Technische Universiteit
Eindhoven
University of Technology

Department of Electrical Engineering
Studentenadministratie

Den Dolech 2, 5612 AZ Eindhoven
P.O. Box 513, 5600 MB Eindhoven
The Netherlands
http://w3.ele.tue.nl/nl/

Author
G.L.L.H. Schmitz

Order issuer
Prof. De Haan

Group/Chair: Electronic Systems

Reference
Series: Master graduation
paper, Electrical Engineering

Date
27 January 2011

Video Camera based Photoplethysmography using Ambient Light

Prof. De Haan

Video Camera based Photoplethysmography using Ambient Light

G.L.L.H. Schmitz
Philips Research Laboratories, Video & Imaging Processing group
Email: g.l.l.h.schmitz@student.tue.nl

***Abstract* — This paper concerns a new photoplethysmography methodology using a digital video camera and ambient visible light. No special dedicated light sources are used. Plethysmographic signals are measured contactless and remotely up to several meters. Although numerous vital body signs can be measured with this method, this paper concerns on detection and monitoring of the heart rate signal. For this purpose facial skin pixels can be selected manually and automatically from the movie frames. Applying spatial, temporal and frequency processing to the pixels shows that the plethysmographic signal is most pronounced in the green channel. The method may be useful for many healthcare and wellbeing solutions. It can also be used for robust presence detection.**

***Index Terms* — Photoplethysmography, Robust presence detection, Unobtrusive monitoring of vital body signs, Video camera based.**

I. Introduction

An essential part in many lifestyle and healthcare applications is monitoring of vital body signs. Three basic vital signs are heart rate (HR), Oxygenation (SpO2) and respiration rate (RR). In order to assess the most basic body functions, these three vital signs have to be detected and guarded. Currently, full-contact vital body signs sensors like finger probes, ear probes or chest belts in combination with a special sport watch are used in fitness, home health management and hospitals [1], [2], [3]. These sensors have to be placed on the appropriate position and are not convenient to wear. Wearing contact sensors also reduces free movement of its user. Also, on-body sensors are hard to keep hygienic. Unobtrusive monitoring of vital body signs eliminates the need for on-body sensors, simplifying the use of many wellbeing and healthcare applications.

Before explaining the proposed method, the principle of photoplethysmography (PPG) is described. A plethysmograph ('plethysmos' is the Greek word for increase) is an instrument for measuring variations in volume within the human body, resulting from fluctuations in the amount of blood it contains. These transient changes occur with every heartbeat [4]. PPG, introduced in the 1930's [5], is an optical obtained plethysmograph for detection of the cardio-vascular pulse wave that propagates through the body. PPG is a non-invasive method that detects the change in volume of the blood vessels by illuminating the skin and measuring the amount of light either transmitted or reflected to a photosensor. It is based on the principle that the absorption of light is dependent on the variations in blood volume in the vessels. This is shown in fig. 1. In the detected PPG signal the desired pulsating (AC) signal is present. Absorption due to pulsating arterial blood is very small compared to the absorption due to non-pulsating (DC) blood. This figure also shows that the absorption due to skin, bone and tissue is dominating over absorption due to blood [6]. Typical PPG methods use dedicated light sources with red or infra-red wavelengths. With these methods, ambient light is considered as a source of noise.

With the proposed method surrounding ambient visible light is not considered as a noise source, but used as a source of illumination. Instead of a special photodiode, a simple digital video camera is used for measuring the intensity of the reflected light from the skin. This video camera based method, called remote PPG, provides remote sensing of vital body signs at larger distances, without obtrusive sensors and dedicated light sources. Compared to the types of plethysmography using dedicated light sources, this method is easy to set up, simple to use and low in cost. Other advantages over on-body sensors are detection and monitoring of different vital signs from multiple persons simultaneous. No user cooperation is needed because evaluation of the condition of the user is done fully automatic. Furthermore, unobtrusive monitoring of vital signs can also be used for robust human presence detection.

Because there is no direct contact between the detector and the skin surface, one of the issues to be considered with this

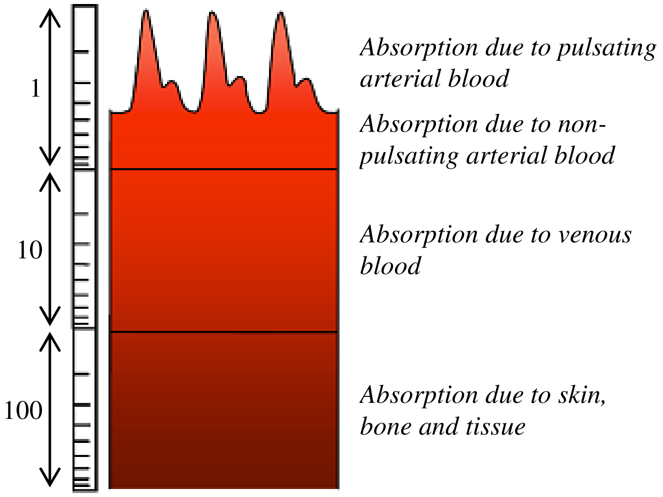

1
10
100
Absorption due to pulsating arterial blood
Absorption due to non-pulsating arterial blood
Absorption due to venous blood
Absorption due to skin, bone and tissue

Fig. 1: Main absorption components for tissue. A factor is added to the logarithmic scales for each component. Absorption due to skin, bone and tissue strongly dominates.

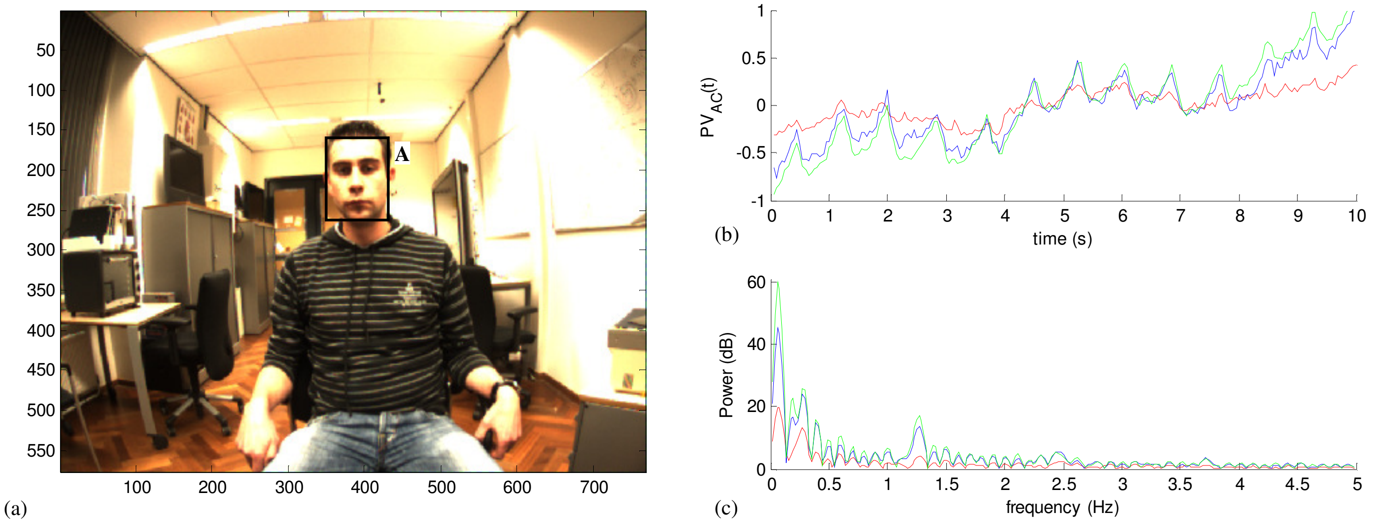


Fig. 2: (a) Still from the movie of volunteer 1 sitting one meter from the camera lens. The selected ROI (A) is indicated by the rectangle. (b) $\text{PV}_{AC}(t)$ signals for the R, G and B channel, indicated by the red, green en blue line respectively. (c) Power spectra indicated by the corresponding color.

method is the effect of direct coupling. Direct coupling refers to direct illumination of the video camera by ambient light without any interaction between the light and skin tissue. Because the direct-coupled ambient light does influence the measured signal, the PPG signal can be modeled as [7]:

$$PPG\ signal = DC_{blood\ \&\ tissue} + AC_{blood\ modulation} + DC_{direct\ coupling} \tag{1}$$

Two important issues are concerned with unobtrusive monitoring of vital body signs. One issue is the change in illumination of the ambient light. The intensity of the reflected and direct coupled light changes and influences the measurements. Another issue, as is with most methods for monitoring vital body signs, is the sensitivity for motion of the user. Due to movement, the illumination angle changes. This affects the amount of reflected light detected by the video camera.

