

Received January 15, 2020, accepted January 28, 2020, date of publication February 13, 2020, date of current version February 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973815

Automatic and Adaptive Signal- and Background-ROIs With Analytic-Representation-Based Processing for Robust Webcam-Based Heart-Rate Estimation

JAMES JOHN¹, SYAM KRISHNA², AND RAMESH R. GALIGEKERE¹², (Senior Member, IEEE)

¹Department of Electrical Engineering, IIT Delhi, New Delhi 110016, India

²Department of Biomedical Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India

Corresponding author: Ramesh R. Galigekere (ramesh.galigekere@manipal.edu)

ABSTRACT Determining a suitable, adaptive region of interest (ROI) *automatically* for extracting information related to cardiac activity (signal-ROI/S-ROI), and another containing information on ambient light-fluctuation (background-ROI/B-ROI, *as close as possible to the signal-ROI*), and robust signal processing are important in webcam based heart-rate (HR) estimation – in real life situations. We describe a novel method of automatically determining both the ROIs. The forehead is the candidate for the S-ROI, due to its uniformity and minimum vulnerability for deformation. We first identify the skin-pixels within the face-region detected by the Viola-Jones (VJ) algorithm. The forehead-region, and a uniform sub-rectangle within it *not containing hair* – determined by using variance as a measure – yields the S-ROI. The B-ROI – consisting of 3 rectangles – each of the same size as that of S-ROI, at the two sides and the top of the VJ-rectangle – is used to generate a reference signal for an adaptive noise-cancellation scheme. The situation arising from (possibly simultaneous) *facial expressions deforming the S-ROI*, is addressed – by extracting the phase sequence associated with the analytic representation of the signal. Experiments conducted with 21 healthy subjects, using this novel array of techniques, have produced good correspondence with the ground truth obtained from a standard finger-pulse transducer – as reflected by the Bland-Altman plot.


INDEX TERMS Webcam, heart-rate, adaptive, region of interest, background, ambience fluctuation, skin detection, color models, facial expression, analytical signal representation.

I. INTRODUCTION

Cardiovascular diseases have been the cause of morbidity and mortality throughout the world [1], leading to efforts towards understanding the human cardiovascular system, in terms of the heart rate (HR), and its variability [2]. Estimating HR by a webcam is a topic of current interest, due to low cost and non-contact nature.

In this context, Verkrussye *et al.* [3] proposed the use of forehead as a region of interest and the green-channel to record the photoplethysmogram, containing information on HR. Poh *et al.* [4], [5] proposed the use of a webcam in ambient light conditions, and independent component analysis (ICA) on the color channels to get information pertaining to HR. They used the central 60% width and the

full height of the rectangle bounding the face, as the ROI for signal-extraction. Kwon *et al.* [6] also used the entire face, but proposed the use of green channel alone (in the absence of subject-motion) over the use of ICA. Similar work by Sun *et al.* [7] also showed that ambient-light-fluctuations affect the signal. The use of sub-regions within the face (*e.g.*, the region below the eyes and above the lips, around the eyes and nose, the mouth) including the forehead, as ROIs, have also been proposed [8]–[14]. However, recently, Stricker *et al.* [15] confirmed that the forehead is the most suitable region for acquiring the signal containing the cardiac component. Therefore, we consider forehead as the candidate (and starting) region for determining a suitable ROI (*i.e.*, S-ROI) for tapping cardiac information. However, as discussed in the sequel and (also addressed in this paper), the presence of hair and involuntary facial expressions can affect the cardiac-component – even while

The associate editor coordinating the review of this manuscript and approving it for publication was Kezhi Li .

tapping the signal from a region on the forehead – the former obstructs information, while the latter affects through dynamic (time-varying) deformations.

Currently, there is a great interest in automating data acquisition and processing. The first step towards automation involves the selection of an ROI for signal acquisition. The Viola Jones algorithm [16], for example, has been used for identifying the face-region (*e.g.*, Poh *et al.* [4], [5] and Monkarese *et al.* [17]). Huang and Dung [18] start by identifying the nose, and outline and use the entire skin-region within the face as the ROI. Stricker *et al.* [15] identified a region on the forehead, based on eye-position and prior geometric considerations, to reduce computation, in the context of robotics. However, they have reported their *method to have become unreliable in the presence of hair*. Tarassenko *et al.* use a ‘face-registration’ algorithm and Bayesian segmentation (to get the entire background), and select square-ROIs – one as S-ROI (‘on the patient’s skin, preferably on the face’) and one as the B-ROI – in the context of monitoring HR during haemodialysis, under strong fluorescence light. However, there is not much clarity on the criterion for the selection of the ROIs. They fit AR-models to both signal and reference waveforms, and employed pole-cancellation to reject aliased version of flicker-frequency. However, while a few waveforms of HR as a function of time are given, the lack of statistical results, in view of the vulnerability of an AR model in the presence of noise and perturbation, leave some concerns not addressed [19]. Finally, note that the work in [12] brought out the *limitations of motion-based methods*, with respect to illumination-variations and subject-motion.

