FITNESS HEART RATE MEASUREMENT USING FACE VIDEOS

*Qiang Zhu**, Chau-Wai Wong**, Chang-Hong Fu*†, Min Wu**

* University of Maryland, College Park, USA
† Nanjing University of Science and Technology, China
*{zhuqiang, cwwong, minwu}@umd.edu, †enchfu@njust.edu.cn

ABSTRACT

Recent studies showed that subtle changes in human's face color due to the heartbeat can be captured by digital video recorders. Most existing work focused on still/rest cases or those with relatively small motions. In this work, we propose a heart-rate monitoring method for fitness exercise videos. We focus on designing a highly precise motion compensation scheme with the help of the optical flow, and use motion information as a cue to adaptively remove ambiguous frequency components for improving the heart rates estimates. Experimental results show that our proposed method can achieve highly precise estimation with an average error of 1.1 beats per minute (BPM) or 0.58% in relative error.

***Index Terms*—** heart rate, photoplethysmography (PPG), fitness exercise, optical flow

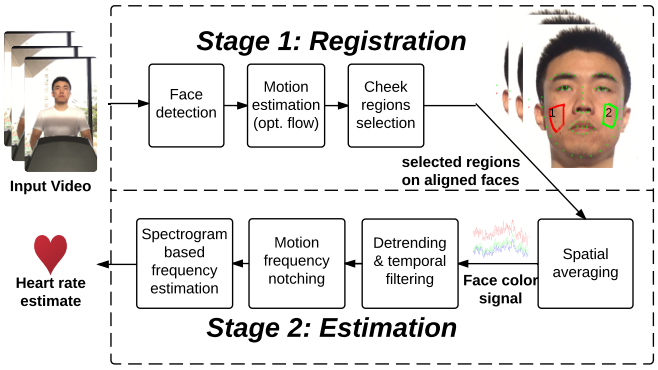


Fig. 1: Flowchart for the proposed heart rate monitoring method for fitness exercise videos.

1. INTRODUCTION

Contact-free monitoring of the heart rate using videos of human faces is a user-friendly approach compared to conventional contact based ones such as electrodes, chest belts, and finger clips. Such monitoring system extracts from a face video a 1-D sinusoid-like face color signal that has the same frequency as the heartbeat. The ability to measure heart rate without touch-based sensors is attractive and gives it potentials in such applications as smart health and sports medicine.

Heart rate from videos was first demonstrated feasible in [1], and since then most work [2–12] has been focusing on still/rest cases or those with relatively small body motions. In contrast, less work [13–15] has been on large motion scenarios such as fitness exercises. In [13], the authors did a proof-of-concept study showing that after using block-based motion estimation for a cycling exercise video, a periodic signal can be extracted from the color in the face area. However, it was not verified against a reference signal and the accuracy of the estimated heart rate was not quantitatively examined. In [8, 14, 15], the authors built resilience against the motion induced illumination changes by finding a particular direction in the 3-D space of the RGB channels that was the least affected. They exploited the fact that illumination changes due to the face motion and the heartbeat have different causes, and analyzed using light reflection characteristics and face-camera geometry. However, their use of the Viola–Jones face detector [16] and/or face-level affine transform did not precisely align faces at the pixel level, which could add noise to the extracted face color signals.

In this work, we aim to examine the best possible performance for fitness exercise videos when the registration error is minimized for the color-based heart-rate monitoring method. A block diagram of our proposed method is shown in Fig. 1. We minimize the registration error using pixel-level optical flow based motion compensation [17, 18] that is capable of generating almost "frozen" videos for best extracting the face color signals. We use the RGB weights proposed in [14] to resist unwanted illumination changes due to motion. We focus on the fitness scenarios that heart rate often wildly vary at different stages of fitness exercises, and present our results in widely adopted metrics [6, 12, 19] for comparison purpose.

The rest of the paper is organized as follows. In Section 2, we propose our video-based heart-rate monitoring method specially designed for fitness exercises. In Section 3, we present the experimental results with comparisons if some modules were otherwise replaced or turned off. Finally, Section 4 concludes the paper.

2. PROPOSED METHOD

Fitness exercise videos may contain large and periodic motions. Our proposed method focuses on a highly precise motion compensation scheme to allow generating a clean face color signal to facilitate the latter analysis steps, and uses the resulting motion cue as the guide to adaptively remove ambiguous frequency components that can be very close to the heart rate.