The goal of the research is automatic detection of the heart rate of a person at a distance of several meters under small motion conditions, using video camera based PPG and the surrounding ambient light. In the following sections it will be shown that measurement of heart rate (HR) and respiration rate (RR) using the proposed method is possible.

Section II describes the basic set-up of the camera based solution. Section III *A* shows how vital signs can be extracted from manually selected skin pixels. Although this paper is mainly focused on monitoring HR, this section also shows the detection of RR. The signal-to-noise ratios of the power spectra are studied, dependant of the method of debayering (*B*), downscaling factor (*C*), and size and position of the region of interest (*D*). In section III *E* the influence of noise due to motion is minimized. With the results, a method for automatic selection of skin pixels is proposed in section IV *A*. In the resulting subsections simultaneous detection of HR's from multiple persons is shown (*B*) and the influence of motion (*C*) and the size of the blocks (*D*) is studied. In section V, human presence detection using the proposed method is compared with the state of the art face detector. Finally, conclusions of the research are described in section VI.

II. Basic Set-up

Plethysmographic signals were measured from three adult male volunteers (volunteer 1: European, Fitzpatrick skin type III, volunteer 2: Eastern Asian, Fitzpatrick skin type IV, volunteer 3: European, Fitzpatrick skin type II), sitting at distances of approximately one, three and five meters from the camera lens. In total 9 test videos have been made. The volunteers were asked to sit down to minimize any movements for 30 seconds. As a light source the surrounding fluorescent lighting was used. The windows in the test lab were blinded to minimize changes of ambient illumination. No special backgrounds were used for the measurements. The Color movies were captured with an USB uEye LE digital camera (1/3" CMOS sensor). The zoom factor of the camera was set to a fixed value, invariant to the distance between the volunteer and the camera lens. The color movies, recorded at 20 frames per second (fps) and pixel resolution of 768x576, were saved in raw format. The raw movie frames were transferred to a PC and converted to sets of full-resolution RGB images. Using the numerical computing environment Matlab® (Mathworks), 3x8 bit (0-255) pixel values (PV) were obtained for the red (R), green (G) and blue (B) channels separately from each movie frame. This results in a set of pixel values $\text{PV}(i, j, t)$, where i and j are the horizontal and vertical position respectively and t corresponding to the time in seconds ($t = n/20fps$, with n the frame number). Fig. 2 (a) shows a still from the test video of volunteer 1 who is sitting one meter from the camera lens.

III. Manual Selection of Skin Pixels

A. Method

A region of interest (ROI) was selected manually using a graphical user interface in an image frame selected from the movie, as indicated by the rectangle in fig. 2 (a). The ROI (A) is selected to maximize the number of facial skin pixels of the volunteer in the rectangle. In this case, the noisy surrounding non-skin pixels cannot influence the results. In principle all visible skin pixels, like arms or hands, can be used for monitoring vital signs. However, in most cases facial skin pixels from people are always visible. Therefore, only these pixels are concerned in this research.

From the recorded movies, windows of 200 frames (10 s) are used for further processing. To improve the signal-to-noise ratio (SNR) spatial and temporal processing is applied to the pixel values $\mathrm{PV}(i,j,t)$. For the R, G and B channels of the movie frames the average pixel value, $\mathrm{PV}_{AV}(t)$ of all pixels within the ROI was calculated as a function of time over the resulting frames. Applying windows of 10 seconds yields satisfying results in HR monitoring.

From these signals the mean value over time is subtracted. The resulting signals $\mathrm{PV}_{AC}(t)$ for the R, G, and B channel are shown in fig. 2 (b), indicated by the red, green and blue line respectively.

Signals $\mathrm{PV}_{AC}(t)$ are filtered over time using a second order low-pass Butterworth filter with cut-off frequency 5 Hz, to obtain the filtered signals $\mathrm{PV}_{LP}(t)$.

To obtain information about the HR, finally frequency processing is performed on the filtered signals $\mathrm{PV}_{LP}(t)$, using Fast Fourier Transformation in Matlab®. The resulting power spectra are plotted in fig. 2 (c), indicated by the corresponding channel color. The filtered signals are padded with trailing zeros to length 1024, achieving a finer discretization of the frequency.

Typical healthy resting heart rate in adults is between 1.00 and 1.67 Hz (60 – 100 bpm). From fig. 2 (c) the heart rate can be determined to be the local maximum in the power spectra at 1.25 Hz. This was found to be in agreement with a commercial heart rate monitor $\mathrm{HRM}_{\mathrm{Pol}}$ as a reference (Polar FS3c heart rate monitor digital watch with chest strap). This proves that the local peak in the power spectrum of fig. 2 (c) indeed indicates the HR of the volunteer. Also notice the second and third harmonic around 2.5 Hz and 3.75 Hz respectively in the power spectra.

In a same way the respiration rate can be determined to be the most left peak around 0.06 Hz (3.6 breaths per minute) in the frequency plot. The RR is very low because the volunteer was asked to hold his breath as long as possible to minimize any movements. In the 30 seconds recording the movie 2 breaths were counted (4 breaths per minute). Although this is not a very accurate reference, this corresponds well with the detected RR of 0.06 Hz.

Fig. 2 (c) shows that HR and RR are most pronounced in the G and B channels respectively. The R channel does not show clear information about these two vital signs. This is due to the absorbance bands of different types of hemoglobin at the particular wavelengths. Furthermore, red light penetrates deeper into human skin as compared to blue and green [6], [8].

From this test can be concluded that it is indeed possible to measure HR and RR remotely, using just a simple video camera and without dedicated light sources.

The remaining of this paper will only be concerned with monitoring of the HR signal. Because HR information is most pronounced in the G channel, this channel will be used mainly in the remaining sections. Instead of the low-pass filter a second order band-pass Butterworth filter with cut-off frequencies 0.5 Hz and 3.5 Hz will be used to obtain the filtered signals $\mathrm{PV}_{BP}(t)$.

B. *Debayering the raw movie data*

The color movies were saved in raw format with the CCD pixels preceded in the optical path by a color filter array (CFA) in a 'rggb' Bayer mosaic pattern. For each 2x2 set of pixels, two diagonally opposed pixels have green filters, and the other two have red and blue filters. The Bayer matrix is shown in fig. 3. The G subimage has twice as many pixels as the R and B subimages.

The process by which the images are decoded from the Bayer matrix, or mosaic, into truecolor images is called debayering or demosaicing. For the G channel this means that the unknown green value *g(i,j)* at a red or blue pixel position needs to be interpolated. In the 'rggb' Bayer mosaic pattern the G value $\hat{g}(i,j)$ for the upper left and bottom right pixels has to be computed for each 2x2 set of pixels. To examine the influence of debayer techniques, the following five methods were used for reconstruction of the raw movie frames [9]:

1) In nearest neighbor interpolation (NNI), for the unknown G value the value of the nearest G pixel is used as:

$$\hat{g}_{NNI}(i,j) = g(i+1,j) \qquad (2)$$

2) With bilinear interpolation (BLI) the surrounding four G values are used to estimate the missing G value as:

$$\hat{g}_{BLI}(i,j) = \frac{1}{4} \sum_{(m,n)=\left\{\begin{smallmatrix}(0,-1),(0,1),\\(-1,0),(1,0)\end{smallmatrix}\right\}} g(i+m,j+n) \qquad (3)$$

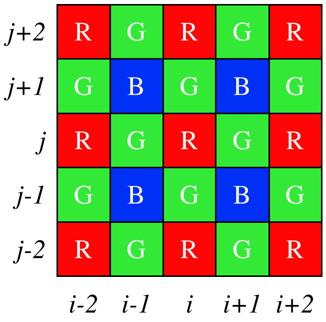
3) Gradient-corrected linear interpolation (GCL) corrects the BLI estimate by a measure of the gradient Δ_R and Δ_B for the known R and B value respectively at the pixel location as:

$$\hat{\mathrm{g}}_{\mathrm{GCL}}(\mathrm{i},\mathrm{j}) = \hat{\mathrm{g}}_{\mathrm{BLI}}(\mathrm{i},\mathrm{j}) + \Delta_{\mathrm{R}}(\mathrm{i},\mathrm{j}) \qquad (4a)$$

$$\Delta_R(i,j) \triangleq r(i,j) - \frac{1}{4} \sum_{(m,n)=\left\{\begin{smallmatrix}(0,-2),(0,2),\\(-2,0),(2,0)\end{smallmatrix}\right\}} r(i+m,j+n) \qquad (4b)$$

For interpolating G values at B pixels, the same formula is used, but corrected with $\Delta_B(i,j)$.