A common but significant issue arising in the context of webcam based HR-computation is that of changes in ambient illumination. It has been addressed to an extent, by adaptive filtering using a reference signal tapped from the ‘background’, to suppress the effects of fluctuations on the signal [12]. The ROI in [12] – a non-rectangular region, was estimated automatically by computing many facial landmarks; feature-points within the ROI were used for tracking the ROI, to counter rigid motion of the head. However, the ROI – the entire face below the eyes – is highly non-uniform and vulnerable to inner facial movements/facial expressions (see Stricker *et al.* [15]). The work in [12] involves segmentation to get the background-region (in each frame), to estimate a reference for the adaptive filter. While a reference from the entire background may be useful in the context considered in [12], *it cannot cater to local fluctuations (i.e., in the vicinity of the S-ROI – at which the signal is actually collected) – the information pertaining to which is crucial; indeed, the actual effect, locally on the S-ROI, may get buried by information extracted by averaging over the entire (global) scene – unless, the fluctuation turns out to be global. On the contrary, a B-ROI close to the vicinity of the S-ROI, will always involve information correlated to the changes in illumination within the S-ROI. After all, it is the fluctuation in the proximity to the S-ROI*

that is important. Finally, it is to be noted that signal segments associated with *errors due to facial expressions were chopped off* in [12] – this is undesirable, and also not amenable to automation.

While much work has been reported on camera-based HR estimation, automation under different challenging circumstances *i.e.*, changes in (i) ambience due to a shadow cast by someone moving in the vicinity – which is dynamic and may occur close to the S-ROI but not present elsewhere/far-off, (ii) involuntary facial expressions, and (iii) their simultaneous occurrence, are realistic problems. The presence of hair on the ROI is also known to affect the signal quality. This paper is the result of our efforts on addressing the aforementioned problems, and involves some novel contributions. Specifically, we propose a novel, automatic method of determining an adaptive, *hair-free sub-region* on the forehead, as the S-ROI. This was motivated by our study of the work of Lewandowska *et al.* [13], who used variance in thermal-images to show that the forehead is the flattest. In addition, we select a B-ROI – consisting of a set of regions – automatically, for estimating the reference signal to be used by an adaptive filter to suppress the effects of fluctuations in ambient light. In addition, the method is computationally simple/inexpensive, and selects the B-ROI *close to the S-ROI* – so that, the reference-signal closely represents the light-fluctuations in the vicinity of the S-ROI (and hence affecting the actual source of the signal). Finally, to combat the effects of facial expression (leading to deformation in the S-ROI and thus affecting the estimate of the cardiac component) – possibly simultaneously during ambience-fluctuation, we exploit the concept of *analytical-representation of signal* to enhance the cardiac-signal, and render the method robust.

It is pertinent to point out that at this juncture that, recently [20] – in the context of using a smartphone for estimating HR – a region of the forehead, based on a pre-determined (fixed) dimensions (geometric parameters) – has been used. Their method, however, is non-adaptive, and does not address the issue of avoiding interference by the presence of hair within the ROI. Indeed, they also stated that further studies are required to understand effects of external lighting and skin color and on the accuracy of the HR-estimate, and also indicate the necessity for *more accurate method for finding the forehead region*. On the other hand, our work is complementary, in that the findings in this paper can be useful in that context also.

The development of the methodologies required for automatically detecting the S-ROI and the B-ROI are explained in the following. The signal processing-steps, including adaptive suppression of the effects of fluctuations in ambient light, and the utility of analytic representation of signals to extract the cardiac-component, are outlined. The experimental set up and the results, validated by measuring HR through the finger pulse transducer, are presented in Section III. The results and some of the limitations are discussed in Section IV. The paper is concluded in Section V.

II. METHODS

A. CAPTURING THE FACE

Of the various possible regions for extracting cardiac-related waveforms, the forehead is the most preferable, as it is relatively uniform and suffers less from facial expressions. The 1st step in arriving at the forehead involves capturing the face, by the well-known Viola-Jones method [16], which provides the ‘V-J rectangle’ containing the face.

B. DETERMINING THE SIGNAL ROI (S-ROI)

Starting from the VJ-rectangle, we devise an algorithm to get an ROI (devoid of hair) within the forehead, lying between the eyebrows at the bottom and hair on the top (and possibly on the sides). The reflected illumination suffers greater change over the indicated border-regions. We use the knowledge of skin to eliminate hair, and local (spatial) variance to get a rectangle corresponding to the flat region over the middle of the forehead. Note that the latter (variance-based procedure) also assists avoiding regions with hair, making the method robust.

1) SKIN-DETECTION

The proposed algorithm starts with skin-detection, based on color. The first step in color-segmentation involves choosing an appropriate color-space. The well-known RGB color space, however, is prone to errors due to possible fluctuations in illumination. Of the various other color-spaces wherein color-information is isolated from that of intensity, the YCbCr model is popular for skin-segmentation. Further, since the color of the skin is influenced by the blood in the superficial vessels [21]–[23], it has higher values of red. Indeed, the expression for the YCbCr model in terms of the RGB model (ITU-R BT601 standard):

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 63.481 & 128.553 & 24.966 \\ -37.797 & -74.203 & 112 \\ 112 & -93.786 & -18.214 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

reveals that Cr is relatively rich in information pertaining to red. Finally, we found that the histogram of the Cr-values of skin pixels constructed from 23 subjects of different complexions, revealed a compact-distribution of Cr over the interval [130,160]. This range is very close to that found in [23]. The ratio Cb/Cr exhibited an even more compact distribution. Figure 1 shows the distribution of the values of Cr, and Cb/Cr, over the skin and non-skin regions within the VJ rectangle. We used the ratio Cb/Cr ∈ [0.6,0.85] for skin pixel extraction. The possible issue of some misclassification arising from the minor overlap between the two histograms, will be alleviated to an extent, by restricting the classification and further processing – to a sub-rectangle of the VJ-rectangle, and scanning/processing from the top – as described in the sequel.

