

# Heartbeat Rate Measurement from Facial Video

Haque, Mohammad Ahsanul; Irani, Ramin; Nasrollahi, Kamal; Moeslund, Thomas B.

*Published in:*
I E E E Intelligent Systems

*DOI (link to publication from Publisher):*
10.1109/MIS.2016.20

*Publication date:*
2016

# Heartbeat Rate Measurement from Facial Video

Mohammad A. Haque, Ramin Irani, Kamal Nasrollahi, and Thomas B. Moeslund

*Visual Analysis of People (VAP) laboratory, Aalborg University, Denmark*
*Email: {mah, ri, kn, tbm}@create.aau.dk*

***Abstract*—Heartbeat Rate (HR) reveals a person's health condition. This paper presents an effective system for measuring HR from facial videos acquired in a more realistic environment than the testing environment of current systems. The proposed method utilizes a facial feature point tracking method by combining a 'Good feature to track' and a 'Supervised descent method' in order to overcome the limitations of currently available facial video based HR measuring systems. Such limitations include, e.g., unrealistic restriction of the subject's movement and artificial lighting during data capture. A face quality assessment system is also incorporated to automatically discard low quality faces that occur in a realistic video sequence to reduce erroneous results. The proposed method is comprehensively tested on the publicly available MAHNOB-HCI database and our local dataset, which are collected in realistic scenarios. Experimental results show that the proposed system outperforms existing video based systems for HR measurement.**

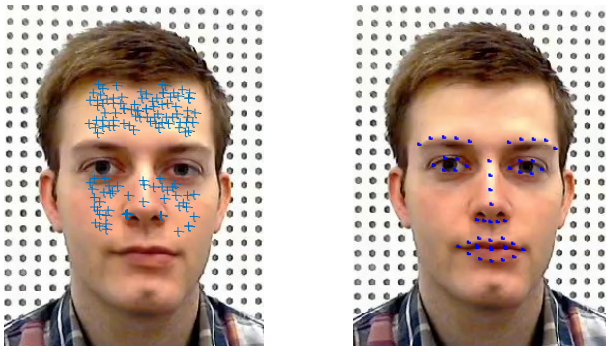
***Index Terms*—Heartbeat rate, facial video, supervised descent method (SDM), good feature to track (GFT), head motion.**

## I. Introduction

Heartbeat Rate (HR) is an important physiological parameter that provides information about the condition of the human body's cardiovascular system in applications like medical diagnosis, rehabilitation training programs, and fitness assessments [1]. Increasing or decreasing a patient's HR beyond the norm in a fitness assessment or rehabilitation training, for example, can show how the exercise affects the trainee, and indicates whether continuing the exercise is safe.

HR is typically measured by an Electrocardiogram (ECG) through placing sensors on the body. A recent study was driven by the fact that blood circulation causes periodic subtle changes to facial skin color [2]. This fact was utilized in [3]–[7] for HR estimation and [8]–[10] for applications of heartbeat signal from facial video. These facial color-based methods, however, are not effective when taking into account the sensitivity to color noise and changes

in illumination during tracking. Thus, Balakrishnan et al. proposed a system for measuring HR based on the fact that the flow of blood through the aorta causes invisible motion in the head (which can be observed by Ballistocardiography) due to pulsation of the heart muscles [11]. An improvement of this method was proposed in [12]. These motion-based methods of [11], [12] extract facial feature points from forehead and cheek (as shown in Fig. 1(a)) by a method called Good Feature to Track (GFT). They then employ the Kanade-Lucas-Tomasi (KLT) feature tracker from [13] to generate the motion trajectories of feature points and some signal processing methods to estimate cyclic head motion frequency as the subject's HR. These calculations are based on the assumption that the head is static (or close to) during facial video capture. This means that there is neither internal facial motion nor external movement of the head during the data acquisition phase. We denote internal motion as facial expression and external motion as head pose. In real life scenarios there are, of course, both internal and external head motion. Current methods, therefore, fail due to an inability to detect and track the feature points in the presence of internal and external motion as well as low texture in the facial region. Moreover, real-life scenarios challenge current methods due to low facial quality in video because of motion blur, bad posing, and poor lighting conditions [14]. These low quality facial frames induce noise in the motion trajectories obtained for measuring the HR.



(a) (b)

Fig. 1. a) Facial feature points extracted by GFT, and b) facial landmarks obtained by SDM.