4) To estimate the G value at R pixels, high-quality linear interpolation (HQL8) adds an extra gradient-correction gain to the GCL estimate as [10]:



j+2	R	G	R	G	R
j+1	G	B	G	B	G
j	R	G	R	G	R
j-1	G	B	G	B	G
j-2	R	G	R	G	R
	i-2	*i-1*	*i*	*i+1*	*i+2*

Fig. 3: Bayer matrix with a 'rggb' Bayer filter mosaic pattern.

$$\hat{g}_{HQL8}(\mathrm{i,j}) = \hat{g}_{\mathrm{BLI}}(\mathrm{i,j}) + \alpha\Delta_{\mathrm{R}}(\mathrm{i,j}) \qquad (5)$$

The same estimation is done for the G value at B pixels, but $\hat{g}_{BLI}(i,j)$ is corrected with $\alpha\Delta_B(i,j)$.

5) With HQL16, instead of 8 bit (0-255) colors, 16 bit high color graphics are stored by the interpolator.

For each type of interpolation the peak signal-to-noise ratio (PSNR) between the maximum peak of the HR frequency and the maximum noise power peak was calculated as:

$$\mathrm{PSNR} = 10 \cdot \log\left({}^{\mathrm{P_{HR}}}/_{\mathrm{P_{noise,max}}}\right) \qquad (6)$$

The first harmonic of the HR signal is not taken into consideration. The PSNR and HR are calculated for each movie. The results are shown in table I and table II.

In table I the upper row shows the method of debayering as mentioned above. The first column indicates the test video. All movies were recorded at pixel resolution of 768x576. Movie *x.y* indicates the recorded data of volunteer *x* at a distance *y* in meters from the camera lens. The bottom row shows the average PSNR over all movies for the specific method. Table II shows for each movie the distance of the volunteer to the camera. In the second column the detected HR is shown. For all movies the resulting HR corresponded to the HR detected by the commercial HR monitor $\mathrm{HRM_{Pol}}$. In the third column the size of the selected ROI's in pixels is shown. For each movie, the ROI was selected to maximize the number of facial skin pixels in the rectangle (ROI A in fig. 2 (a) for movie 1.1) and the same ROI was used for each debayer method. The last column of table II shows the average PSNR over all interpolation methods for the specific movie. Note this is not the same $\mathrm{PSNR_{AV}}$ as in table I.

TABLE I

PSNR OF THE USED INTERPOLATION METHOD FOR EACH MOVIE.

Method	NNI	BLI	GCL	HQL8	HQL16
movie *1.1*	4.9	5.0	5.0	4.9	4.9
movie *1.3*	3.5	3.5	3.4	3.5	3.5
movie *1.5*	5.1	4.9	4.8	4.8	4.8
movie *2.1*	4.4	4.4	4.3	4.4	4.4
movie *2.3*	3.1	3.1	3.1	3.0	3.0
movie *2.5*	3.6	3.2	3.2	3.6	3.6
movie *3.1*	5.9	5.9	5.9	5.9	5.9
movie *3.3*	4.9	5.0	5.0	5.3	5.3
movie *3.5*	5.0	5.0	5.0	5.0	5.0
$\mathrm{PSNR_{AV}}$	**4.5**	**4.4**	**4.4**	**4.5**	**4,5**

TABLE II

DISTANCE FROM THE CAMERA, DETECTED HR, SIZE OF THE ROI AND AVERAGE $\mathrm{PSNR_{AV}}$ FOR EACH RECORDED MOVIE.

	Distance	HR	Size ROI	$\mathrm{PSNR_{AV}}$
movie *1.1*	1 m	1.25 Hz	101x76 p	4.9
movie *1.3*	3 m	1.07 Hz	38x29 p	3.5
movie *1.5*	5 m	1.21 Hz	25x19 p	4.9
movie *2.1*	1 m	1.19 Hz	96x77 p	4.4
movie *2.3*	3 m	1.13 Hz	33x31 p	3.0
movie *2.5*	5 m	1.19 Hz	30x16 p	3.4
movie *3.1*	1 m	0.90 Hz	105x71 p	5.9
movie *3.3*	3 m	0.86 Hz	35x24 p	5.1
movie *3.5*	5 m	0.92 Hz	23x16 p	5.0

As can be seen in table I, the differences between the PSNR's for each debayering method are small. $\mathrm{PSNR_{AV}}$ shows that NNI, HQL8 and HQL16 perform best and equally well. Note that NNI is the simplest method of debayering. Equations (2) and (3) show that NNI and BLI only use native G pixels to interpolate the unknown G pixel. According to (4b) and (5), GCL, HQL8, and HQL16 also use native R and B pixels from the raw movie frames for interpolation. Although the HR signal is most pronounced in the G channel, this does not yield to poorer performance in HR detection. Generally can be concluded that more complex interpolation does not result in better performance in HR detection. Also, using 16 bit colors does not yield better performance in HR detection over 8 bit colors.

In table II can be seen that, despite the distance from the volunteer to the camera lens is large and the size of the ROI is small at larger distances, a good $\mathrm{PSNR_{AV}}$ was achieved for each movie. Note that the ROI was selected independently for each movie. Therefore the selected facial area of each volunteer is not equal for the three movies. This declares why $\mathrm{PSNR_{AV}}$ in table II is not decreasing with larger distances.

The same experiments were also done for the R and the B channels. A summary of the results follows here. Debayering the R and B channel from the raw movie frames is described in [9] and [10]. Let's define HR_R, HR_G, and HR_B as the HR frequencies detected from the R, G, and B channel respectively.

In 78% of the experiments HR_R did not match HR_G, with a maximum deviation of 0.64 Hz and an average deviation of 0.14 Hz from HR_G. For the experiments where HR_B was equal to HR_G an average PSNR of 1.8 was achieved. The total $\mathrm{PSNR}_{AV,R}$ including false detection was also 1.8.

For the B channel, 20% of the detected HR_B did not match HR_G, with a maximum deviation of 1.23 Hz and an average deviation of 0.34 Hz from HR_G. For the matching experiments an average PSNR of 3.0 was achieved for the B channel. The total $\mathrm{PSNR}_{AV,B}$ was 3.3.

Compared to the $\mathrm{PSNR}_{AV,G}$ of 4.5 for the G channel, the average PSNR for the R and the B channel are considerably smaller.

C. Size and position of the ROI

In the previous sections ROI A of fig. 2 (a) is used. This ROI was selected such that as many facial pixels as possible were selected within a rectangle. In fig. 4 (a) five additional ROI's (ROI B - F) are shown in a still frame (zoomed in to the face). The same test video (movie 1.1 BLI) is used as in fig. 2. ROI's A – F are tested to investigate the influence of the size and position of the ROI. Fig. 4 (b) shows the $\mathrm{PV}_{AC}(t)$ signals for the G channel, indicated by the same color as the matching ROI. Fig. 4 (c) shows the corresponding power spectra. The

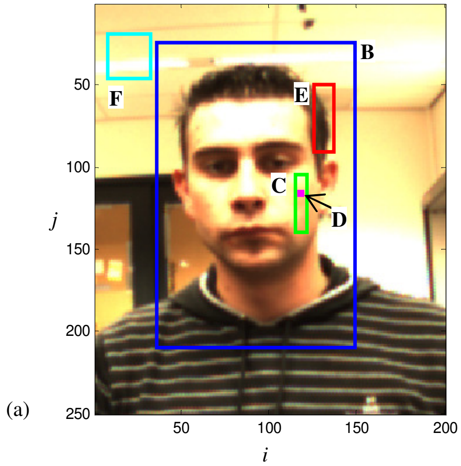
TABLE III

PSNR at pixel location *(i,j)* in fig. 4 (a)

	Pixel location	PSNR
Movie 1.1 (BLI)	(66,71)	1.1
Movie 1.1 (BLI)	(116,74)	1.9
Movie 1.1 (BLI)	(87,79)	2.4
Movie 1.1 (BLI)	(67,120)	3.6
Movie 1.1 (BLI)	(119,133)	3.1

amplitude factors are shown in the y-axis of the plots.

The highest PSNR with a value of 5.0 was achieved with ROI A (101x76 pixels). For ROI B (186x113 pixels) noisy hair, clothing and background pixels were included. The PSNR decreases to 3.3, but still the HR frequency can easily be distinguished from the power spectrum. Selecting an area with skin pixels only, as was done with ROI C (35x7 pixels), results in a clear HR signal with a PSNR of 3.8. The PSNR decreases when the ROI contains less skin pixels. This can be seen from ROI D, where only 1 pixel was used, but still an acceptable PSNR of 3.5 was achieved.



