Serial Fusion of Eulerian and Lagrangian Approaches for Accurate Heart-rate Estimation using Face Videos

Puneet Gupta, Brojeshwar Bhowmick, Arpan Pal

Abstract— **Camera-equipped devices are ubiquitous and proliferating in the day-to-day life. Accurate heart rate (HR) estimation from the face videos acquired from the low cost cameras in a non-contact manner, can be used in many real-world scenarios and hence, require rigorous exploration. This paper has presented an accurate and near real-time HR estimation system using these face videos. It is based on the phenomenon that the color and motion variations in the face video are closely related to the heart beat. The variations also contain the noise due to facial expressions, respiration, eye blinking and environmental factors which are handled by the proposed system. Neither Eulerian nor Lagrangian temporal signals can provide accurate HR in all the cases. The cases where Eulerian temporal signals perform spuriously are determined using a novel poorness measure and then both the Eulerian and Lagrangian temporal signals are employed for better HR estimation. Such a fusion is referred as serial fusion. Experimental results reveal that the error introduced in the proposed algorithm is 1.8±3.6 which is significantly lower than the existing well known systems.**

I. Introduction

Accurate heart rate (HR) is essential to assess pathological and physiological parameters of the circulatory system [1]. These can be subsequently used for analyzing several cardiac diseases, stress monitoring and emotion detection. Existing hemodynamics approaches, electrocardiography (ECG) and photoplethysmography (PPG) involve skin contact and thus these are user uncomfortable and restricted for single user monitoring. Moreover, they are insufficient to analyze: i) sleeping humans; ii) the human while exercising; iii) skin damaged persons; and iv) neonates. This provides the motivation to estimate HR in near real-time using the face videos acquired in a non-contact manner. The underlying principle behind the mechanism is that heart beat introduces blood flow variations and these variations can be observed in the carotid arteries (present in head to neck) by analyzing the face color change and head motion [1].

Well-known face videos HR estimation systems operates in the following manner. Region of interest (ROI) consisting of skin pixels is extracted from the face video. Temporal signals corresponding to the ROI are obtained by Eulerian or Lagrangian approaches. PPG is extracted from them using statistical measures [2]. HR is estimated by analyzing Fast Fourier Transform (FFT) spectrum or R-R intervals of the PPG signal [3].

The temporal signals contain PPG along with the noise due to expression changes, eye blinking, respiration, camera parameters and environmental factors. This paper handles the noise for designing an accurate HR estimation system using face videos. The system performs accurately in near real-time for the face videos. It is observed that neither the Lagrangian temporal signals nor the Eulerian temporal signals are sufficient to provide accurate heart rate estimation in all possible cases. Thus, the proposed system provides a framework for serial fusion of Eulerian and Lagrangian temporal signals. The framework evaluates when the Eulerian temporal signals fail to provide accurate HR using a novel measure known as poorness measure. In such cases, it also employs Lagrangian temporal signals to improve efficacy.

This paper is organized as follows. The background required for better understanding of the proposed system is summarized in the next section. The proposed system is presented in Section IV followed by experimental results in Section V. Eventually, conclusions are mentioned in the Section VI.

II. Preliminaries

A. Lagrangian and Eulerian Signals

Temporal signals can be extracted using Lagrangian perspective where discriminating features are explicitly tracked over time [1]. Such tracking mechanisms are highly time-consuming and they can be spurious when few or less discriminatory features are available for tracking due to poor lighting conditions. Alternatively, Eulerian perspective can be used where temporal signals can be obtained by fixing ROI at some locations and analyzing the variations in it [2]. It avoids time consuming tracking of ROI and works accurately for small variations. Eulerian temporal signals can be failed to provide accurate HR due to improper illumination, inappropriate camera focus or human factors (for example, skin color).

B. Parallel and Serial Fusion

Fusion of multiple classifiers can be obtained by either parallel fusion or serial fusion. In parallel fusion, all the classifiers are simultaneously applied and their evaluated features [4] or scores [5] are consolidated. In contrast, the serial fusion uses one classifier at a time for the evaluation. If the classifier is insufficient to provide the accurate results, other single classifier is used and this procedure of changing the classifier continues till accurate result is obtained or no additional classifier is available [6].

