Remote Heart Rate Measurement From Face Videos Under Realistic Situations

Xiaobai Li, Jie Chen, Guoying Zhao, Matti Pietikäinen
CMV, University of Oulu, Finland
{lxiaobai, jchen, gyzhao, mkp}@ee.oulu.fi

Abstract

Heart rate is an important indicator of people's physiological state. Recently, several papers reported methods to measure heart rate remotely from face videos. Those methods work well on stationary subjects under well controlled conditions, but their performance significantly degrades if the videos are recorded under more challenging conditions, specifically when subjects' motions and illumination variations are involved. We propose a framework which utilizes face tracking and Normalized Least Mean Square adaptive filtering methods to counter their influences. We test our framework on a large difficult and public database MAHNOB-HCI and demonstrate that our method substantially outperforms all previous methods. We also use our method for long term heart rate monitoring in a game evaluation scenario and achieve promising results.

1. Introduction

Heart rate (HR) is an important indicator of people's physiological state. Traditional HR measurement methods rely on special electronic or optical sensors, and most of the instruments require skin-contact which makes them inconvenient and uncomfortable. On the other side, commercial cameras can be found everywhere nowadays such as web-cams, surveillance cameras, and cellphone cameras. The technique of remote HR monitoring using ordinary cameras would have many potential applications. It has been reported that skin color change caused by cardiac pulse can be captured by ordinary cameras for HR measurement [12, 20], but it is still a challenging task since the change caused by the cardiac pulse is very small compared to other numerous factors that can also cause fluctuation of the gray value of local skin. Among all these factors, illumination variations and subjects' motions are two important ones. In this paper, we propose a novel HR measurement framework, which can reduce the noises caused by illumination variations and subjects' motions. Our results show that the framework can achieve promising results under realistic human computer interaction (HCI) situations.

HR measurement research is a conventional topic in the field of biomedical study, but it is seldom concerned in the video (of visible light) processing domain. The latter is known to be good at recognizing and analyzing explicit characteristics like shapes, textures or movements; while implicit bio-signals like the HR are considered 'out of the range' without the help of special optical equipment. In this paper we demonstrate that computer vision methods can help to solve the problem of remote HR measurement.

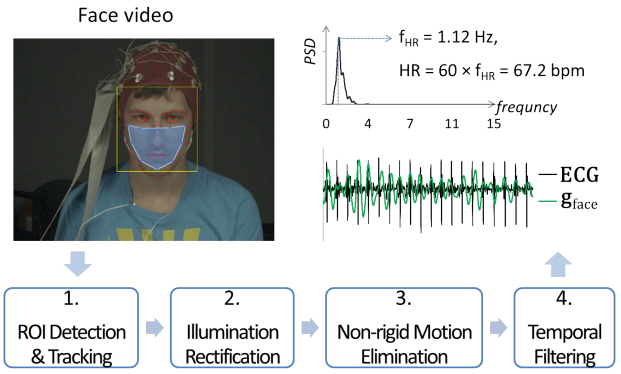

Figure 1. Proposed framework for HR measurement from facial videos in realistic HCI situations.

Remote HR monitoring has many potential applications. With the ability to 'see' inner changes like the heartbeat, video processing research can be broadened in many ways. For example: 1) For remote health care: web-cam can be used for real-time remote medical examinations and support long-term HR monitoring. 2) For affective computing: the focus of the past study is mainly about facial expressions and speech, which are only the out tip of the emotion iceberg; physiological states like the change of HR is inextricably linked with people's emotions, thus should be also integrated to build a comprehensive emotion recognition system. 3) For human behavior analysis: aside from the analysis of explicit behaviors like poses and gestures, inner physiological changes provide additional knowledge for better understanding of people's behavior. 4) For biometrics: the heartbeat could also work as an indicator of vital sign for the anti-spoofing system.

Recently, some investigations [2, 10, 12, 13] have shown that HR can be measured from face videos under well controlled laboratory conditions. But in realistic situations the

task is more difficult because many factors could contaminate the pulse signal measured from face area. For example, in a HCI scenario of video watching or game playing, both environmental illumination variations and subjects' motions can be expected to affect the gray value of the face region. Illumination variations include all forms of noise caused by the change of environment, like the blink of indoor lights, the flash of reflected light from a computer screen and the inner noise of the digital camera. Subject's motion include both rigid movements like head tilt and non-rigid movements like eye blinking and smiling. To our best knowledge, no method has been demonstrated to be able to measure HR successfully under such realistic conditions.

We propose a framework (Figure 1) to reduce the impact of afore-mentioned interferences for remote HR measurement. We use face tracking to solve the problem of rigid head movements; and use the green value of background as a reference to reduce the interference of illumination variations; then segment and shear the signal to reduce the noise caused by sudden non-rigid movements. We demonstrate that our method can significantly reduce the impact of aforementioned interferences and increase the accuracy of HR measurement under realistic HCI situations.

