A Maximum Likelihood Formulation To Exploit Heart Rate Variability for Robust Heart Rate Estimation From Facial Video

Raseena K T[1] and Prasanta Kumar Ghosh[2]

Abstract— **The problem of estimating the heart rate (HR) from a facial video is considered. A typical approach for this problem is to use independent component analysis (ICA) on the red, blue, green intensity profiles averaged over the facial region. This provides estimates of the underlying source signals, whose spectral peaks are used to predict HR in every analysis window. In this work, we propose a maximum likelihood formulation to optimally select a source signal in each window such that the predicted HR trajectory not only corresponds to the most likely spectral peaks but also ensures a realistic HR variability (HRV) across analysis windows. The likelihood function is efficiently optimized using dynamic programming in a manner similar to Viterbi decoding. The proposed scheme for HR estimation is denoted by vICA. The performance of vICA is compared with a typical ICA approach as well as a recently proposed sparse spectral peak tracking (SSPT) technique that ensures that the predicted HR does not vary drastically across analysis windows. Experiments are performed in a five fold setup using facial videos of 15 subjects recorded using two types of smartphones (Samsung Galaxy and iPhone) at three different distances (6inches, 1foot, 2feet) between the phone camera and the subject. Mean absolute error (MAE) between the original and predicted HR reveals that the proposed vICA scheme performs better than the best of the baseline schemes, namely SSPT by -8.69%, 52.77% and 8.00% when Samsung Galaxy phone was used at a distance of 6inches, 1foot, and 2feet respectively. These improvements are 12.13%, 13.59% and 18.34% when iPhone was used. This, in turn, suggests that the HR predicted from a facial video becomes more accurate when the smoothness of HRV is utilized in predicting the HR trajectory as done in the proposed vICA.**

I. INTRODUCTION

Heart Rate (HR) is an important index of a person's health and well-being. HR has conventionally been obtained from the electrocardiograph (ECG) [1], which is contact based and requires several electrodes to be attached to the skin. On the othr hand, for non-contact monitoring of HR, the photoplethysmography (PPG) is well-known as it provides blood volume pulse [2]. In PPG, dedicated source of light is illuminated on subject's skin and the amount of light reflected is measured. Imaging photoplethysmography (iPPG), is a variation of PPG in which the dedicated source of light is replaced by a camera in ambient light conditions. In other words, the HR is estimated from the video of a subject's face captured by the camera. Although iPPG signal can be obtained from various parts of the body, face is preferred since it is easily accessible. It has been shown that iPPG can provide a low-cost and comfortable way to capture human vital signs from the facial video captured by ordinary cameras in the ambient light conditions [3].

In the work by Poh et al. [4], the HR is estimated from facial video using the significant peak in the second source signal obtained after performing independent component analysis (ICA) on the average RGB intensity profiles in the facial video. There have been several attempts to improve upon it in a number of ways including selection of specific regions of the face [5], illumination rectification, non-rigid motion removal [6]. Attempts have been made to use K Nearest Neighbour classifier and linear regression step in addition to ICA [7], as well as Support Vector Machine [8] to improve the HR prediction from the facial videos. Takano et al. [9] performed a first order derivative and autoregressive spectral analysis on facial videos. In all the above works, HR is estimated independently in each analysis window of the data, thus making it less robust to artifacts in the video. Gaonkar et al. [10] proposed sparse spectral peak tracking (SSPT) to address such drawback. In SSPT, a sparse representation of the spectrum of each source signal is obtained using the top few significant peaks and these are used for HR estimation by exploiting the slow-varying nature of HR.

In this work, we propose a maximum likelihood (ML) formulation for HR prediction by exploiting the heart rate variability (HRV) across analysis windows. The proposed formulation optimally selects a source signal in each window such that the predicted HR trajectory not only corresponds to the most likely spectral peaks but also ensures smoothness in the HRV across analysis windows. The likelihood function is efficiently solved using dynamic programming. Experiments with facial videos from 15 subjects captured by two types of phone cameras and three different camera-to-subject distances reveal that the mean absolute error in the HR predicted using the proposed scheme is lower than that using the best of the baseline schemes by ∼16.02% averaged across all recording conditions indicating the benefit of the proposed scheme. We begin with the description of the proposed ML formulation.

