like feature based face detection as used in Ref. 17 on RGB modality and cropped associated faces on the thermal data by using computed homographs. Homographs provided an approximate image registration across modalities. We calculated homography matrices from RGB to thermal using a 8-points homography technique from Ref. 18. The procedure is shown in Figure 2.

2.2 Heartbeat Trace Extraction and HR Estimation

The face region is selected as the ROI in the video frames. This ROI is then processed using two different methods: Method-1 (M1) is a motion-based method, which tracks the trajectories of points identified within the ROI to track the BCG motion traces. Method-2 (M2) is a color-based method which takes the mean of all pixels within the ROI to track the PCG color traces. These two methods are introduced in Figure 3 and described in the following subsections.

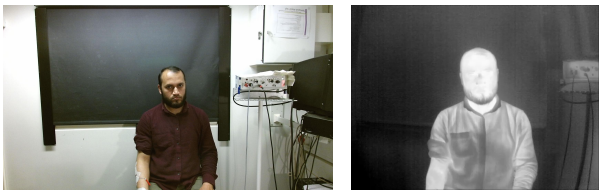


Figure 1. Example of captured video frames in RGB (left) and thermal (right) modalities.

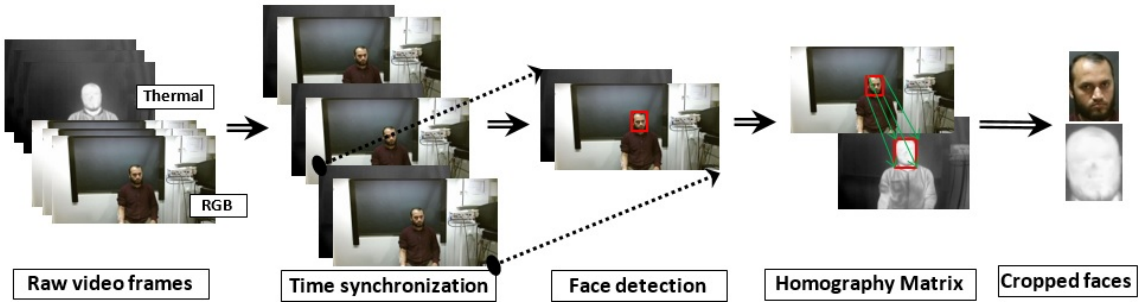


Figure 2. Preprocessing steps employed on the raw video frames of RGB and thermal data.

2.2.1 The Motion-based Method (M1)

We use Balakrishnan's motion-based method Ref. 6 that chooses points within the region of interest of the first image and then track these points throughout the image sequence. This method uses an area around the nose and forehead as their ROI. The trajectory points was chosen by a method called "Good Features to Track" (GFT).[19] The trajectories of these points were saved per frame. These trajectories were passed through a Butterworth band pass filter using cut off frequencies of [0.75, 5] Hz to remove signal components that could not come from heartbeats[16] and a zero phase filter to eliminate phase distortion. The signal traces were evaluated using PCA to single-out the trajectory which had the least noise and the most dominant frequency as described by Ref. 6. The Fourier transform was used to find the frequency domain of chosen trajectory. The dominant frequency was used to estimate the heart beats per minute (BPM).

2.2.2 The Color-based Method (M2)

In this case, we use Poh's color-based method[10] that took the average of all pixels of the facial region of the participant for each frame and produces a signal. Kwon[8] improved Poh's method by only using the green color channel. Thus, in this