The proposed system addresses the aforementioned shortcomings and advances the current automatic systems for reliable measuring of HR. We introduce a Face Quality Assessment (FQA) method that prunes the captured video data so that low quality face frames cannot contribute to erroneous results [15], [16]. We then extract GFT feature points (Fig. 1(a)) of [11] but combine them with facial landmarks (Fig. 1(b)), extracted by the Supervised Descent Method (SDM) of [17]. A combination of these two methods for vibration signal generation allows us to obtain stable trajectories that, in turn, allow a better estimation of HR. The experiments are conducted on a publicly available database and on a local database collected at the lab and a commercial fitness center. The experimental

results show that our system outperforms state-of-the-art systems for HR measurement. The paper's contributions are as follows:

i. We identify the limitations of the GFT-based tracking used in previous methods for HR measurement in realistic videos that have facial expression changes and voluntary head motions, and propose a solution using SDM-based tracking.
ii. We provide evidence for the necessity of combining the trajectories from the GFT and the SDM, instead of using the trajectories from either the GFT or the SDM.
iii. We introduce the notion of FQA in the HR measurement context and demonstrate empirical evidence for its effectiveness.

The rest of the paper is organized as follows. Section II provides the theoretical basis for the proposed method, which is then described in Section III. Section IV presents the experimental results, and the paper's conclusions are provided in Section V.

## II. THEORY

This section describes the basics of GFT- and SDM-based facial point tracking, explains the limitations of the GFT-based tracking, and proposes a solution via a combination of GFT- and SDM-based tracking.

Tracking facial feature points to detect head motion in consecutive facial video frames was accomplished in [11], [12] using GFT-based method. The GFT-based method uses an affine motion model to express changes in the level of intensity in the face. Tracking a window of size $w_x \times w_y$ in frame $I$ to frame $J$ is defined on a point velocity parameter $\boldsymbol{\delta} = [\delta_x \ \delta_y]^T$ for minimizing a residual function $f_{GFT}$ that is defined by:

$$f_{GFT}(\boldsymbol{\delta}) = \sum_{x=p_x}^{p_x+w_x} \sum_{y=p_y}^{p_y+w_y} \big(I(\mathbf{x}) - J(\mathbf{x}+\boldsymbol{\delta})\big)^2 \qquad (1)$$

where $(I(\mathbf{x}) - J(\mathbf{x}+\boldsymbol{\delta}))$ stands for $(I(x,y) - J(x+\delta_x, y+\delta_y))$, and $\mathbf{p} = [p_x, p_y]^T$ is a point to track from the first frame to the second frame. According to observations made in [18], the quality of the estimate by this tracker depends on three factors: the size of the window, the texture of the image frame, and the amount of motion between frames. Thus, in the presence of voluntary head motion (both external and internal) and low-texture in facial videos, the GFT-based tracking exhibits the following problems:

i. Low texture in the tracking window: In general, not all parts of a video frame contain complete motion information because of an aperture problem. This difficulty can be overcome by tracking feature points in

corners or regions with high spatial frequency content. However, GFT-based systems for HR utilized the feature points from the forehead and cheek that have low spatial frequency content.

ii. Losing track in a long video sequence: The GFT-based method applies a threshold to the cost function $f_{GFT}(\boldsymbol{\delta})$ in order to declare a point 'lost' if the cost function is higher than the threshold. While tracking a point over many frames of a video, as done in [11], [12], the point may drift throughout the extended sequences and may be prematurely declared 'lost.'

iii. Window size: When the window size (i.e. $w_x \times w_y$ in (1)) is small a deformation matrix to find the track is harder to estimate because the variations of motion within it are smaller and therefore less reliable. On the other hand, a bigger window is more likely to straddle a depth discontinuity in subsequent frames.

iv. Large optical flow vectors in consecutive video frames: When there is voluntary motion or expression change in a face the optical flow or face velocity in consecutive video frames is very high and GFT-based method misses the track due to occlusion [13].

Instead of tracking feature points by GFT-based method, facial landmarks can be tracked by employing a face alignment system. The Active Appearance Model (AAM) fitting [19] and its derivatives [20] are some of the early solutions for face alignment. A fast and highly accurate AAM fitting approach that was proposed recently in [17] is SDM. The SDM uses a set of manually aligned faces as training samples to learn a mean face shape. This mean shape is then used as an initial point for an iterative minimization of a non-linear least square function towards the best estimates of the positions of the landmarks in facial test images. The minimization function can be defined as a function over $\Delta x$:

$$f_{SDM}(x_0 + \Delta x) = \left\| g\big(d(x_0 + \Delta x)\big) - \theta_* \right\|_2^2 \qquad (2)$$

where $x_0$ is the initial configuration of the landmarks in a facial image, $d(x)$ indexes the landmarks configuration $(x)$ in the image, $g$ is a nonlinear feature extractor, $\theta_* = g(d(x_*))$, and $x_*$ is the configuration of the true landmarks. In the training images $\Delta x$ and $\theta_*$ are known. By utilizing these known parameters the SDM iteratively learns a sequence of generic descent directions, $\{\partial_n\}$, and a sequence of bias terms, $\{\beta_n\}$, to set the direction towards the true landmarks configuration $x_*$ in the minimization process, which are further applied in the alignment of unlabelled faces [17]. The evaluation of the descent directions and bias terms is accomplished by:

$$x_n = x_{n-1} + \partial_{n-1}\sigma(x_{n-1}) + \beta_{n-1} \qquad (3)$$

where $\sigma(x_{n-1}) = g(d(x_{n-1}))$ is the feature vector extracted at the previous landmark location $x_{n-1}$, $x_n$ is the new location, and $\partial_{n-1}$ and $\beta_{n-1}$ are defined as:

$$\partial_{n-1} = -2 \times \mathbf{H}^{-1}(x_{n-1}) \times \mathbf{J}^{\mathrm{T}}(x_{n-1}) \times g(d(x_{n-1})) \quad (4)$$

$$\beta_{n-1} = -2 \times \mathbf{H}^{-1}(x_{n-1}) \times \mathbf{J}^{\mathrm{T}}(x_{n-1}) \times g(d(x_*)) \quad (5)$$

where $\mathbf{H}(x_{n-1})$ and $\mathbf{J}(x_{n-1})$ are, respectively, the Hessian and Jacobian matrices of the function $g$ evaluated at $(x_{n-1})$. The succession of $x_n$ converges to $x_*$ for all images in the training set.

The SDM is free from the problems of the GFT-based tracking approach for the following reasons:

i. Low texture in the tracking window: The 49 facial landmarks of SDM are taken from face patches around eye, lip, and nose edges and corners (as shown in Fig. 1(b)), which have high spatial frequency due to the existence of edges and corners as discussed in [18].

ii. Losing track in a long video sequence: The SDM does not use any reference points in tracking. Instead, it detects each point around the edges and corners in the facial region of each video frame by using supervised descent directions and bias terms as shown in (3), (4) and (5). Thus, the problems of point drifting or dropping a point too early do not occur.

iii. Window size: The SDM does not define the facial landmarks by using the window based 'neighborhood sense' and, thus, does not use any window-based point tracking system. Instead, the SDM utilizes the 'neighborhood sense' on a pixel-by-pixel basis along with the descent detections and bias terms.

iv. Large optical flow vectors in consecutive video frames: As mentioned in [13], occlusion can occur by large optical flow vectors in consecutive video frames. As a video with human motion satisfies temporal stability constraint [21], increasing the search space can be a solution. SDM uses supervised descent direction and bias terms that allow searching selectively in a wider space with high computational efficiency.

Though GFT-based method fails to preserve enough information to measure the HR when the video has facial expression change or head motion, it uses a larger number of facial feature points (e.g., more than 150) to track than SDM (only 49 points). This matter causes the GFT-based method to generate a better trajectory than SDM when there is no voluntary motion. On the other hand, SDM does not miss or erroneously track the landmarks in the

presence of voluntary facial motions. In order to exploit the advantages of the both methods, a combination of GFT- and SDM-based tracking outcome can be used, which is explained in the methodology section. Thus, merely using GFT or SDM to extract facial points in cases where subjects may have both voluntary motion and non-motion periods does not produce competent results.

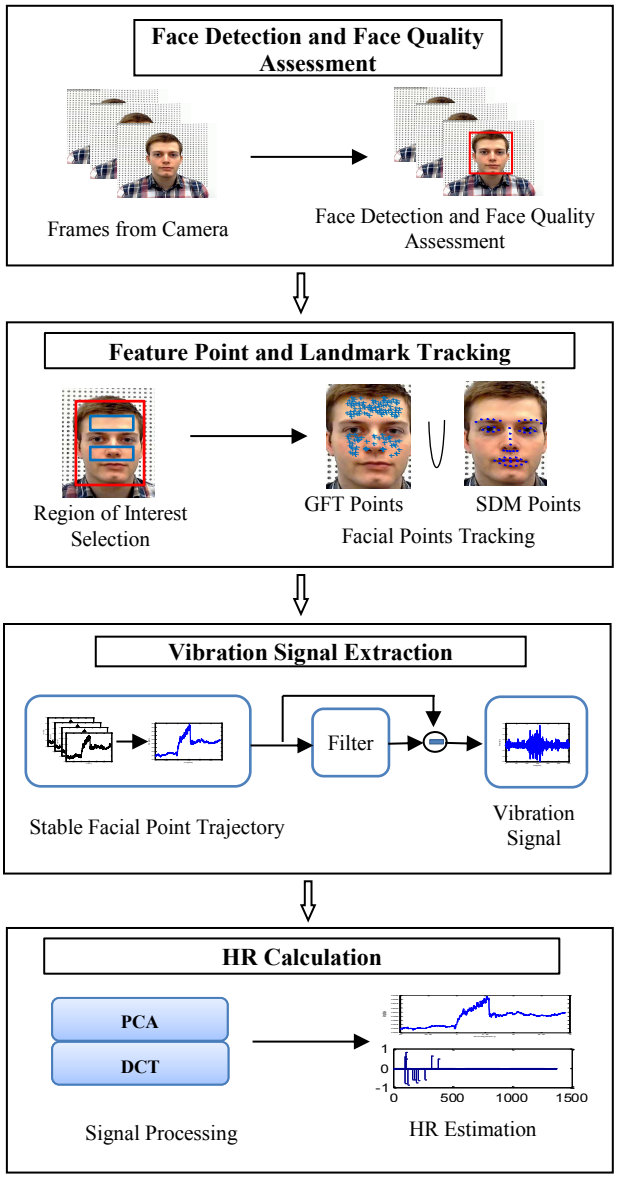



Fig. 2. The block diagram of the proposed system.

