ROBUST ADAPTIVE HEART-RATE MONITORING USING FACE VIDEOS

Puneet Gupta, Brojeshwar Bhowmik, Arpan Pal
TCS Innovation Lab Kolkata,
E-mail:gupta.puneet5,arpan.pal@tcs.com

Abstract

Heart rate (HR) monitoring is indispensable for several real-world scenarios, especially when acquired in a non-contact manner. It can be accomplished using face videos acquired from ubiquitous cameras in an inexpensive, non-invasive and unobtrusive manner. But the HR monitoring can be erroneous when the video contains facial expressions, out-of-plane movements, change in camera parameters (like focus) and variations in environmental factors (like illumination). The proposed system mitigates these problems for improving the HR monitoring. For this, it defines an adaptive temporal signal selection mechanism which identifies and removes the facial areas affected by facial expressions. Moreover, it introduces a novel post-processing mechanism which perform HR monitoring by utilizing face reconstruction and quality. The post-processing is used when the face video contains facial movements. Experimental results reveal that incorporation of adaptive temporal signal selection and post-processing mechanisms can significantly improve the HR monitoring. It depicts that the Pearson correlation between actual and estimated HR is 0.95 while the average absolute error is 1.63 beats per minute, which indicates that the proposed system provides good HR monitoring.

1. Introduction

Doctors extensively utilize heart rate (HR) monitoring for diagnosing the human illness [3]. Applicability of HR monitoring in the human physiological and pathological parameters evaluation, has elicited the attention of the following fields: i) Healthcare, for understanding nervous system, analyzing cardiac diseases and monitoring exercise [22]; ii) psychology, for monitoring the stress and understand the mental state [16]; iii) biometrics, for liveness and spoof detection [15]; and iv) affective computing, for human emotion recognition [10]. HR monitoring can assist these fields provided the monitoring is accurate, the sensor involved in the acquisition is cheap and the acquisition can be carried in a user-friendly manner. Existing HR monitoring approaches based on electrocardiography (ECG) and photoplethysmography (PPG) require skin contact due to which they are not user friendly, cannot perform long-term monitoring and unable to analyze neonates, skin damaged persons and human while sleeping or exercising. Moreover, they require bulky, expensive and dedicated sensors which further limits the applicability. This has motivated to design the HR monitoring system using face videos which allow the acquisition from cheap and portable sensor in a non-contact manner.

The ideology behind HR estimation using face videos is that blood traverse in the human body and varies according to the heart beat. These blood variations can be noticed in terms of change in facial color and head motion. Both the color and motion variations are subtle and hence imperceptible to the human eye, but these can be perceived using camera [19]. Existing face videos based HR estimation systems consist of four stages, viz., Region of interest (ROI) detection, temporal signal extraction, pulse extraction and HR estimation. Facial area containing useful information about blood variations is used to define the ROI. Temporal signals are evaluated by analyzing the color and motion variations introduced in the ROI over time. Noise present in the temporal signal is reduced by applying filtering techniques. The resulting signals are used to extract the cardiovascular pulse signal after applying blind source separation techniques. Statistical measures are applied to the pulse signal for extracting the pulse spectrum [14]. HR is estimated from the spectrum using Fast Fourier Transform (FFT) spectrum or R-R intervals [11].

The temporal signals extracted from face videos contain pulse signal along with the noise originated due to camera sensor noise, inevitable eye blinking; facial poses and expression variations; camera parameters (like, auto-focus can change illumination); face movements; respiration; and environmental factors (like, illumination variations). These issues make correct HR monitoring using face videos challenging. The HR monitoring system that can mitigate these issues is proposed in this paper. It is important to note that most of the existing work in the realm of face video based HR analysis focuses on HR estimation rather than HR monitoring as in the proposed system. As opposed to the HR

978-1-5386-4886-5/18/$31.00 ©2018 IEEE
DOI 10.1109/WACV.2018.00064

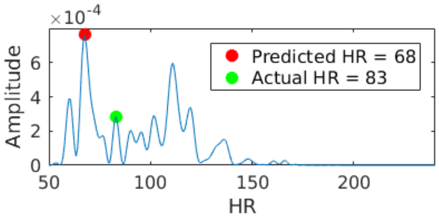


Figure 1. Spurious HR estimation due to noise in pulse Spectrum

estimation which provide one value for the full temporal signal, the HR monitoring provides multiple HR estimations in small time intervals. The monitoring provides more valuable information (like heart rate variability) for medical diagnosis than an HR estimate for the full video. But unfortunately HR monitoring is more vulnerable than HR estimation for the noise affecting few frames like noise due to facial expressions and illumination changes. It is because HR monitoring utilizes small number of frames as opposed to HR estimation. The challenges involved in the HR monitoring are addressed in the proposed system by incorporating the following main research contributions:

1. Variations introduced in the temporal signal due to facial expression and slight movement of facial areas can result in erroneous HR. It is observed that these variations can impact small area like smile usually affect the lip areas, but does not alter the forehead areas. This is leveraged for improving the HR monitoring by proposing an adaptive temporal signal selection mechanism which adaptively identifies and removes the facial areas affected by the facial deformations.
2. HR monitoring is composed of multiple HR estimations which are evaluated in small time intervals. Some of these estimations in HR monitoring can be spurious when the frames utilized in the estimation are affected by noise due to facial expressions and illumination changes. It is observed that in such cases, a large number of frames should be employed to provide better HR estimates. Hence HR monitoring is improved by proposing a post-processing mechanism which determines the erroneous HR estimates and rectify them by analyzing the full temporal signals.
3. HR monitoring can be erroneous due to out-of-plane movements [21]. Face reconstruction can be employed to mitigate the artifacts arise from out-of-plane movements, which in turn provide better HR monitoring. Thus, the proposed system explored the applicability of face reconstruction for improving the HR monitoring. Further, face reconstruction requires large time computation and it can be inaccurate, hence it is utilized in the proposed system only when correct HR cannot be estimated by avoiding the reconstruction.

This paper is organized in the following manner. The background for augmenting the understanding of the proposed system is discussed in the next section. The proposed system is described in Section 3 followed by experimental results in Section 4. Conclusions are given in the last section.

2. Background

Typically, HR estimation using face videos works in the following manner. ROI containing the useful blood flow information, is first extracted from the face video. Usually the full face, forehead region or cheek areas are used to define ROI [6]. Temporal signals are extracted from the ROI using Eulerian [14] or Lagrangian techniques [1]. Discriminating features are explicitly tracked in the face video to extract the Lagrangian temporal signal. Such temporal signals are usually avoided because their tracking is time-consuming and often erroneous when few discriminatory features are available for tracking. On the other hand, temporal signal extraction using Eulerian techniques, examines the variations in a fixed ROI [14] and thus they require less time computations than the Lagrangian techniques. The temporal signal contains pulse signal along-with the noise thus blind source separation techniques are applied to the temporal signal for extracting the pulse signal. The pulse spectrum is obtained by applying FFT on the pulse signal and HR is estimated from it using the observation that the peak containing maximum amplitude in the spectrum corresponds to HR [11].

Noise present in the temporal signal can introduce spurious peaks in the spectrum due to which it may happen that the maximum amplitude peak does not correspond to the actual HR. Figure 1 shows such an example. Filtering techniques can be used to reduce the noise [17]. Few systems for noise reduction are proposed in [5, 9, 8]. The noise present due to motion artifact is reduced by face registration in [5] but it cannot handle facial pose variations and out-of-plane deformations. Illumination changes can be estimated using background variations [9] and brightness [8] but these are often spurious due to background characteristics [7].

3. Proposed System

In this section, the HR monitoring system using face video is proposed. The flow-graph of the system is illustrated in Figure 2. Face video is divided into overlapping windows and each window is analyzed to estimate the HR. The HR pertaining to a window is referred as local HR. For the local HR estimation, several ROIs are detected and temporal signals are subsequently extracted. Pulse signal is extracted from only those ROIs that are least affected by the facial deformations. Local HR and quality are estimated using the pulse signal. Eventually, the local HR is post-processed to obtain better estimate of local HR.

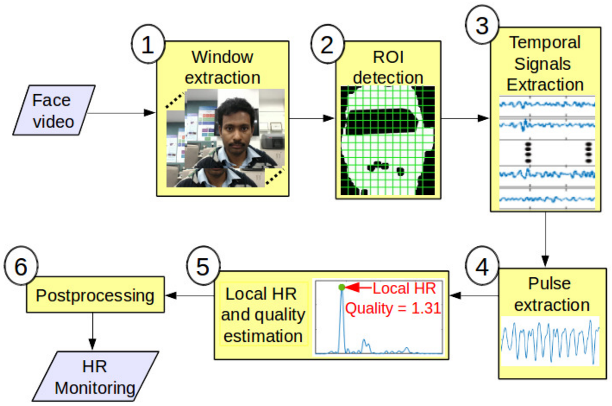


Figure 2. Flow-graph of the Proposed System

3.1. Window Extraction

In HR monitoring, several local HRs are estimated at multiple contiguous time intervals and they are eventually concatenated. For this purpose, face video is divided into multiple overlapping windows such that the width of each window is chosen to be 250 frames and the overlap between successive windows is 150 frames.

3.2. ROI Detection

In this subsection, ROI consisting of skin pixels is detected using the first video frame. The rectangular block consisting of the face is determined by applying Viola-Jones face detector [20] on the frame. Apart from the facial skin pixels, the block also consists of pixels belonging to background, eye areas, hairs and moustaches which are devoid of any HR information. Such non-informative pixels are detected by applying skin detection proposed in [13] and subsequently they are removed from the ROI. It is observed that the eye areas are influenced by inevitable eye blinking which results in spurious HR estimation. Hence, these areas are detected by applying Viola-Jones eyes detector [20] and removed. Another prominent factor for erroneous HR estimation is the unavoidable slight variations in face boundary pixels that tremendously alter the temporal signals. Hence, morphological erosion operation is performed to remove the boundary pixels. The resulting area is divided into multiple non-overlapping square blocks and each block containing only skin pixels is considered as a ROI. The full face area can be considered as one ROI [7] that can result in faster time computation of pulse signal. However, the proposed system avoids this proposition because it results in an erroneous HR estimation due to the irregular effect of emotions and blood flow on different facial regions.