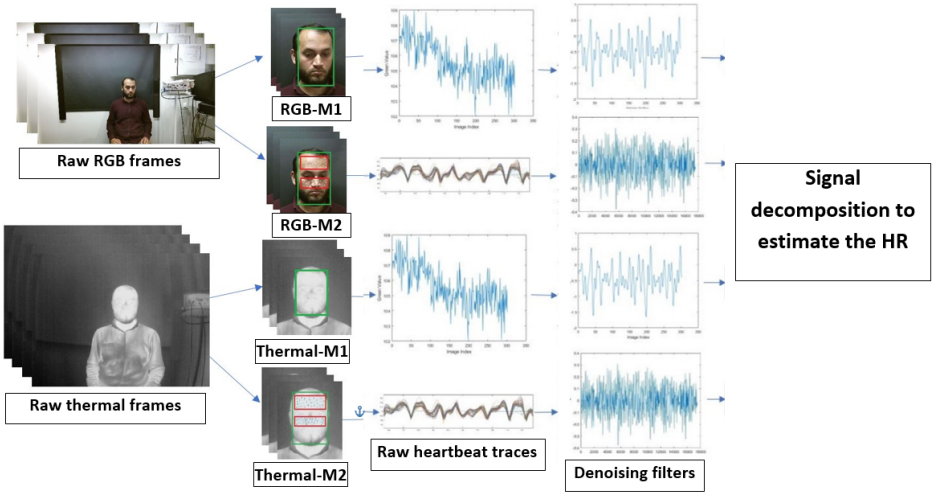


Figure 3. Heartbeat trace extraction and HR estimation from both RGB and thermal modalities using color and motion traces

paper, we also used the green color channel only. After face detection, the ROI was reduced by 20% from each side. This was done because the face detection often took too much data in the x-direction. After the signal was extracted, a moving average filter was applied with a kernel size of 40 to reduce noise that was related to illumination changes. A Butterworth band pass filter was applied using cut off frequencies of [0.75, 5] Hz to remove signal components that could not come from heartbeats.[16] Interpolation of the signal was used to make the signal having roughly the same sampling frequency as the ECG data. Finally, Fourier transform was applied to find the power spectrum of the signal. From the power spectrum the BPM can be derived by multiplying the most dominant frequency by 60.

2.3 Fusion of RGB and Thermal Modalities

After HR estimation by both methods (M1 and M2) on both RGB and thermal modalities, we employ two different late fusion approaches[20] based on two different hypotheses:

- First fusion hypothesis (H1): *Mean fusion of RGB and thermal data will result in a better HR estimation*
- Second fusion hypothesis (H2): *Weighting the output of each modality in the fusion will increase the accuracy*

In the first fusion case (H1), the mean of individual HR estimation were used as a combined estimate of the HR. In the second method (H2) where weighted fusion is proposed, 2 minutes of video was used for training weights and 1 minute of video was used for evaluating these weights for each subject. Weighted fusion was an attempt to utilize the combined accuracy of the different methods, while also minimizing the outlying methods effect on the final estimation.

3. EXPERIMENTAL ENVIRONMENT AND EVALUATION

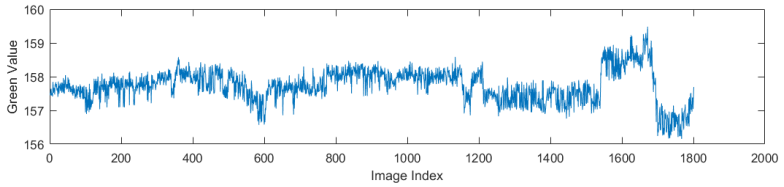
3.1 Experimental Environment

The implementation of the algorithms were done in MATLAB2017. The database was recorded in a laboratory environment from 20 student volunteers. The database contained 2 sessions of 10-11 minutes for each subject and it contained thermal-, RGB-videos and the ground truth ECG data. The RGB videos has a resolution of 1920x1080 and the resolution of the thermal videos was 640x480. The ECG data was collected using Shimmer v2* at 256 Hz sampling rate. The database was split into challenging parts and stable parts. In the stable parts the participants has minor movement and lighting conditions change. This paper and the resulting database focused entirely on this stable data in order to create a baseline benchmark for the heartbeat-rate estimation solutions. There are two subjects' data which are corrupted. Thus, we discarded those two subjects' data and employ the proposed approach on 18 subjects' data. We keep one minute of data from the stable session of each of the 18 subjects while considering *H1*, and 2 minutes for training and 1 minute for testing while considering *H2*. The HR estimation accuracy is evaluated using four statistical parameters which have been used in many of the previous literature.[10,14,15] These are: mean error (M_E), standard deviation (SD_{M_E}), root mean square error (RMS_E) and the error in percentage (M_{ER}).

