Biomedical Optics EXPRESS

Beat-to-beat heart rate estimation fusing multimodal video and sensor data

Christoph Hoog Antink,* Hanno Gao, Christoph Brüser, and Steffen Leonhardt

Chair for Medical Information Technology, Helmholtz Institute, RWTH Aachen University, 52074 Aachen, Germany

*hoog.antink@hia.rwth-aachen.de

Abstract: Coverage and accuracy of unobtrusively measured biosignals are generally relatively low compared to clinical modalities. This can be improved by exploiting redundancies in multiple channels with methods of sensor fusion. In this paper, we demonstrate that two modalities, skin color variation and head motion, can be extracted from the video stream recorded with a webcam. Using a Bayesian approach, these signals are fused with a ballistocardiographic signal obtained from the seat of a chair with a mean absolute beat-to-beat estimation error below 25 milliseconds and an average coverage above 90% compared to an ECG reference.

OCIS codes: (170.0170) Medical optics and biotechnology; (280.0280) Remote sensing and sensors.

References and links

1. M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." Opt. Express **18**, 10762–10774 (2010).
2. S. Kwon, H. Kim, and K. Park, "Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone." EMBS (2012).
3. C. Brüser, S. Winter, and S. Leonhardt, "Robust inter-beat interval estimation in cardiac vibration signals," Physiol. Meas. **34**, 123–38 (2013).
4. T. Wartzek, C. Brüser, M. Walter, and S. Leonhardt, "Robust sensor fusion of unobtrusively measured heart rate," IEEE J. Biomed. Health Inf. **18**, 654–660 (2014).
5. T. G. Vrijkotte, L. J. van Doornen, and E. J. de Geus, "Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability." Hypertension **35**, 880–886 (2000).
6. M. Rovere, G. Pinna, and R. e. a. Maestri, "Short-term heart rate variability strongly predicts sudden cardiac death in chronic heart failure patients," Circulation **107**, 565–570 (2003).
7. C. Hoog Antink, C. Brüser, and S. Leonhardt, "Multimodal sensor fusion of cardiac signals via blind deconvolution: A source-filter approach," IEEE Computing in Cardiology (2014).
8. G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in "Proceedings of IEEE Conference on Computer Vision and Pattern Recognition," (IEEE Computer Society, Washington, DC, USA, 2013), CVPR '13, pp. 3430–3437.
9. T. Wu, V. Blazek, and H. Schmitt, "Photoplethysmography imaging: a new noninvasive and noncontact method for mapping of the dermal perfusion changes." Proc. SPIE **4163**, 62–72 (2000).
10. M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera," Biomed. Opt. Express **6**, 1565–1588 (2015).
11. L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," Physiol. Meas. **35**, 807–31 (2014).
12. M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," IEEE Trans. Acoust. Speech Signal Process. **22**, 353–362 (1974).
13. C. Brüser, J. Kortelainen, S. Winter, M. Tenhunen, J. Parkka, and S. Leonhardt, "Improvement of force-sensor-based heart rate estimation using multi-channel data fusion," IEEE J. Biomed. Health Inform **19**, 227–235 (2015).
14. 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," , vol. 1 (IEEE, Los Alamitos, CA, USA, 2001), vol. 1, pp. 511–518.

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

15. B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in "Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2," (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1981), pp. 674–679.
16. C. Tomasi and T. Kanade, "Detection and tracking of point features," Tech. Rep. (1991).
17. J. Shi and C. Tomasi, "Good features to track," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition pp. 593–600 (1994).
18. R. Irani, K. Nasrollahi, and T. B. Moeslund, "Improved pulse detection from head motions using dct," in "9th International Conference on Computer Vision Theory and Applications," (Lisabon, 2014).
19. W. Verkruysse, L. O Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." Opt. Express **16**, 21434–21445 (2008).
20. M. Lewandowska, J. Ruminski, and T. Kocejko, "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity," Proceedings of the Federated Conference on Computer Science and Information Systems pp. 405–410 (2011).

1. Introduction

Unobtrusive acquisition of biosignals for health- and wellness applications has experienced increasing popularity in recent years [1–4]. In particular, monitoring of the heart rate and its variability outside the classical scenarios such as hospitals and sleep laboratories is an active area of research. It offers great medical potential, as the heart rate variability (HRV) has a wide range of applications from work stress analysis [5] to the prediction of sudden cardiac death in chronic heart failure patients [6].

While unobtrusive measurement modalities greatly increase comfort and the number of application scenarios, they are normally easily disturbed by motion artifacts and have a lower signal to noise ratio (SNR) compared to clinical modalities such as the conductive electrocardiogram (ECG) or the finger-attached photoplethysmogram (PPG). Thus, when analyzing unobtrusively acquired measurement data, episodes that contain no valid information can occur and must be excluded from subsequent processing. At the same time, the more data is excluded, the bigger is the chance that important information is missed. Thus, an unobtrusive measurement system is often evaluated in terms of accuracy and coverage, as one can often only be improved at the cost of the other.

To overcome this, the fusion of biosignals obtained from multiple channels and multiple modalities has proven to be a promising approach. The basic idea is that each acquired biosignal originates from a single mutual source [7], see Fig. 1. This source, the human heart, can be modeled to create a *virtual* signal that triggers several *physical* responses. With different sensors, these multimodal signals can be acquired, for example, a differential electric potential (ECG), a change in optical property (PPG), and a mechanical impulse (ballistocardiography, BCG). While the signal morphology as well as the relative peak location differs, the beat-to-beat interval is the same in each channel.