Temporal and spatial noise due to imperfections of the camera will always be present [11]. The results of ROI A – D show that the noise per pixel is reduced by averaging over more skin pixels. A better PSNR is achieved when the number of skin pixels included in the ROI increases.

PSNR, were also calculated at other facial pixels outside ROI C. The results for five additional ROI's at pixel location (i, j) are shown in table III, where i and j are the horizontal and vertical position in fig. 4 (a) respectively. From this can be concluded that HR information is not pronounced equally in all facial pixels The PSNR at pixel location with a strong luminance is smaller than in less darker areas. This is caused by clipping of (bright) pixel values, due to a high degree of luminous sensitivity of the video camera. Similar results (not shown) were obtained for movie 2.1 (NNI) and movie 3.1 (HQL8).

Within ROI E (41x11 pixels) and F (27x25 pixels) only non-skin pixels were selected. ROI E contains hair and background pixels as ROI F only contains background pixels. In the power spectrum of ROI F can be seen that only noise is present at random frequencies. Because of the flicker noise (also called $1/f$ noise) of the CCD camera the power of the noise decreases at higher frequencies [8].

As stated in the introduction of this paper, two additional

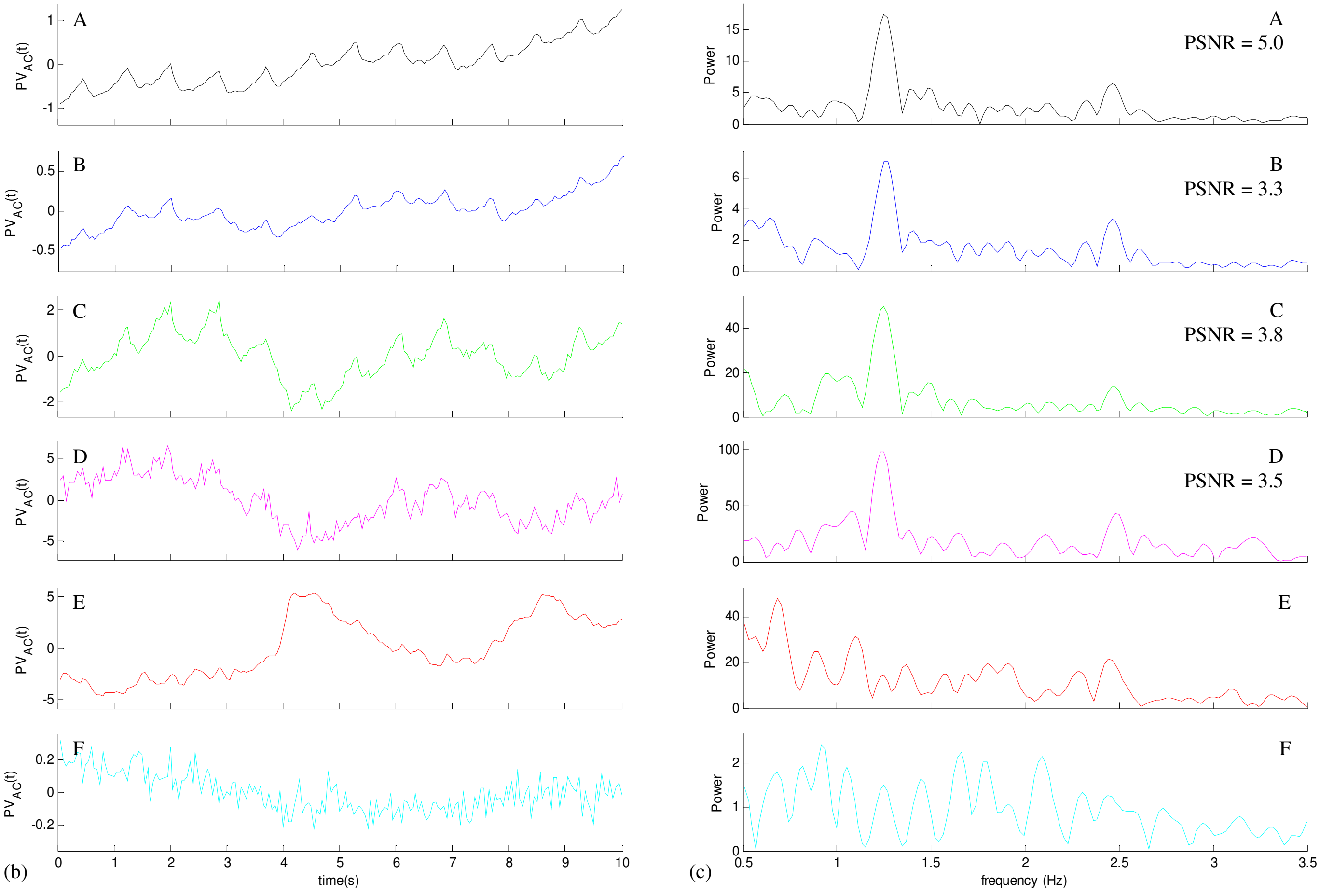


Fig. 4: (b) $\text{PV}_{AC}(t)$ signals and corresponding power spectra of the G channel for ROI A shown in fig. 2 (a) and ROI's B – F shown in (a). (c) Corresponding power spectra. ROI D represents one pixel. The signals and spectra are indicated in the same color as the corresponding ROI's in (a) and in fig. 2 (a). The signals are all scaled to the same size. The PSNR is indicated above the power spectrum if a HR was detected.

noise factors in unobtrusive HR monitoring are noise due change in illumination of the ambient light ($N_{illumination}$) and noise due to motion of the user (N_{motion}). $N_{illumination}$ was minimized by blinding the windows in the test lab. However, because the shading device did not blind the windows completely, still some illumination changes occurred during recordings. However, the effect of $N_{illumination}$ in the HR frequency band is negligible due to the band-pass filter.

In the power spectrum of ROI E can be seen that the effect of N_{motion} in the HR frequency band cannot be neglected. Although only noisy hair and background pixels are selected in ROI E, a high power dominant frequency is present at 0.68 Hz. Due to involuntary movement of the volunteer, periodical signals other than the HR can be present in the HR frequency band. N_{motion} is dominant at the edges in the image. Edges are areas with strong intensity contrast, like (skin or background) pixels adjacent to the hair and eyes of the volunteer. Pixel values at these edges will change (periodically) due to (involuntary) movement. This can also be caused by facial shadings.

The noisy edge pixels are included in ROI B. The power spectrum shows that the signals introduced by slight movement are not dominant over the HR frequency in this particular case. This shows that also N_{motion} per pixel is reduced by averaging over more pixels.

In fig. 4 (c) can be seen that for ROI A, C, and D, the power of the HR frequency increases as the size of the ROI decreases. The power of the signal for ROI E is high due to the high contrast in combination with voluntary movements. No dominant frequency due to movement is present at the HR frequency. This proves that the signals in fig. 4 are true PPG signals and are not introduced by involuntary movement. This is also confirmed by the fact that the HR signal is most pronounced in the G channel. Noise due to (involuntary) movement is not dominant either in the R, G, or B channel of the movie frames. From fig 4 (c) can also be concluded that the optimal size of the ROI is when the number of facial skin pixels is maximum and a minimum number of noisy non skin pixels is selected in the rectangle.

D. *Downscaling the video frames*

To investigate the influence of downscaling on video based HR detection, movie 1.1 (BLI), movie 2.1 (NNI) and movie 3.1 (HQL8) were downscaled by factors K from 2 to 10, using nearest neighbor interpolation (implemented using *imresize* in Matlab®). The concerning ROI's of section *B* were used and kept at the same location in relation to the face of the volunteer for each downscale factor. For each downscaled movie the PSNR and HR are calculated. The PSNR in relation to the downscale factor K is plotted in fig. 5 (a) for movie 1.1, (b) for movie 2.1, and (c) for movie 3.1. The PSNR of the original movies (K = 1) is also shown in fig 5.