## III. THE PROPOSED METHOD

A block diagram of the proposed method is shown in Fig. 2. The steps are explained below.

### *A. Face Detection and Face Quality Assessment*

The first step of the proposed motion-based system is face detection from facial video acquired by a webcam. We employed the Haar-like features of Viola and Jones to extract the facial region from the video frames [22]. However, facial videos captured in real-life scenarios can exhibit low face quality due to the problems of pose variation, varying levels of brightness, and motion blur. A low quality face produces erroneous results in facial feature points or landmarks tracking. To solve this problem, a FQA module is employed by following [16], [23]. The module calculates four scores for four quality metrics: resolution, brightness, sharpness, and out-of-plan face rotation (pose). The quality scores are compared with thresholds (following [23], with values 150x150, 0.80, 0.8, and 0.20, for resolution, brightness, sharpness, and pose, respectively) to check whether the face needs to be discarded. If a face is discarded, we concatenate the trajectory segments to remove discontinuity by following [5]. As we measure the average HR over a long video sequence (e.g. 30 secs to 60 secs) discarding few frames (e.g., less than 5% of the total frames) does not greatly affect the regular characteristic of the trajectories but removes the most erroneous segments coming from low quality faces.

### *B. Feature Points and Landmarks Tracking*

Tracking facial feature points and generating trajectory keep record of head motion in facial video due to heartbeat. Our objective with trajectory extraction and signal processing is to find the cyclic trajectories of tracked points by removing the non-cyclic components from the trajectories. Since GFT-based tracking has some limitations, as we discussed in the previous section, having voluntary head motion and facial expression change in a video produces one of two problems: i) completely missing the track of feature points and ii) erroneous tracking. We observed more than 80% loss of feature points by the system in such cases. In contrast, the SDM does not miss or erroneously track the landmarks in the presence of voluntary facial motions or expression change as long as the face is qualified by the FQA. Thus, the system can find enough trajectories to measure the HR. However, the GFT uses a large number of facial points to track when compared to SDM, which uses only 49 points. This causes the GFT to preserve more motion information than SDM when there is no voluntary motion. Hence, merely using GFT or SDM to extract facial points in cases where subjects may have both voluntary motion and non-motion periods does not produce competent results. We therefore propose to combine the trajectories of GFT and SDM. In order to generate combined trajectories, the face is passed to the GFT-based tracker to generate trajectories from facial feature points

and then appended with the SDM trajectories. Let the trajectories be expressed by location time-series $S_{t,n}(x, y)$, where $(x, y)$ is the location of a tracked point $n$ in the video frame $t$.

### *C. Vibration Signal Extraction*

The trajectories from the previous step are usually noisy due to, e.g., voluntary head motion, facial expression, and/or vestibular activity. We reduce the effect of such noises by employing filters to the vertical component of the trajectories of each feature point. An 8th order Butterworth band pass filter with cutoff frequency of *[0.75-5.0]* Hz (human HR lies within this range [11]) is used along with a moving average filter defined below:

$$S_n(t) = \frac{1}{w}\sum_{i=-\frac{w}{2}}^{\frac{w}{2}-1} S_n(t+i), where\ \frac{w}{2} < t < T - \frac{w}{2} \qquad (6)$$

where $w$ is the length of the moving average window (length is 300 in our experiment) and $T$ is the total number of frames in the video. These filtered trajectories are then passed to the HR measurement module.

### *D. Heartbeat Rate (HR) Measurement*

As head motions can originate from different sources and only those caused by blood circulation through the aorta reflect the heartbeat rate, we apply a Principal Component Analysis (PCA) algorithm to the filtered trajectories ($S$) to separate the sources of head motion. PCA transforms $S$ to a new coordinate system through calculating the orthogonal components $P$ by using a load matrix $L$ as follows:

$$P = S \,.\, L \qquad (7)$$

where $L$ is a $T \times T$ matrix with columns obtained from the eigenvectors of $S^T S$. Among these components, the most periodic one belongs to heartbeat as obtained in [11]. We apply Discrete Cosine Transform (DCT) to all the components ($P$) to find the most periodic one by following [12]. We then employ Fast Fourier Transform (FFT) on the inverse-DCT of the component and select the first harmonic to obtain the HR.