In unobtrusive monitoring, video-based methods play an increasingly important role due to the low cost and ubiquitous availability of sensors in the form of, for example, webcams and smartphones [1,2]. Additionally, since no direct contact to the subject is necessary, no concerns in terms of safety and hygiene arise. At the same time, this makes video-based methods very susceptible to motion artifacts. Moreover, some sort of illumination and a direct line of sight is required, which complicates applications like sleep monitoring. Here, other modalities that can easily be integrated into the mattress, such as capacitive electrocardiography (cECG) [4] or BCG [3], have proven to allow the accurate and robust determination of beat-to-beat intervals (BBI) for HRV analysis. Vast portions of daily life in post-industrial countries are spent sitting, often in front of computers. At the same time, diseases of the cardiovascular system as well as negative stress due to excessive workload pose serious threats to the public health. Thus, constant unobtrusive monitoring of seated subjects has great potential and several concepts have been proposed. Most of those, however, are based on single modalities or do not allow

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

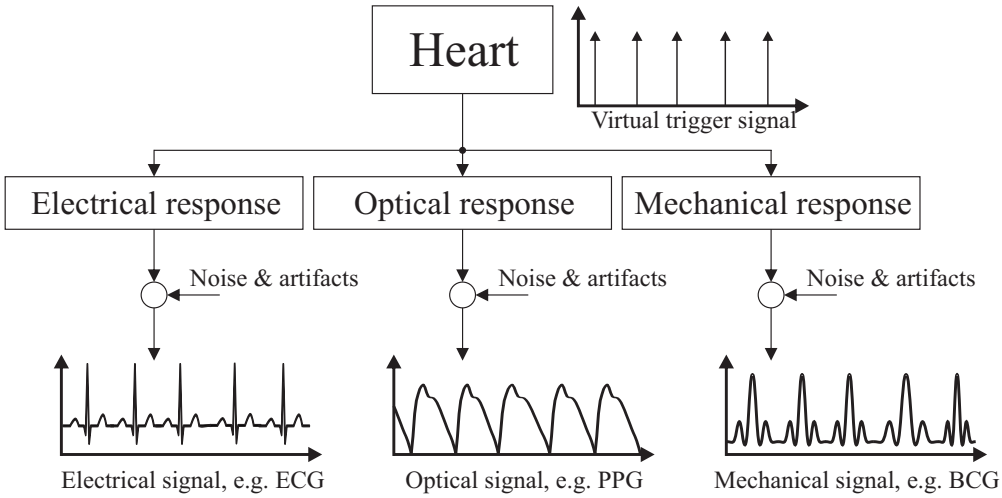

Fig. 1. Model for the generation of multimodal cardiac signals. While the signal morphology as well as the relative peak location differs, the beat-to-beat interval is the same.

HRV analysis with beat-to-beat accuracy.

According to the definition introduced above, the video signal obtained from a subjects head is by itself a multimodal biosignal, as it contains the involuntary head motion [8] that originates from the mechanical impulse of the heart as well as the changes in optical property of the skin [9] caused by the superficial perfusion. Noise and artifacts might influence channels by varying degrees: While the BCG and the pulse-related head oscillations are very susceptible to motion artifacts, advanced video processing methods allow the extraction of PPG signals even if the subject is moving [10]. On the other hand, a remote PPG signal is more difficult to obtain from subjects with a higher melanin concentration, i.e. a darker skin. It is reasonable to assume that there is no influence of melanin content on the head motion tracking and there is obviously none on the BCG signal. Stationary, periodic changes in illumination can be compensated in remote PPG sensing [11], while non-stationary lighting conditions are probably harder to deal with in motion estimation and especially in remote PPG sensing. The signal of a mechanically coupled BCG sensor in the seat of a chair is obviously insensitive to the lighting conditions but might be corrupted by motion of the lower extremities. Thus, it can be assumed that even if no strong artifacts are present, exploiting redundancies in multimodal signals can help improve coverage and accuracy in the procession of unobtrusively acquired biosignals with a low SNR.

In this paper, we bridge the gap between video- and contact-based unobtrusive monitoring modalities by multimodal sensor fusion of video and BCG data. We demonstrate that beat-to-beat intervals can be estimated with an average absolute error below 25 ms and coverage above 90 % when compared to an ECG reference. We further show that even the fusion of only the motion- and photoplethysmographic component of the video data greatly improves coverage while maintaining high accuracy.

The paper is structured as follows: In the next section, the measurement setup as well as the algorithmic details are described. Results and discussion are presented in section 3, the paper is concluded in section 4.

2. Method and materials

The measurement system setup is visualized in Fig. 2. For image acquisition, a consumer grade

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

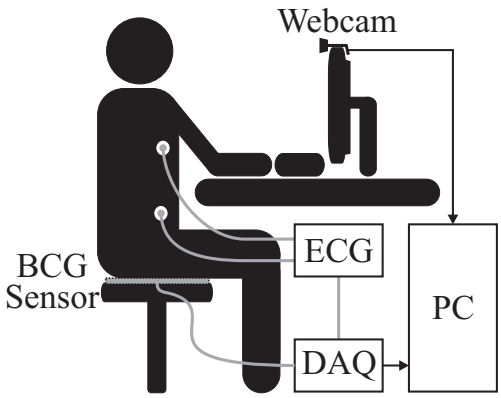

Fig. 2. Overall system setup using a webcam and a BCG mat sensor. For reference, a single channel ECG is recorded. BCG and ECG are digitized in parallel.

webcam *"C260"* by *Logitech international S.A.*, Apples, Switzerland, was used. Images were acquired at $f_{\mathrm{s,video}} = 30$ Hz with a resolution of 800 by 600 pixels. Regular environmental light consisting of a mixture of sunlight and light from fluorescent tubes installed in the ceiling was used for illumination.