For movie 1.1 an incorrect HR of 1.27 Hz was determined for K = 9 and K = 10 (correct HR is 1.25 Hz). For movie 2.1 an incorrect HR of 0.64 Hz, 1.17 Hz, and 1.17 Hz was found for K = 8, K = 9, and K = 10 respectively (correct HR is 1.19 Hz). The PSNR of the incorrect HR's are not plotted in fig 4. The ROI size of the downscaled movie (in pixels) can be calculated as:

$$ROI_{downscaled} = \lceil ROI_{original}/K \rceil \qquad (7)$$

Where $\lceil\cdot\rceil$ indicates that the value between the brackets is rounded to the nearest highest integer. The original PSNR and the size of the ROI for the concerning movies can be seen in

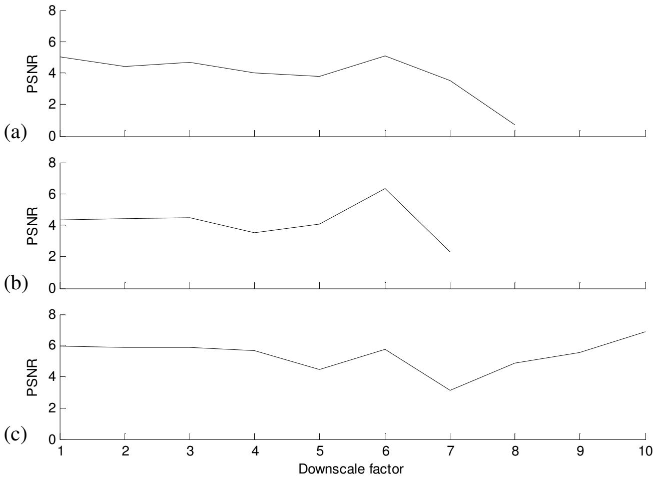


Fig. 5: Plot of the PSNR in relation to the downscale factor for (a) movie 1.1 (BLI), (b) movie 2.1 (NNI) and (c) movie 3.1 (HQL8). For downscaling nearest neighbor interpolation was used.

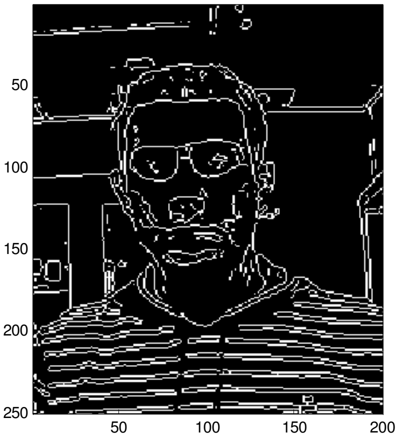


Fig. 6: Output of the edge detector (zoomed in to the face), containing 1's (white pixels) when an edge is detected and 0's (black pixels) elsewhere.

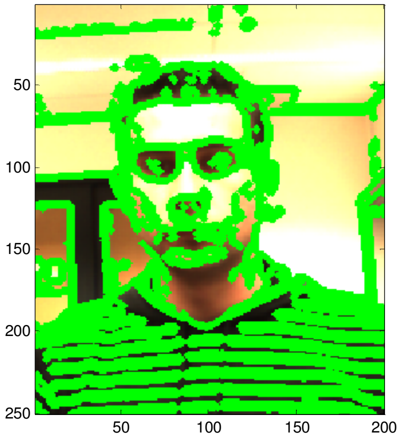


Fig. 7: First frame of movie 1.1 (BLI) (zoomed in to the face) with the pixels adjacent to edges deleted (green pixels).

table I and II.

For movie 1.1 the chart of fig. 5 (a) slightly decreases and peaks at K = 6 to a PSNR of 5.1. From K = 6 the PSNR decreases drastically. For movie 2.1 the PSNR also peaks at K = 6 to value of 6.3 and decreases drastically after K = 6. The chart for movie 3.1 dips at K = 5 and K = 7 to a PSNR of 4.5 and 3.1 respectively. After K = 7 the PSNR increases to a maximum of 6.9 at K = 10.

Downscaling was also done for other debayering methods for the three movies (results not shown). However, similar characteristics were obtained as shown in fig. 5.

From here can be concluded there is no clear relation between the PSNR and the downscale factor. As concluded in the previous section, not all the pixels within the ROI of the original movies contribute equally to the strength of the HR signal. For some pixels a contribution is stronger than for other facial pixels. This explains that the PSNR can increase, even when fewer pixels are selected in the downscaled movie frames.

E. *Minimizing the influence of noise due to motion*

In the previous sections it is shown that, dependent of the size and position of the ROI, the HR signal can easily be distinguished in the power spectrum, even in the presence of noise introduced by involuntary movement. However, N_{motion} can be dominant in the HR frequency band when more intensive motion is present in the video. In this section a method is proposed for minimizing the influence of motion.

N_{motion} is dominant in areas with strong intensity contrast. To find the corresponding pixels an edge detector is used (using *edge* in Matlab®). Edges are found using the Sobel approximation to the gradient [12]. At each pixel position in the input image, the approximate magnitude of the gradient is calculated, by applying two 3x3 masks to estimate the gradient in the horizontal and vertical direction. The edge detector returns an edge location at the pixel points if the gradient exceeds a threshold. The lower the threshold, the more edges will be detected. The threshold is set to a relatively low value, in order that edges due to facial shadings are detected as well.

The input image for the edge detector is the first frame of movie 1.1 (BLI) (fig. 2 (a)). The function returns the binary image shown in fig. 6 (zoomed in to the face). The image contains 1's (white pixels) when an edge is detected and 0's (black pixels) elsewhere.

The 13 closest neighbor pixels (in a diamond shape) for every edge pixel in the binary image are set to zero in the movie frames of the input movie. This is shown in fig. 7. For clarification, the removed adjacent edge pixels are indicated in green. The location of the deleted adjacent edge pixels is identical for all frames of the input movie. Calculating edge pixels for every frame separately and removing the concerning pixels would introduce new temporal frequencies.

With the adjacent edge pixels removed, the power spectrum for ROI A – C of fig. 2 (a) and fig. 5 (a) is recalculated. The PSNR of ROI A, B, and C increases to 5.7 (5.0 without this method), 3.6 (3.3 without this method), and 4.8 (3.8 without this method) respectively. From this can be concluded that with this method, the PSNR improves strongly when the ROI contains regions of relatively high contrast. The influence of N_{motion} is minimized when edge pixels are excluded in the algorithm, even when the noise is caused by facial shading like in ROI C.

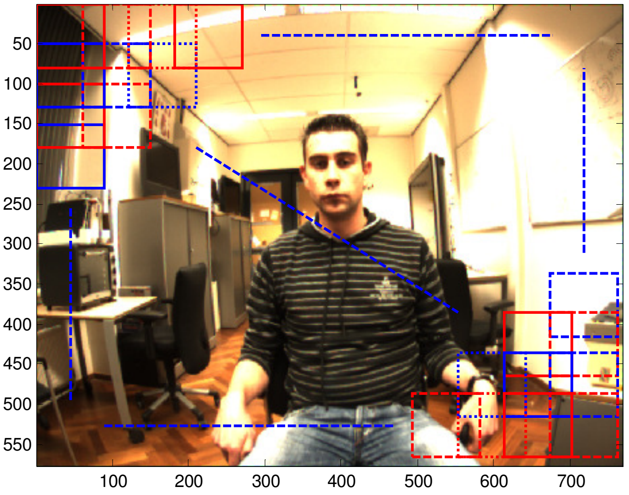


Fig. 8: Still from movie 1.1 (BLI) showing how the movie frames are divided into *NxM* overlapping blocks.

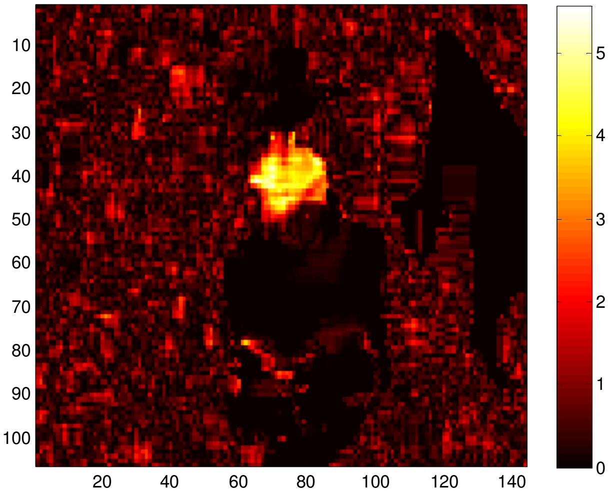


Fig. 9: *NxM* PR map containing the PR of the first and second maximum peak in the power spectrum of the corresponding block.

IV. Automatic Selection of Skin Pixels

A. *Method*

In section III the ROI was selected manually from a still frame of the test video. In this section an algorithm is designed for automatic selection of skin pixels. To evaluate the condition of the user, facial pixels have to be in the line of sight of the camera. With this algorithm, no user interaction for selection of ROI is needed.