3.3. Temporal Signal Extraction

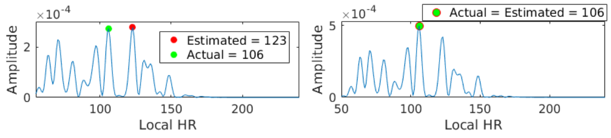
The temporal signals depicting the subtle facial variations due to blood flow, are extracted in this subsection. Eulerian philosophy is used rather than Lagrangian philosophy for the extraction because: i) features available in the ROI are less discriminatory which makes Lagrangian temporal signals error-prone; and ii) signal extraction using Lagrangian philosophy is highly time consuming as compared to the extraction using Eulerian philosophy [14]. Amongst the RGB color channel, the strongest plethysmographic signals is present in the green channel [18]. Hence, the temporal signal is defined by the mean green value of the pixels in a block. Mathematically, the temporal signal corresponding to i^{th} ROI is given by:

$$T_i = \left[\sum_{k=1}^{n} G_{i,1}^{k}, \sum_{k=1}^{n} G_{i,2}^{k}, \sum_{k=1}^{n} G_{i,3}^{k}, ..., \sum_{k=1}^{n} G_{i,f}^{k} \right] \quad (1)$$

where $G_{i,j}^{k}$ denote the green channel intensity of the k^{th} pixel belonging to the i^{th} ROI of j^{th} frame; f is the number of frame; and n is the number of pixels in the ROI.

Each temporal signal contains heart information along with noise due to facial expression, respiration, illumination and facial movements. The following techniques are utilized to mitigate the noise:

1. *Temporal filtering*: Normally the heart beats at a rate of 42 to 240 beats per minute (BPM). Thus, band-pass filter is applied, which removes the frequencies outside the range of 0.7 to 4 Hz. It is useful to remove the noise due to respiration rate, which lies outside this frequency range. Further, it is observed that changes in illumination can introduce non-stationary trends in the temporal signals. Such trends are alleviated by Detrending filter [17].
2. *Adaptive temporal signal selection*: Facial expressions introduce large variations in the temporal signals, but fortunately few temporal signals are affected by the expressions. As an instance, when the person smiles, temporal signals obtained from mouth area are affected while temporal signals obtained from the nose and head are least affected. For correct HR estimation, the proposed system removes those temporal signals that can be affected by the facial expressions. The importance of avoiding the affected temporal signals can be visualized from Figure 3. It can be observed from Figure 3(a) that estimated local HR is significantly deviated from the actual local HR when the temporal signals affected by expressions are considered. But it can be seen from Figure 3(b) that when the affected temporal signals are not considered the estimated and actual HR are same. The affected temporal signals are detected using the intuition that the facial expression tremendously alter the amplitude of the affected temporal signals and thereby increases the standard deviation. This is leveraged by defining the confidence of whether a signal is affected by the expression or not in



(a) (b)

Figure 3. Importance of Adaptive Temporal Signal Selection: Pulse spectrum obtained when adaptive selection is avoided and used are shown in a) and b) respectively.

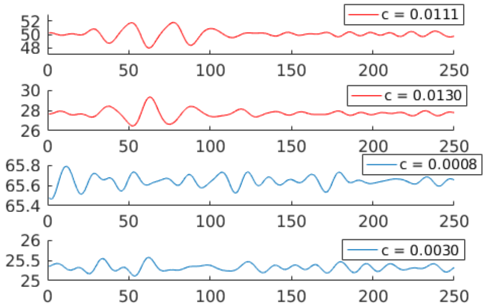


Figure 4. Examples of temporal signals discarded and utilized in the proposed adaptive Temporal Signal Selection according to the confidence, c are shown using red and blue colors respectively.

the following manner:

$$c_i = \frac{std\left(T_i\right)}{mean\left(T_i\right)} \qquad (2)$$

where c_i denote the confidence of T_i temporal signal while std and $mean$ are the standard deviation and mean operators respectively. In Equation (2), confidence is normalized by dividing with their mean because different ROI have different signal strength depending upon the facial structure and the blood flow mechanism. Top 20% of the temporal signals having large confidence values are discarded to reduce the problems due to facial expressions. Figure 4 shows some examples of temporal signals that are discarded and utilized according to the confidence, c.

3.4. Pulse Extraction

Each temporal signal contains pulse information along with noise. In this subsection, blind source separation is used to estimate the source signal, i.e., pulse signal. Blood flow in the face is dependent on the bones and artery structure of the face which results in variable amplitude of temporal signals. Hence, the temporal signals are normalized using z-score normalization to maintain the signal shape. The normalized temporal signal, T_i can be written as:

$$T_i = \mathbf{A}P + \varepsilon \qquad (3)$$

where P is the original pulse signal; ε is noise; and $\mathbf{A}$ is the transformation matrix. The objective of the blind source separation is the pulse signal estimation using the temporal signals, such that original and estimated pulse signals are similar. That is:

$$\bar{P} = \mathbf{B}T_i = \mathbf{C}P + \varepsilon_1 \;\; such\ that \;\; \bar{P} \approx P \qquad (4)$$