The ballistocardiographic signal was obtained by placing an *"EMFi transducer"* mat (L-Series by *Emfit Ltd*, Vaajakoski, Finnland) on the seat of a chair. A custom built analog amplifier / bandpass with a passband of 0.01 to 200 Hz and gain of 24.61 dB was used for analog signal conditioning. Analog to digital conversion was performed using the *"NI USB-6212"* (*National Instruments*, Austin, Texas, USA) data acquisition systems (DAQ) at the sampling frequency $f_{\mathrm{s,DAQ}} = 1000$ Hz.

For reference, a single-channel ECG was recorded using the *"IntelliVue MP70"* patient monitor (*Philips*, Amsterdam, The Netherlands). Its analog output was sampled in parallel to the BCG signal and QRS-peaks were extracted using the *"Open Source ECG Analysis Software"* by *EP Limited* (http://www.eplimited.com/).

Four healthy volunteers with fair skin were asked to sit perfectly still in front of the camera (Trial 1), to perform a reading-task from the computer screen without motion (Trial 2), and to read from the computer screen without further instructions (Trial 3). In every trial, recordings were performed for two minutes. An additional fifth volunteer with darker skin and extremely low resting heart rate participated once to demonstrate the algorithms wide range of detectable intervals. The average heart rates for all participants and trials as determined by the ECG are listed in Table 1.

Table 1. Heart rate for all participants and trials. Note the extremely low heart rate of participant five.

	Average Heart Rate (BPM)				
Trial	Subject #1	Subject #2	Subject #3	Subject #4	Subject #5
1	61.48	74.02	64.05	76.34	43.89
2	65.44	70.33	68.30	75.95	
3	64.02	67.18	71.13	76.62	

In the following, an algorithm that performed very successfully in the beat-to-beat analysis of multichannel BCG-data [3] is recapitulated and augmented by an adaptive Gaussian prior. As argued above, cardiac signals acquired through different modalities might exhibit various

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

waveforms while the underlying periodicity is the same in all channels. It is further reasonable to assume that one heart beat will show great similarity to the one before and that the interval between beats can be determined by analyzing each signals self-similarity. Various metrics for the assessment of self-similarity exist. Here, the short-time autocorrelation (STA) function is widely used. Let $x(n)$ be a time-discrete signal and

$$\omega_i(\nu) = x(n_i + \nu) \tag{1}$$

be an analysis window with index i centered around n_i (Fig. 3(A)). The index is omitted in the following derivation for better readability. Commonly, the STA for each lag η for a window of constant length L is given by

$$S_{\mathrm{STA}}(\eta) = \frac{1}{L} \sum_{\nu=-L/2}^{L/2-\eta} \omega(\nu)\omega(\nu+\eta), \tag{2}$$

see also Fig. 3(B). This implies that for each candidate lag η, $L-\eta$ samples are considered. This approach can be used successfully if the average interval length within the window is of interest. If, however, the interval between *exactly* two consecutive beats is of interest, the lag-adaptive short-time autocorrelation (LASTA)

$$S_{\mathrm{LASTA}}(\eta) = \frac{1}{\eta} \sum_{\nu=0}^{\eta} \omega(\nu)\omega(\nu-\eta) \tag{3}$$

ensures that the exact number of samples necessary for each candidate lag η are considered, see also Fig. 3(C).

Another metric to assess self-similarity is the modified average magnitude difference function (AMDF) used in speech processing for pitch extraction [12]. The modified AMDF

$$S_{\mathrm{AMDF}}(\eta) = \left(\frac{1}{\eta} \sum_{\nu=0}^{\eta} |\omega(\nu) - \omega(\nu-\eta)| \right)^{-1} \tag{4}$$

also uses the lag-adaptive window and is inverted to assume larger values for lags that indicate more self-similarity (Fig. 3(D)).

As a third metric, the maximum amplitude pairs (MAP) function can be considered as indirect peak-detection,

$$S_{\mathrm{MAP}}(\eta) = \max_{\nu \in \{0,\ldots,\eta\}} \left(\omega(\nu) + \omega(\nu-\eta)\right). \tag{5}$$

Like LASTA and AMDF, the MAP function assumes large values for lags that indicate self-similarity, see Fig. 3(E).

To exploit the different noise characteristics of the three estimators and allow multimodal signal fusion, a Bayesian approach is chosen. This can be achieved by interpreting the estimator results as an a-posteriori probability density function (PDF) in a loose statistical sense. Thus, $p(\eta|S_{\mathrm{LASTA}})$, $p(\eta|S_{\mathrm{AMDF}})$ and $p(\eta|S_{\mathrm{MAP}})$ represent the probability that η is the correct interval length according to the respective estimator. Through linear shifting and scaling of each estimator result, the properties of a proper PDF, i.e., positivity and unit area under the curve, can be achieved. If we assume a uniform a-priori distribution $p(\eta)$, we get

$$p(\eta|S_{\mathrm{LASTA}}, S_{\mathrm{AMDF}}, S_{\mathrm{MAP}}) \propto p(\eta|S_{\mathrm{LASTA}})p(\eta|S_{\mathrm{AMDF}})p(\eta|S_{\mathrm{MAP}}) \tag{6}$$

by applying Bayes theorem. Finding the optimal interval η_{opt} is thus reduced to finding the maximum of the multiplication of the scaled estimator outputs, see Fig. 3(F). The extension

