Estimation of the Cardiac Pulse from Facial Video in Realistic Conditions

Arvind Subramaniam[1] and K. Rajitha[2]
[1]*Department of Electrical and Electronics Engineering, BITS PILANI Hyderabad Campus, Telangana, India*
[2]*Department of Civil Engineering, BITS PILANI Hyderabad Campus, Telangana, India*

Keywords: Heart Rate, Faster RCNN, Recursive Least Square Filtering, KLT Feature Tracking, Feature Point Recovery, Independent Component Analysis.

Abstract: Remote detection of the cardiac pulse has a number of applications in fields of sports and medicine, and can be used to determine an individual's physiological state. Over the years, several papers have proposed a number of approaches to extract heart rate (HR) using video imaging. However, all these approaches have employed the Viola-Jones algorithm for face detection. Additionally, these methods usually require the subject to be stationary and do not take illumination changes into account. The present research proposes a novel framework that employs Faster RCNNs (Region-based Convolutional Neural Networks) for face detection, followed by face tracking using the Kanade-Lukas-Tomasi (KLT) algorithm. In addition, the present framework recovers the feature points which are lost during extreme head movements of the subject. Our method is robust to extreme motion interferences (head movements) and utilizes Recursive Least Square (RLS) adaptive filtering methods to tackle interferences caused by illumination variations. The accuracy of the model has been tested based on a movie evaluation scenario and the accuracy was estimated on a public database MAHNOB-HCI. The output of the performance measure showed that the present model outperforms previously proposed methods.

1 INTRODUCTION

The measurement of physiological parameters such as blood pressure, heart rate (HR) and electrocardiogram (ECG) signals have been widely utilized in medical diagnosis. The determination of heart rate is a critical task which is used to indicate the overall health of an individual. Employing conventional techniques to measure HR such as electrocardiography or optical sensors which maintain physical contact with the subject can be uncomfortable, since the pressure can become inconvenient over time. Fortunately, previous papers have demonstrated that it is possible to extract HR by using the facial video of the individual (Verkruysse *et al.* 2008, Poh *et al.* 2010).

In 2007, Garbey *et al.* proposed a novel method to measure the cardiac pulse by analyzing the thermal signals transmitted by the major blood vessels present near the surface of the skin (Garbey *et al.* 2007). The change in temperature occurs due to the variation in the blood flow. Subsequent papers have overcome this drawback by focusing on measuring subtle changes in head motions as a result of the person's heart rate (Cennini *et al.* 2010, Balakrishnan *et al.* 2013). These oscillations are a direct result of the Newtonian reaction of the head to the influx of blood at every heartbeat. One chief drawback of this approach is that it typically requires the user to be stationary throughout the video. This is not practical in a realistic scenario, since the user's movements will include both external motion like head tilt as well as internal motions such as smiling and blinking.

There have been several papers over the years which have solved this problem by concentrating on small variations in skin color with HR (Poh *et al.* 2011, Kwon *et al.* 2012, Yu *et al.* 2013, Mohd *et al.* 2015, Moreno *et al.* 2015). After performing face detection, they have identified a region of interest (ROI) constituting roughly 60% of the face and calculated the mean pixel value of each colour component (RGB) of the ROI for each frame of the video. Next, Independent Component Analysis (ICA) was employed to obtain the corresponding plethysmographic signals from the RGB traces. Finally, they applied FFT to transfer these signals into

Subramaniam, A. and Rajitha, K.
Estimation of the Cardiac Pulse from Facial Video in Realistic Conditions.
DOI: 10.5220/0007367801450153
In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, pages 145-153
ISBN: 978-989-758-350-6

the frequency domain and acquired the frequency of the heartbeat. The output showed that the plethysmographic signal corresponding to the green channel has the maximum power spectral density among the three colour channels.

Although these models are significantly better than their motion-based counterparts, they do not consider real world challenges such as illumination variations into account.

In 2016, Tulyakov et al. introduced a novel optimization framework that estimates HR by using local estimates from multiple regions in the face, by using a chrominance-based method to relax motion constraints (Tulyakov *et al.* 2016).