2. Related works

Remote non-intrusive HR measurement is an attractive topic for both commercial and academic purposes. Many past works that attempt for remote heart rate monitoring include the use of photoplethysmography (PPG) [5, 9]. The blood volume of micro-vascular all over the body changes together with cardiac pulse, so the blood volume pulse (BVP) measured at peripheral body tissues (like palm or fingertip) is usually used as an indicator of cardiac cycle measurement. The principle of PPG method is to illuminate the skin with a light-emitting diode (LED) and then measure the amount of light reflected or transmitted to a photodiode. Since the amount of light absorption is a function of the blood volume, PPG can measure the local blood volume pulse. Although it is possible to use PPG based settings to measure HR without any contact, this method still requires special lighting sources and sensors.

In the past few years several papers proposed color-based methods for remote HR measurement using ordinary commercial cameras [10, 12, 13]. Poh *et al*. [12] explored the possibility to measure HR from face videos recorded by a web-cam. They detected the region of interest (ROI, i.e. the face area) using Viola-Jones face detector and computed the mean pixel values of the ROI of each frame from three color channels. Then Independent Component Analysis (ICA) was applied to separate the PPG signal from the three color traces, and the PPG signal was transferred into frequency domain to find the frequency with the max power within the range of [0.7, 4] Hz as the HR frequency. According to previous findings [20], the green channel trace contains the strongest plethysmographic signal among the three color channels. Poh's results showed that comparing to the raw green trace, ICA separated sources can achieve higher accuracy for measuring HR.

The results in [12] were challenged by Kwon *et al*. [10]. Kwon *et al*. recorded face videos with the built-in camera of a smart-phone, and extracted HR using both the raw green trace and the ICA separated sources. They found that ICA slightly dropped the performance which is contrary to Poh's result. Later Poh *et al*. [13] improved their method by adding several temporal filters both before and after applying ICA. The improved method achieved very high accuracy for measuring HR on their self-collected data.

A motion-based method was proposed by Balakrishnan *et al*. recently [2]. Balakrishnan *et al*. tracked subtle head oscillations caused by cardiovascular circulation, and used PCA to extract the pulse signal from the trajectories of multiple tracked feature points. The method achieved promising performance on their self-collected videos. Since the method relies on motion tracking, subjects must avoid voluntary movement in their experiment. Balakrishnan *et al*. indicated that measuring HR on moving subjects would be a valuable future direction.

All these mentioned methods [2, 10, 12, 13] have the following limitations while considering the adaptability and robustness in general application scenarios:

1) In their testing data, neither illumination variations nor subjects' motions were involved since no task was assigned and subjects were asked to keep still during video recording. Controlled settings lead to simple and almost noiseless data, so all the reported results achieved high accuracy (error rate less than 3%). But in realistic HCI scenarios such as people watching movies from a screen, the reflected light from the screen can change dramatically from time to time; rigid head movements and non-rigid motions like facial expressions are also inevitable. It was not known how these methods would perform on challenging data when illumination variations and subjects' motions are both involved.

Poh *et al*. [12] did report an experimental result of HR measurement during motion, but in their experiment motion only meant performed slow and uniform head swings, which is different from spontaneous movements.

2) None of their data [2, 10, 12, 13] is publicly available, and new methods have to come out with new datasets. Repetitive data collection is a waste of time and most importantly the cross-database difference makes it difficult to make fair comparisons of different methods.

In this paper, we propose a new framework for remote HR measurement which can work under challenging realistic HCI situations, and we evaluate it on a multi-modal database MAHNOB-HCI [17]. MAHNOB-HCI is selected for three reasons: 1) it includes large samples of facial

videos and corresponding ground truth HR signals recorded by Electrocardiography (ECG); 2) the videos were recorded in realistic HCI scenarios that both illumination variations and subjects' motions were involved; 3) it is a public database that can be easily accessed by all researchers which makes fair comparison possible.

We describe our framework in Section 3, and then report the results of three experiments in Section 4. In Experiment 1, we compare our method with others on a simple self-collected database; in Experiment 2 we compare them again on MAHNOB-HCI database to demonstrate the robustness of our framework under realistic situations; in Experiment 3 we show that our method can be used in applications like long-term HR monitoring in game evaluation scenario.

3. Framework

Our framework is composed of four steps as shown in Figure 1. In the first step, we need to get the ROI which includes the raw pulse signal from the face video, and deal with the problem of rigid head movement. We use Discriminative Response Map Fitting (DRMF) method [1] to detect facial landmarks and generate a mask of ROI in the first frame, and then employ Kanade-Lucas-Tomasi (KLT) algorithm [19] to track the location of the ROI. The average green value of the ROI in each frame is computed as the raw pulse signal. The purpose of the second step is to reduce interferences caused by illumination variations. We segment the background region using the Distance Regularized Level Set Evolution (DRLSE) method [11], and use its average green value as a reference to model the illumination variations at the ROI. Normalized Least Mean Squares (NLMS) filter [16] is employed to find the optimized coefficient of the model. The aim of the third step is to reduce interferences caused by sudden non-rigid motions. We divide the pulse signal into segments and discard segments contaminated by sudden non-rigid movements. In the fourth step, several temporal filters are applied for excluding powers of frequencies that are out of HR range, and Welch's power spectral density estimation method [22] is employed to estimate the HR frequency. Details of each step are explained in the following subsections.