2.1. Precise Face Registration

A highly precise pixel-level motion compensation is a crucial step toward generating a clean face color signal. We use an optical flow algorithm to find correspondences of all points on the faces between two frames. Optical flow uses gradient information to iteratively refine the estimated motion vector field [17]. To avoid being trapped in local optima, we introduce a prealignment stage to bring the face images roughly aligned before conducting a fine-grain alignment using optical flow.

We use the Viola–Jones face detector [16] to obtain rough estimates of the location and size of the face. We clip and resize the face region of each frame to 180 pixels in height, effectively generating a prealigned video for the face region.

The prealignment significantly reduces the lengths of motion vectors, which in turn makes results of optical flow more reliable. In our problem, two face images are likely have a global color difference due to the heartbeat. In order to conduct a precise face alignment, instead of using the illumination consistency assumption that is widely used, we assume more generally that the intensity I of a point in two frames are related by an affine model, namely,

$$I(x + \Delta x, y + \Delta y, t + 1) = (1 - \epsilon)\, I(x, y, t) + b \qquad (1)$$

where ϵ and b control the scaling and bias of the intensities between two frames. Both of them are usually small. Traditional techniques tackling the illumination consistency cases such as Taylor expansion and regularization can be similarly applied. Our mathematical analysis showed that omitting the illumination change due to the heartbeat, and applying a standard optical flow method leads to a bias term that is at the same order magnitude compared to the intrinsic error (in terms of standard deviation) of the optical flow system. We therefore use Liu's optical flow implementation [18] in our work.

We divide each video into small temporal segments with one frame overlaping for successive segments. We use the frame in the middle of the segment as the reference for optical flow based motion compensation. This would ensure two frames being aligned do not have significant occlusion due to long separation in time. Fig. 2 shows a couple of face images from a same segment before and after optical flow based motion compensation using the same reference.

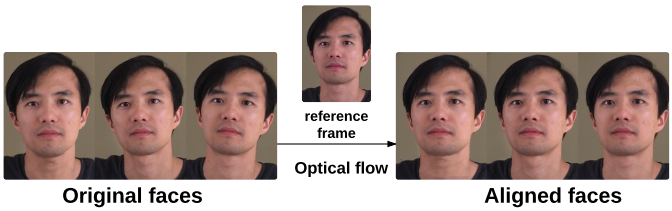


Fig. 2: Face images from a same video segment before and after optical flow based motion compensation using the same reference face.

2.2. Segment Continuity and Cheek Regions Selection

With the precisely aligned face videos in short segments, we can estimate the face color for each frame by taking a spatial average over pixels of the cheek for R, G, and B channels, respectively. We call the three resulting 1-D time signals the *face color signals*.

When concatenating segments into color signals, the last point of the current segment and the first point of the next segment may have different intensities because they correspond to the same frame whose motion compensation were conducted with respect to two different references. To address this problem, the difference of the intensity between the two points is calculated and the resulting value is used to bias the signal of the next segment in order to maintain the continuity.

The face color signals contain color change due to the heartbeat, and illumination change due to face motions such as tilting. The green channel was used because it corresponds to the absorption peak of (oxy-) hemoglobin [1] that changes periodically as the heartbeat, and source separation methods such as the independent component analysis (ICA) were also used to separate the heartbeat component [3]. In [14], the authors proposed using the fixed linear weights $(-1, 2, -1)$ for R, G, B channels to best retain the heartbeat component while compensating the motion induced illumination change. In our experiments, we found that the fixed weights approach outperforms all other approaches, and we therefore adopt it in our proposed method.

To determine the cheek regions for conducting spatial averaging, we construct two conservative regions that do not contain facial structures and are most upfront in order to avoid strong motion-induced specular illumination changes. We use facial landmarks identified by the method proposed in [20] to facilitate the construction of the cheek regions. Each cheek region is constructed to be a polygon that has a safe margin to the facial structures protected by the landmarks. One example for such selected cheek regions and corresponding face landmarks is shown on the face in Fig. 1.

2.3. Detrending and Temporal Filtering

Illumination variation caused by passersby and/or the gradual change of sun light can cause the face color signal to drift, which can be problematic for Fourier-based analysis. Such slowly-varying trend can be estimated and then subtracted

from a raw face color signal, $\mathbf{x}_{\text{raw}} \in \mathbb{R}^L$, where L is the length of the signal. The trend is assumed to be a clean, unknown version of $\mathbf{x}_{\text{raw}}$ with a property that its accumulated convexity measured for every point on the signal is as small as possible, namely,

$$\hat{\mathbf{x}}_{\text{trend}} = \underset{\mathbf{x}}{\text{argmin}} \, ||\mathbf{x}_{\text{raw}} - \mathbf{x}||^2 + \lambda ||\mathbf{D}_2 \mathbf{x}||^2 \qquad (2)$$

where λ is a regularization parameter controlling the smoothness of the estimated trend, and $\mathbf{D}_2 \in \mathbb{R}^{L \times L}$ is a spare toeplitz second-order difference matrix. The closed-form solution is $\hat{\mathbf{x}}_{\text{trend}} = (\mathbf{I} + \lambda \mathbf{D}_2^T \mathbf{D}_2)^{-1} \mathbf{x}_{\text{raw}}$. Hence, the detrended signal is $\mathbf{x}_{\text{raw}} - \hat{\mathbf{x}}_{\text{trend}}$.

