LEARNING-BASED HEART RATE DETECTION FROM REMOTE PHOTOPLETHYSMOGRAPHY FEATURES

YungChien Hsu, Yen-Liang Lin, Winston Hsu

National Taiwan University, Taipei, Taiwan

ABSTRACT

Remote photoplethysmography (rPPG) enables measuring heart rate from recorded skin color variations with consumer cameras. Recent research has aimed to improve the signal strength of color variations caused by heart beat by using independent component analysis (ICA) technique or analyzing chrominance-based model. In this paper, we argue for treating this emerging problem in a novel aspect – proposing a learning-based framework to accommodate multiple and temporal feature and yielding significant and robust improvement. Using support vector regression (SVR) on published chrominance-based feature improves the root mean square error (RMSE) from 22.7 to 7.31 as well as correlation coefficient (CC) from 0.30 to 0.77. With proposed novel multiple feature fusion and multiple segment fusion techniques, we achieved the best estimation result with RMSE 5.48 and CC 0.88. Meanwhile, the proposed framework can be extended to other promising features.

Index Terms— heart rate, photoplethysmography (PPG), regression learning

1. INTRODUCTION

Photoplethysmography (PPG) that detects the optical absorption variation of human skin surface was first described in the 1930s [1] and is popular because it can be used non-invasively. PPG can also detect blood volume variation caused by heart beat and is suitable for monitoring the human heart rate.

Several studies have shown that PPG could measure the variation in a distance [2], and could be used to measure heart rate [3, 4]. This technique is called remote-PPG (rPPG). Other research confirmed that PPG works well under ambient light conditions [5, 6]. A demo that amplifying and visualizing the skin color variation of video taken by consumer camera in ambient light environment has also been showed [7].

However, in the research conducted by Rhee et al. and Poh et al., they found that PPG does not work well when motion-induced signal was involved [8, 9]. Kimura et al. showed that skin color variation caused by blood volume variation is more significant on green channel than on red or blue channels because of optical properties of human skin [10]. Poh et al. used independent component analysis (ICA) technique to separate color variation from the color signal to estimate human heart rate [6]. Another method that measures heart rate based-on skin chrominance-based model analysis was proposed by de Haan and Jeanne [11]. Both Poh et al. and de Haan and Jeanne use different characteristics of color channels and face detection tools to maximize the blood volume variation and minimize the effect of moving human face; moreover, they chose the frequency with the biggest amplitude as the estimated heart rate and discarded the information behind the amplitudes of other frequencies [6, 11].

In this work, we aim to measure human heart rate using videos taken by consumer cameras under ambient light conditions. For robust and scalable heart rate detection, we propose a novel learning-based framework to accommodate more rich features. We compare and leverage the features proposed by several state-of-the-art research to train the model.

In this paper, we make the following contributions:

- To our best knowledge, we are the first to estimate heart rate by multiple rPPG features with a brand-new extendable learning framework.
- We devise novel mechanisms for multiple feature fusion and adaptive normalization schemes.
- We further propose multiple segment fusion by leveraging the temporal redundancy.