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

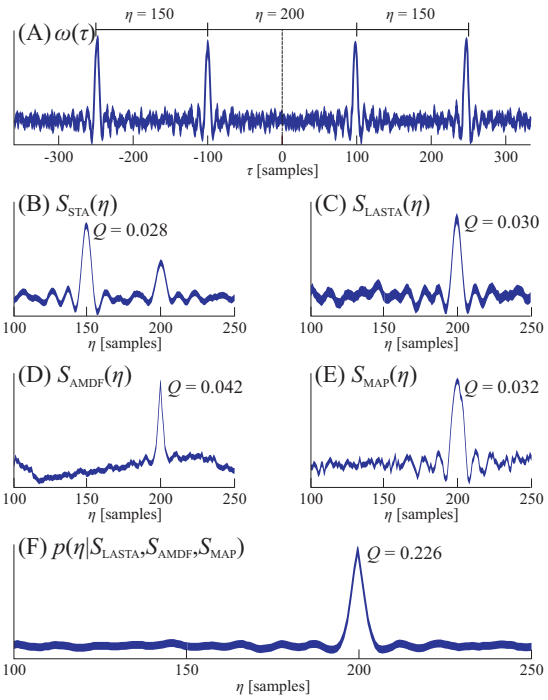

Fig. 3. Visualization of the different self-similarity measurements and their fusion. In (A), a window centered between two beats of an artificial signal with an interval of $\eta = 200$ samples is shown. Since the window contains multiple beats, the STA (B) will be dominated by the two 150-sample intervals surrounding the interval of interest. On the other hand, LASTA (C), AMDF (D) and MAP (E) will show a peak at the correct location. Fusing the three estimators (F) results in an even more distinct peak. All curves are in arbitrary units, the quality metric Q is defined as the ratio of the peak height to the area under the curve.

towards a multimodal, N-channel setting [13], where each estimator is calculated for each of the N channels, is straightforward,

$$\eta_{\text{opt}} = \arg\max_{\eta} p(\eta|S_{\text{Fusion}}) = \arg\max_{\eta} \prod_{l=1}^{N} p(\eta|S^l_{\text{LASTA}}) p(\eta|S^l_{\text{AMDF}}) p(\eta|S^l_{\text{MAP}}). \tag{7}$$

To determine whether or not an estimated interval is reliable, the quality metric

$$Q = \frac{p(\eta_{\text{opt}}|S_{\text{Fusion}})}{\sum_{\eta=1}^{L} p(\eta|S_{\text{Fusion}})}, \tag{8}$$

i.e., the ratio of the peak height to the area under the curve, is calculated. If $p(\eta|S_{\text{Fusion}})$ does not shows a clear maximum, Q is small. Only intervals with $Q > Q_{\text{th}}$ are accepted and thus, the choice of Q_{th} determines the trade-off between coverage and accuracy as introduced above. Here, a fixed $Q_{\text{th}} = 0.3$ is used unless noted otherwise.

The window length L determines the maximum interval, i.e., the minimum heart rate that can be detected. In this approach, it was set to 1500 ms, corresponding to 40 BPM. The shift period of the moving window was 200 ms and a low-pass filter with a passband of 10 Hz was applied to all signals before interval estimation. Moreover, the maximum detectable heart rate was set to 140 BPM. Apart from this, no temporal constraints are put on the algorithm and it is

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

thus capable to process severely arithmetic data. However, applications like stress or workload monitoring exist, where the HRV itself can be expected to lie in a non-pathologically high range. In this case, an interval has a high probability to lie within a certain range determined by the proceeding intervals. Here, we assume an adaptive Gaussian prior,

$$p(\eta) = p(\eta_i|\bar{\eta}_{i-1}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\eta_i - \bar{\eta}_{i-1}}{\sigma}\right)^2\right), \tag{9}$$

with the exponentially weighted moving average of the previous estimated intervals

$$\bar{\eta}_i = \alpha \cdot \bar{\eta}_{i-1} + (1-\alpha) \cdot \eta_i. \tag{10}$$

Here, $\alpha = 0.7$ was used and σ was chosen to be 0.05 seconds, which corresponds to 3 beats per minute. Choosing a smaller σ would make the algorithm more robust towards outliers. At the same time, it would limit its ability to process severely irregular cardiac signals and track fast changes.

The one-dimensional BCG-signal only needs to be bandpass filtered before interval estimation. On the other hand, the two-dimensional video signal needs to be preprocessed more intensely. First, the face is identified using the Viola-Jones face detector [14] in the first frame. In subsequent frames, the KLT feature tracker [15–17] is used to ascertain the face position and determine head motion. The latter consists of a cardiac component and of motion artifacts. To extract the signal of interest, the approach described in [8] is used. Principal component analysis (PCA) is performed and the first five principal components are analyzed via Fast Fourier Transformation (FFT) to find the most harmonic signal, i.e., the signal most likely to contain the cardiac component termed "Video Motion". A more recent approach relying on the discrete cosine transformation [18] was not found to improve the result.

To extract the photoplethysmographic information from the video-stream, several approaches are described in the literature [9, 19]. Recent developments include the use of independent component analysis (ICA, [1]), PCA [20] or a combination of the RGB-channels and were implemented and evaluated on our data. The best results, however, were obtained by the straight forward approach of taking the spatial average of the green channel of a tracked region of interest placed on the subjects forehead as illustrated in Fig. 4. Other methods, for example including ROIs on the subject's cheeks, showed no improvement. To reduce the respiratory component and high frequency noise, a bandpass filter with a passband of 0.7 to 3.6 Hz was applied. This channel is termed "Video PPG" in the following.