After detrending, we use a bandpass filter to reject the frequency components that are outside a normal range of human heart rate. The filter of choice is an IIR Butterworth with a passband from 40 to 240 bpm.

2.4. Motion Frequency Notching

In previous stages, we have designed our method to best remove the impact of face motions: optical flow was used to precisely align the faces, and a color weight vector that is least susceptible to motion was used to reduce impact of the periodic illumination change due to the face tilting. In this part, we further apply a notch operation to remove any remaining trace. We combine motion vectors from the face tracker and the optical flow by addition to generate two time signals, one for the x-direction and the other for the y-direction. For each time bin on the spectrogram, we conduct a notch operation at two frequency locations corresponding to the dominating frequencies of the x- and y- motion components, respectively.

Spectrograms in the first column of Table 1 show that motion traces exist before notching, as highlighted by the arrows. We notice that the motion artifacts can be even stronger than the heart rate (HR) traces. Spectrograms in the second column of Table 1 show that the frequency notching method is effective and the HR traces dominate after notching.

2.5. Robust Frequency Estimation

We design a robust frequency estimator for noisy face color signals from fitness exercises. Instead of directly finding the peak (the mode) of the power spectrum for every time bin that may result in a discontinuous estimated heart-rate signal, we construct a two-step process to ensure the estimated signal is smooth.

We first find a single most probable strap from the spectrogram. We binarize each time bin of the spectrogram image per the 95th percentile of the power spectrum of that bin. We then dilate and erode the image in order to connect the broken strap. We find the largest connected region using such standard traverse algorithm as the breadth-first search and consider it as the most probable strap. A spectrogram and the results of successive steps are shown in Fig. 3.

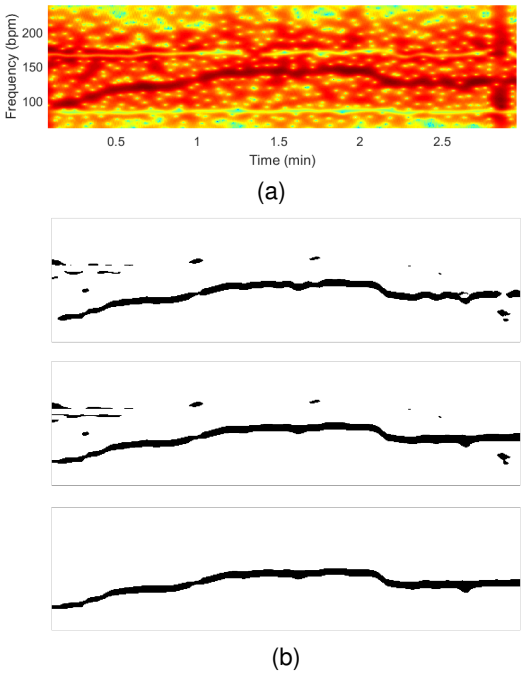


Fig. 3: (a) A spectrogram with weakly connected frequency strap. (b) Results after the following operations (from top to bottom): binarization using 95th percentile, dilation and erosion, and small regions removal.

We finally use a weighted frequency [21] within the the frequency range specified by the strap, $\mathcal{F}_i$, as the frequency estimate for ith time bin. Denoting the frequency estimate as $\hat{f}_{\text{HR}}(i)$, we have

$$\hat{f}_{\text{HR}}(i) = \sum_{f \in \mathcal{F}_i} w_{i,f} \cdot f \qquad (3)$$

