Supervised Learning Approach to Remote Heart Rate Estimation from Facial Videos

Ahmed Osman , Jay Turcot and Rana El Kaliouby
Affectiva Inc.

***Abstract*— A supervised machine learning approach to remote video-based heart rate (HR) estimation is proposed. We demonstrate the possibility of training a discriminative statistical model to estimate the Blood Volume Pulse signal (BVP) from the human face using ambient light and any off-the-shelf webcam. The proposed algorithm is 120 times faster than state of the art approach and returns a confidence metric to evaluate the HR estimates plausibility. The algorithm was evaluated against the state-of-the-art on 120 minutes of face videos, the largest video-based heart rate evaluation to date. The evaluation results showed a 53% decrease in the Root Mean Squared Error (RMSE) compared to state-of-the-art.**

I. INTRODUCTION

The ubiquity of cameras across a myriad of consumer devices has led to an explosion in the number of recorded videos uploaded and shared on social networks and video sharing websites. This has in turn stimulated interest in the field of affective computing as well as the machine learning and computer vision research communities to parse this large video content as well as gauge people's engagement with this wealth of content. The majority of research is focused on analysing features that can be observed by the naked eye, such as facial expressions, head pose, age and gender. Recent work by Verkruysee et al., [3], Poh et al., [4] and [5] demonstrated the feasibility of measuring human heart rate (HR) and other physiological signals from facial videos.

Heart rate is a crucial vital sign that is regularly monitored across a diversity of domains from medical to audience research. HR is used to assess the cardiovascular system functions, e.g., resting HR is used as a predictive factor for cardiovascular diseases [7],[8]. In psychology, HR and HR derivatives, such as heart rate variability **(HRV)** are indicators of underlying human emotional state. Lang et al. [9] compared HRV over a short period of time to HRV over a larger period of time and concluded that HR can be a valid real time and continuous measure for both attention and arousal. Lakens [10] demonstrated that HR is a discriminative feature between participants who relived memories that elicit happiness and participants who relived memories that elicit anger.

Remote measurement of the HR signal with a camera is based on Photo-plethysmography **(PPG)** theory, a non-invasive mean of measuring the cardiovascular BVP signal by analysing the properties of reflected light from the skin [2], where blood absorbs more light than surrounding tissue and variations in the blood volume causes variations in the reflected light.

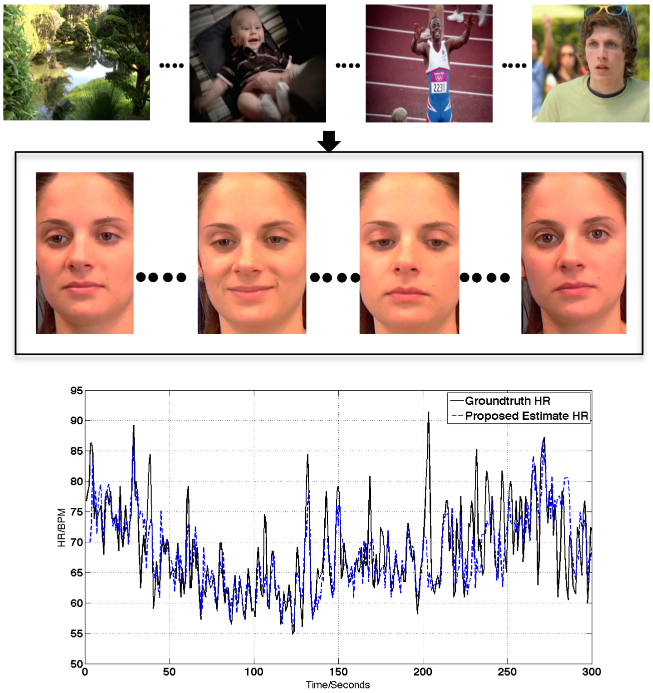


Fig. 1: Remote heart rate estimation from facial videos. Top row: Twenty participant watched 12.5 minutes of emotion-eliciting video content. Middle row: Facial video responses were recorded using an ordinary webcam, as well their BVP signals. Bottom row: Groundtruth HR calculated from the groundtruth BVP signal side by side with the estimate HR computed from the Face.