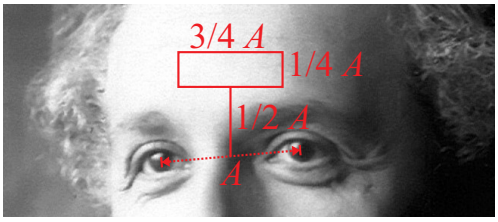

Fig. 4. The region of interest is found by locating the center of the eyes and determining the distance A as proposed by [20].

3. Results and discussion

For all 13 two-minute recordings, six scenarios were evaluated. Interval estimation was performed using

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

1. only the BCG channel (BCG),
2. only the photoplethysmographic information extracted from the video (Video PPG),
3. only the motion information extracted from the video (Video Motion),
4. the PPG and motion video information (Fusion Video),
5. all three signals (Fusion all)

and finally

6. all three signals assuming the adaptive Gaussian prior described above (Fusion all w. Prior).

In Table 2, all results are listed in terms of mean absolute interval estimation error ($\overline{E}_{\text{abs}}$) in milliseconds as well as coverage in percent. The information is presented graphically in Fig. 5.

Table 2. Mean absolute interval estimation error ($\overline{E}_{\text{abs}}$) in milliseconds as well as coverage in percent for all subjects, trials and scenarios.

		$\overline{E}_{\text{abs}}$ [ms] (Coverage [%])					
		BCG	Video	Video	Fusion	Fusion	Fusion
Trial	#		PPG	Motion	Video	all	all w. Prior
1	1	14.1 (77.9)	21.0 (43.4)	17.9 (56.6)	26.5 (94.7)	20.0 (95.6)	18.8 (95.6)
1	2	37.3 (64.2)	23.6 (73.7)	38.4 (20.4)	23.6 (94.2)	14.5 (94.9)	12.5 (94.9)
1	3	288.9 (65.3)	18.3 (64.4)	23.3 (11.9)	25.6 (87.3)	26.7 (94.1)	24.4 (96.6)
1	4	152.8 (29.1)	29.1 (66.0)	31.7 (25.5)	29.5 (75.9)	30.1 (74.5)	29.4 (90.1)
2	1	15.0 (76.7)	27.6 (35.0)	21.5 (49.2)	24.3 (92.5)	16.7 (95.0)	14.6 (96.7)
2	2	20.7 (48.5)	16.0 (80.8)	45.2 (12.3)	21.0 (88.5)	19.4 (90.8)	20.6 (95.4)
2	3	158.8 (26.0)	25.5 (42.5)	75.7 (11.8)	41.2 (73.2)	34.5 (85.8)	32.7 (89.0)
2	4	124.8 (14.9)	40.4 (48.2)	31.1 (21.3)	25.8 (75.9)	33.5 (88.7)	31.5 (93.6)
3	1	20.8 (73.1)	30.1 (51.3)	20.3 (61.3)	24.9 (95.0)	16.0 (93.3)	14.4 (94.1)
3	2	6.1 (69.4)	15.0 (64.5)	17.5 (25.0)	15.5 (74.2)	9.2 (76.6)	17.2 (77.4)
3	3	107.8 (18.9)	60.4 (24.2)	109.0 (9.1)	56.7 (47.7)	58.8 (62.9)	45.6 (87.1)
3	4	120.5 (17.6)	25.8 (52.1)	160.0 (10.6)	26.0 (54.2)	31.2 (62.0)	29.3 (85.2)
-	5	71.3 (48.1)	24.8 (8.6)	414.3 (14.8)	304.7 (24.7)	28.6 (76.5)	23.8 (81.5)
mean		71.1 (48.4)	25.5 (50.4)	45.6 (25.4)	31.8 (75.2)	25.3 (83.9)	24.4 (90.5)

Additionally, a gross analysis is performed, i.e. all 1733 intervals of all recordings are analyzed together. The result is presented in terms of coverage as well as median, 5^{th} and 95^{th} percentile of the absolute interval error in milliseconds in Fig. 6. In Fig. 7, the Bland-Altman-Plot for the scenario "Fusion all w. Prior" displaying all intervals is shown. Finally, for the same scenario, one recording with excellent coverage and error as well as the recording with the highest mean absolute error are displayed in Fig. 8.

Several observations can be made. First, all three modalities show very different results. In the gross statistic (Fig. 6), Video Motion shows the lowest coverage. The coverage of BCG and Video PPG as well as the median error is comparable, while the 95^{th} percentile is more than five times higher for BCG. Comparing the single recordings in Table 2 further reveals that there is no optimal modality per se: The best error and coverage is achieved via BCG for Trial 1 #1 and via Video PPG for Trial 1 #2. For Trial 3 #1, Video Motion achieves the lowest error, BCG the highest coverage.

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

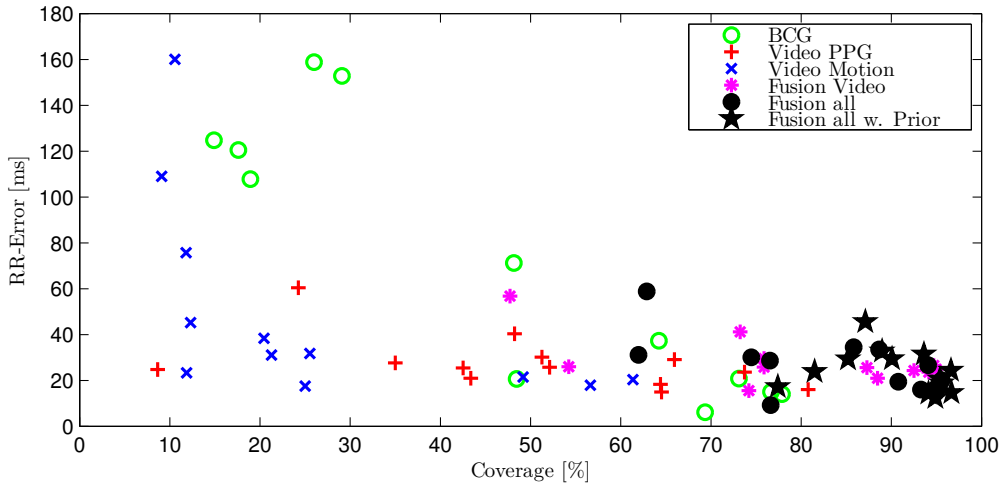

Fig. 5. Average absolute RR-error and coverage for each trial and participant. For better clarity, three outliers (Trial 1 #3 "BCG"; #5 "Video Motion" and #5 "Fusion Video") were excluded, please see Table 2.