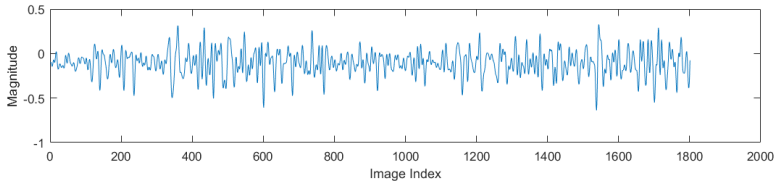
3.2 Experimental Evaluation

The proposed method tracks color change and head motion due to heartbeat in a video. Figure 4 shows an example of obtained raw green color trace from one subject's video and the outcomes of the signal processing steps on that raw heartbeat signal. After the signal was extracted, a moving average filter was applied to reduce noise that was related to illumination changes. On the other hand, a Butterworth band pass filter was applied to remove signal components that could not come from heartbeats. Interpolation of the signal was used to make the signal having roughly the same sampling frequency as the ECG data. Finally, Fourier transform was applied to find the frequency domain of the signal. From the domain the BPM can be derived by multiplying the most dominant frequency by 60 as shown in Figure 4(c).

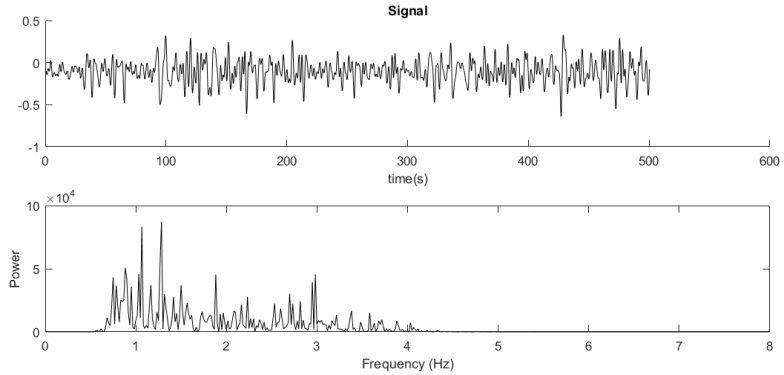
*Online: www.shimmersensing.com/products/ecg-development-kit



(a) Raw green trace



(b) After employing moving average and Butterworth filter



(c) Power spectrum after Fourier transform

Figure 4. Signal evaluation in different steps while estimating the HR.

4. RESULTS AND DISCUSSIONS

4.1 Experimental Results

The results of HR estimation for each of the modalities (unimodal case) for color and motion traces are shown in Figure 5(a), whereas the results after the fusion of modalities using *H1* are shown in Figure 5(b). While fusing the modalities, the median function was applied to all of the outputs of the color-based and motion-based methods in an effort to increase the accuracy. Note that this median function takes the average of the two middle values when given even number of parameters. In Table 1 the accuracy is shown for each of the fusions.

In an effort to increase the accuracy we employed the scenario of *H2* for weighted fusion. We trained weights for each of the outputs and applied to a set of test data. The training data had 2 minutes duration and the test data had 1 minute duration. The training values can be seen in Figure 6(a). The weight were calculated using $w_i = j_i/k_i$ where j_i is the ECG HR and k_i is the estimated HR from a modality for participant i. The resulted weights can be seen in Figure 6(b). Each of the weights were late fused with the output of each modality. The results can be seen in Figure 6(c) in the 'Pulse Fused' category. The accuracy of the weighted test was 10,14%. This is an improvement of 3.15% over the previous best accuracy which was 13,29%. The results for all subjects are summarized in Table 1.

4.2 Discussions

The results from the experiment indicate that the initial assumptions for the RGB and thermal modalities hold true. The reason the fusion of the methods had a lower error rate than individual methods was that when one modality miss some information the other modality serves as a complement. RGB being more susceptible to changes of light that makes it significantly harder for the algorithm to track the selected feature points over time. These drastic changes in the image

cause some of the facial trackers to flat-line and lose track of the feature points. With thermal not being as susceptible to these illumination changes, the accuracy of the motion-based method is significantly higher.