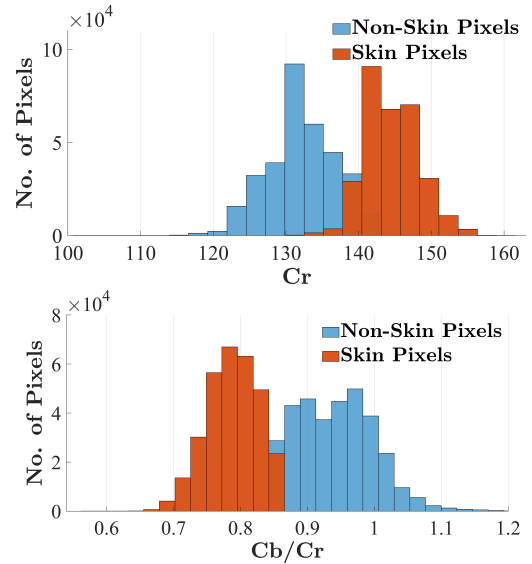


FIGURE 1. Histogram of the values of (Top): Cr, and (Bottom): the ratio Cb/Cr, corresponding to the skin and non-skin regions within the VJ-rectangle.

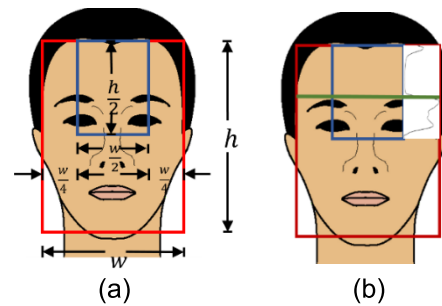


FIGURE 2. (a) A sub-rectangle within the VJ-rectangle, over which row-variances are computed to detect the eyebrow-level. (b) Plot of row-variance, showing the interval over which it is small and flat, which gives the ROI-height. The green-line indicates the detected eye-brow level.

C. UPPER AND LOWER BORDERS OF THE S-ROI

First, a sub-rectangle from the top of the VJ-rectangle, of half the height and half the width, as shown in Fig. 2 (a), is constructed. Starting with the topmost row in the sub-rectangle (which may contain some hair), row-variance (i.e., variance computed over a row) is calculated. This calculation is repeated over the subsequent rows while descending towards the eyebrow-level. The variance will reduce, as we first hit the row involving the forehead only (due to the relative uniformity). The variance will remain small, until we encounter a row close to the eyebrow – after which, the variance will increase again (due to an increase in the non-skin pixels and physical non-uniformity). The rows – at the top and the bottom prior to high values of variance provide the upper and lower borders of the ROI (See Fig. 2 (b)).

D. LATERAL BOUNDS OF THE S-ROI

Starting with the mid-point of the lower-border of the ROI, two small squares i.e., to the left and the right of the preceding

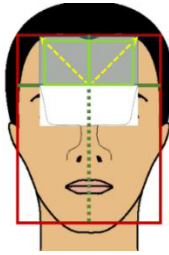


FIGURE 3. Illustration of the procedure for finding the lateral bounds on S-ROI. The green squares show one instance of the “growing squares”, over which local variances are computed. The arrows show the direction of “growth”. The flat intervals in the variance-plots (as a function of the distance from the center-point at the bottom), define the width of the S-ROI.

point but touching each other along the vertical line at the mid-point (dividing the forehead), are created to form a small rectangle. The local variances of the image over these squares are computed. The size of each of the squares is then increased successively in small steps (by the same measure), and the local variances over the two are computed at each step (Fig. 3). This procedure is stopped, when the values of the variances start increasing. Specifically, at every step i , the difference in the local variance *i.e.*, $\text{var}(i) - \text{var}(i - 1)$ is measured. An increase in its magnitude, exceeding a threshold (determined empirically, to be 50), is considered to signify the border of the ROI – *e.g.*, due to the onset of hair-region, physically non-uniform regions, etc.). At that juncture, the rectangle formed by these two squares give the final ROI – and will have minimum perturbation.

E. DETERMINING THE BACKGROUND ROI (B-ROI)

The B-ROI should be close to the face, but not include the face: the VJ-rectangle provides the reference for this purpose. We recommend and use three rectangles (to remove any spatial/directional bias), each of the same size as the signal-ROI: two by the sides of the face and the other over the top, so that fluctuations due to a moving shadow, possibly narrow in width or breadth – are captured.

Specifically, the S-ROI is reflected over to the sides of the face, just outside the VJ rectangle, to get two sub-regions. The other one is defined by reflecting the S-ROI with respect to the top-edge of the VJ-rectangle, by 45% of the height of the face (to avoid the face/head completely). The resulting background-ROI consists of 3 sub-regions (Fig. 4).

F. DATA ACQUISITION AND PROCESSING

1) DATA ACQUISITION

The green channel is used for getting signal related to cardiac activity, as proposed in [3], [6], [12]. The average of green-pixel-values over the S-ROI is computed over every frame, and its value as a function of frame number gives a time-series containing cardiac information *i.e.*, about the pulsatile blood flow. The average of the intensity-values over the three sub-regions of the B-ROI, as a function of frame number, provides the reference signal containing information about changes in the ambient light. The time-series are then de-trended.

