Remote Heart Rate Estimation Based on 3D Facial Landmarks

Yuichiro Maki, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi

***Abstract*— In this paper, we propose a novel video-based remote heart rate (HR) estimation method based on 3D facial landmarks. The key contributions in our method are twofold: (i) We introduce 3D facial landmarks detection to the video-based HR estimation and (ii) we propose a novel face patch visibility check manner based on the face patch normal in the 3D space. We experimentally demonstrate that, compared with baseline methods using 2D facial landmarks, our proposed method using 3D facial landmarks improves the robustness of HR estimation to head rotations and partial face occlusion. We also demonstrate that our visibility check is effective for selecting sufficiently visible face patches, contributing to the improvement of HR estimation accuracy.**

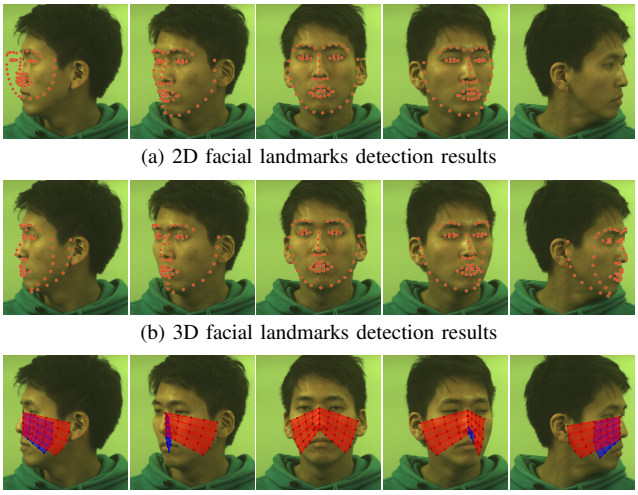
I. Introduction

Video-based heart rate (HR) estimation has attracted increasing attention for its non-contact manner, which enables various remote vital sensing applications such as neonate monitoring [1] and telemedicine [2]. The video-based HR estimation is based on the principles of photoplethysmography (PPG) and attempts to extract subtle temporal skin color change due to blood volume pulse (BVP) under the skin [3], [4]. Face is most commonly used as region of interest (ROI) for video-based HR estimation methods because facial skin is usually exposed. Thus, robustly detecting and tracking the face ROI is of great importance for the video-based methods to accurately estimate HR.

Initial video-based HR estimation methods apply face detection and determine the ROI at the first video frame [5]. Since the ROI is not tracked between successive frames, these initial methods are susceptible to head movements. To improve the robustness, many methods apply 2D facial landmarks detection to track the ROI between frames [6]. Furthermore, some methods divide the face ROI into face patches using the 2D landmarks and locally select suitable patches based on the reliability for each patch [7]–[10].

While 2D facial landmarks detection algorithms are robust to head translations, they are relatively weak to head rotations due to partial face occlusion (see Fig. 1(a)). In contrast, the recent progress of computer vision technology has witnessed that 3D facial landmarks detection algorithms [11], which predict facial landmarks in the 3D coordinate (i.e., 2D image coordinate (x, y) and 1D depth coordinate z), are very robust to head rotations (see Fig. 1(b)). However, to the best of our knowledge, none of the existing studies have investigated the application of 3D facial landmarks detection to the video-based HR estimation.

Y. Maki, Y. Monno, M. Tanaka, and M. Okutomi are with the Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan (e-mail: ymaki@ok.sc.e.titech.ac.jp; ymonno@ok.sc.e.titech.ac.jp; mtanaka@sc.e.titech.ac.jp; mxo@sc.e.titech.ac.jp).