Zhang et al. employed a six channel ICA algorithm to simultaneously detect HR and blink (Zhang *et al.* 2017). Although the algorithm is marginally tolerant to motion artifacts, they have not concretely focused on removing illumination artifacts. As a result, on a challenging dataset comprising of considerable motion as well as illumination artifacts, the accuracy of their method diminishes.

Furthermore, even though a few papers consider illumination variations (Li *et al.* 2014, Lam and Kuno 2015), it is important to note that all of the afore-mentioned models employ the Viola-Jones algorithm for face detection. To the best of our knowledge, no framework has used a different algorithm for face detection.

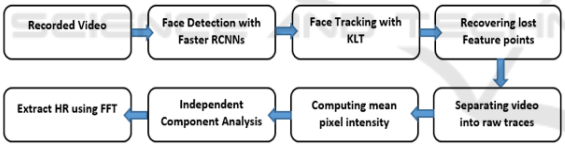

Figure 1: Stages of the proposed approach. The feature point recovery system is triggered when the number of feature points at any given instant falls below a pre-defined threshold value.

The present research focuses on the following objectives:

1. To build a feature point recovery system to get back feature points lost during extreme head movements. The system monitors the total number of feature points at a particular instant reapplies face detection and tracking to obtain lost points.
2. To apply Recursive Least Square (RLS) based adaptive filtering methods to tackle illumination interferences in the video.

In addition, we have employed Faster RCNNs, as opposed to the Viola-Jones face detection algorithm used by previous papers. The advantages of employing Faster RCNNs have been illustrated in Section 3. Although face detection and tracking do not form the crux of this work, we have also optimized this component of the framework, rather than employing the Viola-Jones face detection algorithm.

The rest of the paper is organized as follows: In Section 3, we employ Faster RCNNs to perform face detection and examine the advantages of using Faster RCNNs over the widely used Viola-Jones algorithm. Next, we utilize KLT algorithm to track features of the face and also introduce a feature point recovery system, thus providing greater robustness to motion artifacts. Section 4 introduces a novel approach to rectify changes in illumination which often compromise the accuracy of the HR measurement framework. Section 5 deals with extracting the heart rate using ICA and band limits the frequency in the desired range (0.60 – 4 Hz) with the help of temporal filters. In Section 6, we devise a movie evaluation scenario to verify the accuracy of our framework and determine the percentage error after comparison with a standard heart beat sensor. Additionally, we compare our models with previously proposed models and show that our framework significantly outperforms previous models. Finally, Section 7 draws conclusions.

2 DATASETS

In this study, we have used the WIDER face database to train our Faster RCNN model (Yang *et al.* 2016). The WIDER dataset consists of 12,880 images comprising over 150,000 faces as part of the training set. MATLAB 2015a was used under Windows 10 operating system for the entire implementation of the project.

We have employed two datasets to implement and evaluate our framework. The first dataset consists of 18 videos (9 male and 9 female) recorded by ourselves using a standard laptop webcam, with considerable motion interferences but no illumination variations. Each video has a frame rate of 30 frames per second, pixel resolution of 1280 x 720 and a duration of approximately 50-60 seconds. The subjects ranged from 20-55 years of age and varied in complexion.

The second dataset is the MAHNOB-HCI database (Soleymani *et al.* 2012), a publically available database comprising 527 colour videos of 27 subjects (12 males and 15 females). The videos have a resolution of 780 x 580 pixels recorded at 61 FPS. However, we were able to ultimately test our framework on 487 videos since the rest of the other videos did not have corresponding ECG readings when we downloaded the dataset.

3 FACE DETECTION AND TRACKING

Faster RCNNs, proposed by Ren *et al.* have proven to be over 200x faster than conventional neural networks, in carrying out tasks such as object detection (Ren *et al.* 2017). While previous papers have used the Viola-Jones algorithm for face detection, we have deviated from this approach by employing Faster RCNNs and demonstrate that our face detection framework achieves superior results. Faster RCNNs primarily performs face detection in two steps. The first step consists of a fully convolutional neural network, known as the Region Proposal Network (RPN), to extract the features of an image to generate object proposals and feed it to the subsequent module. Object/Region proposals are made based on the probability of detecting an object in a particular region. In the second module, these region proposals are refined and classified accordingly using ROI pooling. Fig. 2 compares the face detection results obtained from our Faster RCNN model with those obtained from the Viola-Jones face detection algorithm.