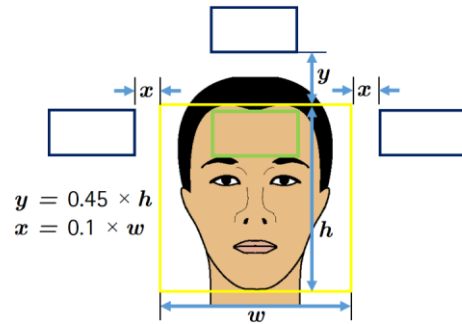


FIGURE 4. Illustration of the signal-ROI (on the forehead), and the background ROI – consisting of three black rectangles – two on the sides and one on the top of the face. The rectangle enclosing the face is the VJ-rectangle.

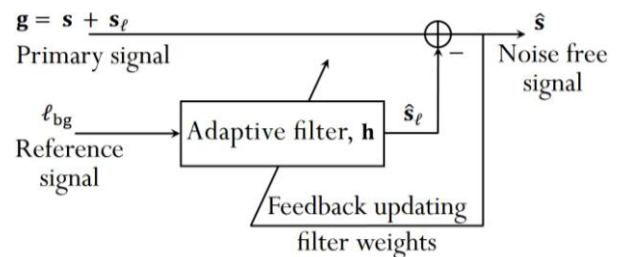


FIGURE 5. Adaptive filtering scheme to suppress the effects of fluctuations in illumination.

2) ADAPTIVE FILTERING

The reference signal is used to suppress the noisy component corresponding to light-fluctuation, from that containing the cardiac-component, by using an adaptive filter (AF) formulated based on the normalized least mean square (NLMS) criterion [24]. While such an adaptive noise cancellation approach was used earlier in a similar context [12], *our contribution in this paper is in determining an adaptive signal-ROI and background ROIs automatically*, for the purpose of providing a good and roust estimate of the cardiac signal – by estimating a time series involving cardiac components along with interference, as well as a good reference time-series – from regions not too far off from the S-ROI – so that it preserves the fidelity of changes in illumination *most relevant to the information in the S-ROI*.

Let $g(n)$ represent the measured signal, modelled as the cardiac component $s(n)$, corrupted by an additive component $s_l(n)$ correlated with the fluctuations in ambient light within the S-ROI. The background signal $l_{bg}(n)$ – containing information about the illumination-fluctuations – provides the reference to the adaptive filter, towards suppressing its effects on the measured signal.

The NLMS adaptive filtering scheme is shown in Fig. 5. It works by estimating that part of the input $g(n)$ (primary) – correlated to the reference waveform $l_{bg}(n)$ (secondary, due predominantly to ambient light only *i.e.*, *without* the cardiac component) – and canceling it from the input, giving an

estimate of the cardiac component.

$$\hat{s}(n) = g(n) - \hat{s}_l(n) \quad (2)$$

where:

$$\hat{s}_l(n) = \sum_{i=0}^{N-1} w(i) l_{bg}(n-i) = \mathbf{w}^T \mathbf{l}_{bg} \quad (3)$$

In eqn. (3), \mathbf{w} and \mathbf{l}_{bg} are N -point vectors. If the estimation error sequence is $e(n) = s_l(n) - \hat{s}_l(n)$, the optimum filter weights obtained by the NLMS criterion *i.e.*, minimizing the normalized mean square error, result in the following update-equations:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu(n)e(n)\mathbf{g}(n) \quad (4)$$

$$\mu(n) = \mu_s / [\varepsilon + \mathbf{g}^T(n)\mathbf{g}(n)] \quad (5)$$

where, ε is a small constant, and μ_s is the “step-size” (which determines the rate of convergence). The filter gradually learns the characteristics of the inputs, and its weights converge to the optimum values as dictated by the NLMS-criterion. If the input statistics change after convergence, the filter would respond by readjusting its weights to the new optimum values, and this process continues. Thus, changes due to illumination (not very abrupt) are tracked, and subtracted from the primary input, to yield the cardiac component.

The adaptive filtered signal is subsequently band-pass filtered [0.5-4] Hz (a standard range used to retain the cardiac component), to yield an enhanced version of $\hat{s}(n)$.

It is to be noted that, the adaptive filter, however, may not be able to completely cope up with (quicker) changes in intensity *i.e.*, it may introduce its own ‘noise’ – including that due to the reference signal which may not be a replica of the fluctuations over the S-ROI, and also due to the *assumed linear model* for the effects of light-fluctuations. Further, *the adaptive filter cannot account for changes in facial expressions*. We propose processing the enhanced signal further – to account for distortions resulting from facial expressions – towards extracting the cardiac component for robust HR-estimation, based on the concept of analytic signal.

3) PROCESSING BASED ON ANALYTIC SIGNAL REPRESENTATION [25]–[28]

Let $x(n)$ represent the enhanced version of the raw signal *i.e.*, band-passed version of the adaptive filter-output obtained in the previous step. Note that band-pass filtering ensures that it is zero-mean and relatively *narrowband*. The analytic representation of $x(n)$, is the inverse of the non-negative part of the Fourier transform of $x(n)$ – a complex sequence of the form:

$$y(n) = a(n)e^{j\phi(n)} \quad (6)$$

The real part of $y(n)$ is $x(n)$ itself, and is of the form $a(n)\cos(\phi(n))$, where $a(n)$ is the *envelop* and $\phi(n)$ is the *instantaneous phase-sequence*. We do not need the magnitude (envelop), but only the cosine of the phase-sequence.

$$s(n) = \cos(\phi(n)) \quad (7)$$

representing the component corresponding to the cardiac frequency. One may also note that since $a(n)$ – being the magnitude of the analytic signal – is positive, $\cos(\phi(n))$ would preserve the sign of $x(n)$. Thus, $s(n)$ would also preserve the zero-crossings of $x(n)$, and hence represent the cardiac component.