(a) 2D facial landmarks detection results

(b) 3D facial landmarks detection results

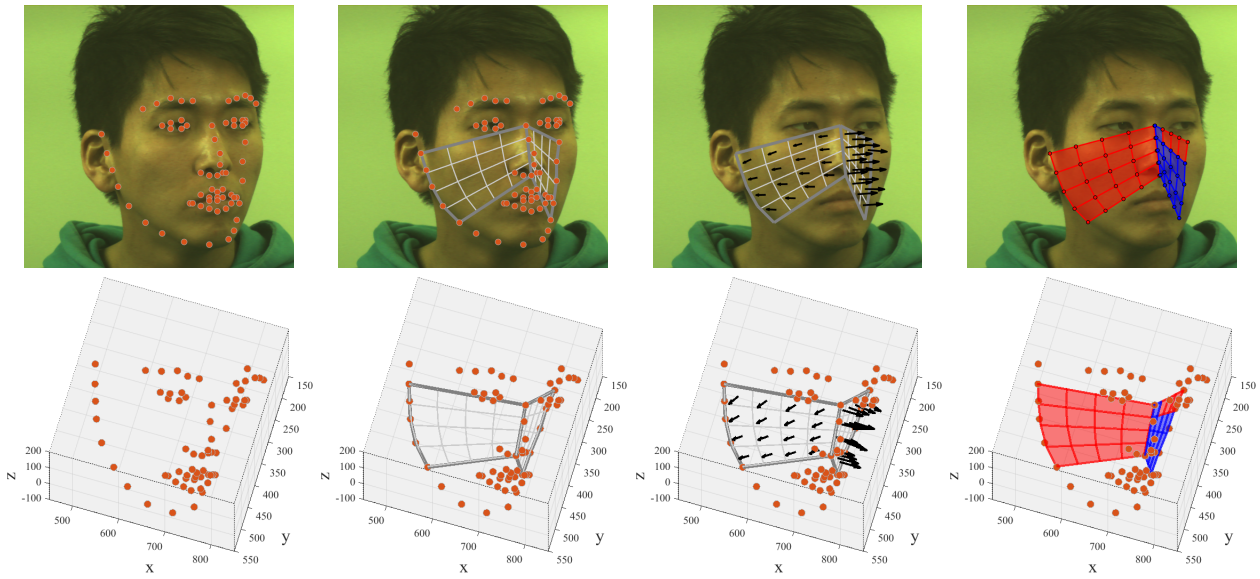
(c) Our visibility check results using the 3D landmarks

Fig. 1. Comparison of (a) 2D facial landmarks detection [12] and (b) 3D facial landmarks detection [13]. The 2D landmarks detection inaccurately estimates or completely misses the landmarks if the head rotation angle is large, while the 3D landmarks detection is much robust to the head rotations. (c) Our visibility check results using the 3D landmarks, where red patches are regarded as visible.

In this paper, we propose a novel video-based HR estimation method based on a new face patch visibility check manner using 3D facial landmarks. In our method, we first detect 3D facial landmarks and divide the face cheek regions into face patches using the 3D landmark positions. We then evaluate the visibility of each patch based on the patch normal, where the patch is considered as visible if the angle between the patch normal and the z-axis (corresponding to the depth direction) is less than a threshold (see Fig. 1(c) for example visibility check results). We introduce the visibility check to disregard the reversed patches (e.g., blue patches in the left and right edge pictures of Fig. 1(c)) and to select the patches parallel to the image plane (according to the threshold), which are expected as more reliable. We finally estimate HR using only the patches that are visible during the whole considered time window. To the best of our knowledge, our method is the first method that incorporates the 3D facial landmark detection and applies the face patch visibility check using the 3D information. We experimentally demonstrate that our proposed method improves the accuracy and the robustness of HR estimation compared with baseline methods using 2D facial landmarks detection.

II. Proposed Method

In this section, we explain our proposed HR estimation method based on the 3D facial landmarks detection and the face patch visibility check using the patch normal.



(a) Detected 3D landmarks (b) Generated face patches (c) Calculated patch normals (d) Visibility check results

Fig. 2. Step-by-step results of our visibility check algorithm: (a) Detected 3D facial landmarks, (b) generated face patches for the left and right cheek regions, (c) calculated normals for each patch, and (d) visibility check results, where red patches are regarded as visible, i.e., the angle between the patch normal and the z-axis is less than the threshold (75 degrees in this case).

A. 3D facial landmarks detection

We first detect 3D facial landmarks using the method in [13]. As shown in Fig. 2(a), for each video frame, this method detects a set of 68 landmarks in the 3D coordinate (x, y, z), where (x, y) corresponds to the 2D image coordinate and z corresponds to the 1D depth coordinate. We experimentally found that the landmark positions detected by the method [13] fluctuate between successive frames, even if the head movement between the frames is small. This is because this method does not use any temporal information. To temporally stabilize each landmark position, we perform temporal smoothing as

$$\tilde{\mathbf{L}}_t(x,y,z) = \frac{\sum_{t=t-N}^{t+N} \mathbf{L}_t(x,y,z)}{2N+1}, \tag{1}$$

where $\mathbf{L}_t(x,y,z)$ is the landmark position of t-th frame, $\tilde{\mathbf{L}}_t(x,y,z)$ is the smoothed landmark position, and N is the range of adjacent frames used for the smoothing ($N = 3$ is used for our experiments).