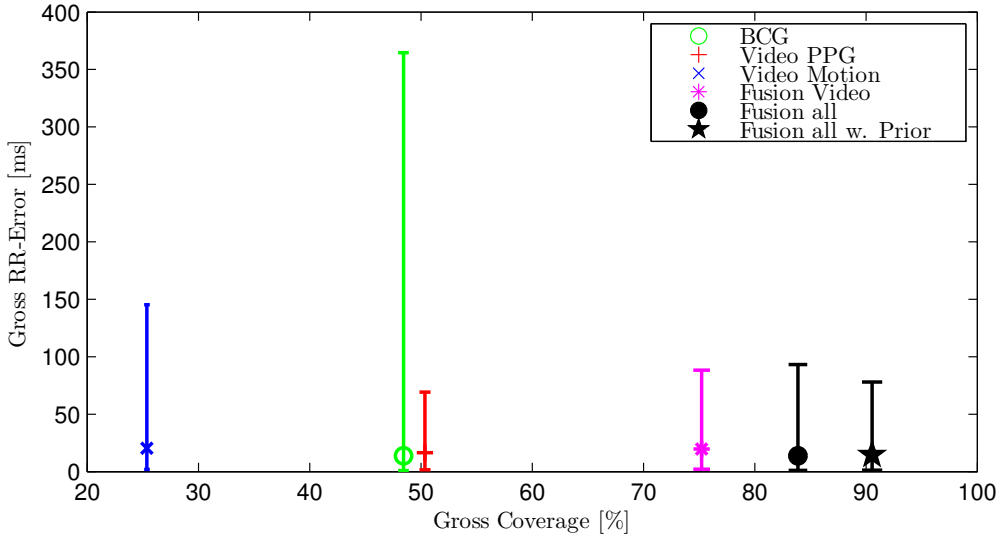

Fig. 6. The gross statistic consider all 1733 intervals of all trials and subjects together. The markers indicate the median value, the bars the 5th and 95th percentile.

If multiple channels are fused, coverage is always improved. The mean absolute error of the fused signal is always improved compared to the worst channel but might be inferior to the optimal signal in some cases, while in others, the fused result might be better than any single channel. When averaging all recordings, the mean error as well as coverage is improved beyond the individual modalities if all are fused together. Additionally, when introducing the adaptive Gaussian prior, coverage as well as mean error could be further improved. Examining this scenario in the Bland-Altman-Plot (Fig. 7), an insignificant bias of $+4.37$ ms can be determined. Additionally, no systematic difference between the estimation of relatively small intervals of

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

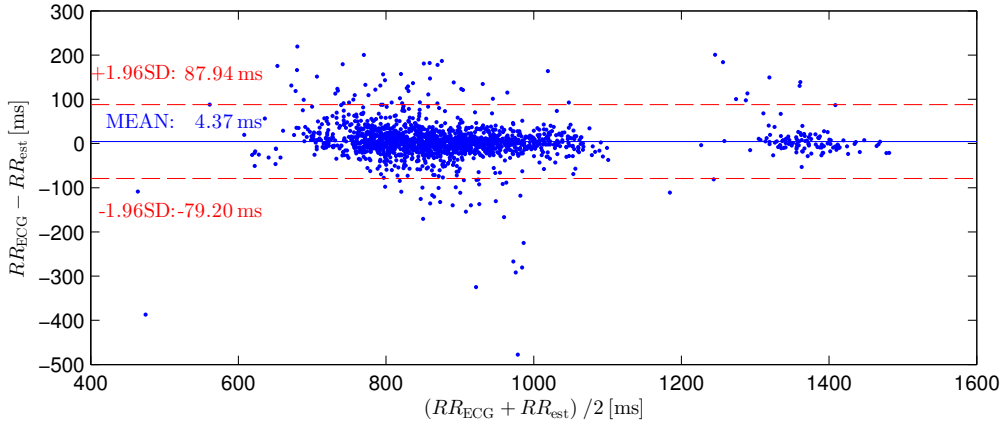

Fig. 7. Bland-Altman-Plot for the scenario "Fusion all w. Prior" comparing 1733 estimated intervals to the ECG reference. Note that the mean error is very small and that no systematic bias between small and large estimated intervals can be observed.