3.1. ROI Detection and Tracking

Previous HR measurement methods use the Viola-Jones face detector [21] of OpenCV [4] to detect faces. It only finds coarse face locations as rectangles, which is not precise enough for HR measurement task since non-face pixels at corners of rectangles are always included. The case becomes even worse when the face rotates. To this end, we first apply Viola-Jones face detector to detect the face rectangle on the first frame of the input video, then use Discriminative Response Map Fitting (DRMF) method [1] to find the coordinates of 66 facial landmarks inside the face rectangle. DRMF is a discriminative regression based approach for the Constrained Local Models (CLMs) framework, which can find precise facial landmarks in generic face fitting scenario. We use $l = 9$ points out of 66 landmarks to define our region of interest (ROI), and generate a mask of the ROI as the blue region showed in Figure 2.

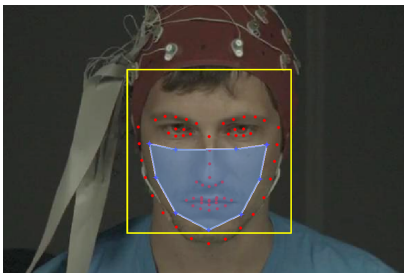

Figure 2. ROI detection and tracking. The yellow line shows the face rectangle, inside which feature points are detected and tracked. The red points indicate 66 landmarks and the light-blue region is the defined ROI.

Two rules are followed for defining the ROI: the first one is to exclude the eye region since blinking may interfere with the estimated HR frequency; the second one is to indent the ROI boundary from the face boundary, otherwise non-face pixels from background might be included during the tracking process.

Then we use tracking to counter the problem of rigid head movement. Poh *et al*. [12] proposed to use face detection on every frame for HR measurement on moving subject, but it is not precise enough as the detected rectangle slightly moves even when the face does not move at all. For our tracking process, feature points are detected inside the face rectangle using the standard 'good feature to track' proposed by Shi *et al*. [15], and are tracked through the following frames using the Kanade-Lucas-Tomasi (KLT) algorithm [19]. For the ith frame, we get the locations of the tracked feature points $\boldsymbol{P}_i$ as $[p_1(i), p_2(i), \ldots, p_k(i)]$, where k is the number of feature points; and the locations of the ROI boundary points $\boldsymbol{Q}_i$ as $[q_1(i), q_2(i), \ldots, q_l(i)]$. We can estimate the 2D geometric transformation of the face between the current and the next frame as: $\boldsymbol{P}_{i+1} = A\boldsymbol{P}_i$, where A is the transformation matrix. We apply transformation A to the current ROI coordinates to get the coordinates of the ROI in the next frame: $\boldsymbol{Q}_{i+1} = A\boldsymbol{Q}_i$.

The tracked ROI contains pixels of facial skin whose color values change with the cardiac pulse. Previous work found that although red, green and blue channels all contains some level of plethysmographic signals, the green channel contains the strongest one among all three [20]. This finding is consistent with the fact that green light is better absorbed by (oxy-) hemoglobin than red light [14], and penetrates deeper into the skin to probe the vasculature as compared to blue light. In our preliminary test we also found the green channel works the best, so we use the green

value in our framework. The raw pulse signal is calculated as the mean green value of pixels inside the ROI of each frame $\boldsymbol{g}_{\text{face}} = [\text{g}_1, \text{g}_2, \ldots, \text{g}_n]$, where n is the frame number.

3.2. Illumination Rectification

In Section 3.1, we tackle the interference caused by rigid head movements. In this section we focus on how to reduce the illumination interference. Let's assume the face video is recorded from a motionless subject, the mean green value of ROI is a function of time. One example of $\boldsymbol{g}_{\text{face}}$ is shown in Figure 3 (the top curve). Two factors affect the values of $\boldsymbol{g}_{\text{face}}$: the first one is the blood volume variations caused by cardiac pulse; the second one is the (temporal) environmental illumination variations during the video recording. We assume the variations of $\boldsymbol{g}_{\text{face}}$ caused by these two factors are additive:

$$\boldsymbol{g}_{\text{face}} = \boldsymbol{s} + \boldsymbol{y}, \tag{1}$$

where $\boldsymbol{s}$ denotes the green value variations caused by cardiac pulse, and $\boldsymbol{y}$ denotes the green value variations caused by illumination changes.

