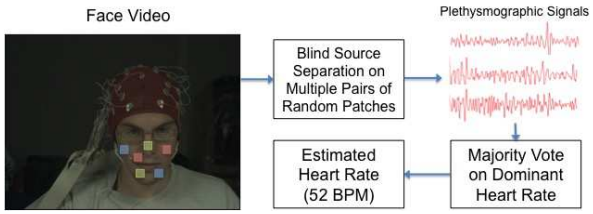
Robust Heart Rate Measurement from Video Using Select Random Patches

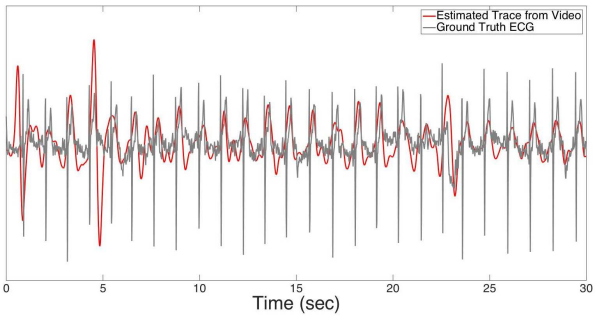
Antony Lam and Yoshinori Kuno
Graduate School of Science and Engineering
Saitama University
{antonylam,kuno}@cv.ics.saitama-u.ac.jp

Abstract

The ability to remotely measure heart rate from videos without requiring any special setup is beneficial to many applications. In recent years, a number of papers on heart rate (HR) measurement from videos have been proposed. However, these methods typically require the human subject to be stationary and for the illumination to be controlled. For methods that do take into account motion and illumination changes, strong assumptions are still made about the environment (e.g. background can be used for illumination rectification). In this paper, we propose an HR measurement method that is robust to motion, illumination changes, and does not require use of an environment's background. We present conditions under which cardiac activity extraction from local regions of the face can be treated as a linear Blind Source Separation problem and propose a simple but robust algorithm for selecting good local regions. The independent HR estimates from multiple local regions are then combined in a majority voting scheme that robustly recovers the HR. We validate our algorithm on a large database of challenging videos.



(a) Proposed Video Heart Rate Measurement System



(b) Estimated Plethysmographic Signal from Majority Vote vs. Ground Truth ECG

Figure 1: Heart Rate Measurement from Video Captured by Consumer Camera

1. Introduction

The ability to measure heart rate (HR) allows us to learn a lot about the physiological and emotional states of people, as well as their overall health. Conventional methods for measuring HR such as electrocardiography (ECG) or photoplethysmography using optical sensors require the sensor to make physical contact with the person. This can be inconvenient and uncomfortable for the person while also limiting the number of applications HR measurement could be used in. Fortunately, it has been shown (under controlled conditions) that it is possible to use a conventional camera to remotely detect small changes in skin color due to a person's cardiac pulse [18].

The ability to use conventional cameras to remotely measure HR would open doors to many possibilities. For example, HR measurement using cameras could provide complementary information in a facial emotion recognition system, which would be of great benefit to affective computing in human-computer and human-robot interaction applications. Another potential application would be in health monitoring. A networked camera system could be set up in a nursing home to continually monitor patient health for the long term without uncomfortable sensors. Realtime monitoring of HR over webcam could also be done by medical professionals. In addition, HR measurement could be used as a biometric to prevent anti-spoofing.

Recent work [1, 8, 10, 13, 14] has demonstrated the viability of video based HR measurement. However, these studies were conducted under controlled settings. In a real setting, many factors could make extraction of HR difficult. For example, there would be motion (rigid and non-rigid) and changes in illumination that would greatly affect the appearance of skin over time. A notable exception is the

recent work of Li et al. [9] where their overall system was able to achieve strong performance on a large collection of challenging videos from the MAHNOB-HCI Database [16].

The approach by Li et al. essentially works by using tracking to overcome rigid motion, using illumination changes in the background to rectify illumination changes on the face, and cropping out "noisy segments" from their estimated plethysmograph (PG) signals that were caused by non-rigid motion (e.g. displays of emotion on the face). While highly effective in some settings, their method has some drawbacks. The background's appearance changes over time cannot always be used to cancel out the effects of illumination change on skin. This is because the spectral reflectance of the background and skin are likely to be different. In addition, complex backgrounds (e.g. outdoors) would make such illumination rectification unreliable. Cutting out time segments that exhibit changes in emotion is also less than ideal as portions of the final estimated PG signal would be missing.

In this paper, we propose a video based face HR measurement method that overcomes the challenges presented in real settings described earlier without requiring use of the background to extract the PG signal and without the need to prune out time segments of the traces. In summary our contributions are:

1. We devise a model of skin appearance over time with the effects of hemoglobin taken into account.
2. Using our model, we present conditions where PG signal extraction from some local regions can be treated as a linear Blind Source Separation (BSS) problem.
3. We present a simple algorithm that can select good local regions to use in PG signal extraction.
4. A majority voting scheme that uses independent HR estimates from different local regions to robustly estimate the final HR (Fig. 1).

We validate the proposed approach by showing state-of-the-art HR estimation on 487 videos from the challenging MAHNOB-HCI Database.