Remote measurement of HR is challenging due to a variety of factors that can add noisy variations in the reflected light from the face e.g. head motion and facial expressions. Current filtering techniques used in the literature add a significant processing delay while not being able to completely remove the noise. In this paper, we propose a novel approach to remote HR estimation using a webcam that accurately estimates the HR signal in close to real time speed.

The main contributions of this paper are:

1) A novel supervised machine learning approach to detect heart beats and estimate the BVP signal from face pixel intensity changes.
2) A confidence score to evaluate the reliability of the HR

estimates.

3) The total processing delay is 0.25 seconds compared to the 30 seconds delay in the current state of the art algorithm - this enables real-time applications of heart rate measurement.

Faster HR estimation will benefit interactive applications such as games that uses the HR signal to guide the game events, an interesting approach explored by the horror game "Nevermind" [11]. The confidence score benefits medical applications such as remote medical care services through video conferencing where a physician could measure the patient HR remotely [12].

The papers is organised as follows, in section II we describe related work in remote HR estimation. In section III we discuss the proposed algorithm pipeline. In section IV we evaluate the proposed algorithm accuracy against state of the art approach, followed by a discussion in section V. Finally the analysis conclusion is summarised in section VI.

II. Related Work

We start this section by defining terminology used through the paper. The time taken between two heart beats is known as the inter beat interval **(IBI)**, and the HR in beats per minute **(BPM)** at time t is the reciprocal of the IBI as shown in eq.1.

$$HR(t) = \frac{1}{IBI(t)} \times 60 \tag{1}$$

.

Pioneering work of Verkruysee et al. [3] investigated measuring BVP signal using ambient light by analysing the variability in the face pixels intensities in the green channel over a period of time as an estimate of the PPG signal and hence the underlying BVP signal. The use of the green channel is motivated by the fact that green light is absorbed by blood better than red and penetrates sufficiently deeper into the skin compared to blue light to probe the vasculature. Verkruysee et al. computed the Fourier transform of the mean face pixel intensities changes in the green channel over a time window and showed that the dominant frequencies were the BVP signal frequencies. The later approach relied on manual inspection of the Fourier transform spectrum and did not address the sources of noise such as motion artifacts that could add frequencies in the same range of the BVP signal frequencies.

The current state of the art approach is based on Poh et al. work [4],[5],[12], where a blind source separation algorithms is used to separate the PPG signal from the noise signals followed by a heuristic to automatically estimate HR from the filtered signal spectrum. The blind source separation algorithm used was *Independent Component Analysis* (ICA) [6], an iterative algorithm that infers the latent source signals from observation signals that are a linear mixture of the source signals. ICA was applied on the mean face pixels intensities changes in the red, green and blue **(RGB)** channels over a time window of length **30** seconds returning three ICA components, one of them is potentially the PPG signal. Spectral analysis was performed on the three ICA components followed by peak detection to identify the frequency with the highest power in the normal mammalian heart rate range (1 Hz - 4 Hz) in each of the three ICA components spectra. Recent work by Monkaseri et al.[14] replaces the latter heuristic by adopting a supervised machine learning model on features extracts from the three ICA components frequency spectra to predict an estimate heart rate.

The drawbacks of the pipeline proposed by Poh et al. are:

1) High computational cost due to computing new ICA basis and performing spectral analysis for every HR estimate returned.
2) Each HR estimate is returned after a 30 seconds delay.
3) If noise was a dominant frequency over a time window, it will manifest itself as a high power frequency in the frequency spectrum hence resulting in a wrong heart rate estimate.

