ACCURATE HEART-RATE ESTIMATION FROM FACE VIDEOS USING QUALITY-BASED FUSION

Puneet Gupta, Brojeshwar Bhowmick, Arpan Pal

Embedded system and Robotics, TCS Research Kolkata, India,

E-mail:{gupta.puneet5, b.bhowmick, arpan.pal}@tcs.com

ABSTRACT

Estimating heart rate (HR) accurately using face videos acquired from a low cost camera in contactless manner is of paramount importance for many real-world applications. Such existing systems perform spuriously due to change in camera parameters, respiration, facial expressions and environmental factors. This paper mitigates the issues for accurate HR estimation. The face video consisting of frontal, profile or multiple faces is divided into multiple overlapping fragments to determine HR estimates. The HR estimates are fused using quality-based fusion which aims to minimize illumination and face deformations. Experimental results demonstrate that the proposed system exhibit better performance than the state of the art systems and establishes the efficacy of the quality-based fusion in HR estimation.

Index Terms— Heart rate, Quality based fusion, Face video, Remote PPG, Color Magnification

1. INTRODUCTION

The heart rate (HR) is important to infer user physiological parameters associated with diseases, like myocardial infarction, diabetic neuropathy, and myocardial dysfunction [1]. Traditional electrocardiography (ECG) and photoplethysmography (PPG) based HR estimation require human skin contact which is not only user uncomfortable, but also infeasible when multiple user monitoring is required or extreme sensitive conditions is a prime concern as in monitoring: i) newborns; ii) sleeping human; and iii) skin damaged patients. Hence, contact-less HR estimation is of paramount importance. It can be accomplished by estimating HR from face videos acquired using web-cams, smart-phone camera or surveillance camera.

The acquisition is based on the principle that there is a change in the blood flow when the heart beats. It introduces variations in the blood flow from heart to head through the carotid arteries that in turn results in the skin color change [2] and head motion [3]. A normal camera can be used to acquire these small color intensity variations and head motion. Such a contact-less and near real-time HR estimation can be used in the following applications: i) remote health care for estimating stress or cardiac diseases [1]; ii) affective computing by analyzing human emotion; iii) biometrics for detecting liveness; and iv) fitness monitoring [4].

Usually, existing face videos based HR estimation systems work in the following manner. Facial skin pixels are determined from the face video and referred as region of interest (ROI). Temporal signals depicting the motion or color variations in the frames across time, are estimated from the ROI using Eulerian or Lagrangian approaches. In a Lagrangian approach, temporal signals are determined by explicitly tracking the ROI or discriminating features over time. Such tracking is computationally expensive hence usually temporal signals are estimated using Eulerian approach, i.e., temporal signals are obtained by fixing ROI and analyzing its variations [5]. The Eulerian approach works accurately for small variations. Noise in the temporal signals is filtered for accurate HR estimation [6]. PPG is extracted from the filtered temporal signals and subsequently it is used to estimate the HR using R-R intervals or Fast Fourier Transform (FFT) spectrum [1]. The confidence in the HR estimation known as quality, provides a useful indicator of the efficacy of estimated HR. In [4], several quality measures have been proposed to evaluate the predicted HR in fitness monitoring environment. Existing systems do not use any quality parameter to improve the HR estimation, but rather to understand the effectiveness of the estimated HR.

Along with the color and motion variations, the cameras acquire noise introduced by respiration, expression changes, camera parameters changes (for eg. focus), eye blinking and environmental factors. Further, the variations in the different face parts vary according to the facial structure, i.e., location of arteries and bones in the face. HR estimation is a challenging problem due to these factors, especially when required in near real-time. This paper proposed a system for accurate HR estimation using face videos which handles these issues. Its major contribution is dividing the video into multiple temporal fragments and eventually consolidating their HR estimates using quality-based fusion. Unlike existing works that require quality to understand the effectiveness of the estimated HR, the proposed system employs the quality to improve the HR estimate. In addition, temporal fragments containing noise due to changes in facial expression, camera parameters and illumination are also pruned using range filters. Moreover,