The component $s(n)$ may therefore be extracted from the analytic signal representation. The analytic signal representation pertaining to a discrete sequence is the inverse DFT of the “right-sided DFT” constructed from the sequence. Computationally, analytical signal representation can be obtained by using the Hilbert Transform (available on MATLAB).

Finally, note that the location of the tallest peak in the power spectrum density (PSD) of $s(n)$ will yield the cardiac frequency.

III. RESULTS

The experimental set up involves a laptop with an integrated webcam (HP True Vision), placed in room with good day light. Subjects were asked to sit still in front of the camera, at a distance between 0.5 m and 1 m.

Changes in illumination were simulated by a person moving around, during the video record – thus affecting illumination dynamically. This was, in fact, a phenomenon which we encountered in real life during this project, which motivated working on the problem itself, and the solution methodology – worked out subsequently – is being presented in this paper. To address the possibilities of involuntary facial expression in the subject simultaneously (in addition to dynamic changes in illumination), the subject was asked to make facial expressions during the record. The video of each subject was recorded for about a minute; the frame rate was 30 fps, at a resolution of 640×480 pixels.

The recorded videos were processed in the sequence presented in the previous section. The results pertaining to each of the steps associated with determination of the ROI, are displayed in Fig. 6. Further, six examples of the detected signal-ROIs – are presented in Fig. 7, demonstrating the efficacy of the automatic ROI-detection technique. Careful observation reveals the adaptability of the ROI, and how hairy regions have been avoided. Clearly, ROIs of predetermined/fixed size cannot be used in all the cases.

The data acquired from the S-ROI is de-trended by two-stage mean-filtering [29]. In the context of adaptive filtering by the NLMS method, the value of the step-size parameter μ_s , was set at 0.5 (estimated empirically), to assist quick convergence. The function *hilbert* in MATLAB was used to construct the analytic signal representation. The periodogram [30] was found to be sufficient to extract the value of HR from the enhanced signal.

A sample of the time-series measured from the signal and background ROIs, is presented in Fig. 8. A sample-sequence of results pertaining to the signal processing steps, is displayed in Fig. 9. Note that the recorded/raw signal (after de-trending), exhibits the effects of fluctuations in the ambience and facial expression, and its PSD is corrupted

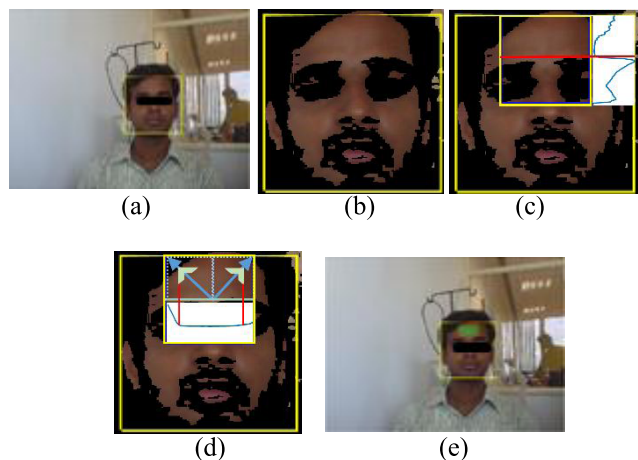


FIGURE 6. Results of steps in S-ROI-detection algorithm: (a) VJ rectangle, (b) Skin-pixel region, (c) Upper & lower bounds on the ROI, (d) Determining the lateral bounds of the ROI in terms of two concatenated square-regions growing in diagonally opposite directions, based on local variance (note its increase as the squares encounter hairy regions), and (e) the final ROI.



FIGURE 7. Examples of signal-ROIs detected automatically, by the algorithm developed in this paper. Note that the ROIs have adapted to the respective face, and have successfully avoided hairy regions. We have superposed a back patch on the eye to conceal the identity of the subjects.

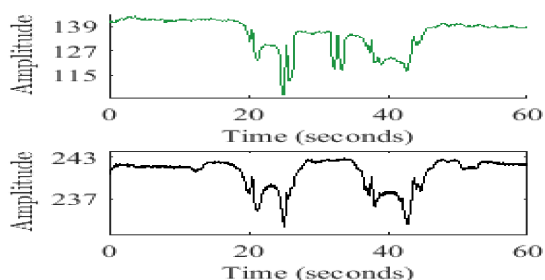


FIGURE 8. (a) The measured signal (green channel) from the signal-ROI, (b) Reference signal from the background ROI.

with noisy peaks. Adaptive filtering removes some of the noise (but not completely). However, the PSD of the cardiac-component extracted using analytic representation, is cleaner, and the location of the tall peak gives the value of HR (93.8232 bpm). A record of the finger pulse transducer, acquired simultaneously (in parallel), is displayed in Fig. 8,

