

# Localised photoplethysmography imaging for heart rate estimation of pre-term infants in the clinic

Sitthichok Chaichulee<sup>a</sup>, Mauricio Villarroel<sup>a</sup>, João Jorge<sup>a</sup>, Carlos Arteta<sup>b</sup>, Gabrielle Green<sup>c</sup>,  
Kenny McCormick<sup>c</sup>, Andrew Zisserman<sup>b</sup>, and Lionel Tarassenko<sup>a</sup>

<sup>a</sup>Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK

<sup>b</sup>Visual Geometry Group, Department of Engineering Science, University of Oxford, UK

<sup>c</sup>Neonatal Unit, John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, UK

## ABSTRACT

Non-contact vital-sign estimation allows the monitoring of physiological parameters (such as heart rate, respiratory rate, and peripheral oxygen saturation) without contact electrodes or sensors. Our recent work has demonstrated that a convolutional neural network (CNN) can be used to detect the presence of a patient and segment the patient’s skin area for vital-sign estimation, thus enabling the automatic continuous monitoring of vital signs in a hospital environment.

In a study approved by the local Research Ethical Committee, we made video recordings of pre-term infants nursed in a Neonatal Intensive Care Unit (NICU) at the John Radcliffe Hospital in Oxford, UK. We extended the CNN model to detect the head, torso and diaper of the infants. We extracted multiple photoplethysmographic imaging (PPGi) signals from each body part, analysed their signal quality, and compared them with the PPGi signal derived from the entire skin area. Our results demonstrated the benefits of estimating heart rate combined from multiple regions of interest using data fusion. In the test dataset, we achieved a mean absolute error of 2.4 beats per minute for 80% (31.1 hours) from a total recording time of 38.5 hours for which both reference heart rate and video data were valid.

**Keywords:** vital sign monitoring, heart rate, infants, deep learning, autoregressive modelling

## 1. INTRODUCTION

Vital signs such as heart rate, respiratory rate and peripheral oxygen saturation ( $\text{SpO}_2$ ) are essential parameters for assessing the well-being of patients in the hospital. Recent studies have demonstrated that these vital signs can be continuously monitored using a video camera under ambient lighting in clinical environments.<sup>1-6</sup> This paper focuses on the continuous heart rate monitoring of pre-term infants in the clinic in which heart rate estimates are derived automatically without human supervision.

In a Neonatal Intensive Care Unit (NICU), pre-term infants are monitored by complex medical equipment depending on their unique needs. They are unstable and have fluctuating vital signs.<sup>7</sup> Conventional vital sign monitoring technologies such as pulse oximetry or electrocardiography (ECG) require the attachment of adhesive electrodes or transducers to the skin surface of the patient. The skin of infants is fragile and very sensitive, especially for those born before 27 weeks of gestation. The bond between the attached sensor and the dermis is stronger than that between the dermis and epidermis,<sup>2</sup> it may damage the skin and increase the risk of developing infections. In addition, the movement of the infant may cause epidermal stripping.<sup>8</sup> Non-contact vital sign monitoring could reduce these patient discomforts as well as reduce physical restrictions on cabling and measurement sites. The system could be used for patients in both hospital and home environments.

Recent work described in [9] presented a convolutional neural network (CNN) model that can be used to detect the presence of a patient and segment the patient’s skin regions for vital sign estimation, thus enabling the continuous monitoring of vital signs to be performed automatically in a challenging hospital environment. This paper begins with the introduction of a clinical study and the description of the data used for heart rate estimation. Subsequently, this paper describes the extension of the previous CNN model to perform the detection of the body parts that are essential for the estimation of vital signs such as the head, torso and additionally the diaper area of the infant. This paper then describes the algorithms used for estimating heart rate and presents

the comparison of heart rate estimates from different body parts with the baseline heart rate computed from the entire patient's skin. This work finally suggests that the combination of heart rate estimates from all body parts and the entire skin yields the lowest estimation error and the highest proportion of estimated time.

## 2. CLINICAL STUDY

### 2.1 Description of the study

This paper used the data from the Neonatal High-Dependency Unit (Neonatal HDU) study which involves the continuous monitoring of 30 pre-term infants using both conventional monitoring and video recording. The study was carried out in the high-dependency area in the Neonatal Intensive Care Unit (NICU) at the John Radcliffe Hospital in Oxford, UK. The pre-term infants were recorded under regular ambient light during daytime for up to four consecutive days. The study was approved by the Medical Research Ethics Committee under the reference number 13/SC/0597 (MONITOR Study).