Firstly, the Faster RCNN model identifies the face even when a portion of the face is hidden from the camera. For instance, our model is able to detect the face without requiring the full frontal profile of the subject, as opposed to the Viola-Jones algorithm, which has primarily been trained and tested on frontal face datasets. Secondly, as illustrated in Fig. 2, it is possible for the Viola-Jones algorithm to falsely detect a face (false positives) if the degree of contrast of the background is similar to the degree of contrast in the person's face. We have used our own dataset to compare the two face detection methods.

Next, the KLT feature tracking algorithm is applied to extract feature points and track the face of the subject. This ensures that our framework is able to consistently detect the face and extract HR even when there is non-rigid head movement. After finalizing the ROI inside the face, the KLT algorithm detects feature points inside the ROI. Since KLT can be executed quickly, we used the feature points to obtain the ROI in the next frame. The feature points of the next frame are given by: $F(t+1) = AF(t)$, where A is the transformation matrix. Next, we apply the transformation matrix A to compute the boundary points of the ROI in the next frame: $R(t+1) = AR(t)$. Since the feature points are employed to compute the coordinates of the ROI in the next frame, the loss of feature points will invariably compromise the overall accuracy of the framework.

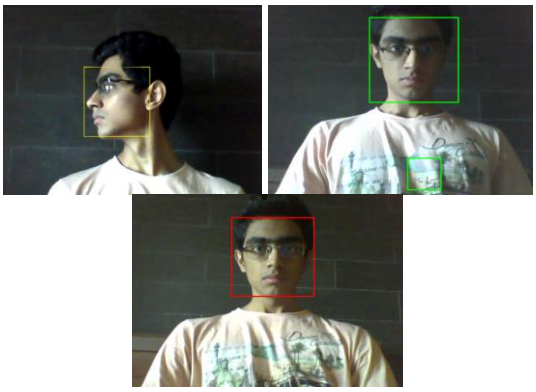

Figure 2: The results after performing face detection (from left): i) The Faster RCNN framework is able to detect faces without requiring the full frontal face. This makes HR extraction more robust to motion artifacts; ii) and iii) are the ROIs obtained after applying Viola-Jones and Faster RCNN respectively, on the same video. There are multiple ROIs detected in (ii) which can lead to erroneous results. This has been solved in (iii), as shown above.

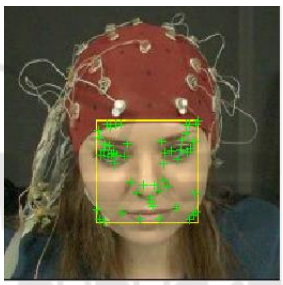

Figure 3: After applying the Face detection and KLT algorithm on the MAHNOB-HCI database.

3.1 Recovering Lost Feature Tracking Points

An important feature of our framework is the recovery of the feature tracking points which are lost during extreme head movements. In case of large head movements, it is possible for parts of the subject's face to get obscured from the camera. As a result, this may lead to the loss of a large number of feature points.

In order to recover the feature points, we monitor the total number of features at any given instant and compare it with a threshold value (T_1). In case the number of tracking points falls below the threshold value, the Faster RCNN and the KLT algorithm are reapplied, exactly 15 frames after the frame where the failure had occurred. We have taken the threshold value to be 60% of the total number of feature points. However, monitoring the total number of feature points is not a sufficient requirement. Although

reapplying KLT would produce the same number of feature points as before, it would not be possible to recover the feature points corresponding to the obscured portion of the face, since the new feature points would have different locations. Hence, we compute the root mean square error (RMSE) of the new feature point centroid ($\mu(t)$) relative to the old centroid ($\sigma(t)$) using

$$\in(t) = \sqrt{\frac{\sum_{i=1}^{N} |\mu(t) - \sigma(t)|^2}{N}}, \quad (1)$$

where N denotes the number of features points at a given frame t and ∈(t) denotes the RMSE. Faster RCNN and KLT are applied once every 15 frames (0.5 seconds) until $\in(t)$ reaches a minima. Fig. 5 illustrates the choice of RMSE and the corresponding frame at which the feature points are recovered. Face detection and tracking are not re-implemented beyond this frame, as shown in Fig.4. Since Faster RCNNs take less than 0.25 seconds per image for face detection, our framework can tackle extreme head rotations without compromising the overall speed of the model.