B. Face patch generation

We next generate local face patches in the 3D space. As shown by the bold gray lines in Fig. 2(b), we use the left and the right cheek regions, where each region is defined by the seven landmarks. We divide each cheek region into 3D patches by the 4×4 uniform grid, as shown by the thin gray lines in Fig. 2(b). A total of 32 patches are generated from both cheek regions.

C. Face patch visibility check

We then check the visibility of each patch. To evaluate the visibility, we calculate the angle between the patch normal and the z-axis as

$$\alpha = \arccos(\vec{n_p} \cdot \vec{n_z}), \tag{2}$$

where $\vec{n_p}$ is the normalized vector representing the patch normal, which is shown by a black arrow in Fig. 2(c), $\vec{n_z}$ is the normalized vector corresponding to the z-axis, i.e., $\vec{n_z} = [0,0,-1]^T$, and α is the angle between the two vectors. We regard that the patch is visible if the angle α is less than a threshold. Figure 2(d) shows the visibility check result using the threshold value of 75 degrees, where red patches represent visible patches. In Section III-C, we will investigate the effect of the threshold value in more detail.

D. HR estimation

Based on the frame-by-frame face patch visibility check as explained above, we estimate HR. We first derive the visible patch set for a considered time window. If a patch is visible at all video frames in the time window, that patch is included to the visible patch set. By checking the visibility of all 32 patches, we obtain the visible patch set, which is used for averaging pixel intensities in the next HR estimation process.

We then follow the widely applied HR estimation pipeline by Poh et al. [5] to estimate HR. Firstly, for each frame, the averaged R, G, and B values are calculated by averaging each of R, G, and B values of all the pixels belonging to the visible patch set. Then, the RGB channel temporal intensity traces are derived by concatenating the averaged values of all video frames. Independent component analysis (ICA) is then applied to the RGB intensity traces to separate the BVP component and the other two components considered as noise. The BVP component is then selected among the three components of the ICA output based on the frequency information, where the component that has the most strong peak power within the range of heartbeat frequency [0.7Hz, 4Hz] is selected as the BVP component. Finally, HR is calculated using the most dominant frequency of the selected BVP component.

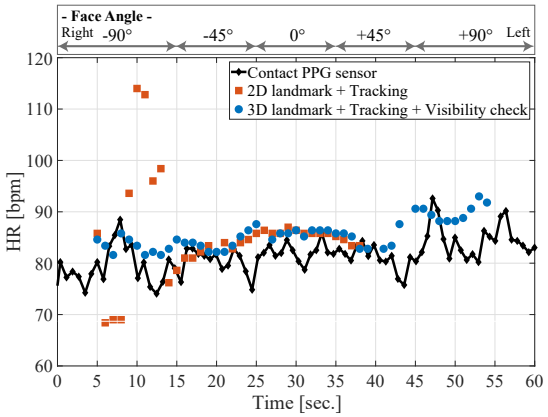


Fig. 3. Comparison of "2D Landmark + Tracking (Baseline)" and "3D Landmark + Tracking + Visibility Check (Proposed)." The baseline method fails to estimate HR in [0, 15] and [45, 60] (sec.) because of the failure of landmark detection due to large head rotations, while our proposed method achieves better estimation accuracy.

III. Experiments

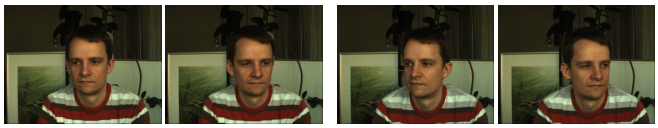
A. Evaluation procedure

We used public PURE dataset [14] and our demonstrative data for evaluating our method. The duration of all used videos is 60 seconds. From each video, we extracted 10-seconds video clips by a sliding time window manner with one-second intervals. As a result, a total of 50 clips was obtained from each 60-seconds video. Then, we estimated HR for every 10-seconds clip, where the timestamp of each estimated HR was assigned to the center of the time window. Ground-truth HR was derived from the contact PPG sensor data in a continuous manner based on inter-beat intervals of the detected peaks. The timestamp of each ground-truth HR was assigned to the later peak timestamp. Then, the ground-truth HR series were linearly interpolated to re-sample to the timestamp of each video-based estimated HR for evaluation.