We address these problem by proposing a new approach that uses a discriminative statistical model as a sliding window to detect individual heart beats from features extracted from the face pixels. The algorithms returns instantaneous HR estimates during noise free segments and conservative HR estimates during segments with high level of noise.

A. Data Collection and Baseline Comparison

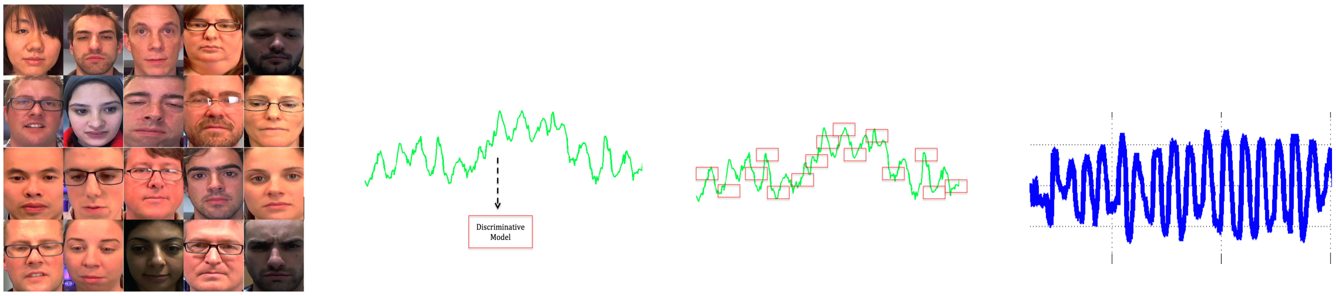
In data collection we recorded 20 participants from a diversity of racial backgrounds and age groups watching a 12.5 minutes video of emotion eliciting content as shown in fig.1, total duration of the collected data was 250 minutes. A Macbook Pro webcam was used to record the participants faces while simultaneously recording there BVP signal using a finger based BVP sensor with a sampling frequency of 128 Hz used as the groundtruth. The participants were not asked to sit still and hence the collected data contained spontaneous movements and facial expressions in response to the media content. The videos were encoded by an H264 encoder at an average bit rate of 2500 kb/s and a resolution of 640x480 at a variable frame rate ranging from 22 - 30 fps. Half of the collected data was used to train the discriminative model in the proposed algorithm and the second half was used for evaluation. The HR estimates from Poh et al. [4] were used as the baseline estimates in evaluation.

III. Methodology

The proposed approach is motivated by an intuition that intensity changes that are due to a gush of blood flow will have unique characteristics that could be learned empirically from the data. We explore the possibility of learning a non-linear decision surface that is discriminative enough between pixel intensity changes that are due to the BVP signal and intensity changes that are due to noise. The algorithm work flow is summarised in fig.2.

A. Face Videos Feature Extraction

The face is localised using a face tracker, followed by extracting the mean of the green channel from the face



(a) A face tracker is used to localise the Face ROI followed by computing the spatial average over the green channel ROI.

(b) The mean of the green channel from the face video is used to train a beat detector model.

(c) The trained model is applied as a sliding window on the mean of the green channel.