Pre-term infants, born before 37 weeks of gestation, are often admitted into the NICU immediately after birth since they are not fully developed and tend to have medical conditions that require specialist care.<sup>10</sup> Constant nursing and medical supervision are provided to the infants until they are strong enough and ready to be discharged.<sup>7</sup> Since pre-term infants are unstable, they provide a wide spectrum of vital sign values<sup>7</sup> which benefits the development and validation of non-contact vital sign estimation algorithms.

Fig. 1 shows the equipment setup for a typical recording session in the Neonatal HDU study. A modification was made to the incubator's canopy by drilling a small hole at the top to allow a video camera to film inside the incubator without an obstruction from the plastic layer of the incubator's canopy. The modified incubator was subjected to a two-week humidity and temperature test period by the hospital's clinical engineering department to verify that the modifications made to the incubator were safe and did not present any risk to the patients.

The video recordings were acquired using a 3-CCD JAI AT-200CL digital video camera (JAI A/S, Denmark). The video camera employs three Sony ICX274AL Charge-Coupled Device (CCD) 1/1.8" image sensors (Sony, Japan) to measure the light intensity of three colour channels red, green and blue separately. Our custom software, which was developed specifically for real-time video recording, acquired 24-bit lossless colour images (8-bit per colour) at a resolution of  $1620 \times 1236$  pixels and a sampling rate of 20 frames per second. No constraints were imposed on the infant's posture, position and orientation.

Reference vital signs (heart rate, respiratory rate and  $\text{SpO}_2$ ) were monitored using a standard Philips IntelliVue MX800 patient monitor (Philips, Netherlands). The Philips monitor was equipped with modules for recording ECG and Impedance Pneumography (IP) signals and a Massimo LNCS Neo  $\text{SpO}_2$  adhesive sensor (Massimo, California, USA) for recording a PPG signal and  $\text{SpO}_2$ . The following vital signs were recorded at 1



Figure 1. A typical data acquisition setup in the Neonatal HDU study. (a) The video camera was positioned over a specifically-drilled hole at the top of the study incubator (red circle). (b) An image obtained from a typical recording session. The filming included the whole infant body. An X-RITE BST-13 colour checker chart (X-RITE, Michigan, USA) was placed inside the incubator during recording for colour correction.

Hz: heart rate from ECG, heart rate from PPG, respiratory rate from IP, and SpO<sub>2</sub>. The following waveforms were recorded: 1-lead ECG signal (at 250 Hz), bipolar IP signal (at 62.5 Hz) and PPG signal (at 125 Hz).

The recruitment of study participants imposed no restrictions regarding infant weight, ethnic groups and skin tones. A total of 90 recording sessions from 30 pre-term infants were made. The recordings were carried out from February 2014 to May 2015. One major objective of the clinical study was not to affect regular patient care. A more detail description of the study can be found in [4].

## 2.2 Selection of recording sessions for heart rate estimation

The study protocol requires the algorithms to be developed, trained and validated on half the study participants only, with the other half to be used for testing and evaluating the performance of the algorithms. In compliance with the protocol, the pre-term infants were divided into two sets of 15 participants each: the training and the test set. The date, time and duration of the video recordings together with patient demographics (ethnicity, gestational age and weight) in both groups were chosen to be balanced.

Since the data of the Neonatal HDU study were stored in a lossless uncompressed format in an encrypted storage cluster, a typical 1-hour video recording produced over 408 GB of data. This presents challenges to

Table 1. Summary of participant demographics for the chosen 20 recording sessions

	Training set	Test set
Total number of subjects	10	10
Total number of recording sessions	10	10
Average recording time (hh:mm)	05:46	05:20
Gestational age <sup>1,2</sup> (weeks)	30.3 ( $\pm 1.7$ )	30.7 ( $\pm 1.5$ )
Weight <sup>1,3</sup> (grams)	1,077.7 ( $\pm 256.0$ )	1,290.0 ( $\pm 237.4$ )
Gender (number of participants)		
Male	4	6
Female	6	4
Ethnicity (number of participants)		
White	7	7
Black	1	0
Mixed – White and Asian	1	2
Mixed – Any Other Mixed	1	1

<sup>1</sup> Values shown as Mean ( $\pm$ SD), <sup>2</sup> On the day of recording, <sup>3</sup> At the beginning of recruitment