II. PROPOSED MAXIMUM LIKELIHOOD FORMULATION FOR HR ESTIMATION

The steps involved in estimating HR from facial video are summarized in Fig. 1. At first, the facial region of interest (ROI) is identified for all the frames in the video. In this work, we follow the ROI as used in the work by Gaonkar et al. [10]. The pixel intensities over the facial ROI are averaged and temporal contours of the average intensity in the red (R),

[1] Raseena K T is with Department of Electrical Engineering, Indian Institute of Science, Bangalore, India, raseenakt1994@gmail.com

[2] Prasanta Kumar Ghosh is with Faculty of the Department of Electrical Engineering, Indian Institute of Science, Bangalore, India prasantg@iisc.ac.in

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

green (G) and blue (B) channels are obtained. Following this, the ICA [11] is performed on the RGB intensity contours and the underlying source signals are extracted. The spectra of all source signals are analysed to estimate HR from each underlying signal. The proposed approach exploits the smooth nature in the HRV for estimating the HR from the spectra of the source signals through a ML based formulation which is described below:

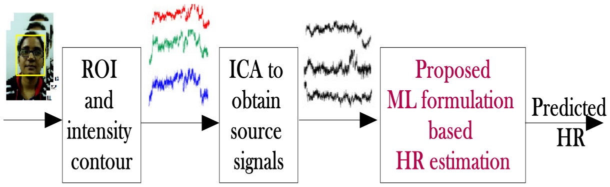
ROI and intensity contour
ICA to obtain source signals
Proposed ML formulation based HR estimation
Predicted HR

Fig. 1. Block diagram describing major steps in the proposed vICA scheme

At first, the confidence value of HR estimated from each source signal is calculated as: Let $\mathbf{c}_k = [c_1^k, c_2^k, c_3^k]^T$ be the HR estimated from the first, second and third source signal respectively in the k-th window. Let $M_j^k,\ j = 1,2,3$ be the spectral magnitude corresponding to c_j^k. And let $S_j^k, j = 1,2,3$ be the sum of magnitude squares of all spectral components at frequencies in between 0.75Hz and 4Hz (frequencies corresponding to normal HR range of 45 bpm to 240 bpm) of j th source signal in k-th window. Then, the confidence value corresponding to c_j^k is given by α_j^k.

$$\alpha_j^k = \frac{(M_j^k)^2}{S_j^k} \tag{1}$$

Let H_k denote the index of the true source signal from which HR is estimated in k-th window. That is $H_k = m, m \in \{1,2,3\}$. We assume $P(\ H_k = m) = \frac{1}{3},\ \forall m$. The probability P_m^k = Prob($H_k = m|\mathbf{c}_k$), $m = 1,2,3$ is calculated from confidence values.

$$P_m^k = P(\ H_k = m|\mathbf{c}_k) \propto p(\mathbf{c}_k|\ H_k = m) \tag{2}$$

$$\triangleq \frac{\text{confidence of } c_m^k}{\text{Sum of confidences of 3 components of } \mathbf{c}_k} \tag{3}$$

$$= \frac{\alpha_m^k}{\sum_{m=1}^{3} \alpha_m^k} \tag{4}$$

We consider maximizing the joint probability of the HR of a participant in all windows instead of maximizing the probability in each window separately. Let K be the total number of windows.

$$P(\ H_1,\ H_2, ...,\ H_K|\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K) \propto$$
$$P(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K|\ H_1,\ H_2, ...,\ H_K) P(\ H_1,\ H_2, ...,\ H_K) \tag{5}$$

We assume that given the HR in all K windows, $\mathbf{c}_k$ in these windows are independent.Thus

$$P(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K|\ H_1,\ H_2, ...,\ H_K) = \prod_{k=1}^{K} p(\mathbf{c}_k|\ H_k) \tag{6}$$

which can be obtained from (2), (3), and (4).

