Using Temporal-Domain Peak Interval Determination for Video-based Short-Term Heart Rate Measurement

Yu-Shan Wu, Gwo-Hwa Ju, Ting-Wei Lee, Heng-Sung Liu and Yen-Lin Chiu
*Chunghwa Telecom Co., Ltd. No. 99, Dianyan Rd., Yangmei City, Taoyuan Country 32601, Taiwan, R.O.C.
E-mail: yushanwu@cht.com.tw Tel: +886-3-4245877
E-mail: jgh@cht.com.tw Tel: +886-3-4244834
E-mail: finas@cht.com.tw Tel: +886-3-4245095
E-mail: lhs306@cht.com.tw Tel: +886-3-4245243
E-mail: lewis32330@cht.com.tw Tel: +886-3-4245695

Abstract— **Video-based Heart Rate Measurement has drawn a lot of attention in recent years. A few papers have been proposed in this work and most of them applied Fourier transform peak interval determination method for heart rate measurement. However, the analyzing duration for one outcome in these methods is usually longer than 15 seconds. This is because that the measurement resolution in Fourier transform is proportional to the number of samples. In this paper, we apply a novel method to combine temporal-domain peak determination method and super resolution method for heart rate measurement. And we further propose a spectrum selection scheme and a data shift scheme to raise the measurement accuracy. The video database taken in our lab is used to evaluate the performance of the proposed method. The experimental results show two important things. The first is that when the analyzing duration of the proposed method is one third of Fourier transform, the precision of the proposed method is only a little lower than that of the Fourier transform method. Furthermore the precision of the proposed method is superior to Fourier transform approach when the analyzing duration is the same for both methods.**

I. INTRODUCTION

The human heart rate is an important healthy indicator used to evaluate the risk of cardiovascular disease, such as cardiomyopathy or hypertensive heart disease. There are many ways to measure the heart rate of human. The electrocardiogram (ECG) is one of the standard techniques. But it requires the subject to wear straps on the chest and this is not comfortable. An FDA-approved finger-based sensor is another technique to detect the changes in fingertip blood volume pulse (BVP). Both methods need to attach something on some parts of the subject's body.

Recently, a contact-free method using a camera is proposed in [1] and has drawn a lot of interests. After that, many works based on this method are also proposed [2] [3] [4]. The method proposed in [1] is described shortly as follows: Firstly, a camera is used to record color movies of a subject's face. Then, the mean values of G and B channels in a preset ROI region are calculated separately in each image of a frame sequence. The mean values of B or G channels of successive frames can be thought as two time signals. Finally, fast Fourier transform (FFT) is applied to these two time signals separately to determine the maximum magnitude spectral component would appear in which frequency and this frequency can be regarded as heart rate. The method proposed in [2] is based on the method in [1] in which a face detection method provided by OpenCV [5] is used to detect and track face region automatically. And all three color channels (R, G and B) are used. And before fast Fourier transform, a blind source separation method is applied to three time signals formed from the mean values of R, G and B channels. The objective of this step is extracting stationary independent component, which can raise the accuracy of heart rate measurement. In [3], a skin color classifier is applied to detect skin pixels precisely and a data adjustment scheme is used after performing fast Fourier transform to increase measurement precision.

In all of the above methods, fast Fourier transform is a critical step to decide the heart rate in frequency domain. However, the bottle neck of fast Fourier transform is that the measurement resolution is proportional to the number of samples under analyzing.

In this paper, instead of peak interval determination in frequency domain, a novel method which combines the cross correlation method and the super resolution method [6] is used for peak interval determination in time domain. And a spectrum selection scheme and a data shift scheme are proposed to improve the measurement accuracy. The experimental results show that when the analyzing duration of the proposed method is one third of fast Fourier transform method, the precision of the proposed method is only a little lower than that of the FFT method with regular duration. On the other hand, experimental results also show that when the analyzing duration is the same for both methods, the precision of the proposed method is much higher than that of the FFT method.

This paper is organized as follows. In section II, the video-based heart rate measurement method will be introduced. In section III, we will explain peak interval determination in time domain, including cross correlation method, super resolution method and the proposed spectrum selection scheme and data shift scheme. In section IV, the experimental results and some discussions will be given. Finally, the conclusion is presented in section V.