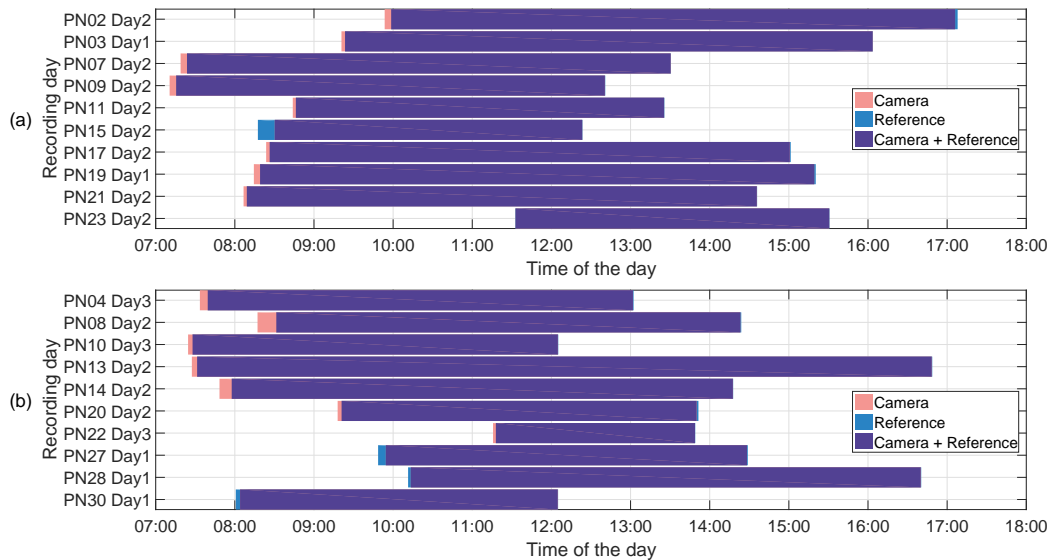


Figure 2. Recording time for each recording session in (a) the training set and (b) the test set.

the developing and testing of new algorithms. This paper, therefore, involved the heart rate estimation of 20 recording sessions, each from a different subject. The first 10 sessions were chosen from the training set, whereas the other 10 sessions were selected from the test set.

The selection was performed according to the following criteria: (1) training and test sets with similar patient demographics (gender, ethnicity, weight and gestational age); (2) no day-time interruption in the recording of video and reference data; (3) patient not under phototherapy treatment during the video recording; (4) video in focus and no colour over-saturation throughout the recording session; and (5) if more than one recording from a subject met all the previous criteria, the longest recording session was chosen.

Tab. 1 summarises the demographics of the chosen participants. Fig. 2 shows the total recording times for each of the chosen sessions in the training and test sets. The recording times were spread throughout the day.

### 3. BODY PART DETECTION

The estimation of heart rate from a specific part of the body requires an algorithm that can automatically locate the specific body part in a sequence of images. The CNN model presented in [9] is capable of identifying the presence of a patient and segmenting the patient’s skin regions. This work extends the CNN model by adding a body part detection capability for locating the patient’s head, torso and diaper. We first discuss our approach for annotating body parts for training the algorithm, followed by a description of our proposed multi-task network and training procedures.

#### 3.1 Body part annotation

The annotation of body parts (head, torso and diaper) was performed on the same set of images as in [9]. The images were taken every 6 minutes or 10 images per hour from each recording in the training set (15 infants); so there was a total of 2,269 images for annotation. Three human annotators were asked to manually label the body parts in order to ensure that high-quality ground truth data were obtained.

For each session, the annotator was required to label each body part in the first image by providing a tight rectangular bounding box with respect to the alignment of the body as shown in Fig. 3. The annotated bounding boxes were retained for the next image such that the annotator could move and modify each bounding box to match the body part in the next image. The annotator was asked to annotate only the body parts that he or she could see, e.g. the diaper hidden inside a blanket was not annotated.

The annotation was performed using the VGG Image Annotator (VIA)<sup>11</sup> software tool which was extended to allow the rotation of the bounding box. The information saved for each bounding box was the (x, y) coordinates of the top-left and bottom-right corners and the rotation angle of the bounding box.