where $\bar{P}$ is the estimated pulse; $\mathbf{B}$ is an appropriate transformation matrix; $\mathbf{C} = \mathbf{BA}$ and $\varepsilon_1 = \mathbf{B}\varepsilon$. It can be observed from Equation (4) that estimated and original pulse can be similar when magnitude of C is approximately 1. Higher order cumulants are used to impose the shape constraints, viz., estimated pulse signal is similar to original pulse signal. It is shown in [12] that cumulant up to fourth order is sufficient for this purpose. Cumulants higher than four are sensitive to outliers and hence avoided. Hence the objective function is defined as:

$$\max_{\mathbf{C}} \; \left|Kurtosis\left[\bar{P}\right]\right| \;\; such\ that \;\; \mathbf{C}^{*}\mathbf{C} = 1 \qquad (5)$$

where *, $|\bullet|$ and $Kurtosis\left[\bullet\right]$ denote conjugate, absolute and Kurtosis operations respectively. Solution of Equation (5) is obtained Kurtosis based maximization proposed in [12] because it provides the global convergence in a time efficient manner.

3.5. Local HR and Quality Estimation

Local HR is estimated using FFT analysis of the estimated pulse signal, $\bar{P}$. That is, FFT is applied to transform $\bar{P}$ into frequency domain using which the local HR, h is evaluated by:

$$h = f \times 60 \qquad (6)$$

where f is the frequency containing the maximum amplitude in the frequency spectrum. The quality of the pulse signal is defined using peak signal to noise ratio, PSNR. Usually, the pulse spectrum contains large amplitudes near HR frequencies and low amplitude at the remaining frequencies that correspond to noise. Thus, signal in PSNR corresponds to the amplitudes of local HR frequency and its few neighbors while noise composed of the amplitudes at the remaining frequencies. That is, the quality, q is given by:

$$q = \frac{\sum_{i=m_h-n_e}^{m_h+n_e} \boldsymbol{S}_e\left(i\right)}{\sum_{i=0.7Hz}^{4Hz} \boldsymbol{S}_e - \sum_{i=m_h-n_e}^{m_h+n_e} \boldsymbol{S}_e\left(i\right)} \qquad (7)$$

where $\boldsymbol{S}_e$ is the frequency spectrum; n_e is the neighborhood size; and m_h provides the position of the local HR frequency. The neighborhood size is chosen such that if h is the local HR then signal constitute of amplitudes corresponding to $[h-5, h+5]$.

Algorithm 1 $Postprocessing(h_p, q_p, t_1, t_2, t_3, V, V_i)$

```
Require: HR and quality that are denoted by $h_p$ and $q_p$ respectively; thresholds $t_1$, $t_2$ and $t_3$; full face video $V$; and the video segment corresponding to $i^{th}$ window, $V_i$.
Ensure: $h_l\,(i)$ contains the local HR in $i^{th}$ window.
Calculate global HR, $h_g$ using full face video, $V$.
if $(|h_p - h_g| > t_1)$ and $(q_p < t_2)$ then
    Reconstruct the face present in $V_i$ using [2]. Use it to re-estimate the HR and quality which is then stored in $h_n$ and $q_n$ respectively. Estimate the local HR in $i^{th}$ window using Equation (8) and $t_3$.
else
    $h_l\,(i) = h_p$
end if
return $(h_l\,(i))$
```

3.6. Post-processing

The proposed system reduces the problems of facial expression, but it can still deteriorate the HR estimation. Moreover, out-of-plane deformations can result in spurious HR estimation. This subsection determines such cases and apply better local HR estimation algorithm that can handle such problems. HR of the full pulse signal referred as global HR is estimated and used to determine the erroneous cases. The parameters, viz., local HR and quality are again estimated for the erroneous cases after face reconstruction. Eventually HR is corrected using the previously estimated parameters and re-estimated parameters. The steps involved are shown in Algorithm 1.

3.6.1 Global HR estimation

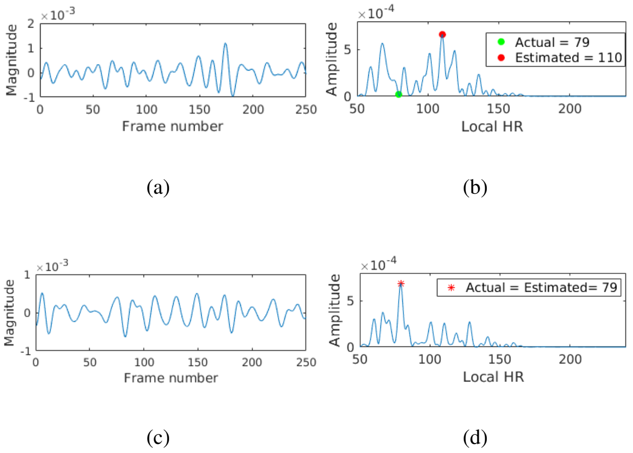
To estimate the global HR, temporal signal (in Equation (1)) is defined for the full video instead of a window. Subsequently, the pulse signal and HR are estimated using Equations (5) and (6) respectively.