2. Related Work

The basic approach from which most of the existing video based HR measurement methods derive is photoplethysmography (PPG). In this approach, a pulse oximeter illuminates the skin and measures changes in light absorption [15]. Extensions to this approach for remote measurement have also been done, however special lighting conditions such as specific narrowband wavelengths and sensors are needed [2, 3].

Verkruysse et al. [18] showed that remote PPG could be done using just a conventional camera (Canon Powershot) and normal ambient light in the visible spectrum (daylight and normal office fluorescent lighting). They showed that the green channel of a conventional camera provided the strongest PG signal since hemoglobin light absorption is most sensitive to oxygenation changes for green light. In addition, the red and blue channels also contained some PG information. The work of Wu et al. [21] also provided impressive visualizations of such color changes through magnification. However, these works do not address hindrances to HR estimation such as arbitrary motion and lighting changes that can occur in everyday settings.

In recent years, other methods to remote PPG have been proposed [8, 10, 13, 14]. Poh et al. [14] used a webcam and Viola-Jones face detection [19] to find the region of interest (ROI). They then computed the mean pixel values of the ROI for each of the RGB channels and frames. From this, a video of one face would give three temporal traces (one for each color channel). These three traces were then treated as signals and run through ICA to isolate the PG signal. In their results, they showed that ICA separation provided better accuracy than using the green channel trace alone. Kwon et al. [8] used a smartphone camera to record videos of human subjects and compared extracting the HR from the raw green channel traces against using ICA on all the RGB channels. Interestingly, they found that in their case, the raw green trace was more accurate for HR measurement. In later work, Poh et al. [13] extended their work to include a series of temporal filters for cleaning up the ICA extracted signals. Moreno et al. [10] also performed remote PPG by extracting the green channel trace and subjecting it to a series of filters.

An interesting departure from the color based methods discussed so far can be found in Balakrishnan et al. [1] where subtle head motions were used to extract a pulse and determined the HR. An advantage of their approach is that even when skin is not visible (e.g. views from back of the head), the pulse could be accurately extracted. The drawback is that their method requires the subject to try to remain as still as possible. More recent work by Irani et al. [5] improved upon Balakrishnan et al. by using stable face regions and the Discrete Cosine Transform. Our paper presents a color based method and is complementary to motion based methods.

The main drawbacks with the aforementioned work [8, 10, 13, 14] is that:

1. Their testing data did not contain illumination variations. In real settings, illumination variations are likely. For example, in an HCI setting, a person watching a movie or playing a video game could have many variations in lighting reflected off the face. In an outdoor setting, there could be numerous sources of ambient light and even shadows.

2. The subjects were asked to remain still. In real settings, we would expect a large range of motion. Considering the face alone, there could be various rigid (scale, translation, rotation) movements of the head. There could also be non-rigid movements in the form of facial expressions.

All these movements would affect the ability to accurately extract a temporal trace for a single point. We note that Poh et al. [13] did consider motion but these consisted of slow uniform head swings.

3. Another issue is that [8, 10, 13, 14] all used their own self-collected datasets, which are not publicly available. This makes fair comparisons against their methods difficult.

The recent work of Li et al. [9], aimed to overcome all these issues. To overcome rigid motion, they first detected face landmarks and then performed tracking. They used the background to perform illumination rectification on the face. For non-rigid motion due to facial emotion, they used a heuristic to cut portions of the illumination corrected traces that corresponded to sudden shifts in facial emotion. Afterward, they used the same bank of temporal filters by Poh et al. [13] to clean up the traces and make determining the HR more accurate. In addition, they chose to test their method on the publicly available MAHNOB-HCI Database, which allows for more fair comparisons.

While effective, their method has some drawbacks. Their illumination rectification step uses illumination changes in the background as an approximation to how the illumination changes would affect the face's appearance. However, the spectral reflectance of various surface points in the background would likely be different from the skin's spectral reflectance. This means that even the same illumination would affect the appearance of the face and background differently. In addition, outside of an HCI setting, the illumination in the background could be completely different from the illumination on the face. Another issue is that non-rigid motion is dealt with by essentially clipping out time segments. This means that cardiac activity in those video segments would be thrown out. Also, for shorter videos, pruning parts of the trace may not be an option.

It should be mentioned, very recent work by Kumar et al. [7] also extracts the PG signal under challenging conditions. We consider some of their ideas complementary to ours. They used a "goodness metric" to adaptively determine a weighted average of bandpass filtered green channel traces from preset face subregions. This average was then considered the PG signal. Their formulation requires illumination to be fixed over the time window of interest. We estimate the PG signal from dynamically chosen subregions in a BSS formulation, which allows for varied lighting over time in terms of both color and brightness. They also did not use the MAHNOB-HCI Database but will be sharing their own dataset soon.

In summary, we propose a video-based HR measurement method that overcomes the challenges presented in real settings described earlier without requiring use of the background, allowing for varied lighting over time, and without the need to prune out portions of the traces. In Sec. 3, we devise a model of skin appearance over time. Using the derived model, we detail in Sec. 4, our robust system. In Sec. 5, we present experiments. Finally, we conclude in Sec. 6 with a brief discussion of future research directions.