We assume that HR in the k-th window is conditionally independent of the $(k-2)$-th window and the ones before given the HR of the $(k-1)$-th window. Then, using the chain rule of probability, we can write $P(\ H_1,\ H_2, ...,\ H_K)$

$$= P(\ H_k|\ H_1,\ H_2, ...,\ H_{k-1}) P(\ H_1,\ H_2, ...,\ H_{k-1}) \tag{7}$$

$$= P(\ H_k|\ H_{k-1}) P(\ H_1,\ H_2, ...,\ H_{k-1}) = ... \tag{8}$$

$$= P(\ H_1) \prod_{k=2}^{K} P(\ H_k|\ H_{k-1}) \propto \prod_{k=2}^{K} P(\ H_k|\ H_{k-1}) \tag{9}$$

$P(\ H_k|\ H_{k-1})$ is the probability of change of HR from H_{k-1} in the $(k-1)$-th window to H_k in the k-th window. We define $P(\ H_k = m|\ H_{k-1} = m')$ to be

$$P(\ H_k = m|\ H_{k-1} = m') \propto \exp^{\left(-\lambda(c_m^k - c_{m'}^{(k-1)})^2\right)} \tag{10}$$

where λ is a hyper parameter. This is to ensure that HR does not change drastically in consecutive windows and change slowly similar to realistic HR variation. Combining (6) and (9) ,

$$P(\ H_1,\ H_2, ...,\ H_K|\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K) \propto$$
$$\prod_{k=1}^{K} p(\mathbf{c}_k|\ H_k) \prod_{k=2}^{K} P(\ H_k|\ H_{k-1}) \tag{11}$$

The HR sequence is obtained by maximizing the probability in (11) as follows:

$$\{\ \hat{H}_k, \forall k\} = \underset{H_1, ...,\ H_K}{\operatorname{argmax}}\ P(\ H_1,\ H_2, ...,\ H_K|\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K) \tag{12}$$

The optimization is solved using dynamic programming algorithm in a manner similar to Viterbi decoding [12]. Hence the proposed scheme is denoted by vICA. The algorithm is as follows: Let $D_m(k)$ be the probability of estimating k heart rates for k windows assuming estimate for k-th window is from m-th underlying source signal. Let the back-tracking pointer in dynamic programming be denoted by $\xi_m(k)$, which underlying signal estimate for $(k-1)$-th window gave maximum probability $D_m(k)$. $D_m(k)$ is computed in a recursive manner and $\xi_m(k)$ is stored in each recursion of the dynamic programming as follows:

1) Initialization: Compute $D_m(1) = P_m^1$
2) Iteration: For $2 \leq k \leq K$ and $1 \leq m \leq 3$, compute the following:

$$D_m(k) = \max_{1 \leq m' \leq 3} \{D_{m'}(k-1) \times \beta(m, m')\} P_m^k$$

$$\xi_m(k) = \underset{1 \leq m' \leq 3}{\operatorname{argmax}} \{D_{m'}(k-1) \times \beta(m, m')\}$$

where $\beta(m, m') = P(\ H_k = m|\ H_{k-1} = m')$ which is the transition probability given by (10).

3) Backtracking:

$$\hat{H}_K = \underset{1 \leq m' \leq 3}{\operatorname{argmax}}\ D_{m'}(K)$$

$$\hat{H}_k = \xi_{\ \hat{H}_{k+1}}(k+1), k = K-1, K-2, ..., 1.$$

$c_{\hat{H}_k}^k$, $1 \leq k \leq K$ are declared as predicted HR.