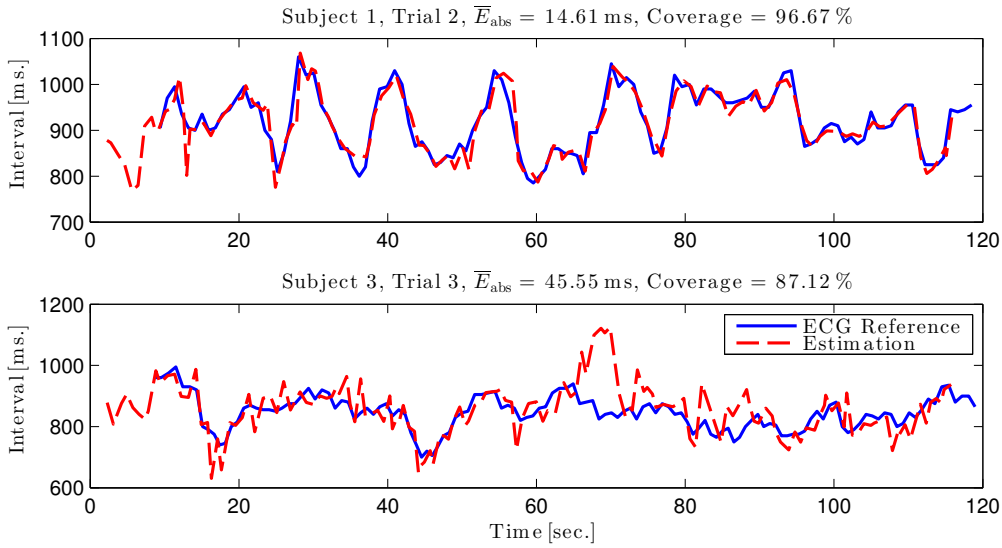

Fig. 8. Interval time course of one recording with excellent coverage and error (top) and the highest mean absolute error (bottom) for the scenario "Fusion all w. Prior".

subjects 1 to 4 (about 900 ms) and relatively large intervals of subject 5 (about 1400 ms) can be observed.

Encouragingly, Fig. 8 reveals that even in the worst recording of this scenario, the average heart rate as well as its trend can be tracked accurately.

Analyzing subject 5 in detail, some interesting observations can be made: While the results of the BCG signal almost perfectly resemble the mean values in terms of error and coverage, the video-results are far worse. Although Video PPG shows a slightly better-than-average error, coverage is only 9% while the average for this modality is 50%. This can be explained by the subjects darker skin tone. It does, however, not explain why the error for Video Motion is almost an order of magnitude worse than the average. Analyzing the morphology of the signal reveals

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

that for this particular subject, one heartbeat is often represented by two very similar peaks. This causes the estimator to wrongly estimate the half interval, although the signal shows the main peak at the location of the true average heart rate in the Fourier spectrum. In pitch estimation, this is known as the "octave error", which leads to a very high mean absolute interval estimation error. Encouragingly, the fusion of all three modalities reduces the error for subject five to 24 ms with a coverage of 82%.

To further analyze the influence of motion on both video-based interval estimations, Fig. 9 shows $\overline{E}_{\text{abs}}$ and the coverage for Video Motion and Video PPG over the mean feature movement. Subject 5 is excluded since it only participated in one trial and the low performances using the video-based channels can be considered outliers as argued above. Feature movement

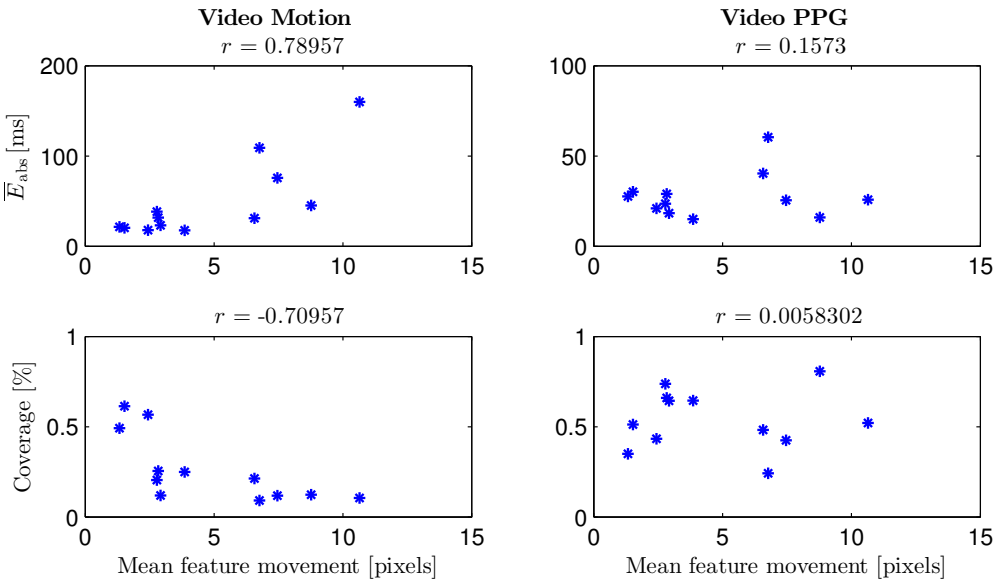

Fig. 9. $\overline{E}_{\text{abs}}$ and coverage for Video Motion and Video PPG over the mean feature movement for subjects 1 - 4. Note that a correlation of mean feature movement and interval estimation quality can only be observed for Video Motion.

is defined as the standard deviation over time of the horizontal coordinate of each feature point. Over all features, the mean is calculated; r indicates the correlation coefficient. One can see that a positive correlation between mean feature movement and $\overline{E}_{\text{abs}}$ as well as a negative correlation between mean feature movement and coverage exists for Video Motion. For Video PPG, no correlation can be observed. Thus, a negative influence of (even arguably small) motions on the tracking of cardiac-related head-oscillations can be shown, while no such influence exists on the video PPG signal.

The influence of the quality threshold Q_{th} is demonstrated in Fig. 10 for BCG, Video PPG and Video Motion. The trade-off between accuracy and coverage is visible. Moreover, the influence of Q_{th} clearly depends on the modality to be analyzed.