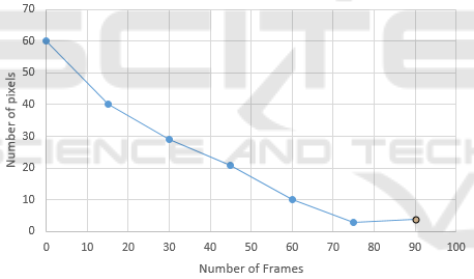

Figure 4: The RMSE represents the deviation of the new centroid $\mu(t)$, from the old centroid $\sigma(t)$ and is given by the difference in the number of pixels between $\mu(t)$ and $\sigma(t)$. The recovery of feature points terminates after the minima is achieved – in this case, 180 frames (3 seconds) after tracking failure.

As shown in Figure 5, after recovering lost feature points, it is possible to consistently track the subject's face and obtain the ROI despite considerable head movement. We found that the plethysmographic signal corresponding to the green channel is the most prominent among the three colour channels. Hence, we have primarily utilized the green channel for HR measurement and illustration purposes.

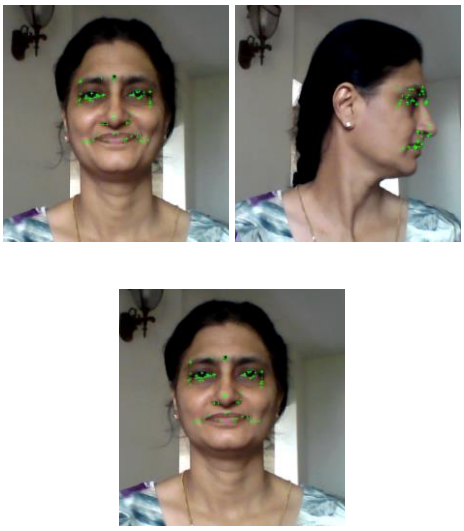

Figure 5: Face tracking using the KLT algorithm and recovery of tracking points in case of tracking failure. Despite getting obscured from the camera, the features in left part of the face are recovered by applying Faster RCNN and the KLT algorithm again, starting 15 frames after the frame of detection failure. From left to right, the frame numbers are 21, 150 and 240 (taken from our dataset).

4 RECTIFYING ILLUMINATION INTERFERENCES

The mean pixel intensity computed from the ROI (x_{orig}(t)) in each frame primarily consists of two signals. The first signal is a result of variations in blood flow caused by the cardiac pulse; the second signal is unwanted illumination variations that occur during the recording (Parra 2007). It is reasonable to assume that these two interferences are additive:

$$x_{orig}(t) = S(t) + N(t), \quad (2)$$

where S(t) denotes the original green channel variations caused by the cardiac pulse, and N(t) denotes the green channel variations caused by illumination variation as shown in Fig. 6.

Our objective is to eliminate the noise signal N(t) and obtain the best possible approximation of the signal S(t). Since the lighting sources are practically the same for the ROI and the background, we assume that the associated illumination changes for the ROI and the background would be the same as well (Basri and Jacobs 2003). Hence, we have employed the mean channel intensity of the background ($x_{bg}(t)$) to remove the noise signal N(t).

We have assumed N(t) to vary linearly with $x_{bg}(t)$:

$$N(t) \approx K x_{bg}(t). \quad (3)$$

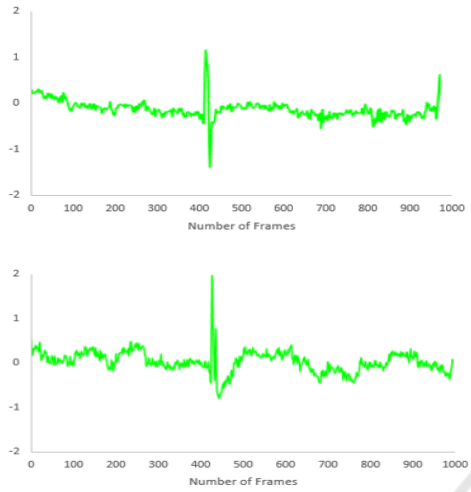

Figure 6: The top curve shows the normalized mean green pixel intensity of the ROI before rectifying illumination interferences. The curve below is a plot of the normalized mean green channel intensity of the background. The spike in the intensity is a result of illumination interference.