When the weighted method were evaluated 3 minutes of video was used. The tests was split into test and training data, where the first 2 minutes of video is used to train a weight based on the accuracy of the method which was evaluated on the last minute of test data. The reason for the split was the to avoid over fitting the weights to the specific video. The error rate was reduced by weighting the different inputs to the fusion. Selecting the weights initially require the use of ECG for training the weights. However, this would only require for a short training and after training, could reduce the amount of time patients spend wearing ECG electrodes.

Fusion of the color-based method improved both thermal and RGB input accuracy but the fused motion-based method had a lower accuracy than motion-based on thermal. The reason why the color-based method improved the accuracy was due to one modality overestimating the HR and the other modality underestimating the HR. These were not the same for each participant so sometimes the thermal modality was overestimating the HR while in the other times the RGB overestimated the HR. Therefore a merge often increased the accuracy. However, the motion-based RGB results often undershot the HR estimate than overshot. The motion-based thermal method did the same as the color-methods and sometimes overshot and sometimes undershot. So if these two were merged the results is worse than the individual methods.

This leads to the question, why not to exclude the motion-based method on RGB images! If we investigate a subset of the participants (1-10) the motion-based method does the same thing as described with the color-based method, where

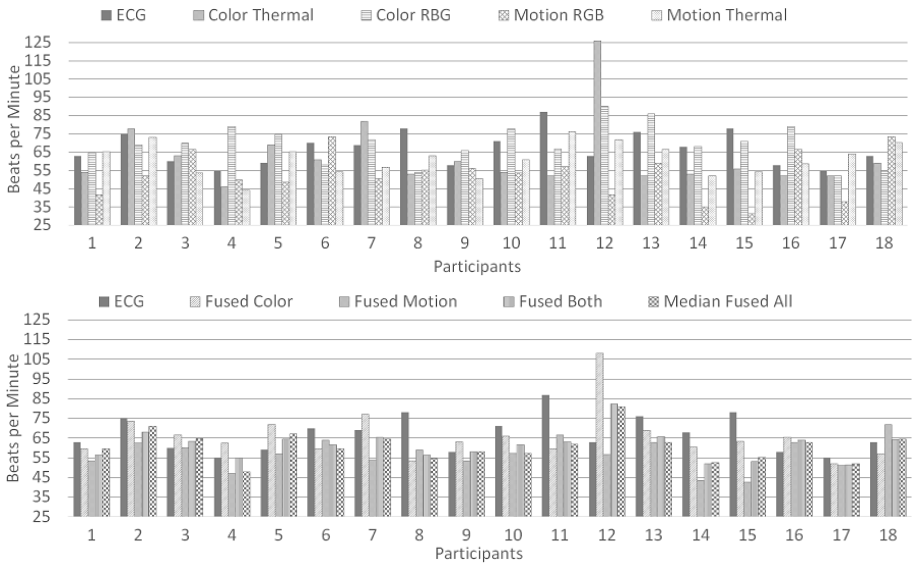


Figure 5. The HR estimation results before (top) and after (bottom) the fusion of modalities for all 18 subjects using *H1*.