Using Video Motion, 405 heartbeat intervals are detected. For Video PPG and Fusion Video, these values are 842 and 1237, respectively. Analyzing the distribution in detail, a large overlap of 239 intervals that are detected with both Video PPG and Video Motion becomes apparent. At the same time, 297 (24%) of the intervals Fusion Video detects are not detected by neither video modality alone. A similar observation can be made when the BCG channel is added. In the Fusion all scenario, a total of 1359 intervals is estimated. Here, still 219 (16%) of the

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

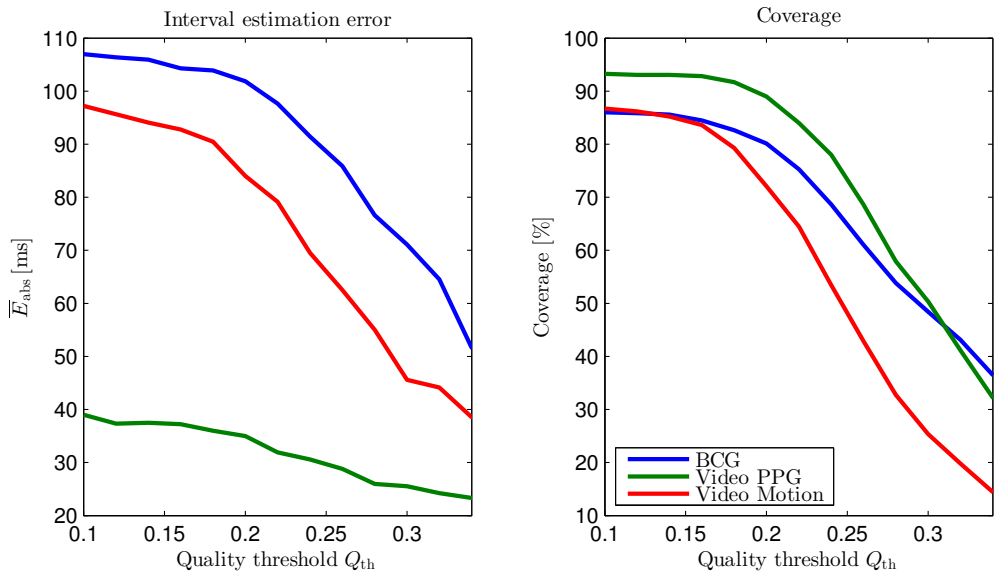

Fig. 10. Interval estimation error and coverage over quality threshold Q_{th}. The choice of Q_{th} determines a modality-dependent trade-off between accuracy and coverage.

intervals could not be estimated by any individual modality. This demonstrates that our interval estimation approach can exploit redundancies in multimodal biosignals even if individual SNRs forbid beat-to-beat interval estimation using each signal separately. This is supported by the mean quality metric of estimated intervals $\overline{Q}$: For the individual modalities, the values are $\overline{Q}_{\mathrm{BCG}} = 0.33$, $\overline{Q}_{\mathrm{Video\ Motion}} = 0.26$ and $\overline{Q}_{\mathrm{Video\ PPG}} = 0.31$. For the fused modalities, they increase to $\overline{Q}_{\mathrm{Fusion\ Video}} = 0.39$ and $\overline{Q}_{\mathrm{Fusion\ all}} = 0.62$, reflecting an increase in SNR of the fused self-similarity signal, see also Fig. 3.

4. Conclusion

In this paper, a novel multimodal sensor fusion approach for unobtrusive vital sign monitoring is presented. The motion and photoplethysmographic component originating from cardiac activity are extracted from a webcam video stream and fused using a Bayesian approach. This improved the coverage of the beat-to-beat interval estimation from 25 % (only motion) and 50 % (only PPG) to 75 % while maintaining a low error of 32 ms compared to an ECG reference. This is a very promising result, considering the sample time of the video signal was 33 ms. By fusing an additional ballistocardiographic signal unobtrusively acquired with a mat placed on the seat of a chair, coverage and error were further improved to 84 % and 25 ms, respectively. As no constraints are put on beat-to-beat intervals except a maximum interval length, severely arrhythmic data could be processed in theory. We could show that our interval estimation approach can exploit redundancies in multimodal biosignals even if individual SNRs forbid beat-to-beat interval estimation. A coverage of 90 % at only 24 ms average absolute error could be achieved by introducing a novel adaptive Gaussian prior. Although this constraints the modified algorithm, it only limits its ability to process cardiac signals exhibiting fast changes in HR, i.e., cases with a *very high* HRV. If the HRV lies in a normal range or is reduced, this constraint does not negatively affect the estimation. This is promising, as it is often a *reduced* HRV that is associated with an unhealthy state [5, 6].

As of now, the approach has only been tested on a small group of healthy volunteers. To

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015

test its ability to correctly estimate beat-to-beat intervals in severely arrhythmic data in practice, a large study containing healthy and non-healthy subjects is necessary. Moreover, several motion scenarios as well as subjects with various skin tones should be included. To further improve numeric results, more advanced algorithms for video preprocessing could be evaluated. Thus far, no improvement could be noticed by implementing several methods proposed in the literature for preprocessing the color information over using the green channel. Hence, we suspect that more room for improvement lies in the advancement of the facial tracking, as this would result in more reliable motion estimation and stable photoplethysmographic information simultaneously. Finally, more unobtrusive measurement modalities such as the cECG should be integrated into the measurement setup to further increase accuracy and coverage.

Acknowledgments

The authors would like to thank Matthias Kramer, Rajib Mondal, Miriam Rimke and Christoph Weiss for participating in the evaluation.

#236013 Received 10 Mar 2015; revised 1 Jun 2015; accepted 1 Jun 2015; published 15 Jul 2015