## IV. Experimental Environment and Datasets

This section describes the experimental environment, evaluates the performance of the proposed system, and compares the performance with the state-of-the-art methods.

### *A. Experimental Environment*

The proposed method was implemented using a combination of Matlab (SDM) and C++ (GFT with KLT) environments. We used three databases to generate results: a local database for demonstrating the effect of FQA, a local database for HR measurement, and the publicly available MAHNOB-HCI database [24]. For the first database, we collected 6 datasets of 174 videos from 7 subjects to conduct an experiment to report the effectiveness of employing FQA in the proposed system. We put four webcams (Logitech C310) at 1, 2, 3, and 4 meter(s) distances to acquire facial video with four different face resolution of the same subject. The room's lighting condition was changed from bright to dark and vice versa for the brightness experiment. Subjects were requested to have around 60 degrees out-of-plan pose variation for the pose experiment. The second database contained 64 video clips by defining three scenarios to constitute our own experimental database for HR measurement experiment, which consists of about 110,000 video frames of about 3,500 seconds. These datasets were captured in two different setups: a) an experimental setup in a laboratory, and b) a real-life setup in a commercial fitness center. The scenarios were:

i. **Scenario 1 (normal):** Subjects exposed their face in front of the cameras without any facial expression or voluntary head motion (about 60 seconds).
ii. **Scenario 2 (internal head motion):** Subjects made facial expressions (smiling/laughing, talking, and angry) in front of the cameras (about 40 seconds).
iii. **Scenario 3 (external head motion):** Subjects made voluntary head motion in different directions in front of the cameras (about 40 seconds).

The third database was the publicly available MAHNOB-HCI database, which has 491 sessions of videos longer than 30 seconds and to which subjects consent attribute 'YES'. Among these sessions, data for subjects '12' and '26' were missing. We collected the rest of the sessions as a dataset for our experiment, which are hereafter called MAHNOB-HCI_Data. Following [5], we use 30 seconds (frame 306 to 2135) from each video for HR measurement and the corresponding ECG signal for the ground truth. TABLE I summarizes all the datasets we used in our experiment.

*B. Performance Evaluation*

The proposed method used a combination of the SDM- and GFT-based approaches for trajectory generation from the facial points. Fig. 3 shows the calculated average trajectories of tracked points in two experimental videos. We included the trajectories obtained from GFT [13], [18] and SDM[16], [17] for facial videos with voluntary head motion. We also included some example video frames depicting face motion. As observed from the figure, the GFT and SDM provide similar trajectories when there is little head motion (video1, Fig. 3(b, c)). When the voluntary head motion is sizable (beginning of video2, Fig. 3(e, f)), GFT-based method fails to track the point accurately and

thus produces an erroneous trajectory because of large optical flow. However, SDM provides stable trajectory in this case, as it does not suffer from large optical flow. We also observe that the SDM trajectories provide more sensible amplitude than the GFT trajectories, which in turn contributes to clear separation of heartbeat from the noise.

TABLE I DATASET NAMES, DEFINITIONS AND SIZES

| No | Name | Definition | Number of data |
|---|---|---|---|
| 1. | Lab_HR_Norm_Data | Video data for HR measurement collected for lab scenario 1. | 10 |
| 2. | Lab_HR_Expr_Data | Video data for HR measurement collected for lab scenario 2. | 9 |
| 3. | Lab_HR_Motion_Data | Video data for HR measurement collected for lab scenario 3. | *10* |
| 4. | FC_HR_Norm_Data | Video data for HR measurement collected for fitness center scenario 1. | *9* |
| 5. | FC_HR_Expr _Data | Video data for HR measurement collected for fitness center scenario 2. | *13* |
| 6. | FC_HR_Motion_Data | Video data for HR measurement collected for fitness center scenario 3. | *13* |
| 7. | MAHNOB-HCI_Data | Video data for HR measurement collected from [24] | *451* |
| 8. | Res1, Res2, Res3, Res4 | Video data acquired from 1, 2, 3 and 4 meter(s) distances, respectively, for FQA experiment | *29x4* |
| 9. | Bright_FQA | Video data acquired while lighting changes for FQA experiment | *29* |
| 10. | Pose_FQA | Video data acquired while pose variation occurs for FQA experiment | *29* |

Unlike [11], the proposed method utilizes a moving average filter before employing PCA on the trajectory obtained from the tracked facial points and landmarks. The effect of this moving average filter is shown in Fig. 4(a). The moving average filter reduces noise and softens extreme peaks in voluntary head motion and provides a smoother signal to PCA in the HR detection process.

The proposed method utilizes DCT instead of FFT of [11] in order to calculate the periodicity of the cyclic head motion signal. Fig. 4(b) shows a trajectory of head motion from an experimental video and its FFT and DCT representations after preprocessing. In the figure we see that the maximum power of FFT is at frequency bin 1.605. This, in turn, gives HR 1.605x60=96.30, whereas the actual HR obtained from ECG was 52.04bpm. Thus, the method in [11] that used FFT in the HR estimation does not always produce good results. On the other hand, using DCT by following [12] yields a result of 52.35bpm from the selected DCT component X=106. This is very close to the actual HR.