3.6.2 Local HR and Quality Re-estimation

The estimated local HR is sufficient for correct HR monitoring except for those few temporal windows that are affected by the noise. The following two observations are used to determine such temporal windows:

1. It is intuitive that HR derived from the full pulse signal (global HR) should not differ significantly from the HR derived from one of its windows (local HR) unless the window contains noise.
2. The pulse signal affected by noise contains multiple peaks thus, its quality (i.e., PSNR) is low.

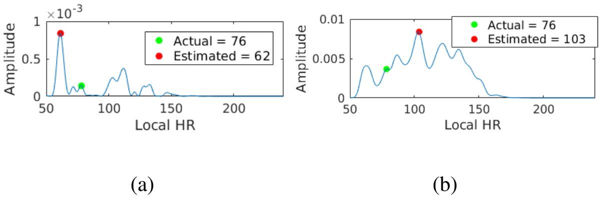
Hence, if the absolute difference between global HR and local HR is greater than a threshold t_1 and quality of local HR is less than the threshold t_2 then it refers to the case of erroneous HR estimation and the parameters are re-estimated. Thresholds t_1 and t_2 are set to 10 and 0.33.



(a) (b)

(c) (d)

Figure 5. Importance of Face Reconstruction: Estimated pulse signal and its spectrum are shown in a) and b) respectively; while pulse signal obtained after face reconstruction and its spectrum are shown in c) and d) respectively



(a) (b)

Figure 6. Example of Incorrect HR Estimation when: a) face reconstruction is avoided; and b) face reconstruction is used.

The prominent reason for the noise in temporal windows is out-of-plane deformations. Most of these deformations are compensated by reconstructing the face using the algorithm proposed in [2]. It provides the registered frontal face images. The temporal signals, pulse, HR and quality are again estimated using the Equations (1), (5), (6) and (7) respectively. An example depicting the improvement in the local HR estimation due to face reconstruction, is shown in Figure 5. The facial reconstruction requires the interpolation of skin pixel intensities which introduce noise. Unfortunately, sometimes, the noise is sufficient enough to deteriorate the temporal signals and thereby result in erroneous HR estimation. Further, the face registration is found to be erroneous in some cases because it depends on the location of eye centers which is error-prone. Apart from the noise, facial reconstruction is a time consuming process. Hence, it is not used unless required in some cases.

3.6.3 HR correction

In this subsection, erroneous local HR estimates are corrected using global HR and the other local HR estimates for better HR monitoring. Assume that the previously estimated local HR and quality are determined by h_p and q_p respectively, while the re-estimated local HR and quality are denoted by h_n and q_n respectively. It is possible that both the h_p and h_n are affected by noise and thus incapable for good HR monitoring. One such example is illustrated in Figure 6. In such cases, it is better to set local HR equal to global HR for better HR monitoring. More clearly, if both the quality estimates, q_p and q_n are less than t_3 then local HR is given by global HR. While if any quality estimate is greater than t_3, local HR is given by the HR estimate corresponding to the larger quality. Threshold t_3 is set at 0.2 in the proposed system. Intuition behind the decision is that the face reconstruction can be erroneous due to which h_n does not provide better HR estimation than h_p in some cases. In essence, the local HR, $h_l\left(i\right)$ for i^{th} window is:

$$h_l\left(i\right) = \begin{cases} h_g & \text{if} \quad \left(\left(q_n < t_3\right) \; and \; \left(q_p < t_3\right)\right) \\ h_n & \text{else if} \quad \left(q_n > q_p\right) \\ h_p & \text{else if} \quad \left(q_n < q_p\right) \end{cases} \tag{8}$$

where h_g is the global HR.

4. Experimental Results

4.1. Dataset Acquisition

The efficacy of the proposed system is evaluated on Intel i5-2400, CPU 3.10GHz in MATLAB 2016a. Total 50 face videos have been acquired from Logitech webcam C270 camera such that each video belongs to a different subject. The camera is placed on the top of a laptop. The subjects have been instructed to sit in front of laptop and camera such that their face is visible in the face video. They are permitted to perform natural facial movements, like slight head tilting, eye blinking and small lip movements. The acquisition has been performed under natural illumination. Each video is 54 seconds long and acquired at 28 frames per second. Ground-truth for the performance evaluation, is simultaneously acquired with the face videos by keeping CMS 50D+ pulse oximeter on the right index fingertip.

4.2. Performance Measurement

Assume that $h_e\left(i,j\right)$ and $h_a\left(i,j\right)$ represent the estimated and actual HR corresponding to the j_{th} window of i_{th} subject. The performance of the proposed HR monitoring system is evaluated using the following metrics:

1. Bland-Altman (BA) plot [4] illustrates the consensus between estimated and actual measurements. Its abscissas and ordinate denote average HR (viz. $\frac{h_e(i,j)+h_a(i,j)}{2}$) and estimation error (viz., $h_e\left(i,j\right) - h_a\left(i,j\right)$) respectively. An accurate HR monitoring system require the estimation error to be close to zero. Alternatively, the mean, μ and standard deviation, σ of estimation error should be close to zero. In addition, the percentage of samples with absolute error less than 5, 10 and 15 BPM denoted by err_5, err_{10} and err_{15} should be as small as possible.
2. The mean average error, MAE is calculated by:

$$MAE = \frac{\sum_{i=1}^{z}\sum_{j=1}^{n_i}\left|h_e\left(i,j\right) - h_a\left(i,j\right)\right|}{\sum_{i=1}^{z}\sum_{j=1}^{n_i}1} \tag{9}$$

   where z is the total number of subjects; $|\bullet|$ is the absolute operator; and n_i is the number of windows for i_{th} subject. Lower value of MAE indicates better HR monitoring.
3. The Pearson correlation coefficient, ρ provides similarity between estimated and actual HR. It is given by:

$$\rho = \frac{\text{cov}(h_e, h_a)}{\sigma_{h_e}\sigma_{h_a}} \tag{10}$$

   where cov and σ denote the covariance and standard deviation operator respectively. High ρ denote high similarity between the predicted and actual HR, i.e, better HR monitoring.
4. Mean and standard deviation of average absolute error, $err_1\left(i\right)$ and average absolute error percentage, $err_2\left(i\right)$ are also used for the evaluation [23]. Consider that n_i represent the number of windows for i^{th} subject, then

$$err_1\left(i\right) = \frac{1}{n_i}\sum_{j=1}^{n_i}\left|h_e\left(i,j\right) - h_a\left(i,j\right)\right| \tag{11}$$

$$err_2\left(i\right) = \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\left|h_e\left(i,j\right) - h_a\left(i,j\right)\right|}{\bar{A}\left(i,j\right)} \tag{12}$$

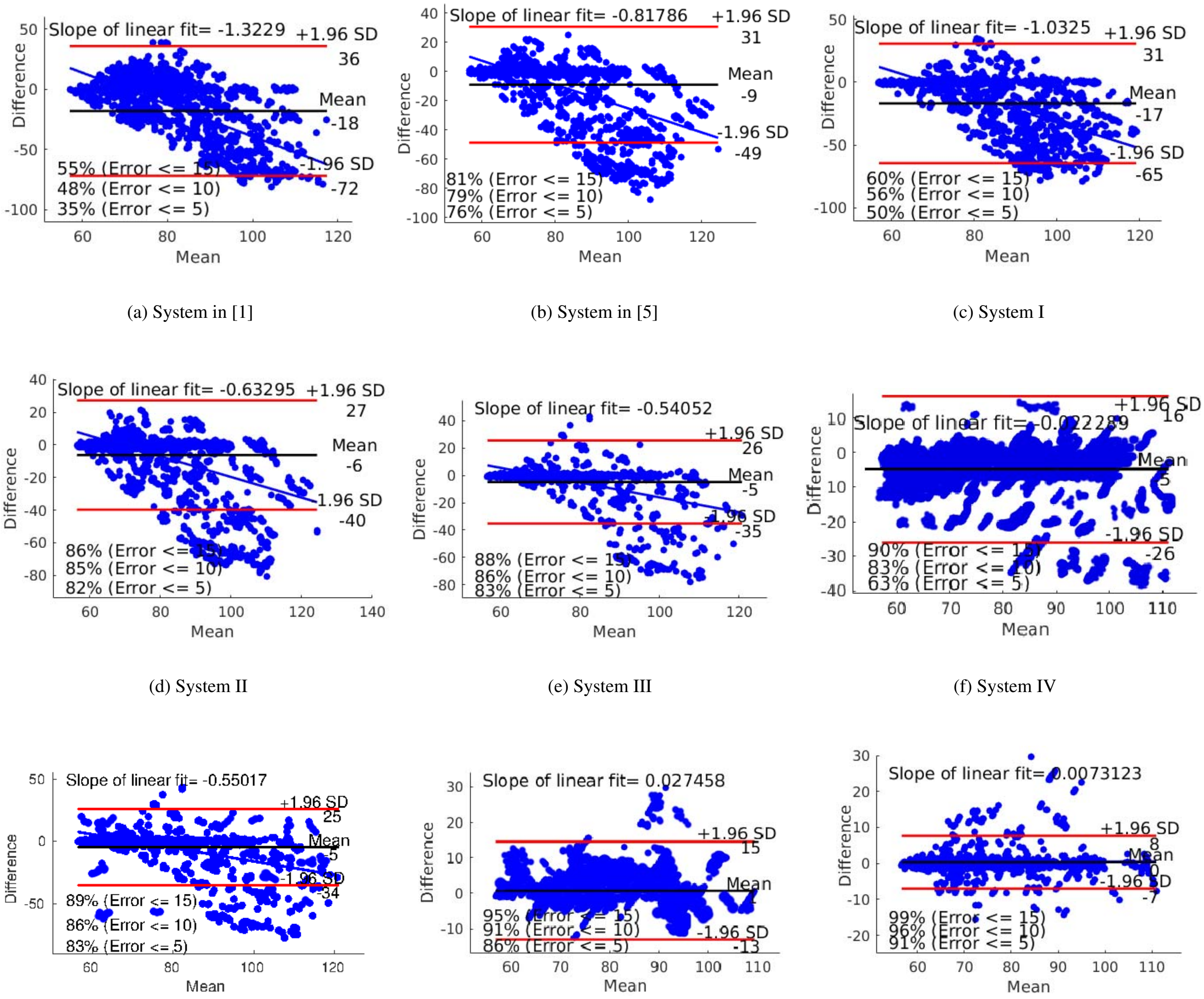
   The mean and standard deviation of err_1 for all the subject are denoted by err_1^m and err_1^s respectively while mean and standard deviation of err_2 are denoted by err_2^m and err_2^s respectively. Lower values of err_1^m, err_1^s, err_2^m and err_2^s indicate better HR monitoring.

