PPG3D: Does 3D head tracking improve camera-based PPG estimation?

Genki Nagamatsu[1], Ewa Magdalena Nowara[2], Amruta Pai[2], Ashok Veeraraghavan[2] and Hiroshi Kawasaki[1]

Abstract— **Over the last few years, camera-based estimation of vital signs referred to as imaging photoplethysmography (iPPG) has garnered significant attention due to the relative simplicity, ease, unobtrusiveness and flexibility offered by such measurements. It is expected that iPPG may be integrated into a host of emerging applications in areas as diverse as autonomous cars, neonatal monitoring, and telemedicine. In spite of this potential, the primary challenge of non-contact camera-based measurements is the relative motion between the camera and the subjects. Current techniques employ 2D feature tracking to reduce the effect of subject and camera motion but they are limited to handling translational and in-plane motion. In this paper, we study, for the first-time, the utility of 3D face tracking to allow iPPG to retain robust performance even in presence of out-of-plane and large relative motions. We use a RGB-D camera to obtain 3D information from the subjects and use the spatial and depth information to fit a 3D face model and track the model over the video frames. This allows us to estimate correspondence over the entire video with pixel-level accuracy, even in the presence of out-of-plane or large motions. We then estimate iPPG from the warped video data that ensures per-pixel correspondence over the entire window-length used for estimation. Our experiments demonstrate improvement in robustness when head motion is large.**

I. INTRODUCTION

The heartbeat is a fundamental and important signal among biological signals, so techniques of its measurement have been researched since ancient times. In the conventional method of measuring the heartbeat, electrodes are put on the skin and the electrocardiogram (ECG) is measured. This method can measure the heartbeat accurately, however it requires direct contact with the skin which often causes discomfort and stress. Recently, methods of measuring the heartbeat without touching the skin have received attention. Among the contactless heartbeat measurements, imaging photoplethysmography (iPPG) has received much attention. iPPG is the technique of measuring the pulse waveform based on the optical properties of the human skin. The principle is that the light absorption in skin changes with changing concentration of hemoglobin in the blood with each heartbeat [13], [15]. In these methods, the heartbeat is usually estimated by measuring the changing intensities on the face. However, it is difficult to track the face accurately, especially in presence of motion. It is known that the accuracy of the iPPG is significantly affected by the quality of the facial regions in the video; cheeks and forehead are regions where iPPG can usually be measured with high accuracy [12]. However, there are few texture features to track in those regions and the intensity information changes by moving the face slightly. Thus, it is difficult to track these regions accurately. To solve the tracking problem, we propose a method of tracking the face with 3D information captured by the RGB-D camera. Using depth images, we aim to measure iPPG more accurately than the state-of-the-art methods using only 2D RGB images. We captured recordings 1) with subjects sitting as still as possible and 2) with subjects moving their head during the measurements. We compared the performance with the proposed 3D method and a benchmark method using only 2D images and demonstrated the effectiveness of our proposed method for heart rate (HR) estimation.

[1]Faculty of Information Engineering, Kyushu University, Japan
[2]Rice University, Houston, TX, USA

II. RELATED WORK

The iPPG is a cheap and simple technique to measure vital signs from video, however it is very prone to motion artefacts. Many iPPG methods have been proposed using RGB cameras. Poh *et al.* [14] proposed a HR measurement method using Independent Component Analysis (ICA) on the RGB intensity sequences obtained from a facial region of interest (ROI). Wang *et al.* [16] proposed a method to improve motion robustness by sampling multiple regions simultaneously using multiple pixels. In this method, the CSK tracking is adopted as an object tracking method. Since these methods use only images for feature tracking, these methods are not robust to changes in image intensity. When subjects move a lot, or occlude a part of their face with their hands, their faces may not be detected correctly and it is hard to measure their HR.