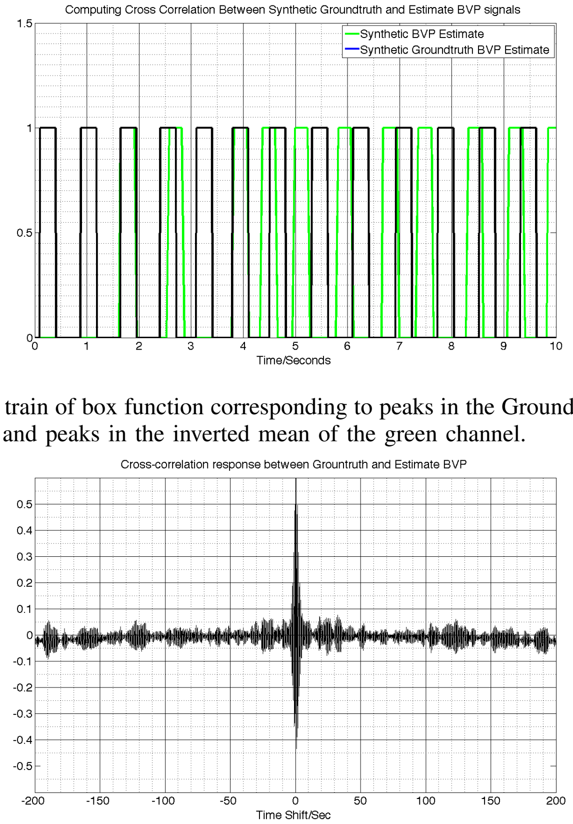
(d) The beat detector response is used to localise beat locations, where locals peaks are potential heart beat timestamps.

Fig. 2: Summary of the proposed pipeline for training a discriminative model to detect heart beats.

region of interest (ROI). The feature at time t is computed as the spatial average over the green channel ROI. The face tracker points are noisy and there subtle fluctuations can cause noisy shifts in the ROI bounding box **(BB)** introducing noise. Therefore the used BB location is only updated when the new location of the BB deviates significantly from it's current position.

B. Signal Synchronisation

The BVP groundtruth signal needs to be aligned with the features extracted from the mean of the green channel in order to successfully train the classifier to learn a one to one mapping between the two signals. The main source of alignment error is the phase difference due to the physiological delay between the time taken by the blood to travel to the finger tip where the BVP sensor is placed and the time taken to travel to the face region [1]. We attempted to align peaks in the groundtruth BVP with the minima in the mean of the green channel, since as the blood volume reaches a maximum in the face pixels the intensity of the reflect light reaches a minimum [3]. For our analysis the mean of the green channel signal was inverted so the peaks in the green channel correspond to peaks in the BVP signal. The alignment error was computed using cross-correlation between the groundtruth BVP and the inverted mean of the green channel. However because the difference in the physical nature of the two signals, the cross-correlation values can be affected by several latent factors, for example the mean of the green channel could change because of a change in the illumination conditions reducing correlation with the groundtruth BVP. A preprocesing step was introduced where two synthetic signals were created consisting of a train of box functions as shown in fig.3a where the box functions were centred around the groundtruth BVP peaks timestamps and the inverted mean of the green channel peaks timestamps, each box function width was set to **0.3** seconds. The cross-correlation was computed between the two synthetic signals as shown in fig.3b and the alignment error is estimated from the point of maximum correlation.



(a) A train of box function corresponding to peaks in the Groundtruth BVP and peaks in the inverted mean of the green channel.

(b) Cross-correlation response between the two synthetics Signals. Point of maximum correlation is used to align the two signals temporally.

Fig. 3: Computing the alignment error between synthetic mean of the green channel and synthetic BVP signal using cross-correlation.

C. Feature Representation

The feature representation used was a temporal representation of the mean of the green channel. A feature at time t corresponds to extracting a window of size w_s seconds centred at t from the mean of the green channel. The variable frame rate of the videos meant that different windows will contain a variable number of samples, therefore to achieve a fixed feature representation each window was discretized into n bins. The first order derivative of the discretized window was then computed since we are primarily interested in the variability in the mean of the green channel. The total processing delay introduced is half the window size $\frac{w_s}{2}$, since the discriminative model will not be applied for the window centred at time t until $t + \frac{w_s}{2}$.

D. Sample Selection for Training and Testing

Features for training and testing were extracted at the minima timestamps and half way between each two successive minima timestamps in the mean of the green channel. A feature extracted at time t is labelled as a positive example if a groundtruth BVP peak exists within $t_{tolerance}$ seconds, while a feature is labelled negative if it is at least $3t_{tolerance}$ seconds away from the nearest groundtruth peak. Features that are neither positive or negative are ignored as they resemble the intermediate state between a heart beating state and a heart resting state. The discriminative model used was a Support Vector Machine **(SVM)** with a Radial Basis Function (RBF) kernel. For testing we used the area under the Receiver Operator Characteristics (ROC) curve as our evaluation metric for the classifier performance.