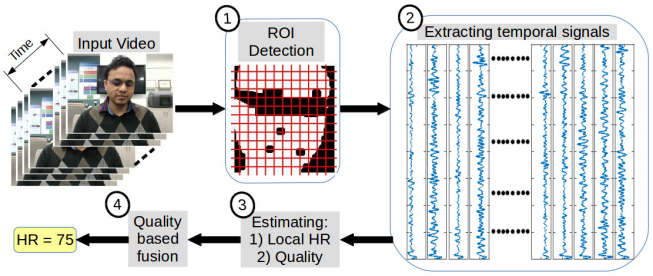


Fig. 1. Flow-graph of the Proposed System

the system is applicable for frontal, profile and multiple faces and performs in near real-time.

The outline of this paper is as follows. The proposed HR estimation system is described in the next section. Subsequently, experimental results are presented in Section 3. Conclusions are given in the last section.

2. PROPOSED SYSTEM

The proposed HR estimation system consists of four stages. In the first stage, blocks consisting of facial skin areas are extracted. Subsequently, temporal signals are extracted and filtered to minimize the noise. In the next stage, smaller parts are extracted from the video. The parts, corrupted by noise are determined using range filter and pruned for further processing. HR of each part is estimated along with its quality. Eventually, quality based fusion is applied to estimate a global HR corresponding to the video. The flow-graph of the proposed system is depicted in Figure 1.

2.1. ROI Detection

The rectangular face region is detected by applying Viola-Jones face detector [7] on the first video frame. Both frontal and profile Voila-jones face detectors are employed for face detection. Model based face detection can be also used for better face detection, for example cascaded deformable shapes [8], Discriminative Response Map Fitting [9], Active shape model [10], and Intraface [2]. But this paper refrains its usage because of its time complexity. The area near the eye is highly sensitive to noise due to eye blinking hence eye regions are also detected using Viola-Jones eyes detector and the area is removed from the face region. Non-face pixels in the detected face regions are detected by applying skin detection proposed in [11]. Moreover, Viola-Jones face detector may provide false face, which is pruned by analyzing the number of skin pixels in each detected face. It has been observed that even a slight motion near face boundaries results in large color variations. This degrades the performance of Eulerian approaches that are applicable for small variations. Hence, morphological operation like erosion is applied to detect boundary pixels and pruned [12]. Assume that the resultant image is referred as skin image. This paper extracts several blocks as ROI from the skin image instead of using full face as one ROI. The reason is that HR estimation can be spurious when full face is used as a ROI because different part of the face undergoes different color variations and non-rigid facial emotion deformations even in the small face area can significantly degrade the performance [8].

For better clarity, henceforth, the system for extracting the HR from a single face is explained. Eventually, the same process is repeated for all the detected faces. Multiple ROI are detected by dividing the detected face region into several square blocks whose block-size is chosen such that there are at-least 10 blocks in the horizontal direction of detected face. The blocks containing all skin pixels are marked as valid blocks while other blocks are removed for further processing. Some of these blocks may suffer from non-rigid deformations due to emotions like smiling.

2.2. Extracting Temporal Signals

It is observed that amongst all the color channels, green channel contains the strongest plethysmographic signals [13]. The reason is that green light infiltrates human skin better than blue light and it is absorbed by hemoglobin much better than red light [9]. Hence, the raw temporal signal for a block is given by the mean green value of pixels inside it for each frame. Mathematically, the temporal signal for a block i, T^i is given by:

$$T^i = \left[t_1^i, t_2^i, \cdots \cdot t_f^i\right] \tag{1}$$

where f denotes the total number of input video frames and t_k^i represent the mean green value of pixels of i^{th} block inside k^{th} frame, i.e.,

$$t_k^i = \frac{\sum_{(x,y)\in b_k^i} I_g\left(x,y\right)}{\sum_{(x,y)\in b_k^i} 1} \tag{2}$$

such that b_k^i denotes the i^{th} block inside k^{th} frame and I_g stores the green channel intensities of the k^{th} frame. Noise in temporal signals is minimized by applying band-pass filter. Since the heart can beat from 42 to 240 beat-per-minute (bpm), frequency range for the filter is 0.7 to 4 Hz.