Our goal is to achieve the target signal $\boldsymbol{s}$ and eliminate noise signal $\boldsymbol{y}$. The problem is that $\boldsymbol{y}$ can not be measured directly. However, in an ordinary HCI environment (e.g. subject watches movies from a screen) the lighting sources for the ROI and other objects in the scene are the same, which are mainly composed of indoor lights and the computer screen. In our framework we use the background region as a reference, and denote the background mean green values of each frame as $\boldsymbol{g}_{\text{bg}} = [\text{g}'_1, \text{g}'_2, \ldots, \text{g}'_n]$ (Figure 3, middle curve). According to the idea of [3], we assume both the face ROI and the background are Lambertian models and share the same light sources. We can use a linear function to estimate the correlation of $\boldsymbol{y}$ and $\boldsymbol{g}_{\text{bg}}$:

$$\boldsymbol{y} \approx h\boldsymbol{g}_{\text{bg}}. \tag{2}$$

Instead of eliminating $\boldsymbol{y}$ directly from $\boldsymbol{g}_{\text{face}}$, we can utilize (2) and define the illumination rectified pulse signal as

$$\boldsymbol{g}_{\text{IR}} = \boldsymbol{g}_{\text{face}} - h\boldsymbol{g}_{\text{bg}}, \tag{3}$$

which according to (1) becomes

$$\boldsymbol{g}_{\text{IR}} = \boldsymbol{s} + (\boldsymbol{y} - h\boldsymbol{g}_{\text{bg}}). \tag{4}$$

Now our goal is to find the optimal h to minimize the error, which is the part of $(\boldsymbol{y} - h\boldsymbol{g}_{\text{bg}})$ in (4).

The optimal h can be found iteratively by using Normalized Least Mean Square (NLMS) adaptive filter [16], which is a variant of the Least Mean Square (LMS) adaptive filter [8]. It is shown that the LMS filter can efficiently reduce motion artifacts in PPG researches [5, 6].

Let's assume at each time point j, $h(j)$ is the currently estimated filter weight. The LMS filter starts from an initial $h(0)$ and updates it after each step with a stepsize μ as

$$h(j+1) = h(j) + \mu\boldsymbol{g}_{\text{IR}}(j)\boldsymbol{g}_{\text{bg}}(j), \tag{5}$$

until $h(j)$ converges to the optimum weight that minimize $(\boldsymbol{y} - h\boldsymbol{g}_{\text{bg}})$ (or the input signal reaches the end).

A problem with the LMS filter is that it is sensitive to the scaling of input signals, which can be solved by normalizing the power of the input signals [16]:

$$h(j+1) = h(j) + \frac{\mu\boldsymbol{g}_{\text{IR}}(j)\boldsymbol{g}_{\text{bg}}(j)}{\boldsymbol{g}^H_{\text{bg}}(j)\boldsymbol{g}_{\text{bg}}(j)}, \tag{6}$$

where $\boldsymbol{g}^H_{\text{bg}}(j)$ is the Hermitian transpose of $\boldsymbol{g}_{\text{bg}}(j)$, and the normalizing quantity $\boldsymbol{g}^H_{\text{bg}}(j)\boldsymbol{g}_{\text{bg}}(j)$ is the input energy.

We use the Distance Regularized Level Set Evolution (DRLSE) method [11] to segment the background region of the video, and achieve $\boldsymbol{g}_{\text{bg}}$ by computing the background mean green value of each frame. With $\boldsymbol{g}_{\text{face}}$ and $\boldsymbol{g}_{\text{bg}}$ as known variables, we can use (6) to obtain the optimal h, which can be put in equation (3) to obtain the illumination rectified signal $\boldsymbol{g}_{\text{IR}}$. One example of $\boldsymbol{g}_{\text{IR}}$ is shown in Figure 3 (bottom curve), in which the illumination variations are reduced and the pulse becomes more visible. The values of the optimal h vary for different input videos, since the distances from the lighting source to the face and the background may vary and the reflectivity of subjects' skin are also different.

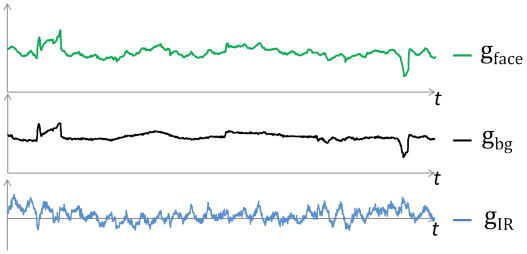

Figure 3. Use NLMS filter to reduce noise caused by illumination variation. The green curve is a raw pulse signal of ROI, the black curve is the corresponding mean green values of background; the blue curve is the filtered signal, of which the noise caused by illumination variations is reduced and the pulse becomes more visible.

3.3. Non-rigid Motion Elimination

One problem remaining unsolved is the non-rigid movements inside the ROI. For example, facial expressions could contaminate the pulse signal and the previous two processes cannot remove it. Figure 4 (top curve) shows one signal which presents the onset of a smiling. The face is neutral in phase 1; and the subject starts to smile in phase 2 which leads to quick and dramatic fluctuations of the signal; then the face reaches to a comparatively stable state in phase 3. If noisy segments like in phase 2 are not excluded, they will end up as big sharp peaks after all the temporal filtering process in the next step. When transfered to the frequency domain, these big sharp peaks will significantly affect the power spectral density (PSD) distribution as they contain

the major part of the power of the whole signal, thus impede the detection of true pulse frequency. It is the remaining part of the signal with smoother changes that contributes to the HR related power spectrum.