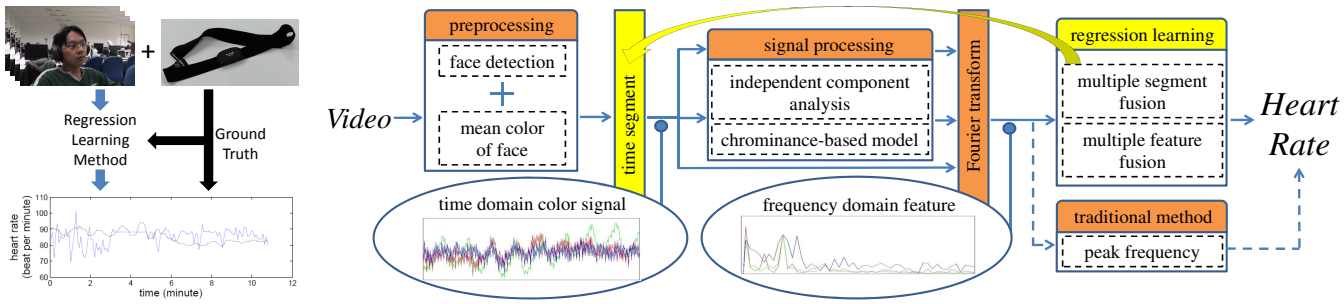
2. METHODS

2.1. Preprocessing

Heart rate detection using rPPG is based on face color variations. We use OMRON OKAO vision [1] face detection tools to locate each face in the video and set the entire face rectangle as our region of interest (ROI). This step also removes possible noises created in the background. In case face detection fails, the latest detected face position would be used. Then we compute the mean value of all pixels in the ROI on R, G, B channels for all frames. We use $RAW_C(t)$ to describe the color signal in time domain. Here $C \in \{R, G, B\}$ stands for

[1] http://www.omron.com/r_d/coretech/vision/okao.html

978-1-4799-2893-4/14/$31.00 ©2014 IEEE



(a) Design of experiment (b) System flow of our method

Fig. 1. (a) is the design of experiment. The participant is asked to wear a heart rate sensor to get the ground truth of heart rate while being video-recorded by the camera. Some of the videos and corresponding ground truths are used to train the model of SVR and others are used in testing. Take one sample video for example, the blue line in the heart rate-time graph in the bottom is the estimation result and the black line represents the ground truth. (b) is the system flow of our method. We first use face detection to extract color signals of human faces. Then dividing color signal into segments if corresponding multiple segment fusion techniques are used. Those time domain signals are processed by the proposed methods to derive frequency domain amplitude features. Different from prior studies, we use frequency domain results as mid-level features for our proposed learning framework, which delivers more robust and extendable results. The procedures with yellow tag are the main differences between previous works and our work.

three color channels, and $t = 1, 2, 3, ...$ stands for different frames. We have tried learning with this time domain color signal and it seems that to learn pattern in such signal is unsuccessful.

2.2. Frequency domain features

We use frequency domain features to find the color variation pattern corresponding to heart beat waveform. Fourier transform is the most used and standard form to transform signal from time domain to frequency domain. Three frequency domain features are used in the following steps. We implement fast Fourier transform (FFT), an approach of discrete Fourier transform (DFT), on $RAW_C(t)$, then computing its amplitude spectrum to get $FT_C(f)$. In this paper, f stands for different frequency, and it is our first low-level feature. For mid-level frequency domain features, we adopt ICA processed feature proposed by Poh et al. We use all three components of ICA processed feature and name them $ICA_N(f)$, where $N \in \{1, 2, 3\}$ stands for different components. Another mid-level feature $X_s - \alpha Y_s$ that proposed by de Haan and Jeanne is adopted and named $CB(f)$. Please note that all features $FT_C(f), ICA_N(f), CB(f)$ are frequency domain amplitude spectrum and only the data with frequency between 0.75 to 4 Hz (45 to 240 bpm) are used for learning. The frequency range is considered as possible human heart rate. Note that the proposed framework can accommodate other promising features as well.

2.3. Model learning

In prior research, in an ad-hoc manner, the frequency with the highest amplitude in the frequency domain features is considered the only heart rate frequency and all of the amplitude information behind non-peak frequency is ignored no matter how big or small the amplitude is. We know that blood volume variations caused by heart beat is not perfect sine wave and the face motion further causes uncertainties. Therefore every value of the features may contain the information of heart beat and noises. For robust detection, we argue for the learning-based methods (e.g., support vector regression (SVR)) over the prior low and mid-level features. Meanwhile, different novel fusion strategies are considered as well.