II. HEART RATE MEASUREMENT METHOD

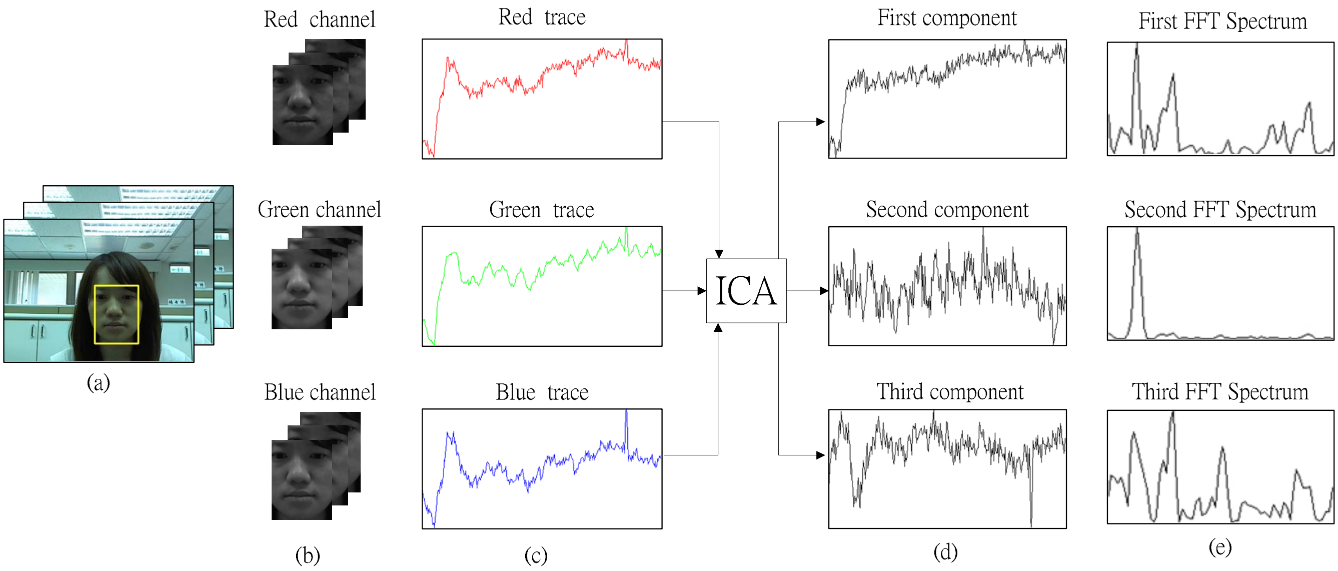

Fig. 1. Video-base heart rate measurement methodology. (a) Face detection. (b) The mean values of R, G and B channels in face region are calculated. (c) The raw RGB time signals. (d) Three independent components obtained from ICA. (e) FFT spectra of three independent components.

Although the video-based heart rate measurement method is originally proposed in [1], the method proposed in [2] is more completed and practical. So in this section, the method proposed in [2] will be introduced briefly. The flow chart of this method is depicted in Fig. 1.

Firstly, a Viola Jones face detector [7] is applied to detect and track face region in consecutive frame sequence, as in Fig. 1(a). Then, the mean values of R, G and B channels of a face region are calculated separately in each frame, as in Fig. 1(b). The mean values of R, G and B channels of consecutive frames can be treated as three time signals, as in Fig. 1(c).

Then, a blind source separation method developed by Cardoso [8] is used to decompose three time signals obtained in above step into three independent components, as in Fig. 1 (d). And this method is called joint approximate diagonalization of eigenmatrices (JADE). Then, fast Fourier transform is performed on these three independent components separately and therefore three corresponding magnitude spectra can be obtained, as in Fig. 1(e).

From many observations, the authors in [2] said that the second independent component obtained from blind source separation is the most closet component to BVP signal. So the second magnitude spectrum is used to decide in which frequency the magnitude spectral component is attaining maximum and this frequency can be regarded as heart rate.

In this paper, the video-based heart rate measurement method is basically based on the method proposed in [2]. The main difference is that after blind source separation, we use a novel method which integrates cross correlation method and super resolution method to perform peak interval determination in time domain.

III. Peak Interval Determination In Time Domain

Peak interval determination of a time signal is an important task in digital signal processing, especially in speech processing. And in speech processing, it is also called pitch determination. Many methods have been proposed for pitch determination [9] [10] [11], and these methods were mainly categorized into two kinds, time domain or frequency domain.