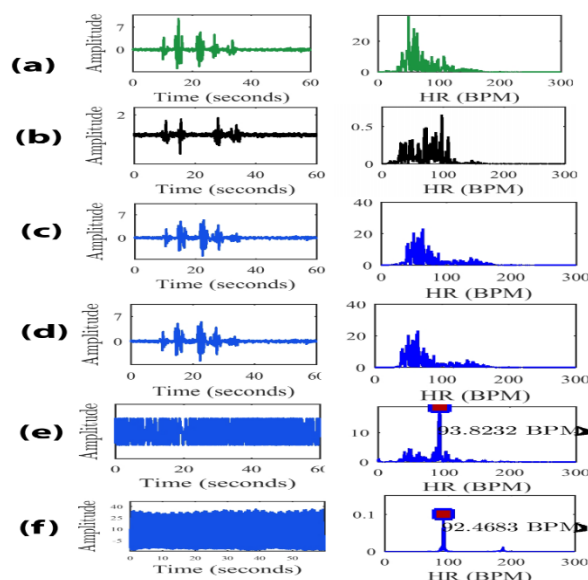


FIGURE 9. Left-column: (a) De-trended raw signal, (b) de-trended reference signal, (c) adaptive filtered signal, (d) band-pass-filtered version of (c), (e) cardiac-component extracted from analytic version of (d), (f) finger-pressure transducer signal. Right column: the respective PSDs.

along with its PSD, the location of the peak in which gives the ground truth (92.4683 bpm). We have added a couple of videos – which demonstrate the acquisition (please see the Supplementary files) of the videos along with the time-series from the S-ROI and B-ROI, recorded simultaneously. They demonstrate the effects of fluctuation in illumination and changes in facial-expression, on the signals.

We tested the algorithm on 21 volunteers, with changes in ambient light and facial expression occurring simultaneously (note that the method worked well also in cases involving either of them alone, or both in sequence). We also investigated the efficacy of B-ROI with one and two (on the sides) sub-regions each, respectively. However, as anticipated, they did not perform as well as that with 3 sub-regions (Fig. 4) – in a few cases. B-ROI with 3-sub-regions worked in all the 21 cases. This proves an important practical aspect that, changes in ambient light need not be uniformly distributed, and therefore, B-ROI must be as close as possible to the S-ROI.

Instead of listing all the values of HR in a table, we provide Bland-Altman analysis [31] – useful in demonstrating the proximity of a set of values to another set of values. The Bland-Altman plot showed the best results in the case of B-ROI with 3 sub-regions, and is displayed in Fig. 10.

IV. DISCUSSION

Although substantial work has been reported so far in the area of webcam-based HR estimation, no one has addressed completely automatic determination of an adaptive, hair-free region on the forehead as S-ROI, nor the concept of using a set of sub-regions comprising the B-ROI close to the S-ROI for robust estimation of HR – in the presence of fluctuations

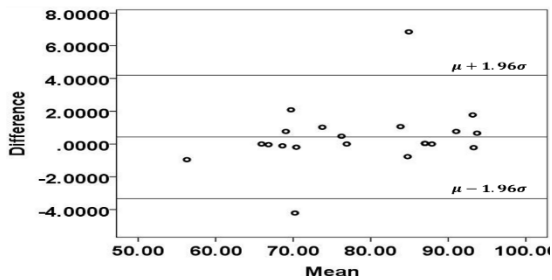


FIGURE 10. Bland-Altman plot (difference “ d ” vs. mean “ μ ”) showing the statistical agreement between the results of the proposed method and the ground-truth obtained through the finger pulse transducer.

in illumination. Further, while the problem arising from a deformation of the S-ROI due to involuntary facial expressions has been indicated in the literature, a solution to that (not by discarding the data itself!) – has not been suggested – to the best of our knowledge. In this paper, a novel set of algorithms for tackling both of them – occurring possibly simultaneously (their occurring over distinct durations is an easier subset of the problem, which our method tackles anyway) – has been described. Deformation due to facial expression produces a non-linear effect on the signal, and we have effectively addressed that by exploiting the analytic signal formulation which allows one to extract the component with sustaining oscillations in the signal. Finally, note that the skin-detection method, based on the $YCbCr$ method, tested by us on people with various degrees of complexions in the Indian scenario (fair to dark), has been shown to work well over a wide range of complexions [22]. Finally, any other method – perhaps, one based on machine learning, can be used instead of the Viola-Jones method – but then, the emphasis is on automatic determination of S-ROI and B-ROI.

The results obtained by implementing our algorithm have been validated with the ground-truth obtained from a finger pulse transducer. We wish to point out that, there were a few instances in which *our method outperformed the finger-print transducer itself* – as observed by us in terms of cleaner PSD than that pertaining to the transducer waveform. A careful examination revealed the source of the problem – to be a slight movement of the transducer. While we do not wish to overstate this, one can say that the proposed method based on the webcam is fairly robust.

Future directions include devising an adaptive scheme for weighting the components from the B-ROI to further-improve robustness to fluctuations in illumination, and compensating for possible subject-motion – although, in our experiments, we did not encounter a problem due to motion (in spite of very mild subject-motion being inevitable) – possibly because of the good choice of ROI over the flat forehead; motion compensation can be introduced if necessary. One may also look into using a band of hue instead of the green channel (as suggested in [20] in a different context *i.e.*, involving the use of a smartphone – which may have its own filters; our work is not about contradicting that). On the other hand, the concepts of automatically and adaptively estimating a

hair-free region of the forehead as the S-ROI and identifying suitable B-ROI for extracting background information, and subsequent signal processing – developed herewith, will be useful to the smartphone scenario. Suitable bandpass filtering can perhaps allow an estimation of the respiratory rate also, as shown in [20] – though that is not the main focus of this paper.