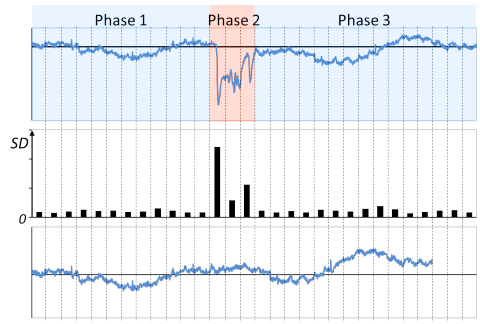

Figure 4. Motion elimination of a pulse signal contaminated by sudden non-rigid motions. The top curve shows the contaminated signal, where a smile was onset in phase 2; the middle bar chart shows the standard deviation (SD) of each segment of the top curve divided by vertical lines; the bottom curve shows the sheared (discard three segments of phase 2) and re-concatenated signal.

Since we are measuring the average HR over a time span (e.g. 30 seconds), such noisy segments can be excluded to achieve more stable results. We divide the $\boldsymbol{g}_{\text{IR}}$ into m segments of the same length $\boldsymbol{g}_{\text{IR}} = [s_1, s_2, \ldots, s_m]$, each segment s is a signal of length $\frac{n}{m}$. The standard deviation (SD) of each segment (Figure 4, middle) is calculated, and 5% (of all testing samples) of segments with the largest SD are discarded. The remaining segments are re-concatenated (Figure 4, bottom curve). The process aims for excluding the noisiest segments contaminated by sudden non-rigid movements, so not all pulse signals need to be cut.

3.4. Temporal Filtering

In this step we apply several temporal filters to exclude frequencies outside the range of interest. We set the frequency range of interest to [0.7, 4] Hz to cover the normal range of HR from 42 beat-per-minute (bpm) to 240 bpm. Several temporal filters have been demonstrated to be helpful for HR measurement in previous research [13]. Here we use three filters: the first one is a detrending filter based on a smoothness priors approach [18], which is used for reducing slow and non-stationary trend of the signal. The second one is a moving-average filter, which removes random noise using temporal average of adjacent frames. The third one is a Hamming window based finite impulse response bandpass filter with cutoff frequency of [0.7, 4] Hz.

After filtering, the pulse signal is converted to the frequency domain and its power spectral density (PSD) distribution is estimated using Welch's method [22]. The PSD estimates the signal's power distribution as a function of frequency. We use the frequency with the maximal power response as the HR frequency f_{HR} (Figure 1 top-right), and the average HR measured from the input video is computed as $HR_{\text{video}} = 60 f_{\text{HR}}$ bpm.

4. Experiments

We evaluate our framework using three experiments. All approaches are implemented using MATLAB of version 2013a under Windows 7 operating system.

4.1. Experiment 1: VideoHR Database

We re-implement previously proposed methods and test them on a simple database 'VideoHR' collected by ourselves, since none of the datasets used in the previously published papers is public. The purpose of Experiment 1 is to demonstrate that we have correctly re-implemented the methods. We refer VideoHR as a 'simple database', because neither ambient illumination variations nor body movement was involved during the video recording.

We use the built-in frontal iSight camera of an IPAD to record videos in a lab with two fluorescent lamps as the illumination sources. All videos are recorded in 24-bit RGB color format at 30 frames per second (fps) with resolution of 640×480 and saved in MOV format. A Polar S810 HR monitor system [7] is used to record the ground truth HR. Ten subjects (two females and eight males) aged from 24 to 38 years were enrolled. During the recording, subjects were asked to sit still on a chair and try to avoid any movement. The IPAD was fixed on a tripod at about 35 cm from the subject's face. Each subject was recorded for about 40 seconds, and 30 seconds (frame 301 to 1200) video of each subject is used for the testing.

We re-implement four previous methods: three color-based methods (Poh2010 [12], Kwon2012 [10], Poh2011 [13]) and one motion-based method (Balakrishnan2013 [2]). In Poh2011 and Balakrishnan2013, they also used customized peak detection functions to find the location of each heart beat for further HR variation analysis. We did not replicate the peak detection process here since we only aim to compare the accuracy of the methods on estimating the average HR. Fourier transformation is applied at the last stage for each method to find the average pulse frequency. The results of all methods on VideoHR database are shown in Table 1. The measure error is computed as $HR_{\text{error}} = HR_{\text{video}} - HR_{\text{gt}}$, where HR_{video} denotes HR measured from video, and HR_{gt} is the ground truth HR obtained from Polar system.