We compared four methods. The first two methods use 2D facial landmarks with or without ROI tracking. In the case without tracking, the ROI pixels are determined at the first frame and fixed during all frames. In the case with tracking, the ROI is tracked using relative 2D landmark positions. The third method applies 3D facial landmarks tracking but does not apply our proposed visibility check. The last one is our proposed method using 3D facial landmarks tracking with visibility check. For fairly comparing all methods, we used the same left and right cheek ROIs, which were determined using the same 2D or 3D landmark indexes, and also applied the same HR estimation pipeline. For the 2D landmarks detection, we used the method of [12] implementation by [15].

B. 2D vs. 3D facial landmarks detection

We first show the effectiveness of applying the 3D facial landmarks detection using one demonstrative video that contains very large head rotations. Figure 1 shows five example video frames, where, from left to right, the face angles are referred to as $(-90, -45, 0, +45, +90)$, respectively. During the 60-seconds video, each face angle was kept as shown in the top of Fig. 3.



(a) Small rotation (b) Medium rotation

Fig. 4. Example frames of PURE dataset [14] in the two situations.

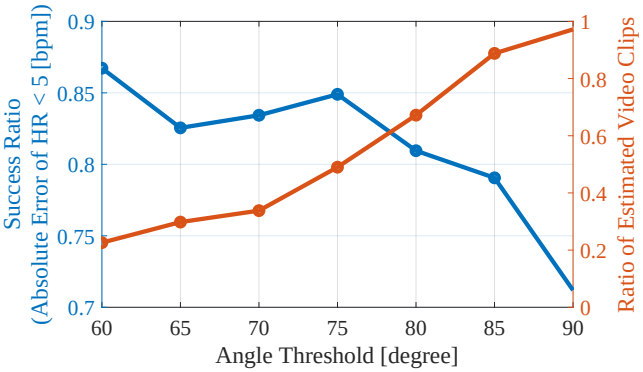


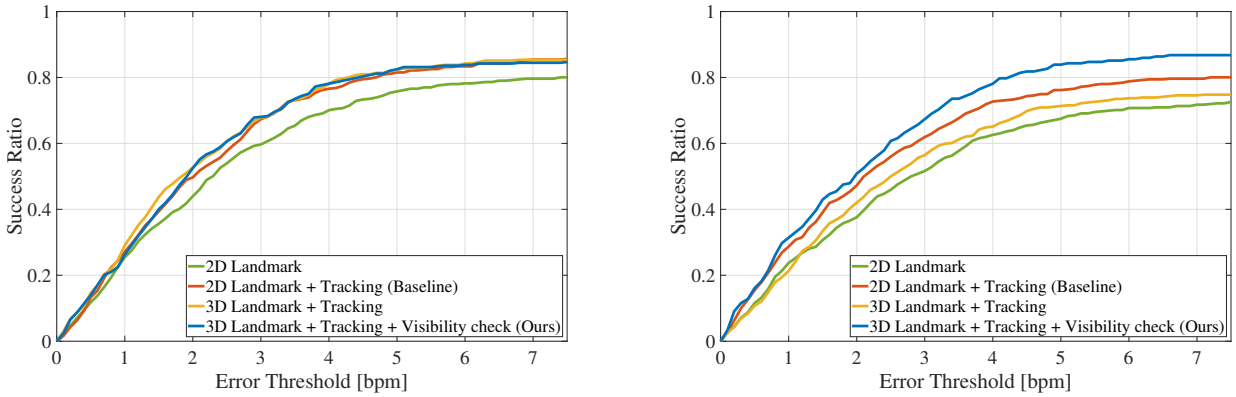
Fig. 5. The effect of angle threshold values in our visibility check. The orange line represents the ratio of estimated video clips among all 500 clips, meaning that at least one face patch is included in the visible patch set and HR can be estimated without the ICA error. The blue line represents the success ratio of the estimated video clips. The result indicates that there is a trade-off between the accuracy and the completeness of the HR estimation.

Figure 3 shows the comparison of our proposed method with the angle threshold of 75 degrees and the baseline method using 2D facial landmarks tracking. The baseline method generates large estimation errors in [0, 15] (sec.) and also cannot estimate HR in [45, 60] (sec.). This is because the 2D landmarks detection algorithm inaccurately estimates or completely misses the landmarks for those sections, as shown in Fig. 1(a). In contrast, our proposed method based on the 3D facial landmarks tracking achieves better robustness to the large head rotations.

C. Effect of angle threshold values

We next investigate the effect of angle threshold values used in our visibility check. For this purpose, we used the most challenging "medium rotation" videos of PURE dataset [14]. Each video consists of the frames with the average face angle of 35 degrees, as shown in Fig. 4(b). Since the 60-seconds videos of 10 subjects are provided, we tested for extracted 500 10-seconds video clips.