In video-based heart rate measurement, since the mean values of R, G and B channels are three respective time signals, peak interval determination method is also used to decide the time lapse between two heart bits. The peak interval determination used in [2] is fast Fourier transform which operated in frequency domain. In this paper, the peak interval determination is performed in time domain and the reason is that by combining the cross correlation method and the super resolution method the analyzing duration for one outcome can reduce to a satisfied range in practical application.

Since the cross correlation used in this paper is actually performed on the same time signal, it is also called autocorrelation. For convenience and reading simplicity, in following paragraph "autocorrelation" is used to replace "cross correlation".

In this section, the autocorrelation method and super resolution method proposed in [6] are introduced. The proposed spectrum selection scheme and data shift scheme are also explained.

A. Autocorrelation

Let $h[1{:}L] = (h_1, h_2, \dots, h_L)$ denote a discrete time signal with L samples. In other words, $h[1{:}L]$ is a L–dimensional vector. And for any starting point i_0, we define two signal segments $x_\tau(i_0)$ and $y_\tau(i_0)$ with the same length m as follows:

$$\begin{cases} x_\tau(i_0) = (h_{i_0}, h_{i_0+1}, \dots, h_{i_0+m-1}) \\ y_\tau(i_0) = (h_{i_0+\tau}, h_{i_0+\tau+1}, \dots, h_{i_0+\tau+m-1}) \end{cases} \quad (1)$$

In other words, $x_\tau(i_0)$ is a m-dimensional vector cut from $h[1\colon L]$ and the starting point for cut is h_{i_0}. And $y_\tau(i_0)$ is also a m-dimensional vector cut from $h[1\colon L]$ but the starting point of the vector is $h_{i_0+\tau}$. For convenience, we rewrite $x_\tau(i_0)$ as $x_\tau = (x_1, x_2, \dots, x_m)$ and $y_\tau(i_0)$ as $y_\tau = (y_1, y_2, \dots, y_m)$. The autocorrelation between x_τ and y_τ is defined as:

$$\gamma_\tau(x_\tau, y_\tau) = \frac{(x_\tau, y_\tau)}{|x_\tau||y_\tau|} \qquad (2)$$

Where $|x_\tau|$ and $|y_\tau|$ are norms of vectors x_τ and y_τ respectively. (x_τ, y_τ) is the inner dot product and is defined as:

$$(x_\tau, y_\tau) = \sum_{j=1}^{m} x_j * y_j \qquad (3)$$

The autocorrelation is operated on a finite range $[R_{min}, R_{max}]$ and the time interval between two peaks is defined as:

$$R_0 = \underset{\tau}{argmax}\, \gamma_\tau(x_\tau, y_\tau), \qquad R_{min} \leqq \tau \leqq R_{max} \qquad (4)$$

B. Super Resolution

The time interval R_0 estimated from (4) is an integer which contains rounding error since the sampling procedure is discrete in real world. The exact peak interval can be denoted as $R = \underline{R} + \alpha$, where $\underline{R}$ is the integer part of R and α is the fractional part of R . Since the exact peak interval is often a fractional number, we need to define the vector $y_{\underline{R}}(i_0 + \alpha)$. However, from the definition of $y_\tau(i_0)$ in (1), we could understand that the elements of $y_{\underline{R}}(i_0 + \alpha)$ can't be sampled directly. So they can only be approximated by using linear interpolation method, as follows:

$$y_{\underline{R}}(i_0 + \alpha) \cong (1 - \alpha) * y_{\underline{R}}(i_0) + \alpha * y_{\underline{R}}(i_0 + 1) \qquad (5)$$

We denote the optimal value of α as α^*. By rewriting (4), it can be defined as:

$$\alpha^* = \underset{\alpha}{argmax}\, \gamma_{\underline{R}}\left(x_{\underline{R}}, y_{\underline{R}+\alpha}\right), \qquad 0 \leqq \alpha < 1 \qquad (6)$$

The optimization process in (6) can be implemented by using the orthogonal projection theorem and it has been derived in [6]. The closed form of α^* is defined as:

$$\frac{\left(x_{\underline{R}}, y_{\underline{R}+1}\right)|y_{\underline{R}}|^2 - \left(x_{\underline{R}}, y_{\underline{R}}\right)\left(y_{\underline{R}}, y_{\underline{R}+1}\right)}{\left(x_{\underline{R}}, y_{\underline{R}+1}\right)[|y_{\underline{R}}|^2 - \left(y_{\underline{R}}, y_{\underline{R}+1}\right)] + \left(x_{\underline{R}}, y_{\underline{R}}\right)[|y_{\underline{R+1}}|^2 - \left(y_{\underline{R}}, y_{\underline{R}+1}\right)]} \qquad (7)$$