Different kinds of statistics were used in previous papers for evaluating the accuracies of HR measurement methods. To comprehensively compare the methods in multiple aspects, we include all five kinds of statistics used in former research works. The first one is the mean of HR_{error} denoted as M_{e}; the second one is the standard deviation of HR_{error} denoted as SD_{e}; the third one is the root mean squared error denoted as $RMSE$; the fourth one is the mean of error-rate percentage $M_{\text{eRate}} = \frac{1}{N}\sum_{v=1}^{N} |HR_{\text{error}}(v)|/HR_{\text{gt}}(v)$, where

N is the number of videos of the database, and the fifth one is the linear correlation between HR_{video} and HR_{gt} accessed using Pearson's correlation coefficients r and its p value. Pearson's r varies between -1 and 1, where $r = 1$ indicates total positive correlation and $r = -1$ indicates total negative correlation. The p value is the probability of the statistical significance test about if the calculated r were in fact zero (null hypothesis). Usually the result is accepted as statistically significant when $p < 0.01$.

Method	$M_{\text{e}}(SD_{\text{e}})$ (bpm)	*RMSE* (bpm)	M_{eRate}	r
Poh2010	0.37(1.03)	1.05	1.07%	0.99*
Kwon2012	-0.16(1.59)	1.52	1.54%	0.98*
Poh2011	0.37(1.50)	1.47	1.65%	0.98*
Balakrishnan2013	-0.14(1.41)	1.35	1.51%	0.99*
Ours	0.72(1.10)	1.27	1.53%	0.99*

Table 1. Performance on VideoHR database. The marker * indicates the correlation is statistically significant at $p = 0.01$ level.

From Table 1 we can see that all four methods performed perfectly with M_{eRate} lower than 2% and correlation r larger than 0.98 on VideoHR database. It shows that all these methods got almost perfect results on this dataset as these methods performed over their own datasets. However, VideoHR dataset is prone to be ideal without illumination variations and subjects' motions. In realistic situations, these challenges always happen. To test the robustness of these methods over these challenges, we carry out an experiment over a difficult database in Experiment 2.

4.2. Experiment 2: MAHNOB-HCI Database

In this experiment we test the four previous methods again on MAHNOB-HCI database. We demonstrate that our proposed method can reduce noises caused by illumination variations and subjects' motions, and substantially outperform previous methods. MAHNOB-HCI is referred as a difficult database here since the videos were recorded in realistic HCI scenarios, both illumination variations and subjects' movements were involved.

MAHNOB-HCI is a public multi-modal database recorded by Soleymani *et al*. [17]. MAHNOB-HCI includes data from two experiments: one is 'emotion elicitation experiment' and the other is 'implicit tagging experiment'. We use the color videos recorded in their 'emotion elicitation experiment' for our testing.

27 subjects (15 females and 12 males) were involved, their ages varied from 19 to 40 years. 20 frontal face videos were recorded for each subject with resolution of 780×580 pixels at 61 fps, while they were watching movie clips from a computer screen. ECG signals were recorded using three sensors attached on participants body, and we used the second channel (EXG2) to obtain the HR_{gt}. Altogether 527 (13 cases lost) intact video clips and their corresponding ECG signals are used in our test. Original videos are of different lengths. We exerted 30 seconds (frame 306 to 2135) from each video and measured the average HR. More details about MAHNOB-HCI database are given in [17].

Method	$M_{\text{e}}(SD_{\text{e}})$ (bpm)	*RMSE* (bpm)	M_{eRate}	r
Poh2010	-8.95(24.3)	25.9	25.0%	0.08
Kwon2012	-7.96(23.8)	25.1	23.6%	0.09
Poh2011	2.04(13.5)	13.6	13.2%	0.36*
Balakrishnan2013	-14.4(15.2)	21.0	20.7%	0.11
Ours				
Step 1+4	-3.53(8.62)	9.31	8.03%	0.69*
Step 1+2+4	-3.46(7.36)	8.13	7.02%	0.79*
All steps	**-3.30(6.88)**	**7.62**	**6.87%**	**0.81***

Table 2. Performance on MAHNOB-HCI database. The marker * indicates the correlation is statistically significant at $p = 0.01$ level.

The results on MAHNOB-HCI are shown in Table 2. The statistics of errors are computed in the same way as we did in Experiment 1. Comparing to the results of Experiment 1, the performance of all four previous methods drops significantly. Poh2011 method performs better than the other three, because it employs several temporal filters (we adopt these filters in step 4 of our framework) to purify the signal. But a correlation of $r_{(527)} = 0.359$ and a M_{eRate} of 13.2% indicate that Poh2011 method is not robust enough to make reliable estimations about the true HRs.

We test our method on MAHNOB-HCI step-by-step and the performance is shown in Table 2. Step 4 is always included as it is a must for achieving the average f_{HR}. For the Step 1 of ROI detection and tracking, the face detector made false detections on 14 of 527 cases. To avoid the false alarms, we discard detected rectangles whose edge lengths are less than 100 (the average face size in MAHNOB-HCI is about 200×200). Then the DRFM method is applied to detect face contours and eye positions according to the correct face rectangles. In some cases the detected landmarks may not be precise, but generally they are good enough for the purpose of defining ROI. The defined ROI covers an area of about 20000 pixels. Compared to the performance of Poh2011, the better selected ROI and tracking process help to achieve a much lower M_{eRate} of 8.03 % and increase the correlation r from 0.36 to 0.69.