Figure 5 shows the result using different angle threshold values. The orange line represents the ratio of estimated video clips, which mean that at least one face patch is included in the visible patch set and HR can be estimated without the ICA error. The result indicates that the stricter threshold value we use, the higher the possibility we obtain no visible patch. In contrast, the blue line represents the success ratio (absolute HR estimation error is less than five beat-per-minute (BPM)) of the estimated video clips. The result indicates that the stricter threshold value tends to make the precision better, which supports our expectation that the face patches more parallel to the image plane are more reliable. Given the results of Fig. 5, we can see the trade-off between the accuracy and the completeness of HR



(a) Small rotation (b) Medium rotation

Fig. 6. Quantitative evaluation on PURE dataset. The vertical axis represents the ratio of estimated 10-seconds video clips whose absolute HR estimation error is less than the threshold of the horizontal axis. The results show that our proposed method achieves higher accuracy for the medium rotation situation, demonstrating the improved robustness of our method to head rotations.

estimation. Considering the trade-off balance, we decided to use the threshold value of 75 degrees in all experiments.

D. Comparison with other methods using PURE dataset

We next compare our method with other methods using the "small rotation" and "medium rotation" situations of PURE dataset (see Fig. 4). Each situation contains the 60-seconds videos of 10 subjects and we used 500 10-seconds clips for the evaluation. For both situations, 2D/3D facial landmarks detection properly works because of no face occlusion.

Figure 6 shows the results for the two situations, where the vertical axis represents the ratio of estimated 10-seconds clips whose absolute HR estimation error is less than the threshold of the horizontal axis. For the "small rotation", the accuracy of our method is comparable with the baseline method because the ROIs of the two methods are almost the same, i.e., almost all face patches are visible in this situation. In contrast, for the "medium rotation", our method outperforms the baseline method. Besides, it is revealed that only replacing the 2D landmarks with the 3D landmarks does not always increase the accuracy. These results validate the effectiveness of our visibility check using the face patch normal information.

IV. Conclusion

In this paper, we have proposed a novel video-based HR estimation method based on 3D facial landmarks detection and face patch visibility check considering the 3D information. We have experimentally demonstrated that our proposed method using the 3D landmarks improves the robustness to head rotations and outperforms the baseline method using 2D facial landmarks detection. Our source code is publicly available at http://www.ok.sc.e.titech.ac.jp/res/VitalSensing/3DfaceHR/

References

[1] L. A. Aarts, V. Jeanne, J. P. Cleary, C. Lieber, J. S. Nelson, S. B. Oetomo, and W. Verkruysse, "Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit – A pilot study," *Early Human Development*, vol. 89, no. 12, pp. 943–948, 2013.

[2] F. Zhao, M. Li, Y. Qian, and J. Z. Tsien, "Remote measurements of heart and respiration rates for telemedicine," *PloS One*, vol. 8, no. 10, pp. e71 384–1–14, 2013.

[3] W. Wang, B. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote-PPG," *IEEE Trans. on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.

[4] Y. Sun and N. Thakor, "Photoplethysmography revisited: From contact to noncontact, from point to imaging," *IEEE Trans. on Biomedical Engineering*, vol. 63, no. 3, pp. 463–477, 2016.

[5] M. Z. Poh, D. McDuff, and R. W. Picard, "Advancements in non-contact, multiparameter physiological measurements using a webcam," *IEEE Trans. on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[6] H. E. Tasli, A. Gudi, and M. den Uyl, "Remote PPG based vital sign measurement using adaptive facial regions," *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, pp. 1410–1414, 2014.

[7] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera," *Biomedical Optics Express*, vol. 6, no. 5, pp. 1565–1588, 2015.

[8] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 3640–3648, 2015.

[9] S. Kado, Y. Monno, K. Moriwaki, K. Yoshizaki, M. Tanaka, and M. Okutomi, "Remote heart rate measurement from RGB-NIR video based on spatial and spectral face patch selection," *Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5676–5680, 2018.

[10] Y. Maki, Y. Monno, K. Yoshizaki, M. Tanaka, and M. Okutomi, "Inter-beat interval estimation from facial video based on reliability of BVP signals," *Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6525–6528, 2019.

[11] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou, "The menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking," *Int. Journal of Computer Vision*, vol. 127, no. 6-7, pp. 599–624, 2019.

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

[13] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 1021–1030, 2017.

[14] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," *Proc. of IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 1056–1062, 2014.

[15] Y. Nirkin, I. Masi, A. T. Trân, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, pp. 98–105, 2018.