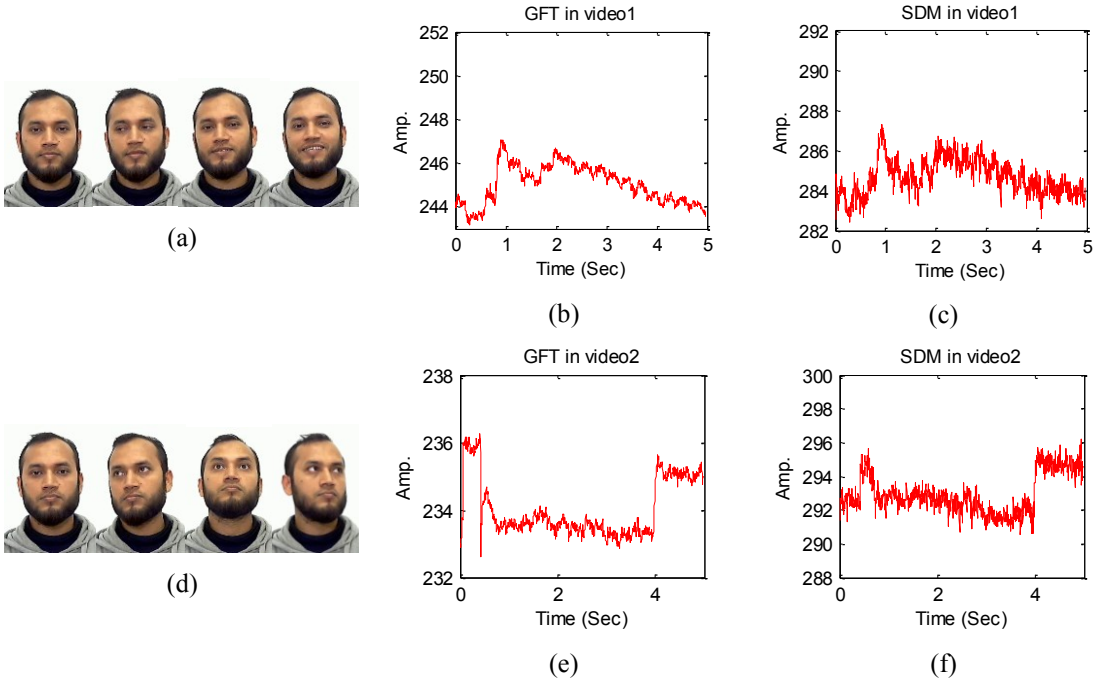



Fig. 3. Example frames depict small motion (in (a)) and large motion (in (d)) from a video, and trajectories of tracking points extracted by GFT [18] (in (b) and (e)) and SDM [17] (in (c) and (f)) from *5* seconds of two experimental video sequences with small motion (video1) and large motion at the beginning and end (video2).

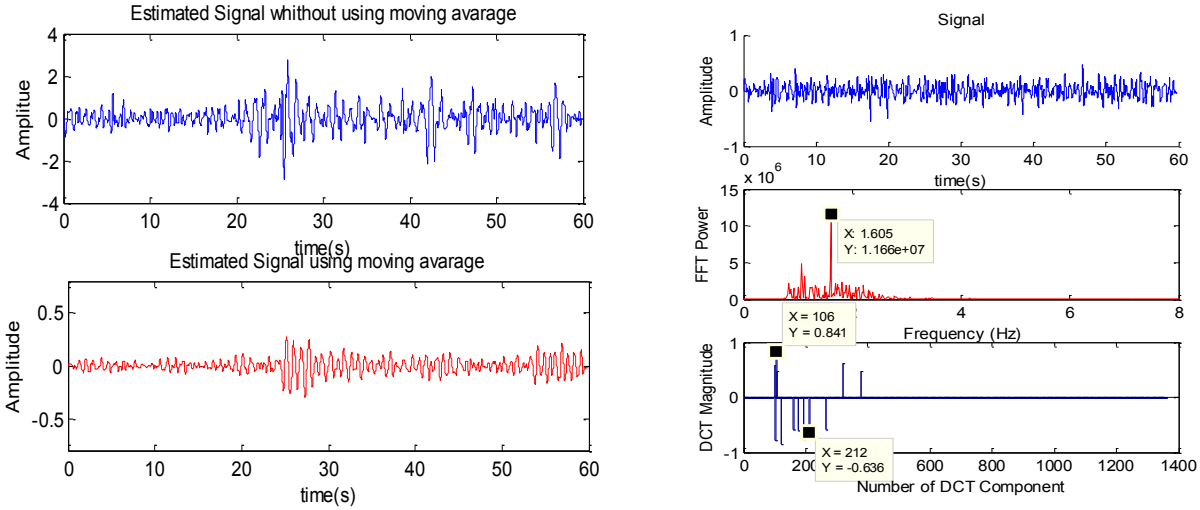



Fig. 4. (a) The effect of the moving average filter on the trajectory of facial points in order to get a smoother signal by noise and extreme peaks reduction and (b) difference between extracting the periodicity (HR) of a cyclic head

motion signal by using FFT power and DCT magnitude.