4.3. Experimental Analysis

Figure 7 and Table 1 depict the BA plots and performance metrics respectively for Systems [1], [5], I, II, III, IV, V, VI and the proposed system. In [1], Lagrangian temporal signals are used. In [5], pulse signal is estimated from only one Eulerian temporal signal extracted after face registration. It does not utilize any time-consuming blind source separation technique. Systems [1, 5] estimate only one HR

Table 1. Comparative Results of HR Monitoring

System	μ	σ	err_5	err_{10}	err_{15}	MAE	ρ	err_1^m	err_1^s	err_2^m	err_2^s	t_s
[1]	-18.04	27.49	35	48	55	22.30	-0.07	22.30	16.44	32.37	25.71	24.86
[5]	-9.07	20.25	76	79	81	10.03	0.35	10.03	10.71	14.41	16.36	30.37
I	-16.85	24.33	82	85	86	18.67	0.11	18.39	16.30	26.05	25.55	32.49
II	-6.18	17.10	82	85	86	7.19	0.46	7.03	8.66	10.05	13.23	9.63
III	-4.80	15.53	83	86	88	6.06	0.54	5.92	7.58	8.12	10.96	9.64
IV	-4.93	10.78	63	86	90	5.75	0.60	6.13	7.41	8.30	10.79	9.63
V	-4.87	15.01	83	86	89	5.95	0.58	5.27	7.32	8.21	10.84	9.63
VI	0.89	7.05	86	91	95	4.46	0.71	4.23	4.88	4.91	5.71	9.64
Proposed	0.34	3.75	91	96	99	1.63	0.95	1.64	2.41	2.05	2.66	16.78



(a) System in [1]

(b) System in [5]

(c) System I

(d) System II

(e) System III

(f) System IV

(g) System V

(h) System VI

(i) Proposed System

Figure 7. BA Plots of the HR Estimation Systems

value using full pulse signal, as opposed to the HR monitoring. To conform with the proposed system, pulse signals of these systems are divided into multiple windows and local HR is estimated. System I is designed by modifying the proposed system such that HR monitoring is performed using temporal signals extracted after face reconstruction. System II is designed by avoiding the post-processing and adaptive temporal signal selection mechanisms in the proposed system, i.e., it uses all the Eulerian temporal signals obtained without any face registration. System III avoids the post-processing, but it uses adaptive temporal signal selection mechanism. The remaining systems employ different post-processing techniques and the proposed adaptive temporal signal selection. Whenever the absolute difference between global HR and local HR is greater than a threshold t_1, the local HR is replaced by previous local HR, average of local HR and global HR in Systems IV, V and VI respectively. It is evident from Figure 7 and Table 1 that:

1. System [1] exhibit the lowest performance. The reason behind this behavior is the time expensive and inaccurate feature tracking required for Lagrangian temporal signal extraction. System [5] demonstrate poor performance as compared to the remaining systems except for time computations. One reason for this is that it considers a full face region as one ROI due to which it is affected by the local noise due to facial expression. Moreover, it does not handle that out-of-plane deformations that further degrade the monitoring.
2. System I can handle the non-linear deformations introduced in the face. Despite this, Systems II and III perform better than System I due to the noise originated from: i) skin pixel interpolation required in face reconstruction; and ii) improper face registration using error-prone location of eye centers. There exist some cases where System I perform better than Systems II and III, as can be seen in Figure 5. This points out that better HR monitoring can be achieved when System I can be used along with System II or III. It can be seen that the System I is the most computationally expensive due to face reconstruction.
3. System III performs better than System II, which indicates that adaptive temporal signal selection mechanism provide better HR monitoring. Moreover, Systems II and III neglects the non-rigid transformations introduced in the face due to which the proposed system provide better HR monitoring than system II. Another reason for better HR monitoring depicted by the proposed system is this that there are few cases where none of the system perform accurately due to noise and in such cases, the proposed system provide better HR monitoring using the global HR. Some examples illustrating the importance of post-processing in HR monitoring are shown in Figure 8. The figure depicts the HR monitoring performed by System III (i.e, avoiding post-processing) and the proposed system.