After obtaining α^*, the cross correlation $\gamma_{\underline{R}}\left(x_{\underline{R}}, y_{\underline{R}+\alpha^*}\right)$ is also derived in [5] and the closed form is:

$$\frac{(1-\alpha^*)\left(x_{\underline{R}}, y_{\underline{R}}\right) + \alpha^*\left(x_{\underline{R}}, y_{\underline{R}+1}\right)}{[|x_{\underline{R}}|^2((1-\alpha^*)^2\left|y_{\underline{R}}\right|^2 + 2\alpha^*(1-\alpha^*)\left(y_{\underline{R}}, y_{\underline{R}+1}\right) + \alpha^{*2}\left|y_{\underline{R+1}}\right|^2)]^{1/2}} \qquad (8)$$

C. The Proposed Spectrum Selection Scheme

Although the authors in [2] said that the second independent component is the most closet component to BVP signal, this is not always the correct case by our observation. In this paper, we apply the spectrum selection scheme proposed in [13] for fast Fourier transform method. This scheme is described shortly as follows. Firstly, we set the heart rate range to [38, 200] bits per minute (bpm). Let E_k denote the entropy of the k-th spectrum $(1 \leqq k \leqq 3)$ and S_f^k denote the normalized energy of the k-th spectrum in the f-th frequency. The calculation of entropy is defined as:

$$E_k = \sum_{f=f_{min}}^{f=f_{max}} -S_f^k * log(S_f^k) \qquad (9)$$

where f_{min} and f_{max} is calculated by:

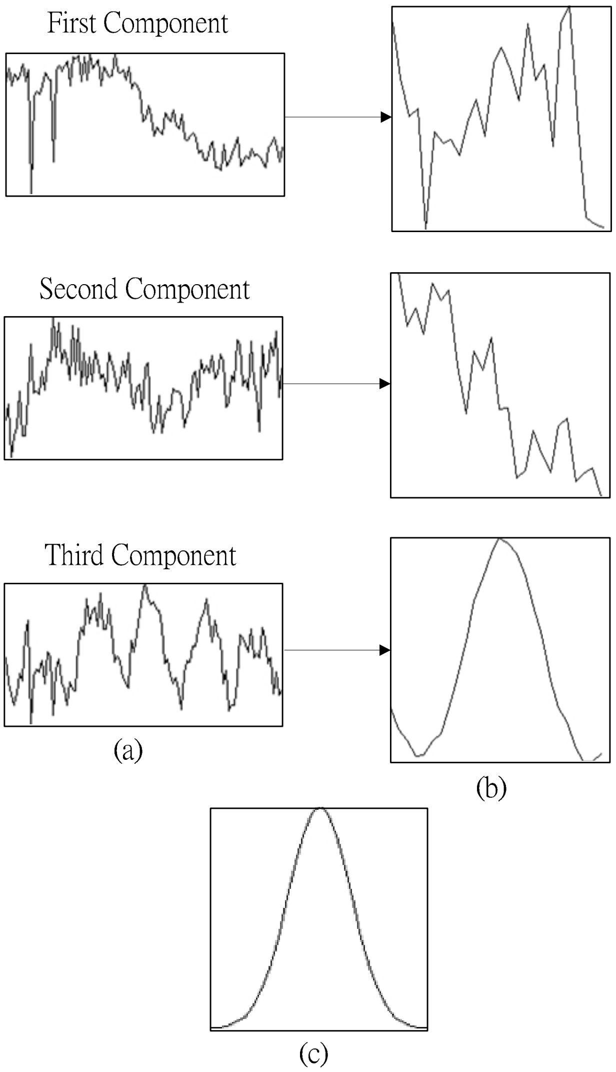

Fig. 2. Illustration for the proposed spectrum selection scheme. (a) Three independent components obtained from ICA operation. (b) Their respective cross correlation results. (c) Simulated Gaussian distribution

$$f_{min} = \frac{38*T}{60*fps}, f_{max} = \frac{200*T}{60*fps} \qquad (10)$$

where T is the total length of FFT and fps is the frame rate, the index of the target spectrum is the spectrum which has minimum entropy and is defined as:

$$S_0 = \underset{k}{argmin}\, E_k, \qquad 1 \leqq k \leqq 3 \qquad (11)$$

For autocorrelation method, we propose a Gaussian distribution matching scheme, as illustrated in Fig. 2. Firstly, if there are three independent components obtained from ICA operation, as in Fig. 2(a). Then, the the autocorrelation is performed on the operation range $[R_{min}, R_{max}]$, as in Fig. 2(b). $[R_{min}, R_{max}]$ in (4) can be calculated by:

$$R_{min} = \frac{60*fps}{200}, R_{max} = \frac{60*fps}{38} \qquad (12)$$

From the observation in Fig. 2 (b) we understood that the correct candidate is the third and it is much like a Gaussian distribution. So we simulate a Gaussian distribution as in Fig. 2 (c) to convolve with these three spectra in Fig. 2 (b) and select a spectrum with the highest score. The simulated Gaussian distribution is denoted as $G(n)$, with mean

$(R_{max}-R_{min})/2$ and deviation $(R_{max}-R_{min})/7$. Let $C_k(n)$ denote the autocorrelation results of the k-th spectrum $(1 \leqq k \leqq 3)$ and $(G * C_k)(z)$ denote the convolution results between $G(n)$ and $C_k(n)$. The index of the target spectrum is the spectrum which has maximum convolution with Gaussian distribution. This scheme can be described as:

$$S_0 = \underset{k}{argmax}(G * C_k)(z), \quad Z_{min} \leqq z \leqq Z_{max} \qquad (13)$$

where $[Z_{min}, Z_{max}]$ is the range for convolution. In this paper, we set $Z_{min} = (R_{max}-R_{min})/4$ and $Z_{max} = (R_{max}-R_{min}) * 3/4$.

D. *The Proposed Data Shift Scheme*

If a time signal is stationary, the autocorrelation is well performed for any i_0 in (1). But this is not the always case in real world. By changing i_0 we can obtain many different autocorrelation results from a single time signal. The proposed data shift scheme is based on this idea. Let $C_k^{i_0}(n)$ denote the cross correlation results of the k-th spectrum for starting point i_0. The data shift scheme can be described as:

$$S_0 = \underset{k, i_0}{argmax}(G * C_k^{i_0})(z), 0 \leqq i_0 < (L - m - R_{max}) \qquad (14)$$

where L is the total number of samples in a time signal and m is the length for cross correlation. In this paper, we set m to R_{max}.

IV. Experimental Results

In this section, we use the video database took in our lab to evaluate the performance of heart rate measurement. There are 11 different people in this database. And there are 7 different shooting scenes for recording color movies and each scene contains 9 or 10 video clips. These 7 scenes are respectively illustrated in Fig. 3. And for convenience, we named these 7 scenes as “glasses”, “noglasses”, “lab”, “meet”, “semi-outdoor1”, “semi-outdoor2”, “office”. There are totally 66 video clips in this database. There are about 600~900 frames in each video clip and the frame rate for the recording video is 20 fps. The resolution in each frame is 320X240 pixels and the recording time is about 30~60 seconds.

During video recording, the subjects were asked to face frontal and the head motion is static. Furthermore, when recording videos, an FDA-approved finger-based sensor is used to record heart rate. And the results obtained from the sensor are treated as ground truth to evaluate the performance of video-based heart rate measurement. Since the number of the results obtained from the finger-based sensor is 256 in one second, there are about 7680~15360 results during one video recording. However we assumed there is only one ground truth for each video. We do the calculation as follows: Firstly, all sensing results obtained during one video recording are sorted. Then, the 25 percent in the front part and the 25 percent in the last part are ignored. Finally, the mean value of the results of the middle part is set as ground truth for this video.