Recently, some iPPG methods using the RGB-D camera have been proposed. Gamni *et al.* [7] proposed a method using MicroSoft Kinect v2 (MS kinect2) which is an RGB-D camera. In this method, the HR was measured using the image sequences obtained from the RGB camera of the MS Kinect2. However, the KLT feature tracking algorithm was used to track the feature points, and tracking was not performed using depth information. Bakhtiyari *et al.* [3] also proposed a method for improving the HR measurement accuracy by measuring respiratory signals using an MS Kinect2. In this method, HR was measured using both RGB and depth image sequences. However, the depth image sequences were only used to estimate the HR signals and respiratory signals, and the depth information was not used to improve the tracking of the feature points. Therefore, these methods using an RGB-D camera did not use 3D information effectively for motion robustness.

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

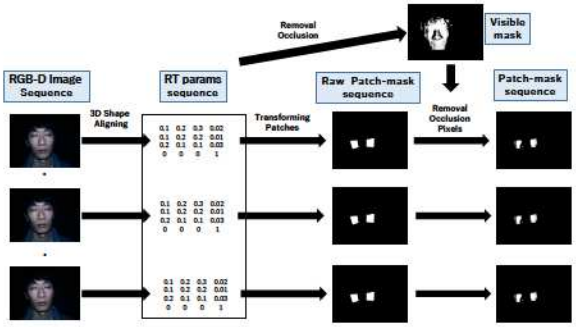

Fig. 1. Overview of the proposed tracking method. First, rigid transformation parameters are obtained by rigid 3D shape alignment. Next, the occlusion area is calculated using the estimated parameters and a visible-mask is created. Patches are tracked using a mask created by rigidly transforming the 3D point cloud of the patch region and projecting it on the image plane.

III. Methods

We propose a method to improve the iPPG robustness to motion by using the depth information to obtain 3D point cloud registration for more accurate face tracking. Fig. 1 shows the overview of the proposed method. In the proposed method, the iPPG is extracted from the RGB-D image sequence. Rigid transformation parameters are estimated by the iterative closest point (ICP) algorithm [4]. The visible mask, which is a mask of pixels that can be observed in all video frames, is generated using the estimated rigid transformation parameters from all frames. The patch mask is created by projecting the rigidly transformed patch onto the plane using the transformation parameters and the visible mask. We use the patch masks to compute the intensity sequences from which we obtain iPPG signals. One frame is defined as a reference frame. The point cloud data of the reference frame is used as a reference point cloud for registration. In the reference frame, a patch mask is set and a reference patch point cloud is also created. The patches can be tracked by rigidly transforming the patches' point clouds using the estimated rigid transformation parameters and projecting them onto the image plane.

A. The 3D tracking method

To align the reference frame point cloud and the point clouds from other frames, we applied the ICP algorithm to the point clouds. The rigid transformation parameters were estimated with the ICP algorithm, where the corresponding point search distance was multi-scale. Multi-scale ICP allows to initially align the global range and then to align the local range. In each stage of the multi-scale ICP, the method of Besl *et al.* [4] was used. The initial rigid transformation parameters in the ICP of each frame used the rigid transformation parameters of the previous frame. Face alignment was implemented with [17].

B. Removal of occlusions in the face area

Because the reference patch used for tracking in the proposed method is fixed, when the face of the subject moves, some parts of the reference patch may be occluded. Occlusions make the intensity sequences noisy, reducing the accuracy of HR estimation. Therefore, it is necessary to remove the occluded areas. After estimating the rigid transformation parameters for each frame, we applied the rigid transformation to the reference frame point clouds using the estimated parameters. If any points in the transformed reference point clouds were hidden by other points as viewed from the camera position, those points were removed because they were occluded. To determine whether a point is occluded with other points, we used a method based on [9]. By using this method to remove the occluded points from the rigid transformation parameters from all frames, only the points that could be observed in all frames were extracted. We obtained the visible mask by projecting the extracted points onto an image plane.