3. Model of Skin Appearance

Using observations from the literature, we formulate and derive a model of how illumination variations and cardiac activity affect the appearance of skin over time. This model will then be used in our method for isolating the PG signal.

Skin pigmentation is primarily determined by the presence of melanin but it is also influenced by hemoglobin absorption [12]. Thus the shape of the spectral reflectance curve over wavelengths for skin is mainly determined by a combination of melanin and hemoglobin.

In deriving our model, we assume the skin to be Lambertian and define the spectral reflectance of skin over wavelengths λ as

$$R_s(\lambda) = a_m R_m(\lambda) + a_h R_h(\lambda) \tag{1}$$

where a_m and a_h are scalars, $R_m(\lambda)$ and $R_h(\lambda)$ are normalized spectral distributions, $a_h R_h(\lambda)$ is the contribution from hemoglobin to the spectral reflectance $R_s(\lambda)$ and $a_m R_m(\lambda)$ is the spectral reflectance of the skin *without* the effects of hemoglobin–that is, having mainly the effects of melanin.[1] Thus we can model the observed pixel value of a point on skin for a given camera channel as

$$\begin{aligned} I &= \int R_s(\lambda) a_l L(\lambda) C_k(\lambda) d\lambda \\ &= a_m a_l \int R_m(\lambda) L(\lambda) C_k(\lambda) d\lambda \\ &+ a_h a_l \int R_h(\lambda) L(\lambda) C_k(\lambda) d\lambda \end{aligned} \tag{2}$$

where a_l is a scalar, $L(\lambda)$ is the normalized illuminant spectrum, and $C_k(\lambda)$ is the camera spectral response for channel k.

So far, we have only considered the appearance of a point on the skin for one instant in time. We are interested in observing appearance changes over time. Thus we add the dimension of time to the observed pixel value I

$$\begin{aligned} I(t) &= a_m a_l(t) \int R_m(\lambda) L(\lambda, t) C_k(\lambda) d\lambda \\ &+ a_h(t) a_l(t) \int R_h(\lambda, t) L(\lambda, t) C_k(\lambda) d\lambda \end{aligned} \tag{3}$$

Since we assume melanin production occurs on time scales longer than the videos, we can assume the spectral reflectance component for melanin remains constant over

[1] Less dominant pigments such as carotenes can play a role. For the sake of brevity, it is implicitly assumed contributions from such factors are included in $R_m(\lambda)$.

time. On the other hand, the spectral reflectance component for hemoglobin (which is influenced by cardiac activity) and the illuminant can change over time in the video. The camera spectral response is of course, constant over time.

4. Robust Estimation of HR

In real settings, we need to consider two primary challenges:

Motion: Human subjects will naturally exhibit rigid (scale, translation, rotation) and non-rigid motion (displays of emotions such as smiling, yelling, and so on).

Illumination Variation: The illumination in the environment could be complex and vary over time. For example, a person could be watching a movie and the screen would illuminate the face with different colors and levels of brightness. In outdoor settings, the ambient light could also vary dramatically over time.

4.1. Rigid Motion

Our main goal is to read HR from faces since faces are prominent in many videos of people. Fortunately, compensating for rigid motion can be accomplished using the appropriate tracker. We find that the pose-free facial landmark fitting tracker by Yu et al. [22] is very effective. This tracker can both localize facial landmarks and track them over a large range of motions.

From the tracker, we obtain 66 facial landmarks for all frames in the video (Fig. 2). Then for the first frame, we select a rectangular ROI relative to the facial landmarks as denoted by the blue area in the figure. Rigid transforms are also applied to the ROI in subsequent frames to maintain the same positioning relative to the tracked landmarks of any given frame. We implement this by determining the transform matrix $\mathbf{A}$ and translation $\vec{b}$ as

$$(\mathbf{A}, \vec{b}) = \underset{\mathbf{A},\vec{b}}{\operatorname{argmin}} \sum_{i=1}^{17} \|\mathbf{A}\vec{v}_{1,i} + \vec{b} - \vec{v}_{n,i}\|^2 \qquad (4)$$

where $\vec{v}_{1,i}$ is the location of the i^{th} landmark in frame 1 and $\vec{v}_{n,i}$ is the location of the i^{th} landmark in frame n. Also, note that we only used the first 17 landmarks because the other landmarks could exhibit non-rigid motions. We solve Eq. 4 via least-squares estimation. Any point $p1$ in the first frame's ROI is then transformed using matrix $\mathbf{A}$ and translation b to determine its coordinates in the n^{th} frame. The pixel values associated with the tracked point $p1$ over all frames then provides us with the trace $I_{p1}(t)$ of the point over time.

4.2. PG Signal Extraction with Illumination Variations Over Time

Having defined the model in Sec. 3, we now describe how we would isolate the PG signal from the effects of illumination variation over time given appearance traces for two points on the skin from one channel in the video. In particular, we use the green channel. This is because past work has found that the green channel in consumer cameras provides the strongest reading of the underlying PG signal [8, 9, 18]. This is not surprising as hemoglobin light absorption is most sensitive to oxygenation changes for green light [12].