2.3.1. Multiple feature fusion

These three features mentioned in section 2.2 are proposed under different assumptions and have promising results on different cases. So we propose multiple feature fusion to leverage their advantages. The amplitude spectrums derived from three kinds of features have different ranges. Before we do the fusion process on multiple types of features, we normalize these features in advance by the following equations:

$$nFT_C(f) = \frac{FT_C(f)}{\mu_{FT}} \quad (1)$$

$$nICA_N(f) = \frac{ICA_N(f)}{\mu_{ICA}} \quad (2)$$

$$nCB(f) = \frac{CB(f)}{\mu_{CB}} \quad (3)$$

where $\mu_{FT}, \mu_{ICA}, \mu_{CB}$ are the mean of $FT_C(f), ICA_N(f)$ and $CB(f)$.

This normalization has some good properties, i.e., retaining the physical meaning of amplitude spectrum, since it does not change the sign. Previous works have shown that green channel has better result than red or blue channel in $FT_C(f)$ [10] and component 2 is usually a better choice than component 1 or 3 in $ICA_N(f)$ [6], so we keep the correlations between different colors and different components. We also want to keep the values of different features comparable, so we decide to divide by mean rather than divide by maximum, since that mean value is more representative than maximum. After normalization, we have three features in similar range and can concate several features together as new input pattern of SVR.

2.3.2. Multiple segment fusion

To achieve the best estimation, a key question here is to determine how long the video should be. The variation of human heart rate over time makes that peak of frequency domain amplitude feature extracted from the long video may not match the designated heart rate. The features extracted from short videos are easy to have a peak value caused by noises but not the color variations from the heart beats. So we divide a long video into several short segments and use the data of segments to get better estimation of the long video. Two approaches are used in this paper. The first one is to learn our model with features extracted from short video segments and the results of several continuous segments are averaged; we called it late fusion method. The second one is to learn our model with the features that concate subfeatures extracted from several continuous short segments together, we called it early fusion method. By these methods, we can avoid the disadvantages of long video features and reduce the deviations made by short video features.

2.4. Evaluation methods

Bland-Altman plot (BA plot) [12] is used for analyzing the agreement between two different assays. The differences between results predicted by our methods and the ground truth recorded by wearing heart rate sensor are plotted against the average of both systems. We calculated the mean of the difference and showed in BA plot. The standard deviation (SD) is also calculated and 95% limits of agreement ($\pm$1.96SD) are shown.

The root mean square error (RMSE) and correlation coefficient (CC) also used to figure out the correlation between our heart rate estimation and the ground truth. It is worthy of mention that RMSE may be small if all estimations are the values of mean heart rate, and CC cannot tell us whether estimation and ground truth changes in 1:1 or not. While analyzing the estimation result, it is preferred to review these two values at the same time.

3. EXPERIMENTS

3.1. Dataset

We use SONY XDR-XR500 video camera to record the videos. All videos are recorded in 24-bit color (R, G, B three channels $\times$ 8-bits/channel) at 29.97 frames per second with resolution 1920 $\times$ 1080 and saved in mp4 format. Since we use mean of ROI to do the very first color signal processing, such high resolution is not necessary for our method.

Four Asian males between the age of 22 and 25 participated in our preliminary study, they wore POLAR WearLink transmitter with Bluetooth to get the heart rate data when being video-recorded. Participants were asked to sit in front of computer monitors and could behave normally except leaving the seat. Some of them wore glasses and earphones.

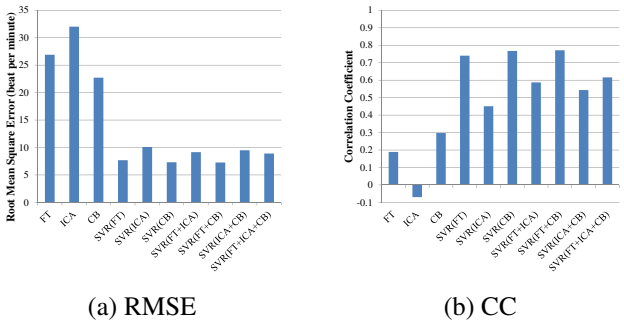
3.2. Results

For each participant we recorded a video of 10 minutes and 50 seconds. Then, we divided the video into two parts, each was 5 minutes and 25 seconds long. In the learning process, we either did training with the first part of the video and testing with the second part or the other way around.