C. Creating observation patches

We averaged the RGB intensity values within each patch for each frame. To create the patches, we used the 68 detected facial landmarks. The position of the patches on the face is very important. Forehead and cheeks are known to be good regions for extracting the iPPG signals [12]. In this study, the forehead regions were covered by the hair for most subjects, so we only used the patches on the cheeks. In the reference frame, the raw reference patch-mask was calculated with facial landmarks. Pixels observed in both the raw reference patch mask and the visible mask were taken as a reference patch mask. The raw patch mask of each frame was calculated by a rigid transformation in the 3D space with estimated parameters and projecting the transformed patch to the image plane. Then, the patch mask was complemented by morphological transformation to fill in the holes in raw patch-mask. To estimate HR we used the spatially averaged intensity values of the pixels inside the patch masks.

D. Estimating HR from iPPG signals

First, we spatially averaged the pixel intensities within each patch on the face for each RGB camera channel. Then, we used ICA to decompose the normalized RGB intensity sequences into three independent signals [14]. We used the FastICA [2] to calculate the ICA. The output independent signals are not ordered, therefore, we manually selected the component that was the most periodic. If the independent components were similar, we computed the Fast Fourier transform (FFT) of each independent component and selected the one with the highest ratio between the first and second maximum peaks.

After selecting the independent component, we applied the FFT to the selected independent component. Since the frequency components obtained by FFT are discrete values, we interpolated them with quadratic spline interpolation. We computed the estimated HR as the frequency with the maximum power spectrum multiplied by 60 to convert it from Hertz (Hz) to beats-per-minute (bpm).

Fig. 2. Experimental setup.

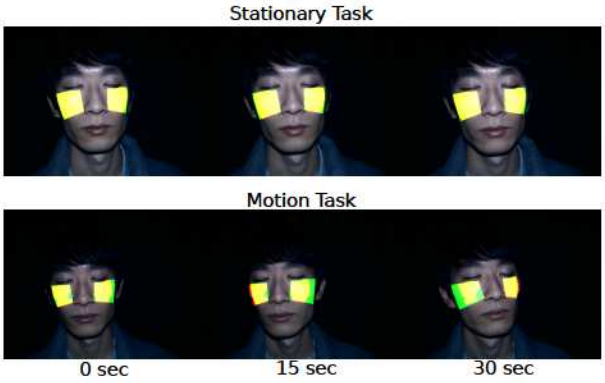

Fig. 3. Examples of RGB images captured during the stationary and motion tasks with the patches tracked with 2D (in green) and 3D (in red) tracking. The overlap area between the 2D and 3D tracking is shown in yellow.

IV. Experiments

We compared the HR estimation accuracy using the proposed 3D tracking of the facial landmarks from depth information to a benchmark method using only 2D images and facial landmark detection. Fig. 2 shows the experimental setup. The RGB-D sequences were recorded using Intel Realsense SR300 [8]. The frame rate of both RGB and Depth cameras was 30 frames per second (fps) with 640 x 480 pixel image resolution. We placed a band-pass filter on the RGB camera with a passband of 400 nm to 700 nm to block any NIR structured light patterns projected by the Intel Realsense SR300 for estimating depth. Because the iPPG signal is a very weak intensity signal, varying ambient light will introduce noise, for example when an AC light source is used. To avoid this problem, we captured the RGB-D sequences in a darkroom and we used illumination with a DC power source. We recorded the ECG signals as the ground truth [1] synchronized with the RGB-D image capture. We bandpass filtered the iPPG signals obtained from RGB sequences in a physiological range of [0.75 Hz, 4.0 Hz].

The comparison method is a method that detects facial landmarks from RGB images. First, we detected the face ROI [11] and facial landmarks [10]. The method of detecting the face ROI is called Max-Margin Object Detection. In this method, the face ROI detection is realized by SVM learning using the HOG (Histogram of Oriented Gradients) features[5]. The method of facial landmarks is using an ensemble of regression trees. In this method, the average face shape is used as the initial position of the feature points, the shift values of the feature point are estimated based on the image feature, and the landmarks are detected by repeatedly shifting the landmark points. These detection methods were implemented in the dlib library [1]. After the facial landmarks are detected in each frame, the patches are created using the aforementioned method and the RGB intensity sequences are computed. We captured two types of experiments, referred to as "stationary" and "motion" tasks with 7 subjects each [2]. During the stationary task the subjects were asked to sit as still as possible. During the motion task the subjects were asked to move their head out of plane horizontally by about 30 degrees (see Fig. 3).