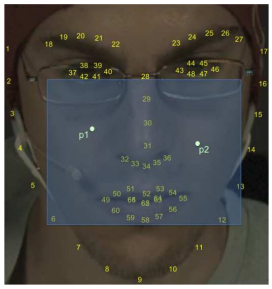


Figure 2: Example of 66 Facial Landmarks. Based on the landmarks, we select the blue area to be our region of interest. The points $p1$ and $p2$ illustrate an example of two points selected for PG signal extraction.

From the tracking procedure in Sec. 4.1, we can obtain two traces $I_{p1}(t)$ and $I_{p2}(t)$ from points $p1$ and $p2$ on the face. Previous work [13, 14] used linear ICA to extract the PG signal using three average traces of the same region from the RGB channels. Inspired by this, we speculate whether the traces $I_{p1}(t)$ and $I_{p2}(t)$ can also be cast into a linear BSS problem for PG signal extraction.[2] From Eq. (3), we have

$$\begin{aligned} I_{p1}(t) &= a_m a_l(t) \int R_m(\lambda) L_1(\lambda, t) C_k(\lambda) d\lambda \\ &+ a_h(t) a_l(t) \int R_h(\lambda, t) L_1(\lambda, t) C_k(\lambda) d\lambda \\ I_{p2}(t) &= b_m b_l(t) \int R_m(\lambda) L_2(\lambda, t) C_k(\lambda) d\lambda \\ &+ b_h(t) b_l(t) \int R_h(\lambda, t) L_2(\lambda, t) C_k(\lambda) d\lambda \end{aligned} \qquad (5)$$

These traces cannot be directly cast into a linear BSS problem as some of the scalars (e.g. $a_l(t)$) vary with time. In addition, the light spectra (L_1 and L_2) could also be different (even for the same time t). Thus to cast this into a linear BSS problem for two signals $\vec{x}_1$ and $\vec{x}_2$ of the form, $\vec{x}_1 = a_1\vec{s}_1 + a_2\vec{s}_2$, $\vec{x}_2 = b_1\vec{s}_1 + b_2\vec{s}_2$, we need to ensure the traces $I_{p1}(t)$ and $I_{p2}(t)$ satisfy some conditions:

1. For any given time t, $L_1(\lambda, t) = L_2(\lambda, t)$ for all wavelengths λ. That is, the normalized light spectra irradiating the two points are the equivalent at the same

[2] Although nonlinear ICA solutions exist, without extra prior information, the problem is ill-posed and has no solution [6].

time. The light spectrum is allowed to vary arbitrarily over time though. Essentially, the lights need to be the same "color" at the same time but this color may change over time. Their relative brightness at the same time can also be different.

2. Defining $\vec{a}_l = [a_l(1)a_l(2)...a_l(n)]^T$ for times $t = 1, 2, ..., n$ and similarly for other scalars, we need the normalized vectors $\frac{\vec{a}_l}{\|\vec{a}_l\|} = \frac{\vec{b}_l}{\|\vec{b}_l\|} = \vec{c}_l$. In other words, we need the relative changes of brightness over time between the two points to be the same but their scales may be different.
3. Similarly, we need $\frac{\vec{a}_h}{\|\vec{a}_h\|} = \frac{\vec{b}_h}{\|\vec{b}_h\|} = \vec{c}_h$. This condition is satisfied because the blood volume is tied to the cardiovascular pulse and for normal cameras,which have frame rates from 30–60 Hz, the delay in the PG signal between the farthest points on the face is negligible [7].
4. The amounts of irradiance, hemoglobin, or melanin are not uniform over the face. That is, we can have cases where $\|\vec{a}_l\| \neq \|\vec{b}_l\|$, $\|\vec{a}_h\| \neq \|\vec{b}_h\|$, or $a_m \neq b_m$. We can certainly expect that irradiance on the face would be non-uniform. In addition, blood perfusion is different for various face regions [7].

Caveat: We do not expect all pairs of points to satisfy all the conditions. For the moment, we assume that the right pairs of points satisfying these conditions can be found. In Sec. 4.4, we provide a more detailed discussion on when the first two conditions would be satisfied and also present an algorithm for selecting good pairs of points to use for HR estimation.

Continuing our discussion, provided the conditions are met, it can be shown that

$$\begin{aligned} I_{p1}(t) &= a_m\|\vec{a}_l\| \int R_m(\lambda)c_l(t)L(\lambda,t)C_k(\lambda)d\lambda \\ &+ \|\vec{a}_h\|\|\vec{a}_l\| \int c_h(t)R_h(\lambda,t)c_l(t)L(\lambda,t)C_k(\lambda)d\lambda \\ I_{p2}(t) &= b_m\|\vec{b}_l\| \int R_m(\lambda)c_l(t)L(\lambda,t)C_k(\lambda)d\lambda \\ &+ \|\vec{b}_h\|\|\vec{b}_l\| \int c_h(t)R_h(\lambda,t)c_l(t)L(\lambda,t)C_k(\lambda)d\lambda \end{aligned} \quad (6)$$

where $c_l(t)$ is the t^{th} element of the vector $\vec{c}_l$ defined earlier. $c_h(t)$ is similarly defined.