To find facial pixels suitable for extraction of HR information, each frame of the G channel is divided into *NxM* overlapping blocks, as indicated in fig. 8. The number of blocks can be calculated with:

$$N = \left\lfloor \frac{Frame_{width} - Block_{width} + BlockShift_{hor}}{BlockShift_{hor}} \right\rfloor \qquad (8a)$$

$$M = \left\lfloor \frac{Frame_{heigth} - Block_{heigth} + BlockShift_{ver}}{BlockShift_{ver}} \right\rfloor \qquad (8b)$$

Where $\lfloor\cdot\rfloor$ indicates that the value between the brackets is rounded to the nearest lowest integer. All variables between the brackets are indicated in pixels.

To minimize the influence of N_{motion}, adjacent edge pixels are removed from the movie frames (see section III *E*). For each block the same spatial, temporal and frequency processing is done as for the manual selected ROI as described in section III *A*. However, instead of a low-pass filter, each signal is filtered with a band-pass filter with cut-off frequencies 0.5 Hz and 3.5 Hz to conserve only HR frequencies. From the resulting power spectra, the power ratio (PR) of the first and the second maximum peak is calculated as:

$$PR = 10 \cdot log\left({P_{max1}}/{P_{max2}} \right) \qquad (9)$$

The goal of this method is to detect large PR in a block with facial pixels and low elsewhere. Note the PR does not have to be identical to the PSNR of (6) when a HR frequency is present in the power spectrum. With the calculation of the PR the first harmonic of the HR signal is not taken into account, as is with the PSNR.

As shown in section III *C*, the best PSNR is achieved when the size of the block is equal to the size of the face of the volunteer. When a block is located in the area containing face pixels, the PR is large as well, as can be seen from the power spectra of fig. 5 (c). With this method, the size of the face is not known in advance. Therefore it is not possible to adjust size of the blocks to match the size of the volunteers face. However, according to section III *D*, a smaller or bigger ROI will also result in a sufficient high PSNR. From this can be concluded that the size of the blocks can be set to a fixed value independent of the size of the volunteers face.

The size of the blocks is set to a size of 50x50 pixels. Each block is shifted 5 pixels horizontally and vertically from the previous block. For each block, the PR is mapped to the corresponding pixel in the *NxM* PR map. The PR map and the corresponding color scale are shown in fig. 9. Here is clearly shown that the PR is large when the corresponding blocks contain facial pixels and relatively small elsewhere. For each block, also the frequency of the maximum peak in the power spectrum is mapped to the *NxM* frequency map. The frequency map and the corresponding color scale are shown in fig. 10. Here can be seen that a lot of unwanted frequencies due to noise are present. However, the HR frequency is mapped correctly to the adjacent pixels with value 1.25 Hz in the facial region.

To extract the HR frequency from the frequency map, a binary mask is obtained from the PR map. The binary mask contains 1's if the PR exceeds a threshold *thr* and 0's elsewhere. This is shown in fig. 11. The threshold is calculated as:

$$thr = (\max_{\mathrm{PR}} - \mathrm{mean}_{\mathrm{PR}})/2 \qquad (10)$$

For the PR map of fig. 9 this results in *thr* = 2.6. The binary mask is used to remove the unwanted frequencies from the frequency map. Pixel values in the frequency map are set to

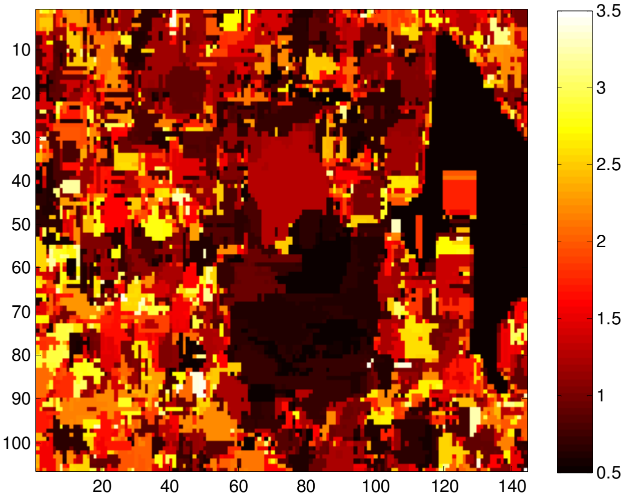


Fig. 10: *NxM* frequency map containing the frequency of the peak in the power spectrum of the corresponding block.

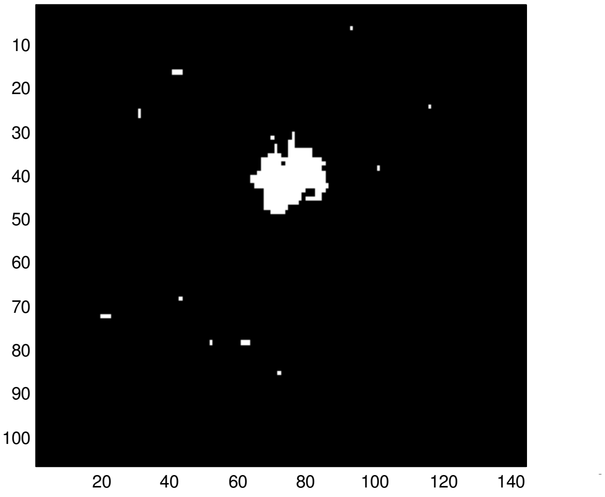


Fig.11: Binary mask obtained from the PR map containing 1's if the PR exceeds 2.6 and 0's elsewhere.

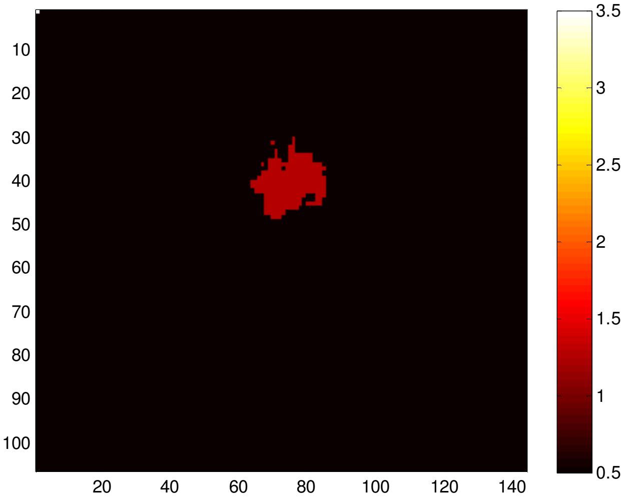


Fig. 12: Thresholded frequency map with the unwanted frequencies removed. Only an area of adjacent pixels at HR frequency is present.

zero if the corresponding pixels in the binary mask are equal to zero.

In the resulting frequency map still some unwanted frequencies are present that passed the threshold. However, in the thresholded frequency map HR frequency pixels is strongly dominant over the resulting frequencies. The dominant frequency can easily be determined by counting the number of pixels per frequency value. The frequency with the largest number of pixels at corresponding value in the frequency map is the HR frequency. The remaining frequencies are set to zero. More dominant frequencies are found when HR's of additional persons are present. The region with neighboring pixels at HR frequency is the facial area. Pixels with frequencies similar to HR that are not in de neighbor of the facial area are also set to zero. The resulting frequency map, containing only the dominant HR frequency and 0's elsewhere is shown in fig. 12.

From the thresholded frequency map the ROI is acquired and mapped back to the original movie frames. The ROI is determined to be the rectangle including the automatically detected facial pixels. The frame with the resulting ROI is shown in fig. 13. The rectangle is drawn from point (ROI_i, ROI_j) having width ROI_{width} and height ROI_{height}, as:

$$ROI_i = (BlockShift_{hor} - 1) \cdot FM_{hor,min} \\ +0.8 \cdot Block_{width} \tag{11a}$$

$$ROI_j = (BlockShift_{ver} - 1) \cdot FM_{ver,min} \\ +0.2 \cdot Block_{heigth} \tag{11b}$$

$$ROI_{width} = (BlockShift_{hor} - 1) \cdot FM_{hor,max} \\ +(0.2 \cdot Block_{width} + 1) - ROI_i \tag{11c}$$

$$ROI_{heigth} = (BlockShift_{verr} - 1) \cdot FM_{ver,max} \\ +\left(0.8 \cdot Block_{heigth} + 1\right) - ROI_j \tag{11d}$$

Where $FM_{hor,min}$, $FM_{ver,min}$, $FM_{hor,max}$, and $FM_{ver,max}$ indicates respectively the first pixel in the columns, the first pixel in the rows, the last pixel in the columns, and the last pixel in the rows containing HR frequency in the thresholded frequency map of fig. 12.