E. BVP and HR Estimation

In run time the trained SVM model is used as a sliding window, where the timestamps of the local peaks in the model response are potential timestamps of heart beat locations as shown in fig.2d. The magnitude of the classifier output is used as a confidence metric to filter out peaks that are potentially false beats introduced by noise. A peak is considered as a confident peak if its magnitude is greater than a threshold $h_{threshold}$. HR values are estimated from the IBI calculated from the duration between two successive confident peaks in the classifier output. We guard against erroneous HR estimates by monitoring the moving average HR_{avg} computed using eq.2 with α set to 0.8. Invalid HR estimates are those estimates that deviate significantly from the moving average $\Delta HR(t)$ = | $HR(t) - HR_{avg}(t)$ | due to missing beats or false positive beats. A conservative heart rate estimate, $HR_{consv}(t)$, is computed from the raw mean of the green channel as a fall back mechanism discussed in section III-F when erroneous estimates are detected. When $\Delta HR(t)$ at time t is greater than the heart rate deviation threshold $\Delta HR_{threshold}$, synthetic heart beats are added to yield HR estimates equal to the value of $HR_{consv}(t)$ at time t.

HR signal is estimated from the IBI of the confident and synthetic beats. As a final post processing step we return a smoothed HR signal by computing the moving average given by eq.2 with α set **0.4**.

$$HR_{avg}(t) = \alpha HR_{avg}(t-1) + (1-\alpha)HR(t) \qquad (2)$$

F. Conservative HR Estimate

A conservative HR signal, $HR_{consv}(t)$, signal is computed from the raw mean of the green channel as a fall back mechanism when HR estimates start to deviate significantly from the moving average. The duration between the minima in the mean of the green channel is used as the IBI estimate to compute raw HR signal. The $HR_{consv}(t)$ is the moving average of the raw HR estimate from the mean of the green channel computed by eq.2 where α is set to 0.8

G. Confidence Score

Intuitively during segments with low levels of noise we expect during a time window w that the total number of detected confident peaks **(NC)** to be greater than the total number of synthetic beats added **(NS)**. The proposed confidence score is measured by calculating the percentage of the number of confident heart beats from the total number of beats detected over a period of time w given by eq.3.

$$Confidence\_Score(w) = \frac{NC}{NC + NS} \qquad (3)$$

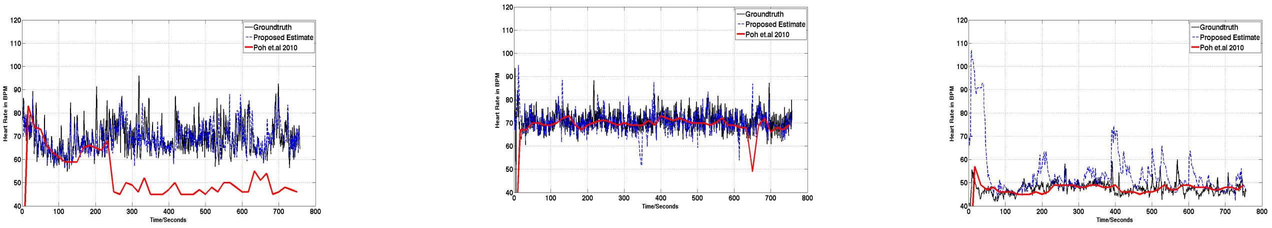
IV. Evaluation

Discriminative model was trained on 5000 positive and 5000 negative examples extracted from 10 sessions, while testing was performed on the features extracted from the remaining 10 sessions which contained 9198 positive examples and 18246 negative examples.