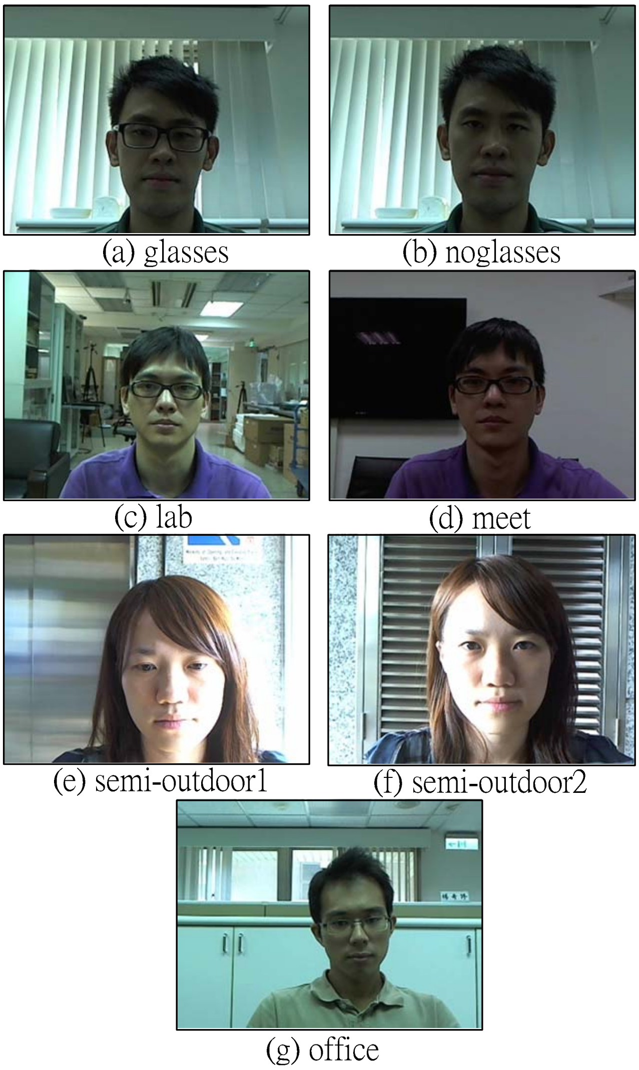
(a) glasses
(b) noglasses
(c) lab
(d) meet
(e) semi-outdoor1
(f) semi-outdoor2
(g) office

Fig. 3 Illustration for shooting scenes.

For all video sequences, Viola Jones face detector [7] was firstly applied to detect and track face region in each frame. Since the facial pose is frontal and the head motion is very lightly in all video sequences, the face region is almost correctly located in every frame. Then, the following steps are all the same as in [2] except the final step for peak interval determination.

There are two main experiments conducted in this section to compare the accuracy of the proposed peak interval determination method with the FFT-based approach proposed in [2]. Since the FFT-based approach is implemented by our own, the implementation details may be different from [2]. And we implement the FFT-based approach proposed in [2] is just for comparing peak interval determination in time domain with peak interval determination in frequency domain. The agreement between video-based heart rate measurement and a finger-based sensor was tested by Bland-Altman analysis [12].

The first experiment shows that the analyzing duration of the proposed method can be shorten to one third of the FFT-based approach and the precision is only a little lower. The experimental setting is as follows. For FFT-based approach, the analyzing duration for one outcome is 15 seconds. The first outcome is appeared in the 15-th second and the second one is appeared in the 17-th second. In other words, the analyzing samples between successive outcomes are overlapped by 13 seconds. For example, if there is a video with 30 seconds, there would have 8 outcomes for this video. For the proposed method, the analyzing duration for one outcome is 5 seconds. The first outcome is appeared in the 5-th second and the second one is appeared in the 7-th second. In other words, the analyzing samples between successive outcomes are overlapped by 3 seconds. And in order to let the number of outcomes for one video is the same for both methods, for a video clip with 30 seconds, the samples in the last 11 seconds will be ignored. Totally we have 638 pairs of measurement results over all 66 video clips and their statics are summarized in Table I. For FFT-based approach, the mean bias was -4.75 bits per minute (bpm) with 95% limits of agreement -29.22 to 19.73 bpm. For the proposed method, the mean bias was -9.96 bmp with 95% limits of agreement -36.14 to 16.22 bpm.

The the second experiment depicts that under the same analyzing duration, the precision of the proposed method is superior to FFT-based approach. The experimental setting is as follows. The analyzing time for both methods is 5 seconds. The first outcome is appeared in the 5-th second and the second outcome is appeared in the 10-th second. In other words, the analyzing samples between successive outcomes are not overlapped. For a video with 30 seconds, there would have 6 outcomes. Totally we have 409 pairs of measurement results over all 66 video clips and their statics are summarized in Table II. For FFT-based approach, the mean bias was 12.96 bits per minute (bpm) with 95% limits of agreement -50.82 to 76.75 bpm. For the proposed method, the mean bias was -9.58 bmp with 95% limits of agreement -39.46 to 20.31 bpm.