B. *Detection of multiple HR's*

With this method HR's of multiple persons can be detected simultaneous. To illustrate that, movie 2 (768x576 pixels, 20 fps, 200 frames, 8 bit colors) including 2 volunteers is used. Any movements were prevented during recording. A (grayscale) still from the movie including the resulting ROI's is shown in fig. 14. Blocks of 25x25 pixels were used. Successive blocks were shifted horizontally and vertically by only 1 pixel. For this movie, two dominant frequencies were present in the thresholded frequency map.

C. *Influence of motion*

To show the influence of (involuntary) movement on automatic detection of skin pixels, adjacent edge pixels were not removed from the movie frames for this test. The resulting frequency map is shown in fig. 15. The neighboring HR frequencies pixels can clearly be seen in the map. HR's were detected at 1.10 Hz for the volunteer on the left and at 1.30 Hz

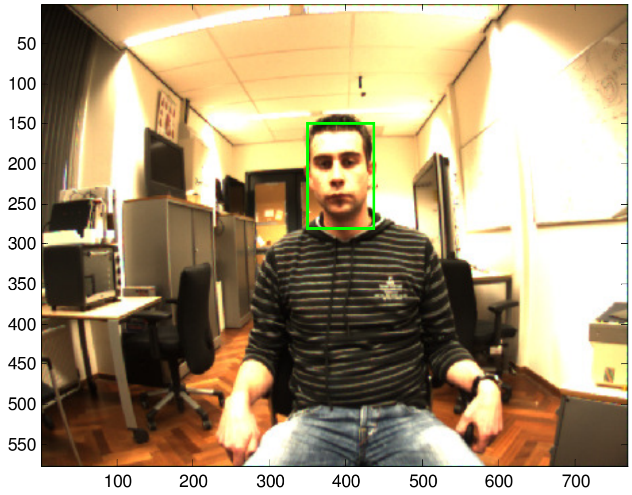


Fig. 13: Resulting ROI acquired from the thresholded frequency map, including the facial pixels detected automatically by the algorithm.

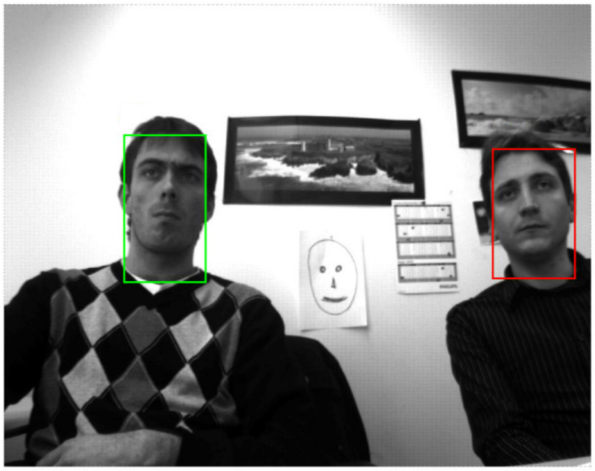


Fig.14: Still from movie 2 (768x576 pixels) including 2 persons and the resulting ROI's detected by the algorithm.

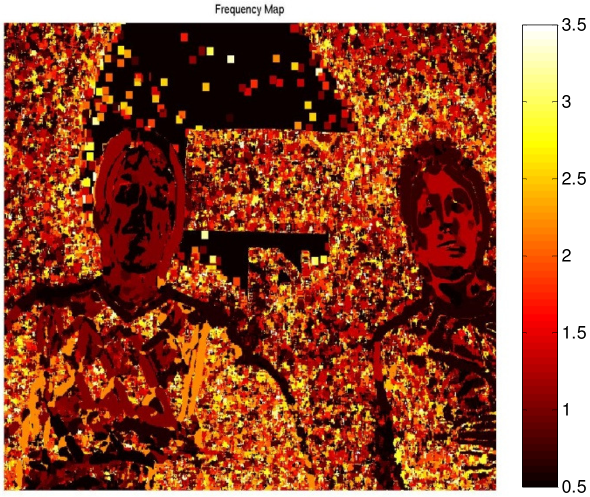


Fig. 15: Frequency map (744x552 pixels) for movie 2. To show the influence of motion, adjacent edge pixels were not removed.

for the volunteer on the right. Besides noise, large regions of frequencies due to movement are present in the frequency map. These are the regions of adjacent pixels at the edges of the faces of the volunteers with frequencies 1.43 Hz and 0.73 Hz in fig. 15. Also, regions of adjacent pixels at 1.43 Hz can be found around the detailed pattern on the shirt of the volunteer on the left. In the PR map, shown in fig. 16, can be seen that PR is largest at edge locations. The largest PR for facial pixels is 2.7. Because the proposed algorithm is based on detection of the largest PR's, automatic selection of skin pixels will fail in this case.

As mentioned in section III *B*, with the proposed method only monochrome videos using the G channel are processed. Earlier tests in this research showed that the relations between the strength of the signals in R, G, and B chrominance components can be used to distinguish between HR and frequencies introduced by motion. However, the proposed method is preferred because of the lower computational load.

D. *Influence of (shifting) blocks*

Fig. 17 shows the thresholded frequency map for movie 2, using the proposed method with the adjacent edge pixels removed from the movie frames. Skin pixels are detected accurately and the shapes of the faces are clearly visible. This was not the case for the thresholded frequency map for movie 1.1 (fig. 12). Due to a smaller block size and especially because successive blocks are shifted only by 1 pixel, skin pixels are detected more accurately. However, the automatically detected ROI's are all determined correctly in fig. 13 as well as in fig. 14. However, the computational load for determining the ROI's of movie 2 is almost 27 times bigger than the computational load for determining the ROI of movie 1.1. Namely, 744x522 versus 144x106 power spectra had to be calculated for movie 2 and movie 1.1 respectively.

V. Robust Presence Detection

Recently, human presence detection using a video camera has become an active research area. The technology enables new application in the security domain and in the consumer electronics domain in for example energy management.

Automatic selection of skin pixels, as described in section IV is a robust method for presence detection. The proposed method, unlike computer vision based approaches, only detects human beings and cannot be fooled by pictures of faces or humans [13].

The state of the art computer vision based technology is the face detector proposed by Viola Jones [14]. The algorithm aims at detecting human faces using object detection. Movie 2 was used as the input for the Viola Jones face detector (using *ObjectDetection*[1] in Matlab®). The output is shown in fig. 18. Here, unlike the proposed method, even the simple drawing on the wall is detected to be a human face. Also notice that the rectangle on the left includes not all the facial pixels and the rectangle on the right in includes background pixels as well. The ROI's acquired by the proposed method do include facial pixels correctly, as illustrated in fig. 14.

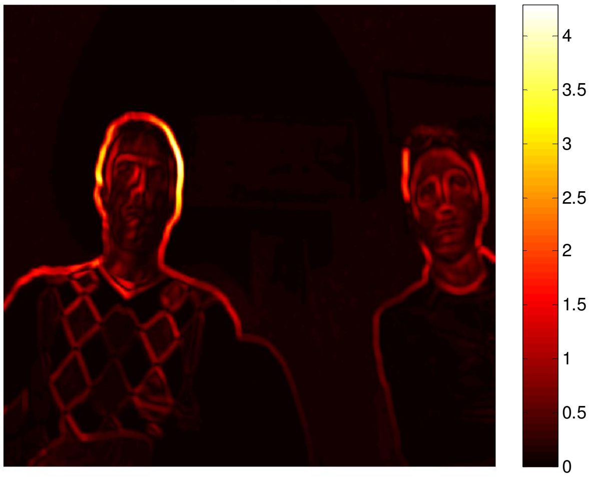


Fig. 16: PR map (744x552 pixels) for movie 2. The adjacent edge pixels were not removed from the movie frames.

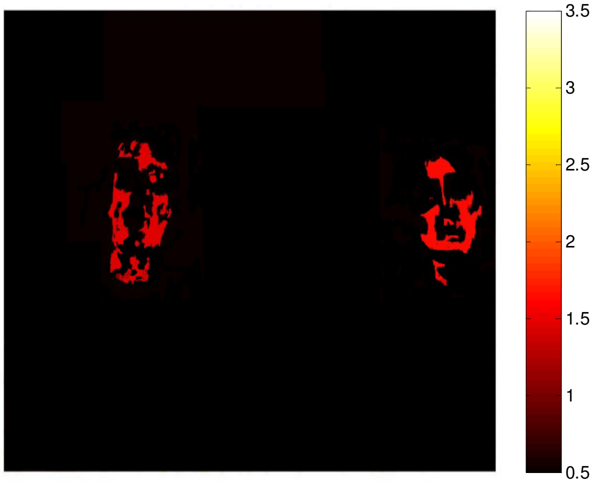


Fig. 17: Thresholded frequency map (744x552 pixels) for movie 2 with the adjacent edge pixels removed from the movie frames.