In our first experiment, we apply 30-second window with 5-second stride and get a total of 480 samples from four videos. Estimating results of three features without learning is to choose the frequency having maximum amplitude between 0.75 to 4 Hz in $FT_G(f), ICA_2(f), CB(f)$. We use libsvm [13] to run SVR learning and every time we train and test with 240 samples respectively. Table.1 and Fig.2 show the result of learning-based and traditional methods by the form of RMSE and CC. The result of SVR with single feature has significant improvements in all three features. We observe that samples seldom have big deviation in learning cases. It implies that our model can distinguish the pattern of noise from the heart rate. With multiple feature fusion techniques, learning with FT+CB feature has best result RMSE 7.28 and

	FT_G	ICA_2	CB	SVR (FT)	SVR (ICA)	SVR (CB)	SVR (FT+ICA)	SVR (FT+CB)	SVR (ICA+CB)	SVR (FT+ICA+CB)
RMSE	26.90	32.01	22.72	7.70	10.09	7.31	9.15	7.28	9.49	8.90
CC	0.19	-0.07	0.30	0.74	0.45	0.77	0.59	0.77	0.54	0.62

Table 1. RMSE and CC of the estimation using traditional methods and SVR learning with multi-feature technique.



(a) RMSE (b) CC

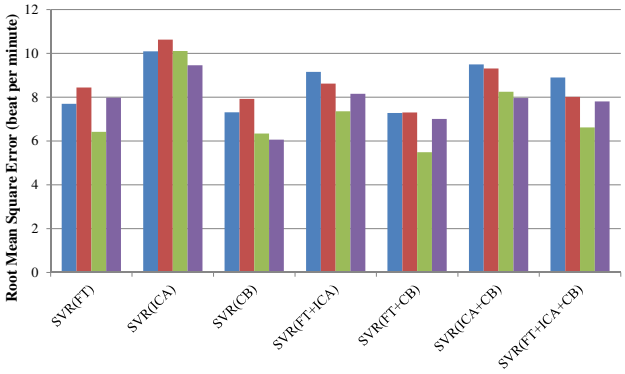
Fig. 2. (a) is the RMSE and (b) is the CC of multiple feature fusion experiment. Numbers could be found in Table.1.

CC 0.77, although further improvement in comparison to single feature learning is not obvious here.

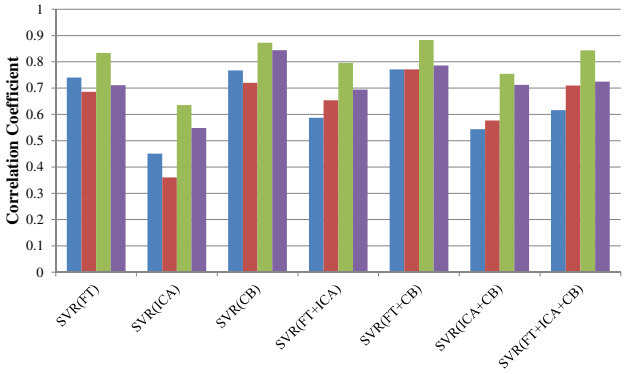
Multiple feature fusion and multiple segment fusion could be used at the same time. Fig.3 shows the result of learning with different feature combination and different setting of multiple segment fusion. Learning with features extracted from 5-second window leads to worse result in single feature learning as comparing with features extracted from 30-second window. But while learning with multiple features, the performance of 5-second window is better than 30-second window in most cases. With multiple features, the model could learn to figure out noise or heart rate signal by comparing them and choosing more reliable ones. The advantage is more helpful while learning from short noisy segments. In the experiment of using single feature and multiple segment fusion, all features get better results comparing to single segment cases. Multiple segment fusion reduces the effect of noisy features by referencing several segments at a time. The CB feature is particularly good under this setting. Learning with CB feature and multi-segment early fusion has RMSE 6.06 and CC 0.84. Both early fusion and late fusion of multiple segment techniques have improved when comparing to the results that learned with single windows. SVR learning using FT+CB features and multi-segment late fusion has the best result. The RMSE is 5.48 and CC is 0.87 respectively. Fig.4 shows a typical BA plot of SVR learning estimation.