III. Experiments and results

A. Description of test data

The FAVIP corpus [10] is used for the experiments in this work. It comprises facial video recordings of one-minute duration taken from each of 15 subjects, of whom 12 are males and 3 are females of varying skin complexions with their age ranging over 23-45 years. The facial videos in FAVIP corpus were captured by Samsung Galaxy (S3) smartphone and iPhone (3GS) at three different distances between the camera and the subject, namely, 6 inches, 1 foot and 2 feet. S3 videos are in mp4 format and 3GS videos are in avi format. Both have a frame rate of 30 frames/second with a resolution 1280×720 pixels. The encoding scheme of S3 is h264(baseline) and that of 3GS is MJPEG. The database also provides the actual heartbeat rate of the subject obtained from the pulse oximeter.

B. Experimental setup

Each one minute video is analysed using 30 seconds window with 1 second shift. HR candidate, c_j^k, (as described in Section II) is obtained by finding the frequency corresponding to the maximum peak amplitude of the spectrum of the j-th underlying source signal in the range 0.75Hz to 4Hz. In each analysis window, a HR value is predicted using vICA. The experiment is carried out in five folds each having a training phase and testing phase. In training phase, the Mean absolute error (MAE) between the ground truth HR and the one estimated using vICA is calculated for three randomly chosen subjects for various values of the hyper parameter λ in Eq. (10). λ is varied from 0 to 0.1 with a step of 0.001. The value of λ which results in the least error is used for evaluation of rest of the subjects. All 15 subjects appear once in the training phase across all folds. In particular, for the five folds, the following subjects are used in the training phase: fold#1 - (10,14,15), fold#2 - (11,12,13), fold#3 - (1,5,7), fold#4 - (2,8,9), fold#5 - (3,4,6). Such an experimental setup is chosen to examine the generalizability of the hyper parameter λ on unseen test cases. If the λ is subject specific and varies significantly across subjects, the proposed vICA would fail to estimate HR for subjects on which the λ has not been optimized.

The HR estimated using vICA is compared with that obtained using two existing methods, namely ICA [4] and SSPT based HR estimation [10]. In the former method, HR is estimated always from the second source signal obtained after performing ICA as proposed by Poh et al [4]. That predicted HR is the frequency corresponding to the maximum peak amplitude of the spectrum of the second source signal in the range 0.75Hz to 4Hz. For SSPT, the best choice of the window is determined on the training set in each fold.

The HR is estimated using vICA, SSPT and ICA for videos obtained using Samsung Galaxy and iPhone separately. In order to examine the accuracy of the estimated HR with varying distance between the camera and the subject, the HR is estimated separately for three different distances. Thus, the five fold experiment is conducted separately for six recording conditions (S3 and 3GS each placed at three different distances).

The performance of HR estimation is evaluated by the MAE between the predicted HR and the ground truth HR in all analysis windows in each test video. The ground truth HR for each window is calculated as the average of HR values in the respective window.

C. Results

The optimized values of λ in different folds are summarized in Table I. It is clear from the table that the optimal value of λ varies across six recording conditions and five folds. However, the optimal λ value is found to be greater than 0.05 in three among thirty cases (six recording conditions × five folds) suggesting that a small value of λ is usually a better choice than its high value in most of the cases. Table II, III, IV show the MAE between the estimated and the ground truth HR in all folds for three different distances 6 inches, 1 foot, and 2 feet respectively. The tables also show the average and standard deviation (SD) of the MAE across all folds for all three schemes in case of both S3 and 3GS. The bold entries indicate the minimum value across three schemes (vICA, SSPT and ICA) in every recording condition. It is clear from the tables that the proposed vICA has the least average MAE (followed by the SSPT) among three schemes considered for both S3 and 3GS for all three distances except for S3 at a distance of 6 inches, where SSPT achieves a lower MAE over that of vICA.