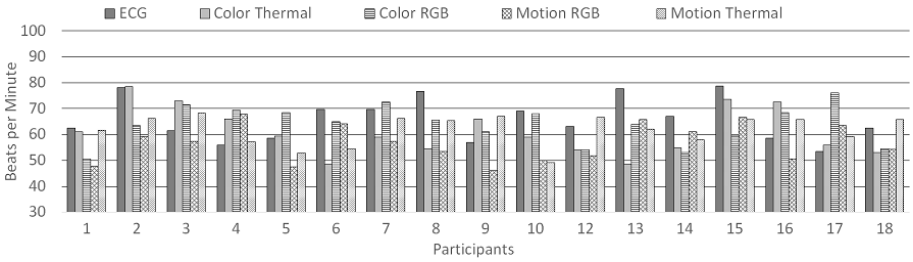
Method	M_E	SD_{M_E}	RMS_E	M_{ER}
RGB_M[6]	14.25	18.46	60.46	25.08
$Thermal_C$	4.70	18.46	19.94	21.78
RGB_C[13]	2.63	10.61	11.16	17.59
$MeanF_C$	1.04	12.37	4.39	16.53
$MeanF_M$	10.05	7.75	42.65	16.31
$MedianF_{C\&M}$	5.54	7.68	25.78	14.39
$Thermal_M$	5.86	8.73	24.84	14.19
$MeanF_{C\&M}$	5.54	7.31	23.52	13.29
$WeightedF_{C\&M}$	3.12	13.44	13.22	10.14

Table 1. Accuracy of all the different methods on the different modalities. Here 'C', 'M' and 'F' stand for 'Color', 'Motion' and 'Fusion' respectively.

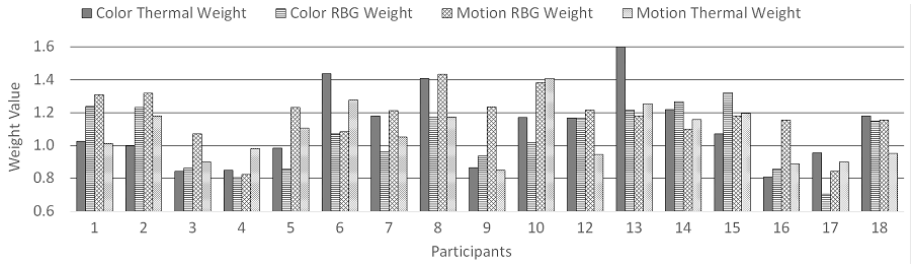
these two correct each other instead of reducing the accuracy. If the results was to be improved, the work effort could be focused on the improvement of color-based and motion-based methods as these had high error rates. Others have reported higher accuracy than what we got: Ref. 16 had 4,65% error rate using motion-based method on RGB images, Ref. 14 had 8,63% error rate of fusing color-based and motion-based method on RGB images, Ref. 10 had 13,2% error rate on color-based method on RGB images. Although these methods used different database than our one, it would be interesting to see the accuracy of the proposed solution with a common database.

5. CONCLUSIONS

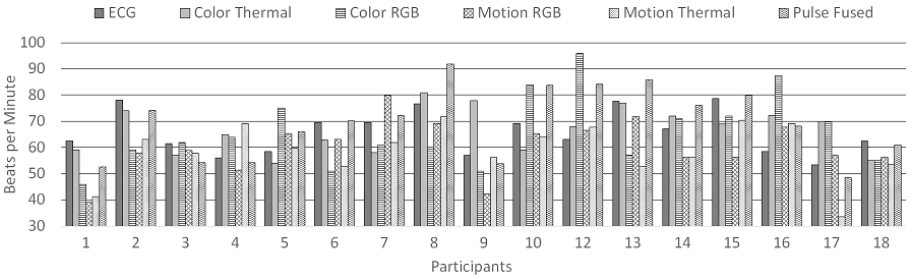
This project investigated 2 hypotheses: (H1) *If RGB and thermal imaging can be combined using mean and median fusion methods to increase the accuracy of estimating HR*, (H2) *If fusing the results could be weighted in such way that the most precise contribute more to the fusion than the others*. To test H1, late fusion was applied to these 4 modalities: color-based HR estimation method on RGB and thermal imaging, and motion-based HR estimation method on RGB and thermal imaging. For H2, weights were trained and evaluated on the 4 modalities. The results show that late fusion increased the accuracy of HR estimation when applied to the 4 modalities. The accuracy was increased even further when the weights were applied. In the future, we will investigate the fusion of RGB and thermal modalities together with the depth videos. Also, the challenging scenarios that include illumination changes and voluntary head motions need to be considered. Early fusion strategies need to be considered as well.