Although the results of several efforts have appeared in the field (estimating HR from webcam), it should be noted that, in practice, there are many different scenarios, not all of which have been addressed, and the corresponding solutions are often application-dependent. The problem we have encountered – in a lab-scenario, is a realistic one, for which an effective array of processing steps has been developed – involving several novel approaches. As far as different types of illumination are concerned, we point out that, in [12], illumination-rectification by adaptive filtering was shown to work under varying types of illumination; further, Basilio [22] showed the effectiveness of the $YCbCr$ model on images of people with different complexion, also captured under a variety of illumination. Thus, [12] and [22] support the robustness of our method, though explicit changes in the colour of illumination were not performed. In conclusion, our work has many novel solutions, tackles challenging situations that arose in our lab (different from those pertaining to databases such as MAHNOB-HCI used in the context of human-computer interaction), and is useful in its own right – the concepts of a B-ROI with three sub-regions close to the S-ROI, and the latter being on the forehead (a well-acknowledged region for tapping cardiac activity, while avoiding hairy regions) and the application of analytical signal-representation, to name a few. Indeed, some of the problems addressed by us (first in 2016 [32], later presented at a conference in April 2019 [33]), has recently been acknowledged as ‘work to be done’, in an interesting paper in the context of a smartphone [20]. With respect to a paper just published [33] – our scenario is different; our idea of automatically determining an S-ROI on forehead (established to be good & robust, unlike the entire skin-pixels on the face in [33]), and a B-ROI for tapping information on fluctuations in illumination explicitly is useful (making the method robust) in general, and the idea of using analytic signal with several interesting possibilities of extracting additional information [34], apart from taking care of the effects of involuntary facial expressions, is a novel contribution. Finally, we do not need a ‘reference’ pulse signal’ (as in [33]) to extract the cardiac signal, and our method worked well on dark-skinned people also. Thus, our techniques in our work are complementary to the existing efforts in this field.

V. CONCLUSION

In the context of extracting heart-rate from facial videos recorded from a webcam, we have described the development of a novel adaptive S-ROI and a special B-ROI, automatically – one for acquiring the signal pertaining to cardiac activity, and the other for acquiring the noise-component

pertaining to light fluctuations. The necessity for having the multiple segments comprising the B-ROI – close to the S-ROI – was highlighted, and its computation is negligible. Estimation of the S-ROI starts from finding the VJ-rectangle, and uses skin-pixel-detection, followed by row and local-area variances to arrive at the final ROI. The time series extracted from the S-ROI and B-ROI were used to obtain an enhanced signal by adaptive filtering and band-pass filtering, and processed further to reject the effects of facial expression, occurring simultaneously (or otherwise), to estimate the cardiac component. The location of the peak in the PSD of the enhanced signal yields the HR. The method, tested in 21 subjects, yielded results close to the ground-truth, as inferred by the Bland-Altman plot.

The concepts of automatically and adaptively estimating a hair-free region of the forehead as the S-ROI and identifying suitable B-ROI for extracting background information, and subsequent signal processing – developed herewith, is expected to be useful even in the context of assessing HR from a smartphone and perhaps other similar applications to come.

ACKNOWLEDGMENT

The authors would like to thank the various subjects who volunteered to have their face-videos recorded for the purpose of this study. They also thank Dr. A. G. Ramakrishnan, Department of Electrical Engineering, IISc, Bengaluru, India, for useful discussions.