The average MAE using vICA drops compared to that using SSPT by -8.69%, 52.77% and 8.00% when S3 is used at a distance of 6inches, 1foot, and 2feet respectively. These drops are 12.13%, 13.59% and 18.34% when 3GS is used. It is interesting to note that the performance of vICA and SSPT are more similar and consistent across different distances when videos using 3GS are considered compared to when videos using S3 are considered. In fact, the percentage drop in MAE from SSPT to vICA in the case of S3 videos varies drastically with three distances. This could be because of the differences in the encoding nature of H.264 in S3 and MJPEG in 3GS. It is known that MJPEG only compresses individual frames of video, while H.264 compresses across frames. Thus, the H.264 may distort the HR related information in the pixel intensities to a greater degree compared to that of MJPEG. From Table III and IV, it is clear that the average MAE for both vICA and SSPT increases when a distance of 2feet is considered compared to that in the case of 1feet indicating that the HR prediction performance drops as the distance between the camera and the subject increases.

vICA and SSPT both use temporal smoothness but their performance vary across subjects. To illustrate this, we present predicted HR on subject where vICA does better than SSPT and vice versa in Fig. 2. It is seen from the figure that vICA performs better on both the phones in case of subject15. However, in the case of subject4, both vICA and SSPT are far from the ground truth value initially, with SSPT prediction being relatively better. This could be due to noisy spectral peaks in the initial part of the recording. But as analysis window position increases, both schemes select the right peaks in case of 3GS whereas in S3, the SSPT fails to do so.

It should be noted that λ=0 in the proposed formulation would correspond to the case where no temporal smoothness is imposed on the predicted HR trajectory. We found that the MAE increases by 18.41%, 89.21% and 25.41% for three distance cases using S3 when λ=0 is used compared to using optimized λ. These increments in MAE are 0.8%, 53.44% and 15.49% when 3GS is used. Increase in MAE due to λ=0 over optimized λ suggests that the predicted HR trajectory becomes more accurate when the smooth nature of the HRV is considered in the proposed formulation.

TABLE I

Optimized value ofλ in different folds

Distance	Phone	Fold1	Fold2	Fold3	Fold4	Fold5
6 inches	S3	0.004	0.001	0.009	0	0.02
	3GS	0.003	0	0.001	0.027	0
1 foot	S3	0.063	0.004	0.025	0.03	0.007
	3GS	0.001	0.046	0.001	0.007	0.002
2 feet	S3	0.041	0.04	0.008	0	0.018
	3GS	0.016	0.005	0.053	0.093	0.001

TABLE II

MAE in bpm for different methods and phones for 6 inches

Fold	S3			3GS		
	vICA	SSPT	ICA	v ICA	SSPT	ICA
1	3.10	**3.01**	11.99	**3.46**	4.75	11.76
2	**4.64**	4.77	12.57	4.72	**4.08**	13.59
3	4.72	**4.5**	12.375	**4.71**	5.89	11.26
4	6.70	**5.43**	14.8	**5.38**	6.01	14.98
5	4.73	**4.27**	12.49	**4.69**	5.4	13.88
avg.	4.77	**4.39**	12.84	**4.59**	5.22	13.09
SD	1.27	**0.88**	1.11	**0.69**	0.81	1.54

TABLE III

MAE in bpm for different methods and phones for 1 foot

Fold	S3			3GS		
	vICA	SSPT	ICA	v ICA	SSPT	ICA
1	**2.54**	4.76	9.45	2.49	**2.38**	15.36
2	**1.91**	5.15	13.55	**2.09**	2.77	14.17
3	**2.91**	5.55	15.09	**3.07**	3.53	15.32
4	**2.67**	5.84	15.47	**3.02**	3.58	16.47
5	**2.89**	6.06	14.68	**2.99**	3.55	15.2
avg.	**2.58**	5.47	13.65	**2.73**	3.16	15.3
SD	**0.40**	0.52	2.45	**0.42**	0.55	0.81