We evaluated our results using two error measures. 1) mean absolute error (MAE) computed as the mean absolute difference between the estimated and ground truth heart rate. The HR was computed from 10 second time windows with one second overlap. 2) signal-to-noise-ratio (SNR) defined as:

$$SNR = 10\log_{10}\left(\frac{\sum_{0.7}^{4}((U_t(f)S(f))^2}{\sum_{0.7}^{4}((1-U_t(f))S(f))^2}\right) \quad (1)$$

where S is the power spectrum of the estimated iPPG signal, f is the frequency in Hz and $U_t(f)$ is equal to one for frequencies around the first and second harmonic of the ground truth HR (HR - 0.1 Hz bpm to HR - 0.1 Hz and 2*HR - 0.1 Hz to 2*HR + 0.1 Hz), and zero everywhere else [6].

A. Results: stationary task

Fig. 3 shows the results of tracking patches with our proposed method and the benchmark method during the stationary and motion tasks. Because there isn't a lot of motion, the estimated HR with the proposed method and the benchmark method are comparable and the MAEs are almost the same. However, the SNR is higher for our proposed method because we are able to obtain cleaner iPPG signals than the benchmark method.

B. Results: motion task

We obtain lower MAE and higher SNR with our proposed method, demonstrating the effectiveness of 3D tracking in presence of motion. However, the head motion in these experiments was small and controlled. Consequently, the improvements offered by 3D over 2D are subtle.

V. Discussion

In the experiment, it took 1000 seconds on average for both the registration calculation and the patch calculation. The maximum memory required for calculation was about

[1] https://store.healthcare.omron.co.jp/category/8/HCG8$01_SET.html$

[2] The experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board.

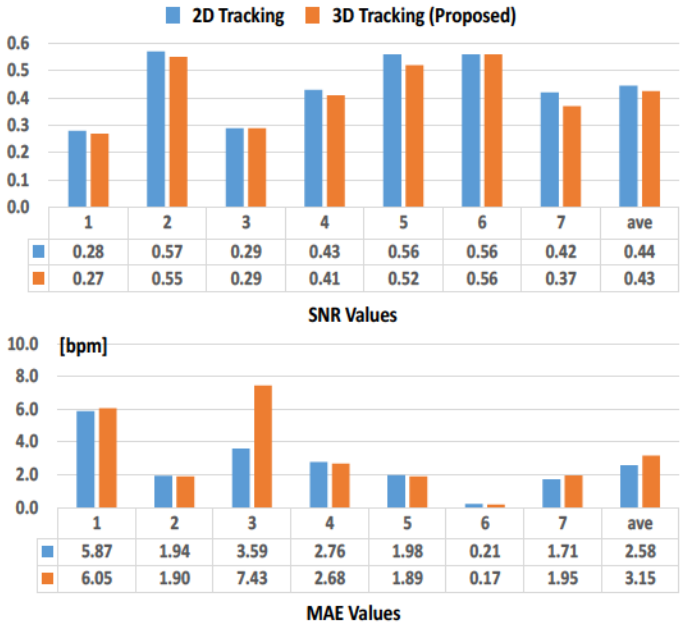

Fig. 4. The SNR values and MAE values during the stationary task.