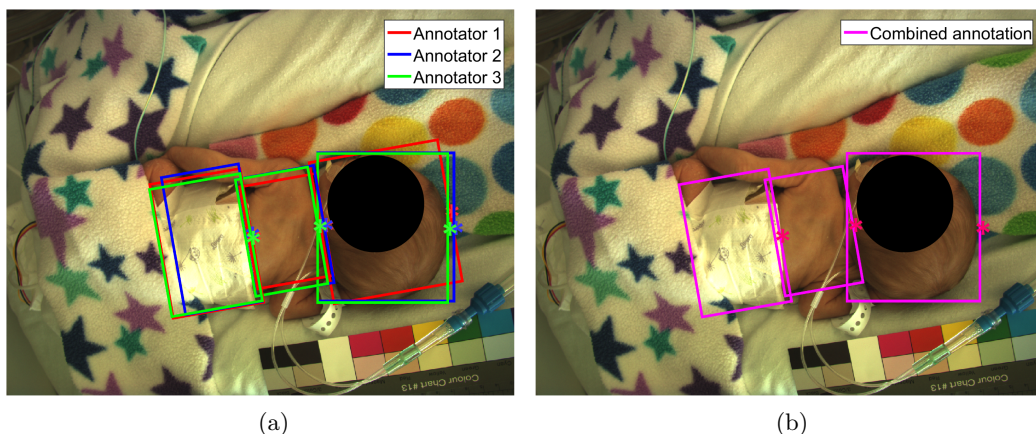


Figure 3. Body part annotation (a) head, torso and diaper labelled from three anotators. Each bounding box is rotated with respect to the alignment of the body. (b) The consensus of the three annotators. The asterisk symbol \* denotes the upper part of the bounding box. The black circles were added to preserve the privacy of the patient.

Once the annotation was finished, bounding boxes from three annotators were combined to compute the ground truth. The bounding boxes were first rotated to be aligned with the horizontal axis of the image frame. The final x-y coordinates were defined as the top-left and bottom-right corners of the rectangular region that at least two annotators agreed. Then, a pairwise agreement between each annotator pairs was calculated by finding the ratio of the intersection to the union between the bounding boxes of each pair of annotators. The final rotation angle was defined as the mean rotation angle of the pair that had the highest agreement score.

The annotation process generated 1,890 hand instances, 1,974 torso instances and 1,651 diaper instances. An inter-annotator agreement was calculated using the ratio of the intersection of the bounding boxes provided by at least two annotators to the union of the bounding boxes provided by all annotators. The mean inter-annotator agreements were 94.8% for the head class, 91.6% for the torso class and 87.1% for the diaper class.

### 3.2 Network architecture

The CNN model presented in [9] consists of a VGG16 shared core network<sup>12</sup> and two branches for patient detection and skin segmentation. The patient detection branch was implemented using global average pooling.<sup>13</sup> The skin segmentation branch was implemented using a fully convolutional network<sup>14</sup> to perform a hierarchical upsampling of feature maps across the shared core network for segmenting skin regions. This work added a body part detection branch to the CNN model for localising body parts (head, torso and diaper) in the image.

The body part detection branch was implemented using the Faster R-CNN network<sup>15</sup> which consists of two modules: a region proposal network (RPN) module that proposes potential regions; and an object detection module that locates an object in the proposed regions and classifies them into object classes (see Fig. 4). The implementation followed that of [15]. The RPN module was composed of a  $3 \times 3$  convolutional layer followed