2.3. Local HR and Quality Estimation

Video length is a crucial parameter for accurate HR estimation. Generally, HR varies with time and there can be noise in some frames due to facial expressions and small head movements. Hence, HR estimated by full video may not be accurate [10]. This provides the motivation to divide the face video into several temporal fragments and extract HR from each fragment which is referred as local HR. The local HR estimates are eventually consolidated for an accurate HR estimate, which is known as global HR. The following steps are performed to extract the local HR estimates:

1. *Extracting fragments*: Each temporal fragment should be sufficiently large otherwise it is possible to observe a peak due to noise rather than the variations introduced by HR. It is observed that HR can be accurately estimated from a fragment spanning at least 180 frames of the video. Moreover, the effect of small temporal changes and head movements in some local HR estimates can be compensated only when a large number of fragments (or local HR) are considered thus temporal signals are divided into overlapping fragments. The overlapping between the consecutive fragments is 60 frames. Some fragments of the temporal signal are corrupted by noise due to small facial movements, focus change or illumination change. The erroneous fragments contain variations due to blood flow and noise, i.e., large intensity discontinuities or large local range [14]. Hence, total 20% of the total fragments are pruned using the range filters [14].
2. *Extracting local PPG*: Each remaining fragment contains multiple signals composed of local PPG but corrupted by noise. This provides the motivation to employ the methodology of blind source separation (BSS). The variations in different signals are different depending upon the facial structure, thus the signals are normalized using:

$$F_j^i = \frac{F_j^i - \mu\left(F_j^i\right)}{\sigma\left(F_j^i\right)} \tag{3}$$

where F_j^i is the i^{th} normalized signal in j^{th} fragment while μ and σ represent the mean and standard deviation operations respectively. Further, F_j^i can be written as

$$F_j^i(n) = AP_a(n) + \eta(n) \tag{4}$$

where P_a is the actual local PPG; A is the channel matrix; n is the time instant; and η is the noise. Remember that local PPG is unknown and needs to be estimated from the fragment signals. Using Equation (4), the estimated local PPG, P_e is given by:

$$P_e(n) = BF_j^i(n) = TP_a(n) + \hat{\eta}(n) \tag{5}$$

where B is the transformation matrix while $T = BA$ and $\hat{\eta} = B\eta$. Since estimated and actual local PPG should be similar (i.e., $P_e \approx P_a$), appropriate constraints on the shape should be imposed to estimate local PPG. Typically, the PPG signal should contain one high frequency component corresponding to the HR pulse and small magnitude of other frequencies. Hence, estimated local PPG should possess high Kurtosis statistic which provides the shape information of the signal both in terms of peakedness and tailness [15]. Mathematically, these constraints are imposed by solving the following objective function:

$$\max_{T} \ |K[P_e]| \ \ subject\ to\ \ T^*T = 1 \tag{6}$$

where $|\bullet|$ and * are used to indicate the absolute value and the conjugate operations respectively while $K[P_e]$ is the Kurtosis for the samples in estimated PPG P_e. For accurate local PPG extraction, the objective function in Equation (6) is solved using [16] that provide globally convergent solution along with the computational tractability and convenience.
3. *Local HR and quality*: Detrending filter [17] is applied to the local PPG signals for reducing non-stationary trends. FFT analysis is used to estimate the local HR from a local PPG signal. For brevity, a local PPG signal, S is transformed into the frequency domain by applying FFT and the frequency corresponding to the maximum magnitude belongs to the heart beat frequency. Hence, the local HR for the j^{th} fragment, h_j is given by $h_j = f_j \times 60$ where f_j is the heart beat frequency in the j^{th} fragment. The frequency spectrum will contain non-zero amplitude for other frequencies due to inevitable noise. Higher amplitude of other frequency components indicates higher noise contents and subsequently less confidence in the estimated local HR. This observation is used to define a quality parameter which indicates the confidence in predicting local HR. Quality q_j for local PPG signal, S_j is given by:

$$q_j = \left(\frac{a_j - b_j}{b_j}\right) \tag{7}$$

where a_j and b_j denote the maximum and the second maximum amplitude of the frequency spectrum of S_j.

2.4. Estimating Global HR using Fusion

Local HR estimates are consolidated using quality based fusion for global HR estimating, i.e., each local HR estimate is weighted by its quality which indicates the confidence in accurate extraction of local HR from its local PPG. Mathematically, local HR and their quality are fused using quality based fusion to estimate the global HR in the following way:

$$H_G = \frac{\sum_{j=1}^{p}(q_j \times h_j)}{\sum_{j=1}^{p} q_j} \tag{8}$$

where H_G and p denote the global HR and total number of local HR estimates respectively. When the input face video contains multiple faces, global HR is detected for each face.

3. EXPERIMENTAL RESULTS

The proposed system has been implemented in MATLAB 2015a on Intel Core i5-2400 CPU at 3.10GHz. Total 45 face videos of 20 subjects are acquired from a Logitech webcam C270 camera in natural lighting conditions at 28 frames per second for a duration of 40 seconds. CMS 50D+ pulse oximeter is also attached to each subject finger to acquire the ground truth.

The efficacy of the proposed system is compared with systems [3, 6], I, II, III and IV in Table 1. The descriptions of

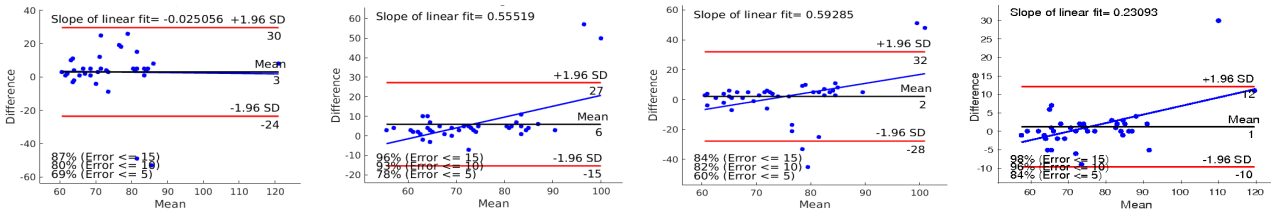
Table 1. Comparative Results of HR estimation

System	Mean	Variance	Time (in sec)	Correct_5^+	Correct_{10}^+	Extracting PPG	Temporal$^\#$ Signal	Signal Usage	Pruning$^\#$	Fusion Type
[3]	2.98	13.78	14.92	69%	80%	PCA*	Lagrangian	Full	NA	NA
[6]	7.77	12.29	6.96	71%	82%	ICA*	Eulerian	Full	NA	NA
I	6.14	10.89	7.25	78%	93%	Proposed	Proposed	Full	NA	NA
II	2.27	15.37	7.65	60%	82%	Proposed	Proposed	Fragments	No	Proposed
III	1.29	9.72	7.61	78%	93%	Proposed	Proposed	Fragments	Yes	Average
IV	5.84	10.63	7.61	76%	91%	Proposed	Proposed	Fragments	Yes	MV*
Proposed	1.27	5.56	7.61	84%	96%	Proposed	Proposed	Fragments	Yes	Proposed

*Abbreviations: Principal component analysis (PCA); Independent component analysis (ICA); and Majority Voting (MV)

$^+$: Correct_5 and Correct_{10} denote percentage of samples with absolute error less than 5 and 10 bpm respectively.