The authors are with embedded system and robotics, TCS Research Kolkata, India and their E-mail are: {gupta.puneet5, b.bhowmick, arpan.pal}@tcs.com

978-1-5090-2809-2/17/$31.00 ©2017 IEEE

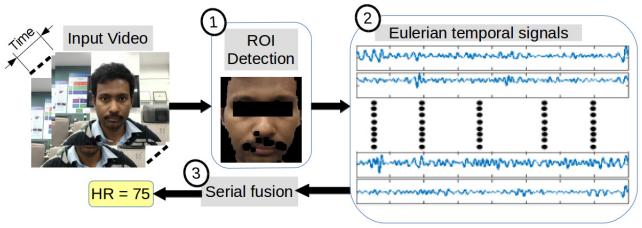


Fig. 1. Overview of the Proposed System

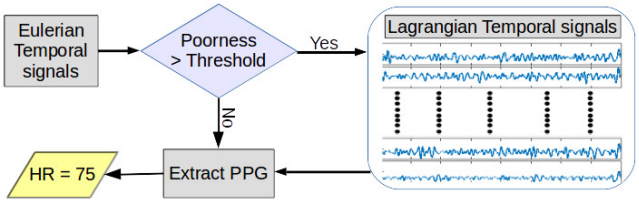


Fig. 2. Flow-graph of the Proposed Serial Fusion

III. Experimental Setup

MATLAB 2015a is used to implement the proposed system on Intel i5-2400 CPU 3.10GHz. For evaluation, 45 face videos were acquired from 20 subjects using a Logitech webcam C270 camera. The subjects were asked to sit in front of laptop mounted with camera and they are allowed to perform natural facial movements, for example eye blinking, small lip movements and slight tilting of the head. Natural illumination was maintained for the acquisition of 40 second long videos at 28 frames per second. The actual PPG was also acquired synchronously by placing CMS 50D+ pulse oximeter on the right index fingertip for ground truth.

IV. Proposed System

In this section, the HR estimation system is proposed. It consists of three stages, viz., ROI detection, extraction of Eulerian temporal signal and serial fusion for estimating HR. The overview of the system is shown in Figure 1.

A. ROI Detection

Viola-Jones face detector [7] is applied to the first frame of the face video to determine the face regions. The non-face pixels and inevitable eye blinking in the face video can deteriorate HR estimation hence they are first determined using skin detection [8] and Viola-Jones eyes detector respectively, and subsequently removed from the facial area. Another reason for noise is the face boundary pixels whose slight color or motion variations can introduce enormous variations in the temporal signals. Thus, the boundary pixels are removed using morphological erosion operation [9]. Considering full face as one ROI can result in erroneous HR estimation because: i) different face regions exhibit different color variations depending on the placement of unevenly distributed face arteries; and ii) the facial emotion introduces noise in some face regions that can result in spurious HR estimation. Thus the remaining area is divided into several square blocks and each block is considered as ROI. The block-size is chosen such that the remaining area should contain 10 blocks in the horizontal direction. For better understanding, HR estimation using single face is discussed. Finally, all the detected faces follow the same methodology for HR estimation.

B. Eulerian Temporal Signals

The green channel exhibit the strongest plethysmographic signals amongst all the color channels [10]. Thus, the mean green value of the pixels in a block is used to define the temporal signal. For better understanding, consider that S^i denote the temporal signal for i^{th} block and the input video consists of f frames. Then,

$$S^i = \left[s_1^i, s_2^i, \cdots s_f^i\right] \tag{1}$$

where s_k^i is the mean green value of pixels of i^{th} block in k^{th} frame. Thus, if B_k^i is the i^{th} block in k^{th} frame, then

$$s_k^i = \frac{\sum_{(x,y)\in B_k^i} F_k^g\left(x,y\right)}{Size\left(B_k^i\right)} \tag{2}$$

where F_k^g contains the green channel intensities of the k^{th} frame and $Size\left(\bullet\right)$ provides the size. Typically the heart beat from the range of 42 to 240 beats-per-minute (bpm) thus band-pass filter for the frequency ranging from 0.7 to 4 Hz is applied to filter the noise. The median filter is further applied to remove the noise [11].