However, since we only have an approximation of the noise signal N(t), the illumination rectified signal x_{rect}(t) will invariably be a close approximation of S(t), and is given by

$$x_{rect}(t) = S(t) + E(t), \quad (4)$$

where

$$\mathrm{E(t)} = N(t) - K x_{bg}(t), \quad (5)$$

is the deviation of the illumination rectified signal from S(t). In order to compute the best possible approximation of $x_{rect}(t)$, we have utilized the Recursive Least Squares (RLS) adaptive filter to calculate the ideal value of K that minimizes the error. The RLS adaptive filter is an algorithm that recursively computes the filter coefficients that minimize a linear cost function related to the input signal.

Let $K(t)$ be the estimated filter weight for each point time point t. After initializing the weights, the RLS filter updates the filter weights as

$$K(t+1) = K(t) + C^{-1}(t) x_{rect}(t) x_{bg}(t) \quad (6)$$

Here, $C(t)$ is the autocorrelation matrix given by

$$C(t) = \sum_{i=0}^{t} x_{bg}(t) x_{bg}^{T}(t)\, \alpha^{t-1}, \quad (7)$$

where $x_{bg}^{T}(t)$ is the transpose of $x_{bg}(t)$ and α is a positive constant smaller than 1. The RLS filter will continue to run its iterations until K(t) converges to a suitable value that minimizes the error/deviation $E(t)$.

We have applied the Local Region Based Active Contour (LRBAC) method to segment the background region of each frame (Lankton and Tannenbaum 2008). Since LRBAC is a region-based approach, it is insensitive to image noise, as opposed to edge-based methods such as the Distance Regularized Level Set Evolution (DRLSE) method. The mean green pixel intensity of the background is computed for every frame and is used to acquire the illumination rectified signal $x_{rect}(t)$ as shown in eqns. 4 and 5.

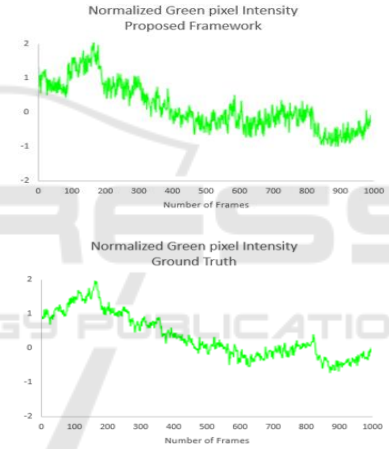

Figure 7: The curve on the left is a plot of the normalized illumination rectified signal while the curve on the right is the ground truth normalized mean green pixel intensity. The ground truth plot is obtained from the recording of the same scene in conditions without any illumination variations.

It can be seen that the illumination rectified signal that we have obtained is almost identical to the ground truth signal.

5 RESULTS

The mean RGB values in the ROI are calculated for each frame of the video. The ROI is then separated into its three constituent RGB channels and the mean value of each raw trace is computed. Each raw trace is then normalized in order to have zero mean and unit standard deviation. This helps us to achieve results

which are independent of the complexion of the subject since all three traces will now have the same statistical metrics.

We used FastICA 2.5 on Matlab 2015a to obtain the three source signals corresponding to each normalized traces. We found that the plethysmographic (PG) signal corresponding to the green trace has the most prominent peak in the frequency spectrum. Let the PG signals corresponding to the red, green and blue traces be $s_1(t)$, $s_2(t)$ *and* $s_3(t)$ respectively. For $i = 1,2,3$

$$s_i(t) = A^{-1}x_i(t), \quad (8)$$

where A is a 3x3 matrix. In order to find each source signal $s_i(t)$, ICA finds an approximation of A^{-1} that minimizes the gaussianity of each source signal.This ensures that the obtained source signals are statistically independent (Breuer and Major 1983).