(a) Training data used to train the weights



(b) Weights used to signify the test data



(c) Individual modalities and after fusion using *H2*

Figure 6. The HR estimation results before and after the fusion of modalities for all 18 subjects using *H2*.

REFERENCES

[1] Klonovs, J., Haque, M. A., Krueger, V., Nasrollahi, K., Andersen-Ranberg, K., Moeslund, T. B., and Spaich, E. G., [*Distributed Computing and Monitoring Technologies for Older Patients*], SpringerBriefs in Computer Science, Springer International Publishing, Cham (2016).

[2] Haque, M. A., Nasrollah, K., and Moeslund, T. B., "Can contact-free measurement of heartbeat signal be used in forensics?," in [*Signal Processing Conference (EUSIPCO), 2015 23rd European*], 769–773, IEEE (2015).

[3] van der Haar, D., "Camera-based heart rate estimation for improved interactive gaming," in [*International Conference on Computer Games, Multimedia & Allied Technology (CGAT). Proceedings*], 75, Global Science and Technology Forum (2015).

[4] Sørensen, L. C., Brage-Andersen, L., and Greisen, G., "Effects of the transcutaneous electrode temperature on the accuracy of transcutaneous carbon dioxide tension," *Scandinavian journal of clinical and laboratory investigation* **71**(7), 548–552 (2011).

[5] Haque, M. A., Nasrollahi, K., and Moeslund, T. B., [*Estimation of Heartbeat Peak Locations and Heartbeat Rate from Facial Video*], 269–281, Springer International Publishing, Cham (2017).

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

[7] Shi, S.-Y., Tang, W.-Z., and Wang, Y.-Y., "A review on fatigue driving detection," in [*ITM Web of Conferences*], **12**, 01019, EDP Sciences (2017).

[8] Kwon, S., Kim, H., and Park, K. S., "Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone," in [*Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*], 2174–2177, IEEE (2012). This is the paper that is using green color channel instead of all RGB channels.

[9] Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J. F., and Sebe, N., "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 2396–2404 (2016).

[10] Poh, M.-Z., McDuff, D. J., and Picard, R. W., "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Optics express* **18**(10), 10762–10774 (2010). this is the paper that uses RBG to estimate HR.

[11] Garbey, M., Sun, N., Merla, A., and Pavlidis, I., "Contact-free measurement of cardiac pulse based on the analysis of thermal imagery," *IEEE transactions on Biomedical Engineering* **54**(8), 1418–1426 (2007).

[12] Pérez-Rosas, V., Narvaez, A., Burzo, M., and Mihalcea, R., "Thermal imaging for affect detection," in [*Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*], 36, ACM (2013).

[13] Takano, C. and Ohta, Y., "Heart rate measurement based on a time-lapse image," *Medical Engineering and Physics* **29**(8), 853–857 (2007).

[14] Haque, M. A., Nasrollahi, K., and Moeslund, T. B., "Estimation of heartbeat peak locations and heartbeat rate from facial video," in [*Scandinavian Conference on Image Analysis*], 269–281, Springer (2017).

[15] Li, X., Chen, J., Zhao, G., and Pietikäinen, M., "Remote heart rate measurement from face videos under realistic situations," in [*Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*], *CVPR '14*, 4264–4271, IEEE Computer Society, Washington, DC, USA (2014).

[16] Haque, M. A., Irani, R., Nasrollahi, K., and Moeslund, T. B., "Heartbeat rate measurement from facial video," *IEEE Intelligent Systems* **31**(3), 40–48 (2016).

[17] Haque, M. A., Nasrollahi, K., and Moeslund, T. B., "Real-time acquisition of high quality face sequences from an active pan-tilt-zoom camera," in [*2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*], 443–448 (Aug 2013).

[18] Hartley, R. and Zisserman, A., [*Multiple View Geometry in Computer Vision*], Cambridge University Press, New York, NY, USA, 2 ed. (2003).

[19] Shi, J. et al., "Good features to track," in [*Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*], 593–600, IEEE (1994).

[20] Xu, Y. and Lu, Y., "Adaptive weighted fusion," *Neurocomput.* **168**, 566–574 (Nov. 2015).