Eq. (6) can be expressed more compactly as

$$\begin{aligned} I_{p1}(t) &= a_1 I_m(t) + a_2 I_h(t) \\ I_{p2}(t) &= b_1 I_m(t) + b_2 I_h(t) \end{aligned} \quad (7)$$

where $I_m(t) = \int R_m(\lambda)c_l(t)L(\lambda,t)C_k(\lambda)d\lambda$, $I_h(t) = \int c_h(t)R_h(\lambda,t)c_l(t)L(\lambda,t)C_k(\lambda)d\lambda$, $a_1 = a_m\|\vec{a}_l\|$, $a_2 = \|\vec{a}_h\|\|\vec{a}_l\|$, similarly for the terms in $I_{p2}(t)$, and $a_i \neq b_i$ because of the fourth condition. Then assuming the I_m and I_h components are independent, we can solve for them (up to some scale) as a linear BSS problem using a standard algorithm such as FastICA [4].[3]

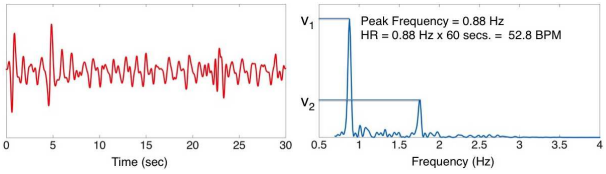


Figure 3: Example of an extracted PG signal (left) and its power spectral density distribution (right). The HR is calculated from the peak frequency. In this case, the confidence ratio is computed as $r = v_1/v_2 = 3.16$. Higher ratios provide us with more confidence the estimated HR is correct.

After solving for the components, there is still an ambiguity as to which signal corresponds to the PG signal. From the observation that I_m is the dominant component (since melanin primarily determines skin pigmentation), we devised a heuristic to solve for the ambiguity. We observe which of the two separated signals has the nearest average Euclidean distance to the raw traces (I_{p1} and I_{p2}) and determine that signal to be the I_m signal. The remaining signal is then chosen to be the PG signal (I_h). Once the PG signal is determined, the HR can be estimated.

4.3. Estimating HR from an Extracted PG Signal

Like in Li et al. [9], we run a set of temporal filters over the extracted PG signal. We first run a detrending filter [17] to reduce slow and non-stationary trends of the signal. We then run a moving average filter to smooth out noise. The signal is then converted to the frequency domain and its power spectral density (PSD) distribution is computed via Welch's method [20]. The PSD is set to range from 0.7 to 4 Hz to correspond with normal human HR [9, 13]. The frequency with the highest peak can then be multiplied by 60 to determine the HR in beats per minute (BPM). (Fig. 3.)

We also devise a heuristic for indicating how confident we are in the HR estimate. We take the ratio of the amplitude at the highest peak to the amplitude at the second highest peak ($r = v_1/v_2$ in Fig. 3). We call this the "confidence ratio" and it provides an indication of how dominant the peak frequency is relative to other frequencies in the PG signal. A high ratio would indicate that the PG signal had a single dominant frequency, thus providing more confidence that the estimated HR is correct. Note that the confidence ratio by itself is sensitive to cases where the entire PSD is close to zero as noise could make the ratio high. However, such cases are uncommon and the majority voting scheme in Sec. 4.4 mitigates this issue.

[3] Note that the lighting effects from $L(\lambda, t)$ are still in I_h. This does not affect our HR results because $L(\lambda, t)$ is not likely to have a steady pulse like the heart. The temporal filters used in Sec. 4.3 also mitigate the effects of $L(\lambda, t)$.

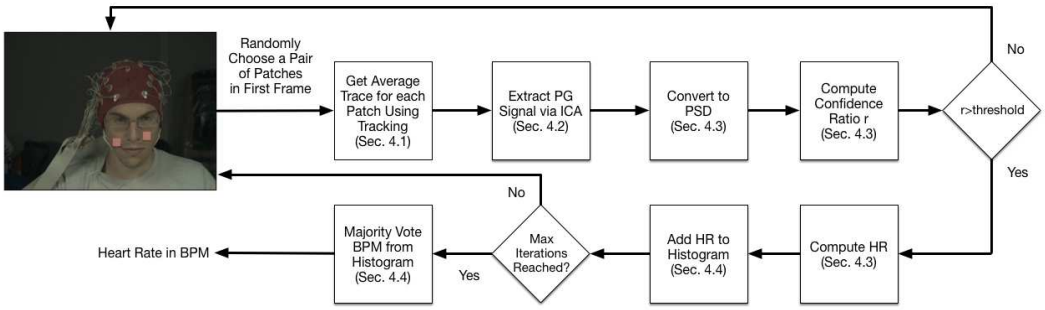


Figure 4: Flowchart of System

4.4. Robustly Estimating HR by Majority Vote

In Sec. 4.2, we showed that provided two point traces that satisfy our conditions, we can use ICA to extract the PG signal. Not all pairs of point traces are guaranteed to satisfy our conditions but we posit that there are a reasonable number of pairs of point traces which do satisfy our conditions (or approximately satisfy them).