For the Step 2 of illumination rectification, we use the background as the reference. Since the videos of MAHNOB-HCI were recorded in a relatively dark environment, the illumination variations caused by the computer screen can be captured by the background. By using NLMS filter we further increased the correlation r to 0.79 and lowered the M_{eRate} by 1.01%. The value of the stepsize μ can affect the performance of NLMS filter, which was also no-

ticed in previous papers [6]. The response curve of the correlation coefficient r versus stepsize μ is shown in Figure 5 (a). When the stepsize is small, the input signal might not be long enough for the filter to reach convergence. Here the best r was achieved when μ is bigger than 0.003. The weight of the NLMS filter was initialized as zero.

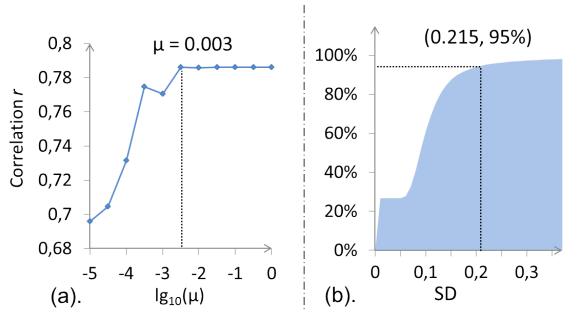

Figure 5. (a). Correlation coefficient r computed at different stepsizes μ. The filter reaches convergence when μ is bigger than 0.003. (b). The cumulative distribution function of SDs of all sample segments from MAHNOB-HCI. Threshold of $SD = 0.215$ is used for shearing the top 5% segments with the largest SD values.

For the Step 3 of non-rigid motion elimination, we divide signals into segments of one second, and the cumulative distribution function (CDF) of SDs of all sample segments are shown in Figure 5 (b). A cutoff threshold of $SD = 0.215$ is used for the shearing, and 228 out of 527 signals are sheared. The improvement made by Step 4 is not as big as two former steps since not all cases are affected.

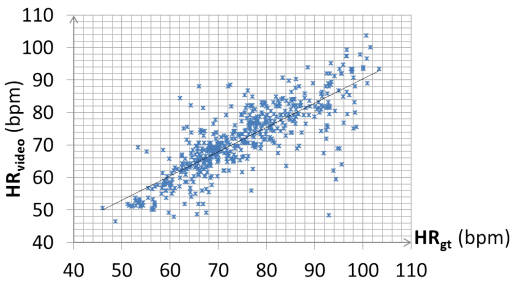

Figure 6. The scatter plot comparing the HR_{video} measured by our method with the ground truth HR_{gt} from ECG.

The HR_{video} measured with all four steps of our framework are plotted against the HR_{gt} in Figure 6. It can be seen from this figure that overall our predicted HR is well correlated with the ground truth. On a wide range of HR from 46 bpm to 103 bpm, good HR estimations are made in most cases. There are some out-lier points falling far from the correlation line which indicate poor estimations. We check these poorly estimated cases and find that in some of these videos head rotations of more than 60 degrees were involved, which caused errors in the tracking of the face ROI. For application scenarios like detecting the vital signs of an emergency situation, HR measurement with error less than 5 bpm is likely to be acceptable [12]. In order to check how many cases were well estimated, we further compared the distributions of HR_{error} of our method with Poh2011 method, which got the best performance among four previous methods. As shown in Figure 7, we estimate the HR of 403 cases (76.5%) with errors less than 5 bpm, while for Poh2011 method the number is only 296 (56.2%).

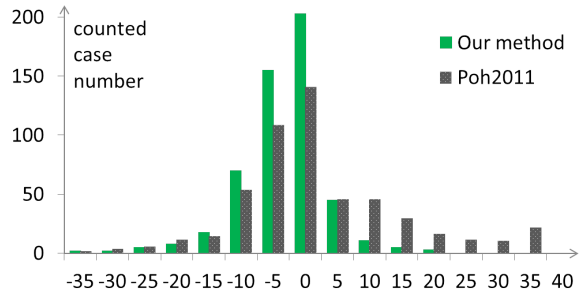

Figure 7. Comparing the distributions of HR_{error} of our method with Poh2011 method.

4.3. Experiment 3: HR monitor for game evaluation

Without the restriction of motion and illumination changes, our method can be applied for long-term HR monitoring when subjects are performing some tasks. Here we test it on one subject in a game playing scenario. In game research, user tests are usually carried out for analyzing users' experiences of game playing. Results of user tests will help game developers in their future designing work.

We record the face video of one subject for 10 minutes while the subject is playing a video game. The same setting is used for data recording as in Experiment 1. The distance between subject's face and the camera is about 50 cm. The average HR of every 10 seconds is computed from both the face video and the ground truth, and the results of HR_{video} and HR_{gt} are plotted in Figure 8. It can be seen that the subject's HR changes as the content of the game progresses, which can be used for later analysis about the player's experience. HRs measured by using our framework has a mean error rate of 1.89%.