Although the precision of the proposed method is lower than that of FFT-based approach in first experiment, the difference is not much. In second experiment, the precision of the proposed method is obviously superior to FFT-based approach. Therefore, in real time application, our method is very useful for short-term analyzing duration and can also fuse with FFT-based approach for later outputs.

V. Conclusions

In this paper, we applied a novel method which combines autocorrelation and super resolution to perform peak interval determination in time domain. And we further proposed a spectrum selection scheme and data shift scheme to improve the performance. The experimental results showed that when analyzing duration of the proposed method is one third of FFT method, the precision of our method is only a little lower than FFT method. And the experimental results also demonstrated that when analyzing time is 5 seconds for both approaches, our method is superior to FFT method, which would be very useful in real time application.

TABLE I
THE EXPERIMENTAL RESULTS OF THE FIRST EXPERIMENT

	Methods	
Statistics	FFT-based approach	The Proposed Method
No. of measurement pairs	638	
Mean bias (bpm)	-4.75	-9.96
Mean absolute bias (bpm)	6.93	12.17
SD of bias (bpm)	12.49	13.36
Upper limit (bpm)	19.73	16.22
Lower limit (bpm)	-29.22	-36.14
RMSE	13.36	16.66

TABLE II
THE EXPERIMENTAL RESULTS OF THE SECOND EXPERIMENT

	Methods	
Statistics	FFT-based approach	The Proposed Method
No. of measurement pairs	409	
Mean bias (bpm)	12.96	-9.58
Mean absolute bias (bpm)	19.12	12.84
SD of bias (bpm)	32.54	15.25
Upper limit (bpm)	76.75	20.31
Lower limit (bpm)	-50.82	-39.46
RMSE	35.03	18.01

References

[1] W. Verkruysse, L.O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optics. Express,* 16(26), pp. 21434-21445, 2008.

[2] Poh, Ming-Zher, Daniel J. McDuff, and Rosalind W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics. Express,* vol. 18, issue. 10, pp. 10762-10774, 2010.

[3] Kual-Zheng. Lee, Pang-Chan. Hung and Luo-Wei Tsai, "Contact-Free Heart Rate Measurement Using a Camera," *2012 Ninth Conference On Computer and Robot Vision(CRV),* 10.1109/CRV.2012.27, pp. 147-152, 2012.

[4] T. Pursche, J. Krajewski and R. Moeller, "Video-based heart rate measurement from human faces," *2012 IEEE International Conference On Consumer Electronics(ICCE),* 10.1109/ICCE.2012.6161965, pp. 544-545, 2012.

[5] A. Noulas and B. Krose, "EM detection of common origin of multi-modal cues," in *Proceedings of ACM Conference On Multimodal Interfaces(ACM),* pp. 201-208, 2006.

[6] Y. Medan, E. Yair and D. Chazan, "Super Resolution Pitch Determination of Speech Signals," *IEEE Transactions On Signal Processing,* vol. 39, issue. 1, pp. 40-48, 1991.

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of IEEE Conference On Computer Vision and Pattern Recognition(CVPR),* pp. 511, 2001.

[8] J.-F. Cardoso, “High-order contrasts for independent component analysis,” *Neural Comput.* 11(1), pp. 157-192, 1999.

[9] M. J. Ross et al., “Average magnitude difference function pitch extractor,” *IEEE Transactions On Acoust., Speech, Signal Processing,* vol. ASSP-22, pp.353-362, 1974.

[10] W. Hess, “Pitch Determination of Speech Signals,” New York: Springer, 1983.

[11] S. Kadambe and G. F. Boudreaux-Bartels, “Application of the wavelet transform for pitch detection of speech signals,” *IEEE Transactions On Information Theory,* vol. 38, issue. 2, pp. 917-924, 1992.

[12] J. M. Bland and D. G. Altman, “Statistical methods for assessing agreement between two methods of clinical measurement,” *Lancet,* 327:307-310, 1986.

[13] Ting-Wei Lee, Gwo-Hwa Ju, Heng-Sung Liu and Yu-Shan Wu, “Liveness detection using frequency entropy of image sequences,” *IEEE International Conference On Acoustics, Speech and Signal Processing(ICASSP),* 2013.