The first two conditions essentially require the two points to be irradiated by the same light source at any given instant in time (but the brightness of irradiation at the two points can be different). At first glance, this appears somewhat restrictive. However, it is reasonable to expect that most videos would have many pairs of points satisfying this requirement. There may be many sources of light in a real setting but the same light source is likely to affect entire subregions of the face. In other words, it is unlikely that *almost all* individual surface points would each be illuminated by completely separate light sources. The conditions also allow for this same light source to arbitrarily change its spectrum over time, which affords a degree of flexibility (e.g. a monitor is a single light source that can vary its light spectrum). The last two conditions are also reasonable and were justified in Sec. 4.2. The next question is, how we can find pairs of points that satisfy the conditions.

Unfortunately, finding good pairs of points is non-trivial since whether those conditions are satisfied depend on a number of factors such as, where the light sources are located for any given time in the scene, what the spectral distributions of the light sources are, the geometry of the face, and so on.

Fortunately, it is possible to avoid explicitly determining which pairs of points are good. Assuming a pair of point traces at least approximately satisfies the conditions presented in Sec. 4.2, we expect the extracted PG signal would have a PSD that clearly shows a strong frequency corresponding to the HR of the person in the video. We also expect *most* pairs of point traces which violate the conditions to yield extracted PG signals that would not clearly show a dominant frequency. From these ideas, we devise a simple algorithm that can be used to select the good points.

We randomly select many pairs of points in the ROI and perform PG extraction (Sec. 4.2) on each pair. Afterwards, we use the confidence ratio heuristic described in Sec. 4.3 to determine which of the many extracted PG signals provides a confident estimation of the HR. From each extracted PG signal with a high confidence ratio, we compute the HR and add it to a histogram. A majority vote can then be performed on the histogram to obtain the final HR estimate for the human subject in the video. How we set our histogram and other implementation details are described in Sec. 5.

Fig. 4 shows the flowchart of our complete system. We note that rather than taking just the trace of a single point, it is better to determine the average trace of a patch centered on the chosen point as it removes noise. This is done by taking the average of all pixel values in the patch for each frame independently. The resultant trace is then filtered by a moving average.

This simple procedure has some benefits. There is no need to know anything about the scene a priori, we do not need complex models to find good pairs of points, and as experiments will show, it is quite robust. When enough iterations are set, the algorithm will make use of as many good pairs of point traces as it can.

Another point is that since we use local regions on the face, our method mitigates the issue with non-rigid facial emotions. Provided most of the face does not exhibit non-rigid motion, the majority voting scheme can essentially ignore poor HR estimates from difficult parts of the face.

5. Experiments

We compare our method to recent algorithms in the literature on a challenging publicly available database. To be fair, the same face tracker is used in all cases.

5.1. Database of Videos

We used the MAHNOB-HCI Database [16] for our experiments. This database consists of 61 FPS RGB videos of human subjects exhibiting facial emotions in response to visual stimuli (from a computer monitor). In addition to video footage, the subjects' ECG readings were also recorded at

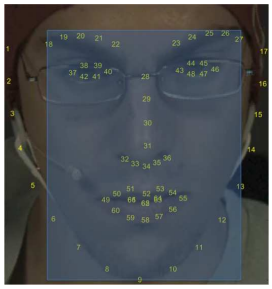


Figure 5: Other ROI Tested in Experiments Fig. 2 shows the primary ROI used in our system.

Method	MAE	RMSE	%Absolute Error <5 BPM
Baseline (ROI in Fig. 2)	36.7	42.7	4.9%
Poh2011	13.6	21.3	46.2%
Li2014	7.8	15	68.1%
Ours (ROI in Fig. 5)	5.2	9.4	73.1%
Ours (ROI in Fig. 2)	**4.7**	**8.9**	**75.1%**

Table 1: Errors on Heart Rate Estimation. The Baseline method uses traces taken from the entire ROI defined in Fig. 2 as input to the ICA algorithm. All videos were at 61 FPS and ground truth ECGs at 256 Hz.

256 Hz with synchronization timings to the videos provided. To be consistent with previous work [9], we chose videos from the emotion elicitation portion of the database for the first 27 human subjects. Each subject was recorded for 20 sessions however, not all videos could be used because some did not have corresponding ECG readings. We ultimately were able to use 487 videos.[4]

Since the videos are of different lengths, like [9], we chose 30 seconds from each video (frames 306 to 2135) for our tests. The database has three separate ECG readings recorded from different parts of the body and we chose to use the second channel (EXG2, upper left corner of the chest). We then took the portion of the ECG corresponding to the chosen video frames and ran a QRS-detection algorithm [11] to detect peaks. The average time between the peaks determined the ground truth HR for the session.

5.2. Algorithms Compared and Parameter Settings

We implemented the recent algorithms by Li2014 [9] and Poh2011 [13] for our comparison. These algorithms were chosen based on their high performance as reported by Li et al. [9]. Li2014 uses a different face tracker from ours and Poh2011 uses the Viola-Jones face detector. To make a fair comparison of each approach in terms of its HR estimation accuracy (and not tracking performance), we used the same facial landmark tracker [22] for all algorithms. The tracked landmarks were then used to track the ROIs as defined in their respective papers. In Poh2011, the ROI is a crop of the entire face. In Li2014, it is a polygonal region from below the eyes to above the chin. In our work, the ROI is illustrated in Fig. 2. To test our algorithm's sensitivity to the choice of ROI, we also defined another ROI (Fig. 5) and observed HR estimation performance with it.