Finally, The power spectra of the PG signals is obtained in order to determine the most prominent frequency component. Next, several temporal filters are applied to exclude frequencies outside the desired range. Since, the normal heart rate ranges from 40 bpm to 240 bpm, the desired frequency range was set to [0.6, 4] Hz. Finally, the heart rate of the video is calculated as $HR = 60f_{HR}$ bpm. The prominent peak in the power spectrum (1.27 Hz) corresponds to 76 bpm.

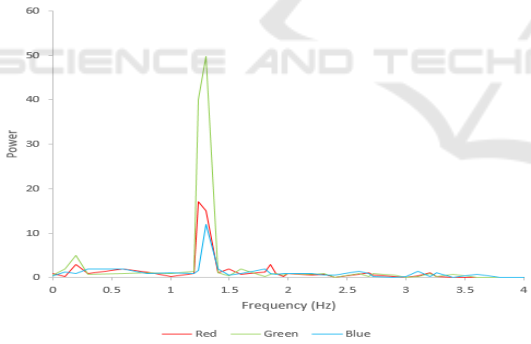

Figure 8: The power spectra displaying the source signals corresponding to the red, green and blue components of the video. The red, green and blue plots correspond to $s_1(t)$, $s_2(t)$ and $s_3(t)$ respectively.

6 PERFORMANCE ANALYSIS OF THE PROPOSED FRAMEWORK

6.1 Testing the Framework

We have devised an experiment to evaluate the robustness and accuracy of our framework in realistic conditions. We record the facial video of an individual for approximately 10 minutes while the individual is watching a horror movie. The settings we have installed are the same as those used for recording all previous videos (mentioned in Section 1). We have also monitored the heart rate by attaching a heart rate sensor to the user in order to compare our results with the ground truth results and measure the percentage error of our framework.

From the figure, it can be observed that the HR of the subject changes with time depending upon the scene that the subject is watching. To make it more interesting, we have chosen scenes which are likely to elicit a higher HR from the subject. It is clear from the plot that the HR measured by our framework is in agreement with the ground truth data, barring slight differences. The HR computed has a mean error percentage of 1.71%.

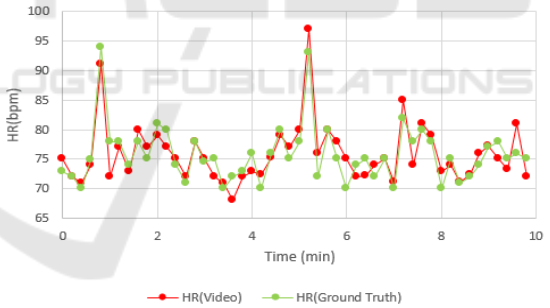

Figure 9: The heart rate of the subject while watching the horror movie 'The Conjuring'. Specific scenes from the movie have been chosen in order to obtain a higher HR and test the accuracy of our framework.

6.2 Datasets for Error Analysis

In order to ascertain the efficiency of our method, we implement our framework as well as previously proposed methods on two different sets of videos we have collected, since previous papers have not made their datasets public. The first dataset contains 18 videos having an average duration of 50-70 seconds, collected from a group of 18 subjects (9 male and 9 female). We have recorded videos with considerably high motion artifacts (head movements > 60 degrees),

but very few illumination interferences. For the second set of videos, we have used the MAHNOB-HCI database (Soleymani *et al.* 2012), in which we tested our framework on a set of 487 colour videos of 27 subjects (12 males and 15 females). The videos have a resolution of 780 x 580 pixels recorded at 61 FPS. As a result, our implementation of Li2014 and SAMC has a different result than the original papers, since both SAMC2016 and Li2014 were tested on all 527 videos of the MAHNOB-HCI dataset.

6.3 Determination of Mean Error and RMSE

A Polar H10 HR sensor is used to monitor and provide the ground truth HR. We have chosen four colour-based methods (Poh2011 (Poh *et al.* 2011), Kwon2012 (Kwon *et al.* 2012), Li2014 (Li *et al.* 2014) and Zhang2017) and one motion-based method (Balakrishnan2013 (Balakrishnan *et al.* 2013)) for comparison. Since every paper had different statistical measures to evaluate the performance of their model, we have employed four widely used metrics to measure the performance of each framework: the mean absolute error (μ_{error}), the root mean-square error (RMSE), percentage of Absolute error less than 5 bpm and the correlation coefficient r.