We used a window of size $w_s = 0.5$ seconds and divided it into $n = 10$ bins. The maximum deviation from the running average heart rate allowed was $\Delta HR_{threshold} = 20$ BPM. The threshold used for trained model output peak confidence $h_{threshold} = 0$. Grountruth BVP signal was smoothed prior to peak detection where peak timestamps correspond to the groundtruth heart beats. The IBI was computed form the duration between two successive groundtruth peaks and was used to compute the groundtruth HR by eq.1.

A. Results

The mean of the green channel was synchronised with the groundtruth BVP signal before training the SVM model, the average alignment error was **1.15 seconds**. The output ROC curve from the trained model evaluated on the test set is presented in fig.6 with an area under the curve of **0.93**. The trained model was applied as a sliding window on the test videos as discussed in section III-E. A sample plot of the groundtruth BVP signal and the classifier sliding window response side by side with the mean of the green channel is shown in fig 5. The plot shows almost a one to one correspondence between peaks in the classifier response and peaks in the BVP groundtruth signal. Moreover the classifier response decreased and was almost flat at $t = 2$ and $t = 25$



(a) Proposed estimates have a stronger agreement with the groundtruth HR.

(b) HR estimates from the proposed algorithm and Poh et al. have a strong agreement with the groundtruth.

(c) Poh et al. HR estimates have a stronger agreement with the groundtruth.

Fig. 8: Side by side plots of the HR estimates from the proposed algorithm and Poh et al. and the groundtruth HR.

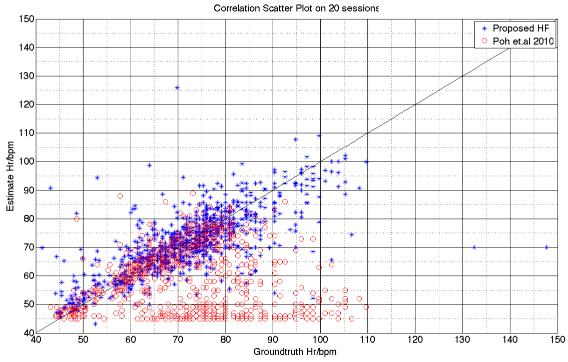


Fig. 4: Correlation Scatter plot for hr estimates produced by Poh et al. and the proposed algorithm.

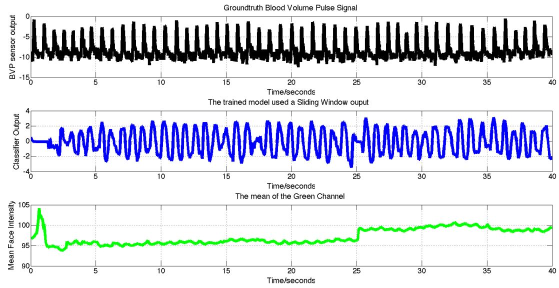


Fig. 5: Example Side by Side plot of the groundtruth BVP signal, the beat detector sliding window response and the corresponding mean of the green channel.

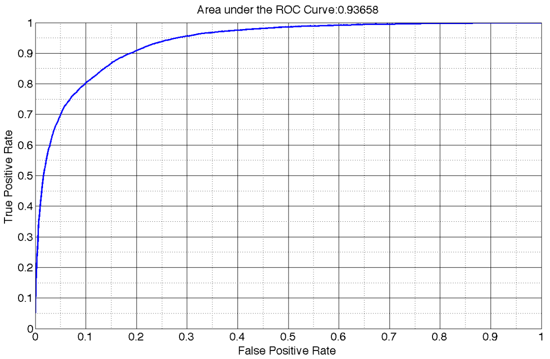


Fig. 6: Beat Detection classifier ROC curve.