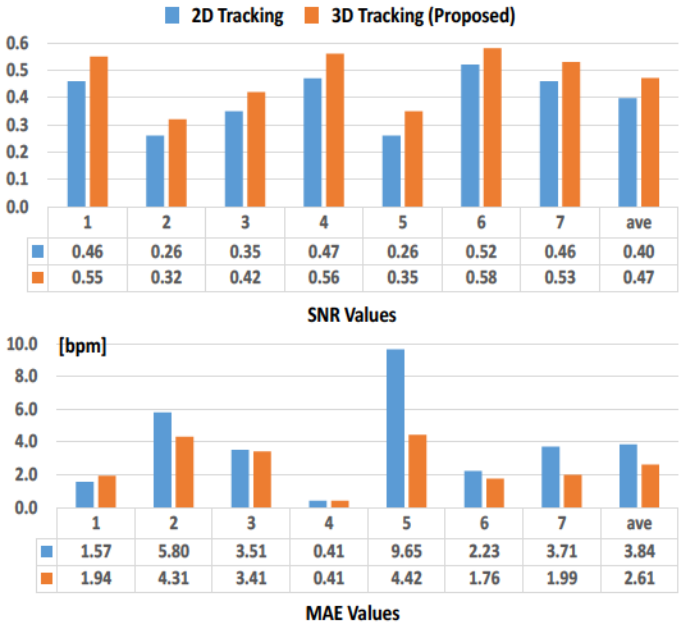

Fig. 5. The SNR values and MAE values during the motion task.

160 MB. The CPU used in the experiment was Intel Core i9-9900k CPU. Since the proposed method uses 3D information, it takes a lot of time for calculation. Thus, it is impossible to run this method in real time. It is a future task to reduce the calculation time and memory consumption to make it a practical method.

VI. Conclusions

We presented a face tracking method using 3D information to improve HR estimation from video in presence of motion. We compared the MAE and SNR of our 3D tracking method to the state-of-the-art 2D tracking method. Our 3D tracking performs modestly better than the 2D tracking. However, the dataset that we collected only had subjects sitting still or moving only slightly, making it difficult to demonstrate the advantages of the proposed 3D tracking. More experiments are needed to conclude how much improvement in motion robustness is offered by 3D tracking over 2D tracking. Nonetheless, the 3D tracking approach may open up possibilities of faithfully computing iPPG signals in presence of large motion, where current methods often fail, for example in a driving or a fitness context.

VII. Acknowledgement

This work was supported by JSPS/KAKENHI 20H00611, 18K19824, 18H04119, 16KK0151 in Japan.

References

[1] "Dlib library," http://dlib.net/.

[2] H. A. and O. E., "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411–430, may 2000. [Online]. Available: https://ci.nii.ac.jp/naid/10008962170/en/

[3] K. Bakhtiyari, N. Beckmann, and J. Ziegler, "Contactless heart rate variability measurement by ir and 3d depth sensors with respiratory sinus arrhythmia," in *ANT/SEIT*, 2017.

[4] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, Feb 1992.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.

[6] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[7] E. Gambi, A. Agostinelli, A. Belli, L. Burattini, E. Cippitelli, S. Fioretti, P. Pierleoni, M. Ricciuti, A. Sbrollini, and S. Spinsante, "Heart rate detection using microsoft kinect: Validation and comparison to wearable devices," in *Sensors*, 2017.

[8] Intel, "Intel realsense camera sr300," https://software.intel.com/en-us/realsense/sr300.

[9] S. Katz, A. Tal, and R. Basri, "Direct visibility of point sets," in *ACM Transactions on Graphics*, vol. 26, 07 2007.

[10] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, 2014.

[11] D. E. King, "Max-margin object detection," *ArXiv*, vol. abs/1502.00046, 2015.

[12] S. Kwon, J. Kim, D. Lee, and K. S. Park, "Roi analysis for remote photoplethysmography on facial video," *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4938–4941, 2015.

[13] M. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, Jan 2011.

[14] M.-Z. Poh, D. McDuff, and R. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics express*, vol. 18, pp. 10 762–74, 05 2010.

[15] H. Rahman, M. U. Ahmed, S. Begum, and P. Funk, "Real time heart rate monitoring from facial rgb color video using webcam," in *The 29th Annual Workshop of the Swedish Artificial Intelligence Society*, June 2016. [Online]. Available: http://www.es.mdh.se/publications/4354-

[16] W. Wang, S. Stuijk, and G. de Haan, "Exploiting spatial redundancy of image sensor for motion robust rppg," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 415–425, Feb 2015.

[17] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.