C. Serial Fusion

The proposed serial fusion is described in this section. It is based on the observation that in some cases Eulerian temporal signals fail to provide accurate HR due to improper illumination, inappropriate camera focus or human factors (for example, skin color). A poorness measure is defined to identify such cases, i.e., measure the efficacy of Eulerian temporal signals. If it is large, temporal signals extracted using Lagrangian approach can be used to improve the HR estimation. The flow-graph of the serial fusion is shown in Figure 2.

Ideally, the HR varies continuously with time, but the variations are small. Such HR variability in the consecutive time intervals is used to define the poorness measure. To evaluate it, the video is first divided into several intervals and HR is calculated using each interval. The length of the intervals should not be small otherwise peak due to noise can be observed while it should not be large otherwise few HR variability are available. Time intervals of 6 seconds are used in this paper to define each interval. The amplitudes of the temporal signal in an interval vary according to facial structure, thus these are normalized using z-score normalization which keeps the signal shape intact. Mathematically, i^{th} signal of j^{th} interval, $T_{(i,j)}$ is normalized using:

$$T_{(i,j)} = \frac{T_{(i,j)} - mean\left(T_{(i,j)}\right)}{std\left(T_{(i,j)}\right)} \tag{3}$$

where $mean\left(\bullet\right)$ and $std\left(\bullet\right)$ denotes the mean and standard deviation respectively. Moreover, each temporal signal in an

interval contains PPG signal corrupted by noise thus

$$T^t_{(i,j)} = MP^t + \varepsilon^t \tag{4}$$

where P^t and ε^t denote the PPG signal and noise respectively at time instant t while M represents the transformation matrix. The aim of recovering 1-D PPG signal in an interval using all the temporal signals, is achieved by:

$$O^t_e = QT^t_{(i,j)} \tag{5}$$

where O^t_e is the estimated PPG at time t and Q represents an appropriate transformation matrix. Hence, it can be seen from Equations (4) and (5) that

$$O^t_e = ZP^t + \hat{\varepsilon}^t \tag{6}$$

where $Z = QM$ and $\hat{\varepsilon}^t = Q\varepsilon^t$. Moreover, an accurately estimated PPG should be mostly equal to the actual PPG, i.e., $P_e \approx P_a$. Hence magnitude of Z should be 1 to avoid scaling and making the shape of estimated and actual PPG similar. Normally, local PPG spectrum is peaked at one frequency that corresponds to HR frequency and it possesses small amplitude at other frequencies. Essentially, it contains high Kurtosis statistics that measures peakedness and tail of a signal [12]. Hence, the following objective function which maximizes the Kurtosis statistics, is used for PPG estimation:

$$\max_{Z} \quad |Kurtosis\,[O_e]| \quad such\ that \quad Z^*Z = 1 \tag{7}$$

where $|\bullet|$ and $Kurtosis\,[\bullet]$ represent the absolute value and Kurtosis operators respectively while * denote the conjugate. The global solution of Equation (7) is attained using [12]. The HR of an interval is estimated from O_e using FFT analysis. The number of samples required in FFT depends on the total number of frames per seconds. O_e is transformed to frequency domain using FFT and the HR for the i^{th} interval, h_i is given by:

$$h_i = f_i \times 60 \tag{8}$$

where f_i is the frequency corresponding to the maximum amplitude in the i^{th} interval.

There can be a small change in HR estimates in the consecutive time intervals which should not impact the poorness measure and the poorness measure keeps on increasing as the change increases. To leverage this, the poorness measure for the full Eulerian temporal signal, P_m is defined as:

$$P_m = \sum_{j=2}^{p} G\left(|h_j - h_{j-1}|\right) \tag{9}$$

where p and $|\bullet|$ denote the total number of intervals and the absolute value respectively while function G is given by:

$$G\left(x\right) = \begin{cases} 0, & \text{if } x < \alpha \\ \frac{\left(1-e^{(\alpha-x)}\right)}{\left(1+e^{(\alpha-x)}\right)}, & \text{otherwise} \end{cases} \tag{10}$$