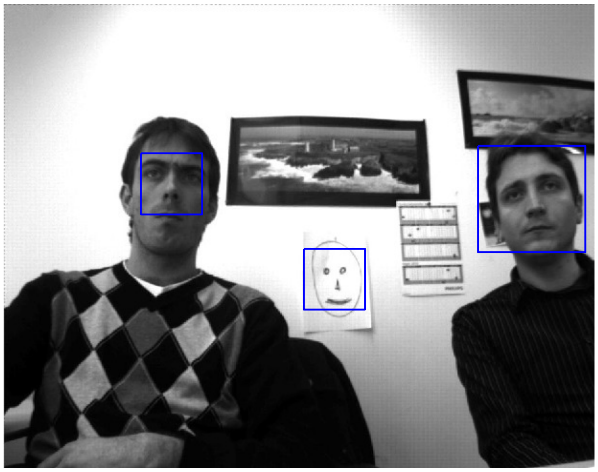


Fig. 18: Output of the state of the art Viola Jones face detector. Even the simple drawing on the wall is detected to be a human face.

[1] The Matlab® function *ObjectDetection* is part of the code *Viola Jones Object Detection* (Author: D.J. Kroon, 18-11-2010) and is available at: http://www.mathworks.com/matlabcentral/fileexchange/

From this can be concluded that automatic selection of skin pixels with the proposed method is a robust and an accurate approach for human presence detection. Unlike the state of the art object detector, the proposed method does not provide false detection of drawings or pictures.

VI. Conclusions

A. General

This paper described the results of the research on video camera based photoplethysmography, for measuring vital body signs remotely under small motion conditions. It was shown that HR and RR are detected correctly up to several meters, using a simple video camera and the surrounding ambient light as a light source. With the proposed method, a high PSNR was achieved, even when the distance between the user and the camera lens was about 5 meter.

The remaining of the paper was only concerned with monitoring of the HR. Only the G channel from the color movies was used, because HR was most pronounced in this channel.

Using a more complex method for debayering the raw movie frames did not result in significantly better PSNR. Also, no increase in performance is achieved when the HR was calculated from high color movies (16 bit) instead of truecolor movies (8 bit).

ROI's were selected manually from a movie. The best PSNR was achieved when all facial skin pixels were included in the rectangular ROI. Decreasing the size of the ROI in relation to the size of the face deteriorated the PSNR. However, HR could still be distinguished clearly in the power spectrum, even when only 1 pixel was selected. Not in all facial pixels the HR information has the same strength. HR signal was more pronounced in less bright pixels as in pixels with a strong luminance level, due to clipping of bright pixel values.

The assumption has been verified that downscaling the movie frames could increase the PSNR, although no clear relation was present between the PSNR and the downscale factor.

Volunteers were asked to sit still to minimize any movements, however still some noise due to involuntary movements were present. N_{motion} was dominant at the edges of the volunteers faces. With the use of an edge detector, these edge pixels were set to zero in the movie frames. The threshold for the edge detector was set relatively small, so edges due to facial shadings were also detected. Deleting the adjacent edge pixels did improve the PSNR significantly.

The obtained results were used to design a method for automatic detection of skin pixels. Adjacent edge pixels were deleted and movie frames were divided into *NxM* overlapping blocks. Because no information about the size of the volunteers face was available in advance, the size of the blocks was set to a fixed value.

The values of HR frequencies and PR's were mapped to the frequency and PR map respectively. The PR map was used to select the facial region from the frequency map. In the facial region PR's were largest. Selecting the corresponding pixels in the frequency map and mapping them back to the movie frames resulted in the automatically detected ROI. The ROI was detected very accurately and included all the facial pixels in the movie frames.

The proposed method fails when adjacent edge pixels were not removed. PR's due to movement were much larger than PR's due to the HR signal. When subsequent blocks are shifted only by one pixel from each other, HR valued pixels in the frequency map are determined more accurate. However, the computational load is much larger and the resulting ROI in the movie frames was also detected accurately when the blocks were shifted over more pixels. It was shown, that with the proposed method HR's of two volunteers were detected simultaneously.

Automatic selection of skin pixels can be used as a robust and accurate presence detector. It was shown that, unlike the state of the art Viola Jones face detector, the proposed method detected only human faces as it could not be fooled with drawings or pictures representing them.

B. Future research

This paper showed that the proposed method performed very well under small motion conditions due to involuntary movements of the user. However, when large intensive motion is present, the proposed method fails. The face should be tracked when the user moves. This can be achieved using motion estimation. Instead of the temporal axis, as is done with the proposed method, motion vectors can be used to track pixels containing HR information. Signals along the motion vectors can then be used to calculate the power spectra. Within the Philips research laboratories a lot of research is done in motion estimation and motion compensation. In future research it can be investigated how motion estimation can be combined with the proposed algorithm.

Acknowledgment

I want to thank Ihor Kirenko and Vincent Jeanne for their help during my graduation project and Marek Bartula for his help on high-quality linear interpolation[2]. I especially want to thank Gerard de Haan[3] for his help and personal interest during my graduation period.

To my mom Mia Schmitz-Kurstjens
10 January 1948 – 20 December 2008

References

[1] Philip O. Isaacson, David W. Gadtke, and Timothy L. Johnson, "Finger clip pulse oximeter," United States Patent 5,490,523, Feb. 13, 1996.

[2] Jeffrey M. Haynes, "the Ear as an Alternative Site for a Pulse Oximeter Finger Clip Sensor," in *Respiratory Care*, vol. 52, no. 6, pp. 727-729, June 2007.

[3] Matthew J. Banet, Michael J. Thompson, and Zhou Zhou, "Chest strap for measuring vital signs," United States Patent Application Publication US 2007/0142715 A1, Jun. 21, 2007.

[4] Peck Y S Cheang and Peter R Smith, "An Overview of Non-contact Photoplethysmography," in *ELECTRONIC SYSTEMS AND CONTROL DIVISION RESEARCH 200*, (Department of Electronic and Electrical

[2] Ihor Kirenko, Vincent Jeanne and Marek Bartula are research scientists at the Video & Imaging Processing group of Philips Research Eindhoven.

[3] Gerard de Haan is a research scientist at the Video & Imaging Processing group of Philips Research Eindhoven and part-time professor at the Eindhoven University of Technology.

Engineering, Loughborough University, LE11 3TU, UK), pp. 57-59.

[5] A. B. Hertzman and C. R. Spealman, "Observations on the finger volume pulse recorded photoelectrically," Am. J. Physiol. 119, pp. 334-335 (1937).

[6] F.P. Wieringa, “Pulse Oxigraphy, and other new in-depth perspectives through the near infrared window (Thesis),” Erasmus University Rotterdam, May 9, 2007.

[7] M.J. Hayes and P.R. Smith, “Artifact reductions in photoplethysmography,” in *Applied Optics*, vol. 37, no. 31, pp. 7437-7444, 1998.

[8] Wim Verkruyse, Lars O Svaasand, and J Stuart Nelson, “Remote plethysmographic imaging using ambient light,” Optics express, vol. 16, No. 26, 21434-21445.

[9] Rajeev Ramanath, Wesley E. Snyder, and Griff L. Bilbro, “Demosaicking methods for Bayer color arrays,” Journal of Electronic Imaging, vol. 11, No. 3, 306-315.

[10] Henrique S. Malvar, Li-wei He, and Ross Cutler, “High-quality linear interpolation for demosaicing of bayer-patterned color images”, Microsoft Research.

[11] “CCD Image Sensor Noise Sources (Application note),” Revision 2.1, Kodak Image Sensor Solutions, Jan 10, 2005.

[12] O.R. Vincent, and O. Folorunso, “A Descriptive Algorithm for Sobel Image Edge Detection,” in *Proceedings of Informing Science & IT Education Conference (InSITE),* pp. 97-107, 2009.

[13] V. Jeanne, F.J. de Bruijn, G. Cennini, and G. Schmitz, “Robust Presence Detection (Paper ID),” unpublished.

[14] P. Viola and M.J. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition*, 2001, Vol. 1, pp. 216-221.