$^\#$: Extracting temporal signal using Lagrangian or Eulerian approach or pruning of erroneous fragments using range filters.



(a) System in [3] (b) System I (c) System II (d) Proposed System

Fig. 2. BA Plots of Some HR Estimation Systems

these systems are provided in the table. Other well known fusion strategies (like product, min, max-based fusion) are not shown because they are highly affected by noise [18]. For better visualization, Bland-Altman (BA) plot [19] of some systems are also depicted in Figure 2. The table and the figure indicates that:

1. System [3] depicts the lowest performance because of the time consuming and erroneous feature tracking. Moreover, Systems I and [6] requiring full signal exhibit lower performance than the remaining systems employing the fragments for HR estimation. It is because multiple fragments minimizes the temporal variations in some frames due facial movements. However Systems I and [6] require less time computation than the remaining systems because they extract PPG once while the remaining systems require PPG extraction for each fragment, i.e., multiple times. Fortunately, the time difference is low for the face videos of 40 seconds as compared to the improvement in accuracy, hence fragment based analysis is preferred over full signal analysis.
2. System II exhibits poor performance than the proposed system because it prunes the erroneous fragments detected using range filter. Likewise, the proposed system outperforms Systems III and IV, which indicates the proposed quality-based fusion exhibit better performance than the other fusion strategies.
3. The performance of the proposed system can be improved if Model based face detector is applied instead of Viola-Jones face detector, but these are avoided in this paper because of its time complexity and incapability to perform when faces are not frontal. Moreover, in [4], several quality parameters, especially based on lighting parameters are avoided in this paper because they are mostly unchanged in different fragments of a face video.
4. The proposed system require only 7.61 seconds for a video of 40 seconds, which indicates that it can be used in near real-time scenarios.

4. CONCLUSION

This paper has proposed an accurate HR estimation system using face videos acquired in contact-less manner. It has performed in near real-time and is applicable for the face videos consisting of frontal, profile and multiple faces. It has minimized the noise due to changes in facial expressions, respiration, camera parameters and environmental factors. For this, several local HR have been estimated at different fragments and they are eventually fused using quality-based fusion. Experimental results have depicted that the proposed system outperforms the well known systems.

5. REFERENCES

[1] Task Force 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.

[2] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2396–2404.

[3] Guha Balakrishnan, Fredo Durand, and John Guttag, "Detecting pulse from head motions in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3430–3437.

[4] Wenjin Wang, Benoît Balmaekers, and Gerard de Haan, "Quality metric for camera-based pulse rate monitoring in fitness exercise," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2430–2434.

[5] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William T. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, 2012.

[6] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[7] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2001, pp. 511–518.

[8] Antony Lam and Yoshinori Kuno, "Robust heart rate measurement from video using select random patches," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3640–3648.

[9] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 4264–4271.

[10] Chong Huang, Xin Yang, and Kwang-Ting Tim Cheng, "Accurate and efficient pulse measurement from facial videos on smartphones," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–8.

[11] Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai, "A novel skin color model in ycbcr color space and its application to human face detection," in *Proceedings of the International Conference on Image Processing (ICIP)*. IEEE, 2002, vol. 1, pp. I–289.

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

[13] H Emrah Tasli, Amogh Gudi, and Marten den Uyl, "Remote ppg based vital sign measurement using adaptive facial regions," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 1410–1414.

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

[15] Lawrence T DeCarlo, "On the meaning and use of kurtosis.," *Psychological methods*, vol. 2, no. 3, pp. 292, 1997.

[16] Constantinos 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.

[17] Mika P Tarvainen, Perttu O Ranta-Aho, Pasi A Karjalainen, et al., "An advanced detrending method with application to hrv analysis," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 2, pp. 172–175, 2002.

[18] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

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