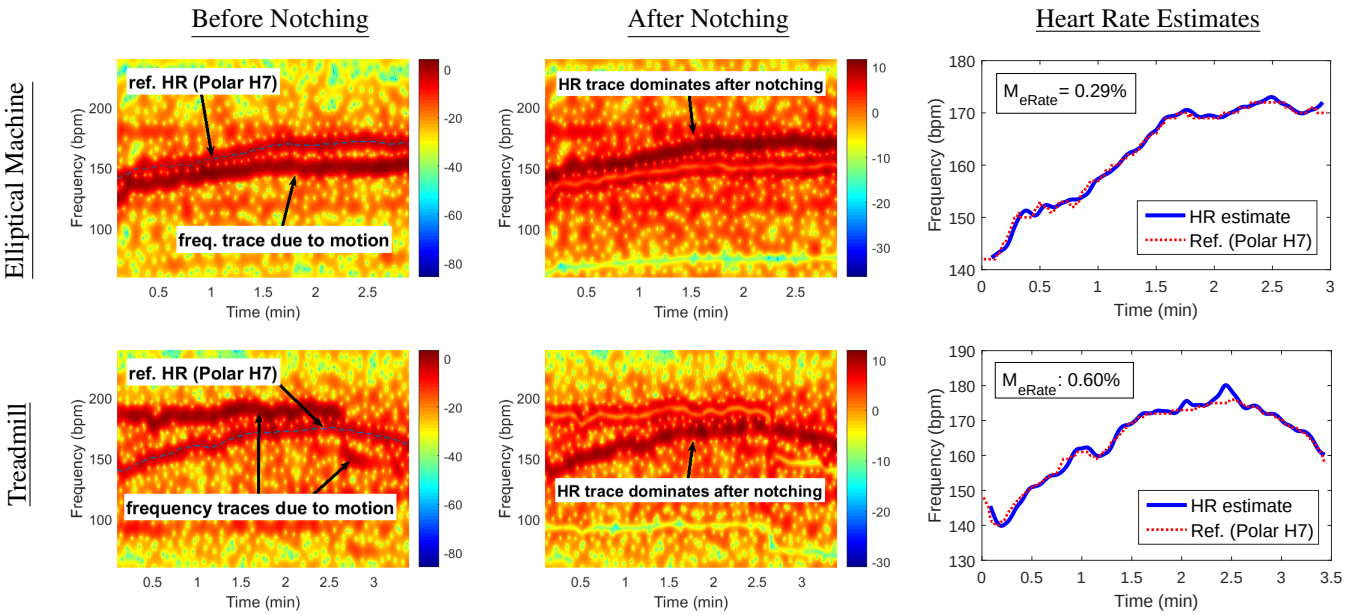
where $w_{i,f} = |S(i,f)| \, / \sum_{f \in \mathcal{F}_i} |S(i,f)|$, and $S(i,:)$ is the power spectrum at the ith bin.

3. EXPERIMENTAL RESULTS

Our proposed method was evaluated on a self-collected fitness exercise dataset to demonstrate the efficacy on dealing with fitness motions, and results were presented in widely adopted metrics [6, 12, 19] in the field of heart rate monitoring from videos.

The fitness exercise dataset has 9 videos in which 6 contain human motions on an elliptical machine and the other 3 contain motions on a treadmill. Each video is about 3 minutes long in order to cover various stages of a fitness exercise. Each video was captured in front of the face by a commodity mobile camera (iPhone 6s) affixed on a tripod or held by the hands of a person other than the test subject. The gym was well-lit with several over-the-top florescent lights and with diffuse daylight passing into the gym through glass walls. The heart rate of the test subject was simultaneously monitored by

Table 1: Contrast of spectrograms before and after notching the frequencies of fitness motions (cols. 1–2). Heart rate estimates and the ECG-based reference measurement using Polar H7 chest belt (col. 3). Video demos: http://www.mast.umd.edu/project/heart-rate.



an electrocardiogram (ECG)-based chest belt (Polar H7) for reference.

Each video was divided into segments of 1.5 secs in order to guarantee small scene changes within each segment for optical flow's best performance. The regularization parameter for the detrending on the face color signal was set to $\lambda = 20$ for 30Hz videos used in this experiment. The window length for spectrogram was set to 10 secs with 98% overlap.

Representative results from two videos are shown in Table 1. Column 1 and column 2 show the spectrograms for the detrended and filtered face color signal before and after motion information guided notching. Column 3 show the HR estimates obtained using the robust frequency estimation algorithm. We plotted the HR estimates with the reference HR, and found that the estimates are almost unbiased and are fluctuating around the reference. The relative error (M_{eRate}) are as low as 0.29% and 0.60% for the two videos, respectively. We included two demos, each of which contains a raw video, a motion compensated video, and a synchronized HR estimate and a reference HR.

We summarize the mean and standard deviation of the error measures for all of our videos and the results are listed in Table 1. The averaged error for the proposed method is 1.1 bpm in root mean-squared error (RMSE) and 0.58% in relative error. The performance is slightly reduced if a more complicated joint blind-source separation (JBSS) approach is used to search for the weights for R, G, B channels, instead of using the more theoretically grounded fixed weights $(-1, 2, -1)$.