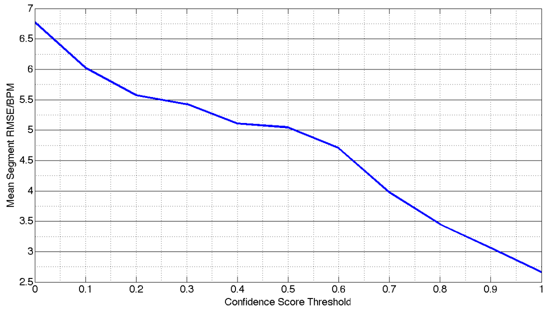


Fig. 7: The plot is generated by varying a confidence score threshold (x-axis) and computing the average RMSE across 1 minute segments that have a confidence score greater than or equal to the threshold (y-axis).

when there was a sudden change in the mean of the green channel due to motion artifacts. Each test video was divided into 1 minute segments, a total of 13 segments per video, and the confidence score and the RMSE for each segment was computed. Fig.7 presents the relationship between confident score and corresponding average segment RMSE. As the confidence score increased the corresponding average RMSE per segment decreased.

The output heart rate estimates from the proposed algorithm are compared against Poh et al. heart estimates. Example HR estimate plots for multiple sessions are shown in fig.8. Poh et al. algorithm was applied every 30 seconds on the test face videos the common HR samples between the groundtruth HR signal and HR estimate from Poh et al. and HR estimates from the proposed algorithm that share a common timestamp were selected for evaluation. The summary evaluation metrics used were RMSE, Pearson Correlation Coefficient and the Mean Error and Error standard deviation (Error STD). The output results are summarised in table I. The correlation scatter plot in fig.4 shows estimates from Poh et al. and the proposed algorithm against the groundtruth HR values. The HR estimates from the proposed algorithm were mainly clustered along the diagonal line with few outlier points off the diagonal. Estimates from Poh et al. were clustered around the diagonal and a second cluster below the diagonal line where the groudtruth HR is consistently under estimated.

	Training Data		Testing Data	
	Proposed Est.	Poh et al.	Proposed Est.	Poh et al.
RMSE	8.12	33.54	9.52	19.94
Correlation	0.77	-0.162	0.65	0.153
Mean Error	0.38	25.96	-0.74	9.79
Error STD.	8.13	21.29	9.50	17.38

TABLE I: Evaluation results of the proposed algorithm estimates against Poh et al.

V. Discussion

The proposed algorithm estimates the underlying BVP signal by detecting the individual heart beats. The time delay for a window centred at time t is equal to half the window size. For a window of size $w_s = 0.5$ seconds the delay is 0.25 seconds, while the processing delay in Poh et al. algorithm is 30 seconds, hence the proposed algorithm decreased the processing delay 120 times. The short delay allowed faster estimation of the HR beats and therefore faster estimation of the instantaneous HR. However the proposed estimates are prone to error since HR calculation is based on detecting individual HR beats, a single false positive beat or a missed beat will introduce significant error. It was crucial to introduce a confidence metric to filter out false heart beats complemented with strategies to guard against significant changes in the HR estimates as discussed in section III-E. The second challenge in estimating HR from individual beats is the heart beat timestamp localisation accuracy. The time window limits the localisation accuracy to be within a ”bin width” seconds from the groundtruth. For example the experiments above used a window of size $w_s = 0.5$ seconds divided into 10 bins, the size of each bin was $b_s = 0.05$ seconds, therefore the timestamp of a detected peak in the classifier output will at least be approximately within **0.05** seconds from the groundtruth beat timestamp. If the groundtruth IBI between two groundtruth beats was **1.16** seconds (HR=70 BPM) and the corresponding IBI was overestimated by just **0.05**, will result in an error of magnitude **3.85 BPM** and underestimating the IBI by **0.05** seconds results in an error of magnitude **4.33 BPM**. Therefore it was crucial to have a final post processing step to smooth the estimated HR signal by computing the moving average described in sec.III-E to reduce errors due to beat timestamp localisation.