where the parameter α represents the permissible variations in HR and it is set to 5 bpm. It should be noted that domain of function G is non-negative due to the use of absolute value operation in Equation (9). If P_m is lower than the threshold, th then estimated Eulerian temporal signals are used for HR estimation. The threshold th is given by $th = \frac{p}{3}$ where p is the number of intervals. While $P_m > th$ indicates large HR variations which in turn implies that Eulerian temporal signals are incapable for accurate HR estimation. In such a case, another HR and poorness measure are estimated using the temporal signals derived from the Lagrangian approach proposed in [1]. The Lagrangian approach in [1] explicitly tracks the distinguishable facial features in the face videos and the variations in the vertical direction are considered to generate the temporal signals. The another HR is estimated by solving Equations (7) and (8) while poorness measure is defined using Equation (9). The HR corresponding to the minimum poorness measure is denoted as the actual HR. HR corresponding to each face is detected when multiple faces are present in the face video.

TABLE I

COMPARATIVE RESULTS OF HR ESTIMATION

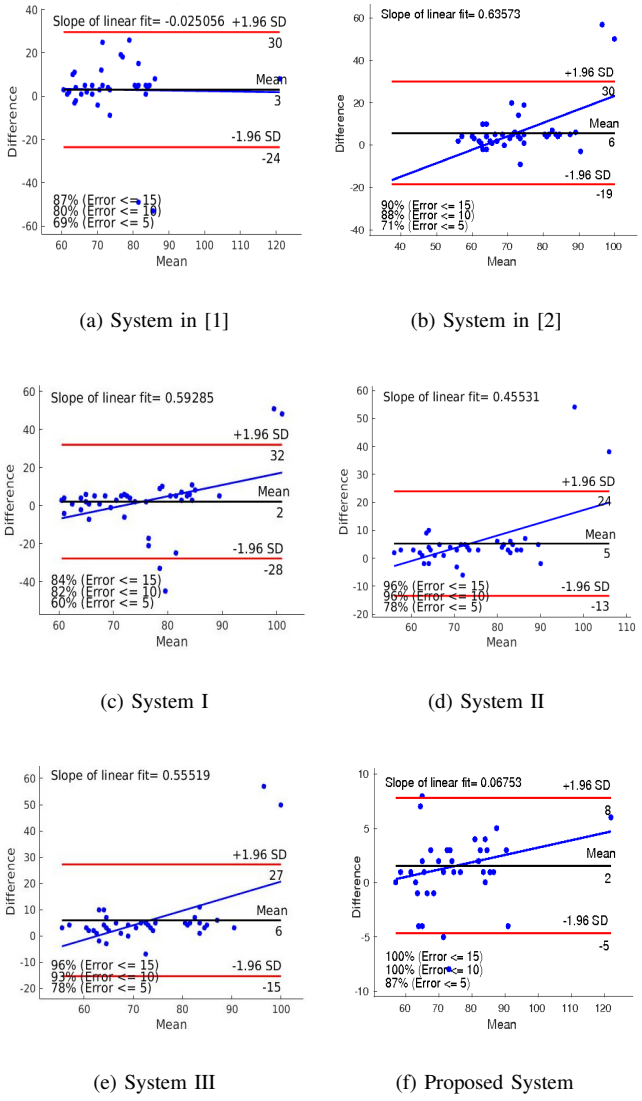
System	Mean*	Variance*	Time (sec)#	Correct$_5^+$	Correct$_{10}^+$
[1]	2.98	13.78	14.92	69%	80%
[2]	5.77	12.29	6.96	71%	88%
I	2.45	15.30	18.89	60%	82%
II	5.43	9.63	5.97	78%	96%
III	6.48	10.72	20.36	78%	93%
Proposed	1.80	3.06	6.43	87%	100%

$^+$: Correct$_5$ and Correct$_{10}$ denote percentage of samples with absolute error less than 5 and 10 bpm respectively.
*: Mean and variance of the error shown in BA plots, viz. Figure 3.
#: Time required in seconds for the face videos of 40 seconds.