We note that Poh2011 originally includes a custom peak detection step to find the location of each heart beat for additional analysis. Since we are only interested in the HR, we did not replicate their peak detection. We determined HR from their extracted PG signals by observing the peak frequency of its PSD.

For our algorithm, there are a few parameters to set. We first note that more stable tracking of points was achieved by applying a 0.2 second (12 frames) moving average to the tracks (these smoothed tracks were used with all algorithms compared). As mentioned earlier, we also found that taking average traces from a patch centered on a point was better than just the trace from the point because it would remove noise. The size of the patch centered on a given point was chosen to be 41x41 pixels. We averaged the patch's pixels in each frame independently and then filtered the resultant trace with a moving average window of 0.2 seconds. For the temporal filters used to postprocess the extracted PG signal from each pair of traces, we set $\lambda = 100$ for the detrending filter [17] and the moving average window was 0.2 seconds. For the confidence ratio threshold (Fig. 3), we set it to 2. Thus only PSDs whose confidence ratio was greater than 2 would have its HR estimate counted in the majority vote. For the number of random pairs of patches chosen, we set that to 500[5]. (This is the number of iterations in Fig. 4.) For the histogram to perform the majority vote, we simply rounded the estimated HR from each accepted PSD to the nearest integer. The mode of all those HR estimates was then chosen as the final HR estimate.

To illustrate the advantage of our trace pair selection approach, we also implemented an algorithm we call Baseline. This algorithm first creates a grid of non-overlapping 41x41 patches in the ROI defined in Fig. 2. It then computes all the average traces from the patches as is done in our system. ICA is then performed on all the traces for finding 2 independent components. Deciding which of the components is the PG signal and computing the HR is then the same as with our proposed method.

5.3. Results

We ran each of the algorithms on the same data. We report, the mean absolute error (MAE) and root-mean squared

[4]Previous work [9] reported that they used 527 videos from the same portion of the database. However, at the time we downloaded the data, we only found 487 videos with corresponding ECG readings.

[5]Tests on smaller subsets of the dataset suggest that performance actually plateaus after 100 pairs. Future work will investigate the effects of the number of pairs on accuracy.

error (RMSE) from all 487 videos in Table 1. For applications in detecting vital signs during emergencies, an error of less than 5 BPM is likely sufficient [14], so we also show the percentage of videos where the absolute error was less than 5 BPM.

Our method had the best performance out of all the algorithms. Vastly outperforming Baseline shows that there is a definite advantage to selecting only good traces to use in linear ICA based PG signal extraction. The poor performance of Baseline is likely due to its inclusion of traces which violate our conditions for linear BSS (Sec. 4.2). Baseline may have also included poorly tracked traces.

Poh2011 takes the average of the entire ROI from each frame for each of the RGB channels and constructs three traces (one from each color channel). They then use ICA to extract the PG signal. Again, we attribute our better performance to the selection of good pairs of traces to use for PG signal extraction, whereas Poh2011 takes the entire ROI.

Another point is that although Poh2011 solves a linear BSS problem, the three traces from the color channels do not strictly correspond to a linear mixing of signals in all cases. This is because the pixel value under channel k for time t is $I_k(t) = \int R_m(\lambda)L(\lambda,t)C_k(\lambda)d\lambda + \int R_h(\lambda,t)L(\lambda,t)C_k(\lambda)d\lambda$. (Here, the spectra are not normalized.) Then the three color channel traces are generated by changing the camera spectral response C_k. If lighting stays constant over time, the filtering from C_k has basically the same effect as linear mixing with coefficients. However, if the lighting changes arbitrarily over time, the resultant effect would be that of mixing coefficients that also vary with time. This no longer fits the linear BSS formulation and in an unconstrained setting, would affect accuracy.

Our method also outperforms Li2014. In Li2014, illumination variation over time is first observed in the background. Then this information is used to rectify the illumination changes on the face. Using the background is not always an option because the environment may not have a background with reflected light that is the same as that from the face. Also, the spectral reflectance of the background is not likely to be the same as the face. Non-rigid motions caused by facial expressions also caused their extracted PG signals to have anomalous segments. These time segments were clipped heuristically, which is fine if the videos are sufficiently long. Our method works by only looking at the face and so we do not need to consider the background. Our confidence ratio heuristic rejects PG signals extracted from face regions that are problematic. As a result, we do not need to clip out time segments.

To test the robustness of our proposed algorithm to the choice of ROI, we decided to define another ROI (Fig. 5) and compare against the originally chosen ROI (Fig. 2). The results indicate that including non-skin areas such as the eyes, lowers performance slightly but our method still outperforms previous work. This indicates that the intended purpose of our random patch algorithm–to select good subregions within a given ROI, worked well.