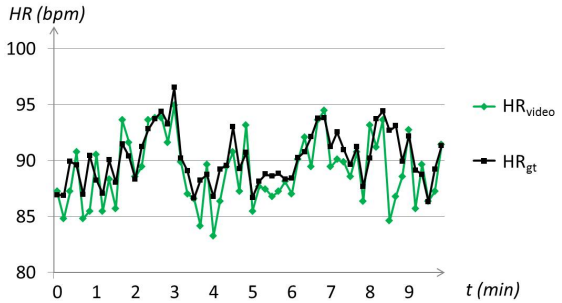

Figure 8. HR monitoring of one subject while playing a video game. The black curve is the ground truth HR measured by Polar system; the green curve is HR measured from video by using our framework.

5. Conclusions

Previous methods of remote HR measurement from ordinary face videos can achieve high accuracies under well controlled situations, but their performance degrades when environmental illumination variations and subjects' motions are involved. Our proposed framework contains three major processes to reduce these interferences: first, we employ DRMF to find the precise face ROI and use tracking to address the problem caused by rigid head movement; second, NLMS adaptive filter is employed to rectify the interferences of illumination variations; third, signal segments with big SD values are discarded in order to reduce the noise caused by sudden non-rigid movements. We have demonstrated that all three processes help to improve the accuracy of HR measurement under realistic HCI situations.

MAHNOB-HCI database is used for the testing since all afore-mentioned interferences are involved in the videos. For the ROI detection, we include the mouth region to cover more skin pixels in the ROI since talking is seldom involved in MAHNOB-HCI videos. But in conversation scenarios the mouth region could be excluded to avoid motion noise. For the step of illumination rectification, we use the background as the reference, because the videos of MAHNOB-HCI are dark and no other object is present in the scene. In scenarios when the background is not suitable as the reference, other static objects can be used as the reference; for example a gray board can be placed aside the subject's head.

Our proposed method substantially outperformed the four previous methods and achieved an average error rate of 6.87% on all 527 samples of MAHNOB-HCI. One factor that impacts our results is head rotations of an large angle, especially in the yaw direction. In such situations feature points on half of the face are lost and the tracked ROI location may be erroneous thus degrade the accuracy of HR measurement. In the future work, the ROI tracking could be improved to tolerate more extreme head movements.

In the current study averaged HR is estimated for one input video as the first trial to make the system work under realistic situations. HR monitoring during game playing is shown as an example of applications of this system. In future, more efforts will be devoted to detecting the peak of each heartbeat, so that sophisticated analysis about the heart rate variation (HRV) can be made which will help us to get more information about the subject's physiological status.

Acknowledgement

This work was sponsored by the Academy of Finland, and Infotech Oulu.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, 2013.

[2] G. Balakrishnan, F. Durand, and J. Guttag. Detecting pulse from head motions in video. In *CVPR*, 2013.

[3] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 2003.

[4] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[5] G. Cennini, J. Arguel, K. Akşit, and A. van Leest. Heart rate monitoring via remote photoplethysmography with motion artifacts reduction. *Optics express*, 2010.

[6] K. Chan and Y. Zhang. Adaptive reduction of motion artifact from photoplethysmographic recordings using a variable step-size lms filter. In *Proceedings of IEEE on Sensors*, 2002.

[7] F. X. Gamelin, S. Berthoin, and L. Bosquet. Validity of the polar s810 heart rate monitor to measure rr intervals at rest. *Medicine and Science in Sports and Exercise*, 2006.

[8] M. H. Hayes. 9.4: Recursive least squares. *Statistical Digital Signal Processing and Modeling*, 1996.

[9] K. Humphreys, T. Ward, and C. Markham. Noncontact simultaneous dual wavelength photoplethysmography: a further step toward noncontact pulse oximetry. *Review of scientific instruments*, 2007.

[10] S. Kwon, H. Kim, and K. S. Park. Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone. In *EMBS*, 2012.

[11] C. Li, C. Xu, C. Gui, and M. D. Fox. Distance regularized level set evolution and its application to image segmentation. *IEEE Trans. on Image Processing*, 2010.

[12] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 2010.

[13] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Trans. on Biomedical Engineering*, 2011.

[14] S. Prahl. Optical absorption of hemoglobin. `http://omlc.ogi.edu/spectra/hemoglobin/`, 1999.

[15] J. Shi and C. Tomasi. Good features to track. In *CVPR*, 1994.

[16] H. Simon. *Adaptive filter theory*. Prentice Hall, 2002.

[17] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. on Affective Computing*, 2012.

[18] M. P. Tarvainen, P. O. Ranta-aho, and P. A. Karjalainen. An advanced detrending method with application to hrv analysis. *IEEE Trans. on Biomed. Eng.*, 2002.

[19] C. Tomasi and T. Kanade. *Detection and tracking of point features*. CMU, 1991.

[20] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 2008.

[21] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[22] P. Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans. on Audio and Electroacoustics*, 1967.