V. EXPERIMENTAL RESULTS

Bland-Altman (BA) plot [13] is used to understand the efficacy of the proposed system. Figure 3 and Table I depict the BA plots and performance measures respectively of Systems [1], [2], I, II, III and the proposed system. For more rigorous experimentation, the mean and variance are explicitly shown along with time comparisons and percentages of samples with absolute error less than 5 and 10 bpm. The lower value of these parameters indicates better performance. The Lagrangian temporal signals followed by Principal Component Analysis is used in [1] while Eulerian temporal signals followed by independent Component Analysis are used in [2] for PPG estimation. Moreover, the Eulerian and Lagrangian temporal signals are fused in System I by parallel fusion and only Eulerian temporal signals are utilized in System II for HR estimation. HR is evaluated in System III by utilizing both Eulerian and Lagrangian temporal signals followed by the proposed kurtosis maximization. It is evident from the figure and the table that:

1) The lowest performance is depicted by System [1]. It is mainly due to the time expensive and inaccurate facial feature tracking. Likewise, System [2] depicts poor performance because it considers a full face region as ROI. Moreover, Systems [1] and [2] employ different PPG extraction that results in spurious HR estimation.
2) Fusion does not always guarantee the performance improvement and instead performance can be degraded



(a) System in [1]

(b) System in [2]

(c) System I

(d) System II

(e) System III

(f) Proposed System

Fig. 3. BA Plots of Some HR Estimation Systems. Its abscissas denote the average HR while the ordinate denotes the error in the HR.

if inappropriate features are fused [6]. Similar behavior can be observed by comparing System I and System II. The peculiar behavior is because Lagrangian temporal signals are erroneous in some cases where better HR can be obtained using Eulerian temporal signals alone. Another fusion strategy is employed in System III which performs inaccurately when Lagrangian temporal signals saturate after some time due to tracking error. The proposed system avoids the evaluation of Lagrangian temporal signals when not required. Hence, it performs better than System III.

3) System I and III are computationally expensive because they use Eulerian and Lagrangian temporal signals in all the cases. Lower time computation is observed in the proposed system because it sometimes avoids time consuming Lagrangian temporal signal due to the serial fusion. The least computational time is shown by System II because it completely avoids the Lagrangian temporal signals. The time difference between the proposed system and System II is 0.54 seconds that can be considered insignificant for the face videos of duration of 40 seconds. Keeping in view the performance improvement, the proposed system advocates the serial fusion.

VI. Conclusion

An accurate and near real-time HR estimation system has been proposed in this paper. It employs face videos acquired in a non-contact manner. The artifacts introduced due to facial expressions, respiration, eye blinking and environmental factors have been mitigated. The Eulerian temporal signals have been failed to estimate accurate HR in some cases. Such cases have been identified by defining a novel measure, poorness measure and both the Eulerian and Lagrangian temporal signals are employed in these cases for better HR estimation. Experimental results have demonstrated the superiority of the proposed system over existing well known systems. Moreover, it has shown that better performance is expected when the proposed serial fusion is used instead of parallel fusion.

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)*, 2013, pp. 3430–3437.

[2] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

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

[4] P. Gupta and P. Gupta, "An accurate finger vein based verification system," *Digital Signal Processing*, vol. 38, pp. 43–52, 2015.

[5] P. Gupta, S. Srivastava, and P. Gupta, "An accurate infrared hand geometry and vein pattern based authentication system," *Knowledge-Based Systems*, vol. 103, pp. 143–155, 2016.

[6] A. Uhl and P. Wild, "Parallel versus serial classifier combination for multibiometric hand-based identification," in *International Conference on Biometrics (ICB)*. Springer, 2009, pp. 950–959.

[7] 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)*. IEEE, 2001, pp. 511–518.

[8] 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)*, vol. 1. IEEE, 2002, pp. I–289.

[9] P. Gupta and P. Gupta, "An efficient slap fingerprint segmentation and hand classification algorithm," *Neurocomputing*, vol. 142, pp. 464–477, 2014.

[10] 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)*. IEEE, 2014, pp. 1410–1414.

[11] P. Gupta and P. Gupta, "Multi-modal fusion of palm-dorsa vein pattern for accurate personal authentication," *Knowledge-Based Systems*, vol. 81, pp. 117–130, 2015.

[12] C. B. Papadias, "Globally convergent blind source separation based on a multiuser kurtosis maximization criterion," *IEEE Transactions on Signal Processing*, vol. 48, no. 12, pp. 3508–3519, 2000.

[13] J. M. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The lancet*, vol. 327, no. 8476, pp. 307–310, 1986.