The accuracy of the HR estimates measured remotely can be affected by multiple factors such as motion artifacts, skin pixels occlusion, face pixels saturation, lack of illumination, noise from the video encoder and many more. It is infeasible to predict invalid HR estimates by attempting to detect all the scenarios that add noise to the mean of the green channel. The proposed confidence score in section III-G provides a convenient metric to autonomously assess the reliability of the HR estimates.

The supervised approach allows in future research to explore extending the proposed feature space with more predictive features to complement the information in the mean of the green channel. Recently Balakrishnan et al. [13] showed that each heart beat causes subtle head motion as a result of the wave of blood flow from the heart to the brain. Head motion can complement the mean of the green channel in scenarios where the skin pixels are occluded, for example by a layer of cosmetics or blood due to an injury. Future direction of research could also explore statistical models that encode the periodic nature of the BVP signal to complement the beat detection process.

VI. Conclusion

The HR signal is a vital sign monitored across a diversity of domains. The ubiquity of cameras provide the necessary infrastructure to measure HR remotely by analysing the subtle changes in the face pixel intensities. In this work we demonstrated the feasibility of training a discriminative model to detect heart beats that are used to accurately estimate the underlying HR signal. The proposed algorithm is faster than state of the art and returns a confidence score to evaluate the reliability of the returned HR estimates. Successful estimation of HR from the mean of the green channel reflects the strong presence of the PPG signal. Accurate estimation of HR was made possible by using a a discriminative model to return instantaneous HR during noise free segment and conservative HR during high level noise segments.

References

[1] Kushki, Azadeh, et al. ”Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites.” Physiological measurement 32.10 (2011): 1529.

[2] Hertzman, A. B., and C. R. Spealman. ”Observations on the finger volume pulse recorded photoelectrically.” Am. J. Physiol 119.334 (1937): 3.

[3] Verkruysse, Wim, Lars O. Svaasand, and J. Stuart Nelson. ”Remote plethysmographic imaging using ambient light.” Optics express 16.26 (2008): 21434-21445.

[4] Poh, Ming-Zher, Daniel J. McDuff, and Rosalind W. Picard. ”Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.” Optics Express 18.10 (2010): 10762-10774.

[5] Poh, Ming-Zher, Daniel J. McDuff, and Rosalind W. Picard. ”Advancements in noncontact, multiparameter physiological measurements using a webcam.” Biomedical Engineering, IEEE Transactions on 58.1 (2011): 7-11.

[6] Hyvrinen, Aapo, Juha Karhunen, and Erkki Oja. Independent component analysis. Vol. 46. John Wiley Sons, 2004.

[7] Cook, Stphane, et al. ”High heart rate: a cardiovascular risk factor?.” European heart journal 27.20 (2006): 2387-2393.

[8] Fox, Kim, et al. ”Resting heart rate in cardiovascular disease.” Journal of the American College of Cardiology 50.9 (2007): 823-830.

[9] Lang, Annie. ”Involuntary attention and physiological arousal evoked by structural features and emotional content in TV commercials.” Communication Research 17.3 (1990): 275-299.

[10] Lakens, Daniel. ”Using a Smartphone to measure heart rate changes during relived happiness and anger.” IEEE Transactions on Affective Computing (2013): 1.

[11] ”Nevermind Game.” Nevermind. Web. 28 Sept. 2014. http://www.nevermindgame.com/.

[12] Poh, Ming-Zher, Daniel McDuff, and Rosalind Picard. ”A medical mirror for non-contact health monitoring.” ACM SIGGRAPH 2011 Emerging Technologies. ACM, 2011.

[13] Balakrishnan, Guha, Fredo Durand, and John Guttag. ”Detecting pulse from head motions in video.” Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013.

[14] Monkaresi, Hamed, R. Calvo, and Hong Yan. ”A Machine Learning Approach to Improve Contactless Heart Rate Monitoring Using a Webcam.” (2013): 1-1.