4. CONCLUSION

In this paper, we propose a novel method for learning-based framework for heart rate detector by leveraging the mid-level rPPG based features. We also investigate different fusion strategies (along with the normalization schemes from different modalities) for utilizing feature and temporal redundancies. The experiment achieves significant improvements over consumer videos in the ambient light environment. Comparing with prior state-of-the-art, we can reduce the detection error from 22.72 to 5.48 in terms of root mean square error.



(a)RMSE



(b)CC

Fig. 3. (a) is the RMSE and (b) is the CC of estimation of SVR learning with different combination of features and/or different multi-segment fusion strategies. Blue, red, green and purple bars show the results of 30-second window, 5-second window, multi-segment late fusion and multi-segment early fusion. For multi-segment experiments, we use 30-second window and divide them into six 5-second segments.

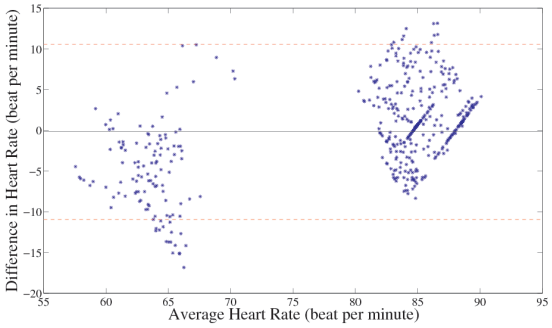


Fig. 4. BA plot of heart rate detected from wearing sensor and heart rate estimation of SVR learning with FT+CB features and multi-segment late fusion technique. The experiment has mean bias -0.18, standard deviation 5.49, RMSE 5.48 and CC 0.87.

5. REFERENCES

[1] A.B. Hertzman, "Photoelectric plethysmography of the fingers and toes in man," *Exp. Biol. Med. 37(3)*, pp. 529–534, 1937.

[2] Markus Huelsbusch and Vladimir Blazek, "Contactless mapping of rhythmical phenomena in tissue perfusion using ppgi," *Proc. SPIE*, vol. 4683, pp. 110–117, 2002.

[3] C. Takano and Y. Ohta, "Heart rate measurement based on a time-lapse image," *Med. Eng. Phys. 29(8)*, pp. 853–857, 2007.

[4] M. Garbey, N. Sun, A. Merla, I. Pavlidis, M. Garbey, N. Sun, A. Merla, and I. Pavlidis, "Contact-free measurement of cardiac pulse based on the analysis of thermal imagery," *IEEE Trans. Biomed. Eng. 54(8)*, pp. 1418–1426, 2007.

[5] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson, "Remote plethysmographic imaging using ambient light.," *Optics express*, vol. 16, pp. 21434–45, 2008 Dec 22 2008.

[6] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics Express*, vol. 18, 2010.

[7] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William T. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)*, vol. 31, no. 4, 2012.

[8] Sokwoo Rhee, Boo-Ho Yang, and Haruhiko Asada, "Artifact-resistant power-efficient design of finger-ring plethysmographic sensors.," *IEEE Trans. Biomed. Engineering*, vol. 48, no. 7, pp. 795–805, 2001.

[9] Ming-Zher Poh, Nicholas C. Swenson, and Rosalind W. Picard, "Motion-tolerant magnetic earring sensor and wireless earpiece for wearable photoplethysmography.," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 786–794, 2010.

[10] Y. Kimura Y. Kanzava and T. Naito, "Human skin detection by visible and near-infrared imaging," in *MVA2011 IAPR Conference on Machine Vision Applications*, 2011.

[11] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *Biomedical Engineering, IEEE Transactions on*, vol. 60, pp. 2878–2886, 2013.

[12] Martin J. Bland and Douglas G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 1, no. 8476, pp. 307–310, Feb. 1986.

[13] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.