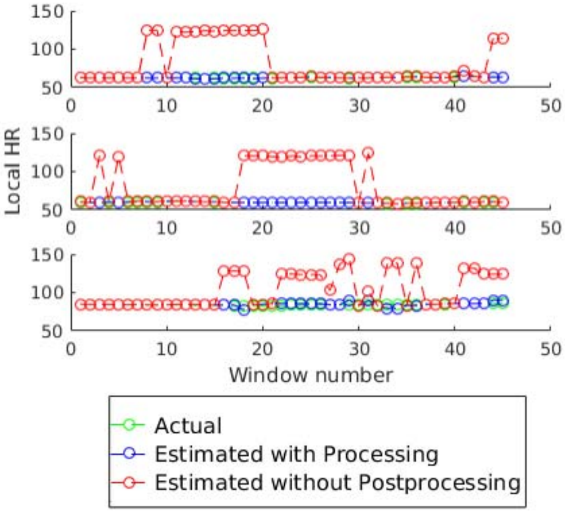


Figure 8. Importance of Post-processing in HR Monitoring

4. System VI performs better than Systems IV and V which indicate that it is better to replace the local HR from global HR rather than average of local HR and previous local HR. But the best performance is achieved when the proposed post-processing is employed, which incorporates local HR, global HR and face reconstruction.
5. The time taken by the proposed system is 16.78 seconds, which can be considered negligible for 54 seconds long face video. Further, the proposed system provides significantly better HR monitoring than the other systems. Thus, it can be inferred that the proposed system can be used for HR monitoring.

5. Conclusion

The proposed system has alleviated the problems due to facial expressions and out-of-plane movements for improving the HR monitoring. The adaptive temporal signal selection mechanism has been utilized to identify the facial areas affected by expression changes. The affected areas has been removed for better monitoring. Further, a novel post-processing mechanism has been employed, which reconstructs the face in case the face video contains noise mainly out-of-plane movements. It has also defined several constraints to provide correct HR monitoring.

Experimental results have revealed HR monitoring can be improved significantly after incorporating the proposed adaptive temporal signal selection and post-processing mechanisms. The proposed system has provided good HR monitoring with the Pearson correlation of 0.95 and average absolute error of 1.63 BPM, which is significantly better than the existing well known HR estimation systems.

References

[1] G. Balakrishnan, F. Durand, and J. Guttag. Detecting pulse from head motions in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3430–3437, 2013.

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[3] G. G. Berntson, J. Thomas Bigger, D. L. Eckberg, P. Grossman, P. G. Kaufmann, M. Malik, H. N. Nagaraja, S. W. Porges, J. P. Saul, P. H. Stone, et al. Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology*, 34(6):623–648, 1997.

[4] J. M. Bland and D. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.

[5] C. Huang, X. Yang, and K.-T. T. Cheng. Accurate and efficient pulse measurement from facial videos on smartphones. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016.

[6] S. Kwon, J. Kim, D. Lee, and K. Park. Roi analysis for remote photoplethysmography on facial video. In *IEEE International Conference of the Engineering in Medicine and Biology Society (EMBC)*, pages 4938–4941. IEEE, 2015.

[7] A. Lam and Y. Kuno. Robust heart rate measurement from video using select random patches. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3640–3648, 2015.

[8] D. Lee, J. Kim, S. Kwon, and K. Park. Heart rate estimation from facial photoplethysmography during dynamic illuminance changes. In *IEEE International Conference of the Engineering in Medicine and Biology Society (EMBC)*, pages 2758–2761. IEEE, 2015.

[9] X. Li, J. Chen, G. Zhao, and M. Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4264–4271, 2014.

[10] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo. Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Transactions on Affective Computing*, 6(4):385–394, 2015.

[11] T. F. of the European Society of Cardiology et al. Heart rate variability standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17:354–381, 1996.

[12] C. B. Papadias. Globally convergent blind source separation based on a multiuser kurtosis maximization criterion. *IEEE Transactions on Signal Processing*, 48(12):3508–3519, 2000.

[13] S. L. Phung, A. Bouzerdoum, and D. Chai. A novel skin color model in ycbcr color space and its application to human face detection. In *International Conference on Image Processing (ICIP)*, volume 1, pages I–289. IEEE, 2002.

[14] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2011.

[15] K. B. Raja, R. Raghavendra, and C. Busch. Video presentation attack detection in visible spectrum iris recognition using magnified phase information. *IEEE Transactions on Information Forensics and Security*, 10(10):2048–2056, 2015.

[16] L. Salahuddin, J. Cho, M. G. Jeong, and D. Kim. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 4656–4659. IEEE, 2007.

[17] M. P. Tarvainen, P. O. Ranta-Aho, P. A. Karjalainen, et al. An advanced detrending method with application to hrv analysis. *IEEE Transactions on Biomedical Engineering*, 49(2):172–175, 2002.

[18] H. E. Tasli, A. Gudi, and M. den Uyl. Remote ppg based vital sign measurement using adaptive facial regions. In *IEEE International Conference on Image Processing (ICIP)*, pages 1410–1414. IEEE, 2014.

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

[20] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 511–518. IEEE, 2001.

[21] N. Wang, X. Gao, D. Tao, and X. Li. Facial feature point detection: A comprehensive survey. *arXiv preprint arXiv:1410.1037*, 2014.

[22] W. Wang, B. Balmaekers, and G. de Haan. Quality metric for camera-based pulse rate monitoring in fitness exercise. In *IEEE International Conference on Image Processing (ICIP)*, pages 2430–2434. IEEE, 2016.

[23] Z. Zhang, Z. Pi, and B. Liu. Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Transactions on biomedical engineering*, 62(2):522–531, 2015.