The results have been tabulated below. Table 1 illustrates the results that have been obtained on our dataset, which consists of challenging facial videos recorded with considerable motion artifacts, but no illumination variations.

Table 1: Performance of previously proposed frameworks on our database (extreme motion interferences without illumination variations).

Framework	μ_{error}	RMSE(%)	% Absolute Error<5bpm	r
Poh2011	9.1	15.9	56.4	0.39*
Kwon2012	8.8	15.6	46.1	0.15
Bala2013	11.5	18.2	33.2	0.06
Li2014	7.2	11.5	71.3	0.78*
Zhang2017	7.6	11.1	72.1	0.68*
SAMC2016	4.7	9.2	76.2	0.85*
Ours	**3.4**	**6.4**	**77.9**	**0.89***

* - Indicates that the correlation is significant at p = 0.01

From Table 1, it can be seen that under ideal illumination conditions, all of the proposed methods perform satisfactorily. The higher error rate in Balakrishnan2013 (here, Bala2013), a motion-based method, arises due to large head movements in the video. We have included extreme motion variations such as head rotations higher than 60 degrees. Since our framework is able to recover feature tracking points which are lost during the subject's head movement, our model outperforms other recently proposed models and is robust to head rotations of more than 60 degrees. As a result, our mean percentage error and RMSE are lower than Li2014 and SAMC2016.

Table 2: Performance on the MAHNOB-HCI dataset (consists of motion as well as illumination interferences).

Framework	μ_{error}	RMSE(%)	% Absolute Error<5bpm	r
Poh2011	13.5	21.1	46.2	0.32*
Kwon2012	24.1	25.5	41.2	0.21
Bala2013	16.1	22.9	39.1	0.26
Li2014	7.8	15	68.1	0.72*
Zhang2017	8.7	15.7	63.2	0.65*
SAMC2016	5.1	7.2	73.2	0.75*
Ours (i)+(iv)	28.7	32.2	33.8	0.41
Ours (i)+(ii)+(iii)+(iv)	**4.7**	**6.8**	**78.3**	**0.79***

* - Indicates that the correlation is significant at p = 0.01

Table 2 shows the results of implementing the above models on the MAHNOB-HCI dataset.
Here, (i), (ii), (ii) and (iv) represent Faster RCNN+feature tracking, recovery of feature tracking points, illumination rectification and temporal filtering respectively.

As it can be seen, our method provides the best results out of all the afore-mentioned algorithms. Our framework significantly outperforms Poh2011, Kwon2012 and Balakrishnan2013 since these models do not account for changes in illumination. Our model also outperforms Li2014, SAMC2016 and Zhang2017. This is due to the recovery of feature tracking points in our model, as a result of which the HR is accurately measured despite large head rotations by the subject. Moreover, Zhang2017 does not account for illumination variations as a result of which the accuracy is diminished on the MAHNOB-HCI dataset.

7 CONCLUSIONS

Over the years, there have been several methods to

extract HR from facial video such as motion based methods, color-based methods as well as approaches that have employed thermal imaging techniques. To the best of our knowledge, all of the previously proposed approaches have employed the Viola-Jones algorithm for face detection. We have deviated from this approach and employed Faster RCNNs for face detection. The Faster RCNN employed in the present study was able to detect faces without requiring the full frontal profile of the face, thus making it more robust. Secondly, depending upon nature of the background, the Viola-Jones algorithm may detect multiple ROIs, which may lead to confusion. This is not the case with our face detection algorithm, since it is independent of the background in the video.

An important feature of our framework is the ability to recover feature points which may have been lost during extreme head rotations. This makes our model robust to extreme motion artifacts and is able to measure HR even the subject performs a complete rotation (360 degrees). Next, while some of these papers have reduced the problem of head movements, all of them have a degradation in performance in the presence of illumination interferences. In our framework, we have accounted for this artifact by using RLS adaptive filtering methods and the local region-based active contour method (LRBAC) to segment the background and remove the noise signal in the video arising from changes in illumination.