REFERENCES

- [1] J. V. Fuster and B. B. Kelly, Eds., *Promoting Cardiovascular Health in the Developing World: A Critical Challenge to Achieve Global Health*. Washington, DC, USA: Academic, 2010.
- [2] "Heart rate variability: Standards of measurement, physiological interpretation and clinical use. Task force of the European society of cardiology and the north American society of pacing and electrophysiology," *Circulation*, vol. 93, no. 5, pp. 1043–1065, Mar. 1996.
- [3] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Opt. Express*, vol. 16, no. 26, pp. 21434–21445, Dec. 2008.
- [4] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Opt. Express*, vol. 18, no. 10, pp. 10762–10774, May 2010.
- [5] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a Webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, Jan. 2011.
- [6] 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 *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2012, pp. 2174–2177.
- [7] Y. Sun, C. Papin, V. Azorin-Peris, R. Kalawsky, S. Greenwald, and S. Hu, "Use of ambient light in remote photoplethysmographic systems: Comparison between a high-performance camera and a low-cost Webcam," *J. Biomed. Opt.*, vol. 17, no. 3, 2012, Art. no. 037005.
- [8] T. Pursche, J. Krajewski, and R. Moeller, "Video-based heart rate measurement from human faces," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2012, pp. 544–545.
- [9] Y.-P. Yu, R. Paramesran, and C.-L. Lim, "Video based heart rate estimation under different light illumination intensities," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Dec. 2014, pp. 216–221.
- [10] T. Kitajima, S. Choi, and E. A. Y. Murakami, "Heart rate estimation based on camera image," in *Proc. 14th Int. Conf. Intell. Syst. Design Appl.*, Nov. 2014, pp. 50–55.
- [11] D. McDuff, S. Gontarek, and R. W. Picard, "Improvements in remote cardiopulmonary measurement using a five band digital camera," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2593–2601, Oct. 2014.
- [12] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4264–4271.
- [13] M. Lewandowska, "Measuring pulse rate with a Webcam—A non-contact method for evaluating cardiac activity," in *Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, Szczecin, Poland, 2011, pp. 405–410.
- [14] L. Feng, L.-M. Po, X. Xu, and Y. Li, "Motion artifacts suppression for remote imaging photoplethysmography," in *Proc. 19th Int. Conf. Digit. Signal Process.*, Aug. 2014, pp. 18–23.
- [15] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *Proc. 23rd IEEE Int. Symp. Robot Hum. Interact. Commun.*, Aug. 2014, pp. 1056–1062.
- [16] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.
- [17] H. Monkaresi, R. A. Calvo, and H. Yan, "A machine learning approach to improve contactless heart rate monitoring using a Webcam," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, pp. 1153–1160, Jul. 2014.
- [18] R.-Y. Huang and L.-R. Dung, "Measurement of heart rate variability using off-the-shelf smart phones," *BioMed. Eng. OnLine*, vol. 15, no. 1, p. 11, Jan. 2016.
- [19] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiol. Meas.*, vol. 35, no. 5, pp. 807–831, May 2014.
- [20] S. Sanyal and K. K. Nundy, "Algorithms for monitoring heart rate and respiratory rate from the video of a user's face," *IEEE J. Transl. Eng. Health Med.*, vol. 6, pp. 1–11, 2018.
- [21] L. A. Brunsting and C. Sheard, "The color of the skin as analyzed by spectro-photometric methods: III. The role of superficial blood," *J. Clin. Invest.*, vol. 7, pp. 593–613, 1929.
- [22] J. A. M. Basilio, "Explicit image detection using YCbCr space color model as skin detection," in *Proc. Amer. Conf. Appl. Math., 5th WSEAS Int. Conf. Comput. Eng. Appl.*, Puerto Morelos, Mexico, 2011.
- [23] M. A. Rahman, I. K. Edy Purnama, and M. H. Purnomo, "Simple method of human skin detection using HSV and YCbCr color spaces," in *Proc. Int. Conf. Intell. Auto. Agents, Netw. Syst.*, Aug. 2014, pp. 58–61.
- [24] S. J. Orfanidis, *Optimum Signal Processing*, 2nd ed. New York, NY, USA: McGraw-Hill, 1988.
- [25] A. Papoulis, *Signal Analysis*. New York, NY, USA: McGraw-Hill, 1977.
- [26] L. Marple, "Computing the discrete-time 'analytic' signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, Sep. 1999.
- [27] M. Elfataoui and G. Mirchandani, "A frequency-domain method for generation of discrete-time analytic signals," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3343–3352, Sep. 2006.
- [28] J. Dugundji, "Envelops and pre-envelops of real waveforms," *IRE Trans. Inf. Theory*, vol. 4, no. 1, pp. 53–57, Mar. 1958.
- [29] A. E. Awodeyi, S. R. Alty, and M. Ghavami, "Median filter approach for removal of baseline wander in photoplethysmography signals," in *Proc. Eur. Modelling Symp.*, Nov. 2013, pp. 261–264.
- [30] A. E. Oppenheim and R. W. Schaffer, *Digital Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2015.
- [31] J. M. Bland and D. G. Altman, "Statistical method for assessing agreement between two methods of clinical measurements," *Lancet*, vol. 1, no. 8476, pp. 307–310, Feb. 1986.
- [32] J. John, "Robust non-contact heart-rate estimation using Webcam," M.S. thesis, Dept. Biomed. Eng., Manipal Inst. Technol., Manipal Acad. Higher Edu., Manipal, India, Jun. 2016, Art. no. 576104.
- [33] L. Qi, H. Yu, L. Xu, R. S. Mpanda, and S. E. Greenwald, "Robust heart-rate estimation from facial videos using Project_ICA," *Physiol. Meas.*, vol. 40, no. 8, Sep. 2019, Art. no. 085007.
- [34] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals," *Proc. IEEE*, vol. 80, no. 4, pp. 520–538, Apr. 1992.



processing, and signal processing and machine learning.

JAMES JOHN received the B.Tech. degree in electronics and biomedical engineering from the TKM Institute of Technology, Kollam, Kerala, and the M.Tech. degree in biomedical engineering from the Department of Biomedical Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. He is currently working as a Project Assistant with the Department of Electrical Engineering, IIT Delhi. His areas of interest include image/video



video processing. He is a Life Member of the Biomedical Engineering Society of India.

SYAM KRISHNA received the B.E. degree from Anna University, Chennai, and the M.Tech. degree in biomedical engineering from the Manipal Institute of Technology. He is currently pursuing the Ph.D. degree with the Department of Biomedical Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. His research interest includes non-contact estimation of heart rate and its variability, in humans and in animal models, through



the Robarts Research Institute (RRI), and the Neurovascular Research Laboratory at the School of Kinesiology, University of Western Ontario (UWO) – both in London, Ontario. He worked briefly as a Guest Scientist at the Medical Engineering Group, Siemens AG, Erlangen, Germany. He also served as an Adjunct Professor for the Department of Electrical Engineering, UWO, Canada. Currently, Dr. Galigekere is a Professor at the Department of Biomedical Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. He has published in various international journals and conferences, and has been serving as a reviewer for many. Two of his invention-disclosures (with co-inventors) were awarded by Intellectual Ventures, and subsequently awarded U.S. patents. Current interests of Dr. Galigekere are in biomedical signal processing, and imaging/image processing (contact/non-contact methods of estimation and analysis of vitals in humans and animal models – particularly zebrafish larvae, analysis of naadi signals (three pulse-signals from the wrist), image processing, and visualization in microscopy). He is a Life Member of the Biomedical Engineering Society of India.

• • •