Furthermore, we conducted an experiment to demonstrate the effect of employing FQA in the proposed system. The experiment had three sections for three quality metrics: resolution, brightness, and out-of-plan pose. The results of HR measurement on six datasets collected for FQA experiment are shown in TABLE II. From the results, it is clear that when resolution decreases the accuracy of the system decreases accordingly. Thus, FQA for face resolution is necessary to ensure a good size face in the system. The results also show that the brightness variation and the pose variation have influence on the HR measurement. We observe that when frames of low quality, in terms of brightness and pose, are discarded the accuracy of HR measurement increases.

TABLE II ANALYZING THE EFFECT OF THE FQA IN HR MEASUREMENTS

| Exp. Name | Dataset | Average percentage (%) of error in HR measurement |
|---|---|---|
| Resolution | Res1 | *10.65* |
| | Res2 | *11.74* |
| | Res3 | *18.86* |
| | Res4 | *37.35* |
| Brightness | Bright_FQA before FQA | *18.77* |
| | Bright_FQA after FQA | *17.62* |
| Pose variation | Pose_FQA before FQA | *17.53* |
| | Pose_FQA after FQA | *14.01* |

TABLE III PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND THE STATE-OF-THE-ART-METHODS OF HR MEASUREMENT ON OUR LOCAL DATABASE

| Dataset name | Average percentage (%) of error in HR measurement | | |
|---|---|---|---|
| | Balakrishnan et al. [11] | Irani et al. [12] | The proposed method |
| Lab_HR_Norm_Data | *7.76* | *7.68* | *2.80* |
| Lab_HR_Expr_Data | *13.86* | *9.00* | *4.98* |
| Lab_HR_Motion_Data | *16.84* | *5.59* | *3.61* |
| FC_HR_Norm_Data | *8.07* | *10.75* | *5.11* |
| FC_HR_Expr_Data | *25.07* | *10.16* | *6.23* |
| FC_HR_Motion_Data | *23.90* | *15.16* | *7.01* |

*C. Performance Comparison*

We have compared the performance of the proposed method against state-of-the-art methods from [3], [5], [6], [11], [12] on the experimental datasets listed in TABLE I. TABLE III lists the accuracy of HR measurement results of the proposed method in comparison with the motion-based state of the art methods [11], [12] on our local database. We have measured the accuracy in terms of percentage of measurement error. The lower the error generated by a method, the higher the accuracy of that method. From the results we observe that the proposed method showed consistent performance, although the data acquisition scenarios were different for different datasets. By using both GFT and SDM trajectories, the proposed method gets more trajectories to estimate the HR pattern in the case of HR_Norm_Data and accurate trajectories due to non-missing facial points in the cases of HR_Expr_Data and HR_Motion_Data. On the other hand, the previous methods suffer from fewer trajectories and/or erroneous trajectories from the data acquired in challenging scenarios, e.g. Balakrishnan's method showed an up to 25.07% error in HR estimation from videos having facial expression change. The proposed method outperforms the previous methods in both environments (lab and in a fitness center) of data acquisition, including all three scenarios.

TABLE IV PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND THE STATE-OF-THE-ART-METHODS OF HR MEASUREMENT ON MAHNOB-HCI DATABASE

| Method | RMSE (bpm) | Mean error rate (%) |
|---|---|---|
| Poh et al. [3] | *25.90* | *25.00* |
| Kwon et al. [7] | *25.10* | *23.60* |
| Balakrishnan et al. [11] | *21.00* | *20.07* |
| Poh et al. [6] | *13.60* | *13.20* |
| Li et al. [5] | *7.62* | *6.87* |
| Irani et al. [12] | *5.03* | *6.61* |
| The proposed method | *3.85* | *4.65* |

TABLE IV shows the performance comparison of HR measurement by our proposed method and state-of-the-art methods (both color-based and motion-based) on MAHNOB-HCI_Data. We calculate the Root Mean Square Error (RMSE) in beat-per-minute (bpm) and mean error rate in percentage to compare the results. From the results we can observe that Li's [5], Irani's [12], and the proposed method showed considerably higher results than the other methods because they take into consideration the presence of voluntary head motion in the video. However, unlike Li's color-based method, Irani's method and the proposed method are motion-based. Thus, changing the illumination condition in MAHNOB-HCI_Data does not greatly affect the motion-based methods, as indicated by the results. Finally, we observe that the proposed method outperforms all these state-of-the-art methods in the

accuracy of HR measurement.