TABLE IV

MAE in bpm for different methods and phones for 2 feet

Fold	S3			3GS		
	vICA	SSPT	ICA	v ICA	SSPT	ICA
1	7.2	**5.65**	21.63	**4.21**	4.24	23.02
2	**5.08**	9.75	22.39	**7.57**	10.82	25.68
3	**9.72**	11.55	22.04	**8.97**	10.49	27.11
4	13.7	**11.28**	26.99	**8.92**	10.91	22.67
5	**9.94**	11.38	25.33	**9.57**	11.59	27.02
avg.	**9.12**	9.92	23.67	**7.84**	9.61	25.1
SD	3.24	**2.49**	2.35	**2.16**	3.02	2.13

IV. Conclusions

We propose a ML formulation for estimating HR from facial video by exploiting the smooth nature of HRV and show that it performs better than when the smoothness in HRV is not exploited. The proposed scheme also performs better than the best baseline scheme which performs spectral peak tracking for HR estimation. While the proposed scheme works on all the analysis windows together in the entire video, it can be made real time by solving the optimization in Eq. (12) in a manner similar to the online Viterbi decoder. This is part of our future work.

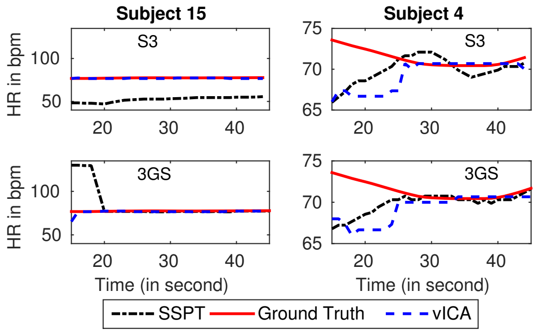

Fig. 2. Illustration of original and estimated HR using vICA and SSPT for distance of 1 foot for subject15 and subject4.

References

[1] Y. Sun, S. Hu, V. Azorin-Peris, R. Kalawsky, and S. Greenwald, “Non-contact imaging photoplethysmography to effectively access pulse rate variability,” *Journal of biomedical optics*, vol. 18, no. 6, pp. 061 205–061 205, 2013.

[2] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Advancements in non-contact, multiparameter physiological measurements using a webcam,” *IEEE transactions on biomedical engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[3] D. McDuff, S. Gontarek, and R. W. Picard, “Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 12, pp. 2948–2954, 2014.

[4] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.” *Optics express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[5] T. Kitajima, S. Choi, and E. A. Y. Murakami, “Heart rate estimation based on camera image,” in *Intelligent Systems Design and Applications (ISDA), 2014 14th International Conference on*. IEEE, 2014, pp. 50–55.

[6] X. Li, J. Chen, G. Zhao, and M. Pietikainen, “Remote heart rate measurement from face videos under realistic situations,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 4264–4271.

[7] H. Monkaresi, R. A. Calvo, and H. Yan, “A machine learning approach to improve contactless heart rate monitoring using a webcam,” *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1153–1160, 2014.

[8] A. Osman, J. Turcot, and R. El Kaliouby, “Supervised learning approach to remote heart rate estimation from facial videos,” in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–6.

[9] C. Takano and Y. Ohta, “Heart rate measurement based on a time-lapse image,” *Medical Engineering and Physics*, vol. 29, no. 8, pp. 853–857, 2007.

[10] P. A. Gaonkar, R. Bhuthesh, D. Gope, and P. K. Ghosh, “Robust real-time pulse rate estimation from facial video using sparse spectral peak tracking,” in *Signal Processing and Communications (SPCOM), 2016 International Conference on*. IEEE, 2016, pp. 1–5.

[11] J.-F. Cardoso, “High-order contrasts for independent component analysis,” *Neural Computation*, vol. 11, no. 1, pp. 157–192, Jan. 1999.

[12] J. Omura, “On the Viterbi decoding algorithm,” *IEEE Transactions on Information Theory*, vol. 15, no. 1, pp. 177–179, 1969.