We conducted additional experiments to check the impact when the optical flow algorithm was disabled. In this case, the face images were roughly aligned using the face tracker.

Table 2: Performance (in terms of mean and standard deviation in parentheses) of proposed method and cases when some modules were otherwise replaced or turned off.

Module combinations	RMSE in bpm	M_{eRate}
tracker + JBSS (no op)	7.6 (5.7)	3.60% (2.87%)
tracker + fixed (no op)	5.6 (3.4)	2.61% (1.45%)
tracker + op + JBSS	1.3 (0.7)	0.65% (0.30%)
tracker + op + fixed (proposed)	**1.1 (0.6)**	**0.58% (0.33%)**

The reported errors in Table 2 show significant performance reduction from 1.1 bpm to 5.6 bpm or from 0.58% to 2.61%. The estimation error has increased about four times, which shows that a precise alignment is a crucial step for the video-based heart-rate monitoring method for fitness scenarios.

4. CONCLUSION

In this paper, we proposed a heart rate monitoring method for fitness exercise videos. We focused on building a highly precise motion compensation scheme with the help of the optical flow, and used motion information as a cue to adaptively remove ambiguous frequency components for improving the heart rates estimates. Experimental results show that our proposed method can give precise estimates at an average error of 1.1 bpm in RMSE or 0.58% in relative error.

Acknowledgment: We thank Jiahao Su for his contributions to the initial phase of this project. We thank Prof. James M. Hagberg for the enlightening discussion on chest-strap based heart rate monitoring in sports medicine.

5. REFERENCES

[1] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optices Express*, vol. 16, no. 26, pp. 21 434–45, Dec. 2008.

[2] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Optics express*, vol. 18, no. 10, pp. 10 762–74, May 2010.

[3] ——, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[4] M. Lewandowska, J. Rumiski, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam – a non-contact method for evaluating cardiac activity," in *Federated Conference on Computer Science and Information Systems (FedCSIS)*, Szczecin, Poland, Sep. 2011, pp. 405–410.

[5] 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 EMBS Annual International Conference*, San Diego, CA, Aug. 2012, pp. 2174–2177.

[6] 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)*, Columbus, OH, Jun. 2014, pp. 4264–4271.

[7] Y. Cui, C.-H. Fu, H. Hong, Y. Zhang, and F. Shu, "Non-contact time varying heart rate monitoring in exercise by video camera," in *International Conference on Wireless Communications & Signal Processing (WCSP)*, Nanjing, China, Oct. 2015.

[8] L. Feng, L. M. Po, X. Xu, Y. Li, and R. Ma, "Motion-resistant remote imaging photoplethysmography based on the optical properties of skin," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 879–891, May 2015.

[9] S. Yu, X. You, X. Jiang, K. Zhao, Y. Mou, W. Ou, Y. Tang, and C. L. P. Chen, "Human heart rate estimation using ordinary cameras under natural movement," in *IEEE International Conference on Systems, Man, and Cybernetics*, Hong Kong, Oct. 2015, pp. 1041–1046.

[10] 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.

[11] S. Fernando, W. Wang, I. Kirenko, G. de Haan, S. Bambang Oetomo, H. Corporaal, and J. van Dalfsen, "Feasibility of contactless pulse rate monitoring of neonates using Google Glass," in *EAI International Conference on Wireless Mobile Communication and Healthcare (MOBIHEALTH)*, ICST, Brussels, Belgium, Belgium, 2015, pp. 198–201.

[12] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, Jun. 2016, pp. 2396–2404.

[13] Y. Sun, S. Hu, V. Azorin-Peris, S. Greenwald, J. Chambers, and Y. Zhu, "Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise," *Journal of Biomedical Optics*, vol. 16, no. 7, pp. 077 010:1–9, Jul. 2011.

[14] 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, Oct. 2013.

[15] G. de Haan and A. van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," *Physiological Measurement*, vol. 35, no. 9, p. 1913, Aug. 2014.

[16] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[17] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, vol. 2, Vancouver, Canada, Aug. 1981, pp. 674–679.

[18] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[19] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Heartbeat rate measurement from facial video," *IEEE Intelligent Systems*, vol. 31, no. 3, pp. 40–48, May 2016.

[20] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, Dec. 2013, pp. 1944–1951.

[21] R. Garg, A. L. Varna, A. Hajj-Ahmad, and M. Wu, "'Seeing' ENF: Power-signature-based timestamp for digital multimedia via optical sensing and signal processing," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 9, pp. 1417–1432, Sep. 2013.