## V. CONCLUSIONS

This paper proposes a system for measuring HR from facial videos acquired in more realistic scenarios than the scenarios of previous systems. The previous methods work well only when there is neither voluntary motion of the face nor change of expression and when the lighting conditions help keeping sufficient texture in the forehead and cheek. The proposed method overcomes these problems by using an alternative facial landmarks tracking system (the SDM-based system) along with the previous feature points tracking system (the GFT-based system) and provides competent results. The performance of the proposed system for HR measurement is highly accurate and reliable not only in a laboratory setting with no-motion, no-expression cases in artificial light in the face, as considered in [11], [12], but also in challenging real-life environments. However, the proposed system is not adapted yet to the real-time application for HR measurement due to dependency on temporal stability of the facial point trajectory.

## REFERENCES


[1] J. Klonovs, M. A. Haque, V. Krueger, K. Nasrollahi, K. Andersen-Ranberg, T. B. Moeslund, and E. G. Spaich, *Distributed Computing and Monitoring Technologies for Older Patients*, 1st ed. Springer International Publishing, 2015.

[2] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian Video Magnification for Revealing Subtle Changes in the World," *ACM Trans Graph*, vol. 31, no. 4, pp. 65:1–65:8, Jul. 2012.

[3] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Opt. Express*, vol. 18, no. 10, pp. 10762–10774, May 2010.

[4] H. Monkaresi, R. . Calvo, and H. Yan, "A Machine Learning Approach to Improve Contactless Heart Rate Monitoring Using a Webcam," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, pp. 1153–1160, Jul. 2014.

[5] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote Heart Rate Measurement From Face Videos Under Realistic Situations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 4321–4328.

[6] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[7] 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 *2012 Annual International Conference of the IEEE Engineering in Medicine and*

*Biology Society (EMBC)*, 2012, pp. 2174–2177.

[8] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Heartbeat Signal from Facial Video for Biometric Recognition," in *Image Analysis*, R. R. Paulsen and K. S. Pedersen, Eds. Springer International Publishing, 2015, pp. 165–174.

[9] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Can contact-free measurement of heartbeat signal be used in forensics?," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015, pp. 1–5.

[10] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Efficient contactless heartbeat rate measurement for health monitoring," *Internatinal J. Integr. Care*, vol. 15, no. 7, pp. 1–2, Oct. 2015.

[11] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting Pulse from Head Motions in Video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3430–3437.

[12] R. Irani, K. Nasrollahi, and T. B. Moeslund, "Improved Pulse Detection from Head Motions Using DCT," in *9th International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014, pp. 1–8.

[13] J. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker," *Intel Corp. Microprocess. Res. Labs*, 2000.

[14] A. D. Bagdanov, A. Del Bimbo, F. Dini, G. Lisanti, and I. Masi, "Posterity Logging of Face Imagery for Video Surveillance," *IEEE Multimed.*, vol. 19, no. 4, pp. 48–59, Oct. 2012.

[15] K. Nasrollahi and T. B. Moeslund, "Extracting a Good Quality Frontal Face Image From a Low-Resolution Video Sequence," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 10, pp. 1353–1362, Oct. 2011.

[16] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Quality-Aware Estimation of Facial Landmarks in Video Sequences," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015, pp. 1–8.

[17] X. Xiong and F. De la Torre, "Supervised Descent Method and Its Applications to Face Alignment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 532–539.

[18] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.

[19] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[20] A. U. Batur and M. H. Hayes, "Adaptive active appearance models," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1707–1721, Nov. 2005.

[21] Y. Feng, J. Xiao, Y. Zhuang, X. Yang, J. J. Zhang, and R. Song, "Exploiting temporal stability and low-rank structure for motion capture data refinement," *Inf. Sci.*, vol. 277, pp. 777–793, Sep. 2014.

[22] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int J Comput Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[23] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Real-time acquisition of high quality face sequences

from an active pan-tilt-zoom camera," in *10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2013, pp. 443–448.

[24] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A Multimodal Database for Affect Recognition and Implicit Tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

## AUTHORS' BIOGRAPHIES

**Mohammad A. Haque** is a PhD fellow at the Visual Analysis of People (VAP) Lab, Aalborg University (AAU). He is interested in vision-based patient monitoring, biometrics, and decision support systems.

**Ramin Irani** is a PhD fellow at the VAP Lab, AAU. His research interests include facial expression recognition, social cue recognition, and soft biometrics.

**Kamal Nasrollahi** is an associate professor at the VAP Lab, AAU. His research interests include facial analysis systems, biometrics recognition, soft biometrics, and inverse problems.

**Thomas B. Moeslund** is the head of the VAP Lab and Media Technology Section, AAU. His research focuses on all aspects of the automatic analysis of images and video data.