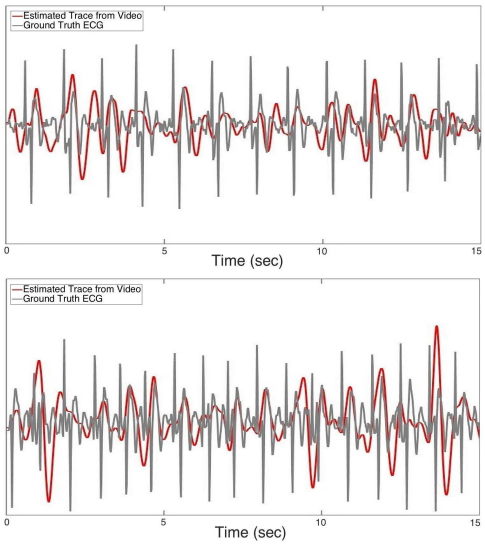


Figure 6: Sample Estimated Traces from Video vs. Ground Truth ECGs.

5.4. PG Signal from Majority Vote

While not the main goal of this paper, an interesting result is that after estimating the final HR for the person, we can go back to the histogram and see which PG signals contributed to the majority vote. All these PG signals can then be averaged to obtain a PG signal that seems to match the ground truth ECG reasonably well (Figs. 1b and 6). This is interesting because ECGs are from electrical activity as opposed to optical information. Interesting future work would be to improve the accuracy of the recovered PG signal itself. This would be useful for estimating heart rate variability, which is also an important indicator of health [10].

6. Conclusion and Future Work

We derived a model of appearance that considers hemoglobin effects on the spectral reflectance of skin over time. With this model, we showed that given certain conditions, the PG signal could be extracted from two raw point traces in the green channel despite illumination variations. An algorithm for selecting pairs of point traces from the face that likely satisfy our conditions was presented. This involved randomly choosing pairs and accepting those that exceeded our confidence ratio threshold. The proposed method outperformed previous work. A drawback of our method is speed. Our current Matlab implementation takes 7 minutes to process 30 seconds of video (not including tracking time). To reduce runtime, in the future, we hope to choose pairs of traces more intelligently. We also believe relaxing the conditions in Sec. 4.2 could provide more pairs of traces that would be usable for HR estimation, which would reduce the required number of iterations in Fig. 4.

References

[1] G. Balakrishnan, F. Durand, and J. Guttag. Detecting pulse from head motions in video. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3430–3437, June 2013. 1, 2

[2] G. Cennini, J. Arguel, K. Akşit, and A. van Leest. Heart rate monitoring via remote photoplethysmography with motion artifacts reduction. *Optics Express*, 18(5):4867–4875, Mar 2010. 2

[3] K. Humphreys, T. Ward, and C. Markham. Noncontact simultaneous dual wavelength photoplethysmography: A further step toward noncontact pulse oximetry. *Review of Scientific Instruments*, 78(4):–, 2007. 2

[4] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000. 5

[5] R. Irani, K. Nasrollahi, and T. B. Moeslund. Improved pulse dectection from head motions using DCT. In *9th International Conference on Computer Vision Theory and Applications*. Institute for Systems and Technologies of Information, Control and Communication, 2014. 2

[6] C. Jutten, M. Babaie-Zadeh, and J. Karhunen. Chapter 14 - nonlinear mixtures. In P. Comon and C. Jutten, editors, *Handbook of Blind Source Separation*, pages 549–592. Academic Press, Oxford, 2010. 4

[7] M. Kumar, A. Veeraraghavan, and A. Sabharwal. DistancePPG: Robust non-contact vital signs monitoring using a camera. *Biomedical Optics Express*, 6(5):1565–1588, May 2015. 3, 5

[8] 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 *IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2174–2177, Aug 2012. 1, 2, 3, 4

[9] X. Li, J. Chen, G. Zhao, and M. Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 4264–4271, June 2014. 2, 3, 4, 5, 7

[10] J. Moreno, J. Ramos-Castro, J. Movellan, E. Parrado, G. Rodas, and L. Capdevila. Facial video–based photoplethysmography to detect HRV at rest. *International Journal of Sports Medicine*, 36(6):474–480, 2015. 1, 2, 3, 8

[11] J. Pan and W. J. Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3):230–236, Mar. 1985. 7

[12] E. J. Parra. Human pigmentation variation: evolution, genetic basis, and implications for public health. *Yearbook of Physical Anthropology*, 50:85–105, 2007. 3, 4

[13] M.-Z. Poh, D. McDuff, and R. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, Jan 2011. 1, 2, 3, 4, 5, 7

[14] 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*, 18(10):10762–10774, May 2010. 1, 2, 3, 4, 8

[15] K. Shelley and S. Shelley. *Clinical monitoring: practical applications for anesthesia and critical care*. Harcourt Brace, 2001. 2

[16] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3:42–55, April 2012. Issue 1. 2, 6

[17] M. Tarvainen, P. Ranta-aho, and P. Karjalainen. An advanced detrending method with application to hrv analysis. *IEEE Transactions on Biomedical Engineering*, 49(2):172–175, Feb 2002. 5, 7

[18] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, Dec 2008. 1, 2, 4

[19] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004. 2

[20] P. D. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, Jun 1967. 5

[21] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph. (Proceedings SIGGRAPH 2012)*, 31(4), 2012. 2

[22] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1944–1951, 2013. 4, 7