We also performed an experiment where we monitored a subject while watching specific scenes of a horror movie for a period of 5-10 minutes, and extract the HR of the subject. The average HR of approximately every 20s is plotted and compared with the ground truth data. Upon comparison with the polar H10 HR monitoring sensor, we found that our framework achieved a mean error percentage of 1.71%.

Moreover, we also implemented previously proposed approaches on our database of 18 videos and found that our framework outperformed the previous four methods and attained a root mean-square error of 8.28% on the MAHNOB-HCI database.

One principle source of error might be the difference in sampling rates of the HR sensor and our webcam. While our webcam had a sampling rate close to 30 Hz, the H10 polar HR sensor had a higher sampling rate of 256 Hz. Also, in case of extremely low illumination where the face is not visible, it would be useful to combine motion and region-based methods to better solve motion and illumination interferences.

A direction for future research would be to focus on the integration of motion as well as colour-based methods to estimate HR. The complementary nature of these methods would enable a more robust approach to simultaneously tackle motion and illumination artifacts in the video.

REFERENCES

Balakrishnan, G., Durand, F., and Guttag, J., 2013. Detecting pulse from head motions in video. *In*: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 3430–3437.

Basri, R. and Jacobs, D.W., 2003. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 (2), 218–233.

Breuer, P. and Major, P., 1983. Central limit theorems for non-linear functionals of Gaussian fields. *Journal of Multivariate Analysis*, 13 (3), 425–441.

Cennini, G., Arguel, J., Akşit, K., and van Leest, A., 2010. Heart rate monitoring via remote photoplethysmography with motion artifacts reduction. *Optics Express*, 18 (5), 4867.

Garbey, M., Sun, N., Merla, A., and Pavlidis, I., 2007. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE Transactions on Biomedical Engineering*, 54 (8), 1418–1426.

Kwon, S., Kim, H., and Park, K.S., 2012. Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone. *In*: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. 2174–2177.

Lam, A. and Kuno, Y., 2015. Robust heart rate measurement from video using select random patches. *In*: *Proceedings of the IEEE International Conference on Computer Vision*. 3640–3648.

Lankton, S. and Tannenbaum, A., 2008. Localizing region-based active contours. *IEEE Transactions on Image Processing*, 17 (11), 2029–2039.

Li, X., Chen, J., Zhao, G., and Pietikäinen, M., 2014. Remote heart rate measurement from face videos under realistic situations. *In*: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 4264–4271.

Mohd, M.N.H., Kashima, M., Sato, K., and Watanabe, M., 2015. A non-invasive facial visual-infrared stereo vision based measurement as an alternative for physiological measurement. *In*: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 684–697.

Moreno, J., Ramos-Castro, J., Movellan, J., Parrado, E., Rodas, G., and Capdevila, L., 2015. Facial video-based photoplethysmography to detect HRV at rest. *International Journal of Sports Medicine*, 36 (6), 474–480.

Parra, E.J., 2007. Human pigmentation variation: evolution, genetic basis, and implications for public health. *American journal of physical anthropology*.

Poh, M.-Z., McDuff, D.J., and Picard, R.W., 2010. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18 (10), 10762.

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

Ren, S., He, K., Girshick, R., and Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (6), 1137–1149.

Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M., 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3 (1), 42–55.

Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J.F., and Sebe, N., 2016. Self-Adaptive Matrix Completion for Heart Rate Estimation from Face Videos under Realistic Conditions. *In*: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Verkruysse, W., Svaasand, L.O., and Nelson, J.S., 2008. Remote plethysmographic imaging using ambient light. *Optics Express*, 16 (26), 21434.

Yang, S., Luo, P., Loy, C.C., and Tang, X., 2016. WIDER FACE: A face detection benchmark. *In*: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 5525–5533.

Yu, Y.P., Kwan, B.H., Lim, C.L., Wong, S.L., and Raveendran, P., 2013. Video-based heart rate measurement using short-time Fourier transform. *In*: *2013 International Symposium on Intelligent Signal Processing and Communication Systems*. 704–707.

Zhang, C., Wu, X., Zhang, L., He, X., and Lv, Z., 2017. Simultaneous detection of blink and heart rate using multi-channel ICA from smart phone videos. *Biomedical Signal Processing and Control*, 33, 189–200.