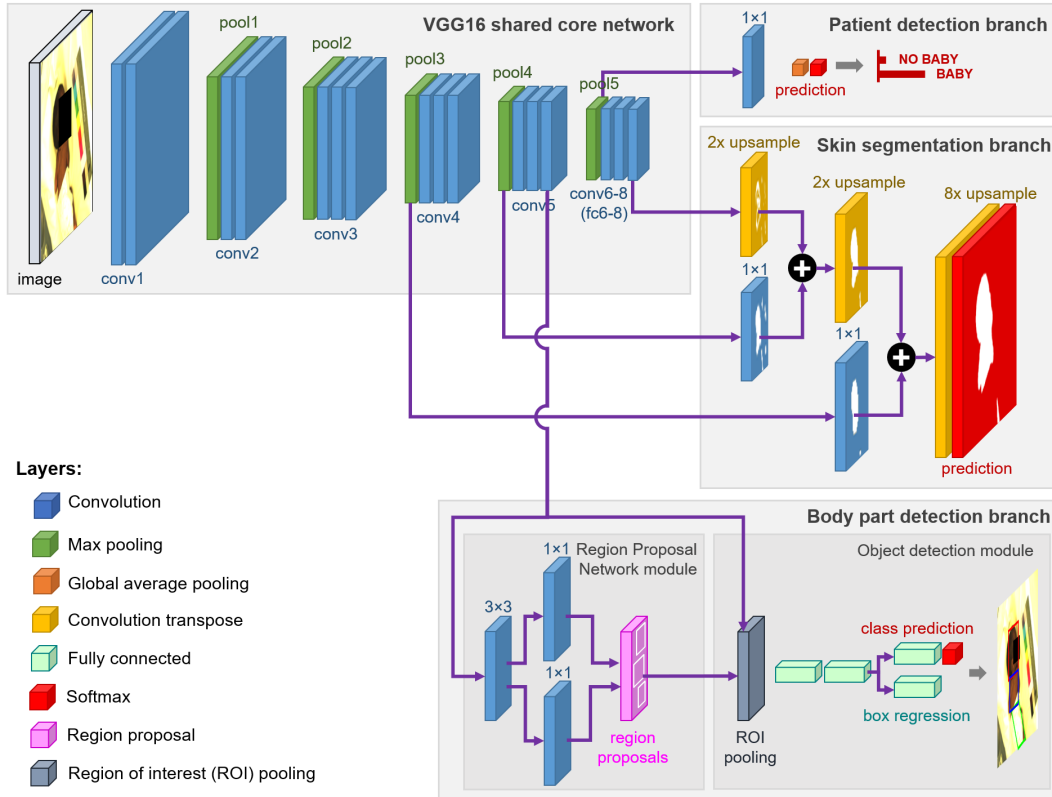


Figure 4. Our proposed CNN model consists of a VGG16 shared core network<sup>12</sup> and three branches for patient detection, skin segmentation and body part detection respectively. This unified network can simultaneously perform all video analysis pre-processing steps required prior to the estimation of vital signs.

by two  $1 \times 1$  convolutional layers for the classification of the proposal regions (object or not object) and the regression of proposal region coordinates (four x-y coordinates) respectively. The RPN module was slid over the last convolutional layer of the `conv5` group in the shared core network to generate region proposals. The object detection module, which receives the proposed regions from the RPN module, consists of two fully-connected layers with 4,096 outputs, similar to the `fc6` and `fc7` layers in the VGG16 network,<sup>12</sup> followed by two fully-connected layers for the classification of each proposed region into an object class (background, head, torso or diaper) and the regression of per-class bounding-box offsets (5 outputs per class – four x-y coordinate offsets and a rotation angle) respectively. The bounding-box offsets specify a translation and rotation with respect to the proposed region. More detail on the implementation of the Faster R-CNN network is given in [15,16].

The body part detection branch normally returns a large number of overlapping bounding boxes. A non-maximum suppression technique was applied to remove repeated detections by selecting a bounding box with the highest score while skipping the bounding boxes overlapped with the selected one by at least 50%.

It is important to note that, in addition to the original Faster R-CNN network, this work included the rotation offset parameter which could prevent the overlapping of the bounding box to other body parts if the infant was not positioned in vertical or horizontal directions (see Fig. 1b in which the subject is positioned at  $60^\circ$ ). The use of the axis-aligned bounding box (no rotation offset parameter) could result in the estimation of heart rate from the torso to also include a large part of the face.

The proposed extensions use the shared convolutional feature maps in the network, thus the body part detection task was performed at a small additional run-time cost. Our proposed unified network can simultaneously perform all video analysis steps required prior to the estimation of vital signs.

### 3.3 Network training

All new layers were initialised using a Gaussian distribution with zero mean, except for the first two fully-connected layers of the object detection module which were initialised using weights from the VGG16 network.<sup>12</sup> In the body part detection branch, the classification layers were equipped with log loss and the regression layers are equipped with smooth  $L_1$  loss.<sup>16</sup> Since our proposed network was extended from [9], the network was re-trained jointly for all patient detection, skin segmentation and body part detection tasks. Only the images that have all three annotations were used for training and validation. All the images and their ground truths were resized to  $512 \times 512$  pixels. Their aspect ratio was maintained with black spaces at the top and bottom of the image. The mean of each colour channel computed over all the images was subtracted from every image frame.

The following data augmentation techniques were applied to the images randomly while training with the aim to reduce overfitting and improve the network’s generalisation: horizontal and vertical flipping, rotation by a random value between  $-180^\circ$  and  $180^\circ$ , zooming by a random scale between 0.75 and 1.25, scaling the lightness component in the HSL (Hue, Saturation, Lightness) colour space by a random scale between 0.50 and 1.50 and scaling the saturation component in the HSL colour space by a random scale between 0.50 and 1.50.

The network was implemented on the MatConvNet framework.<sup>17</sup> The training was performed end-to-end using a standard Stochastic Gradient Descent (SGD) with a mini-batch size of one image, similar to [15]. The learning rates were scheduled to start at 0.0001 and reduced by a factor of 10 for every 50k mini-batches. The training used a momentum of 0.9 and a weight decay of 0.005.

### 3.4 Evaluation protocol

The same evaluation protocol as in [9] was used. The images were separated into two independent sets,  $D_1$  and  $D_2$  such that one set had 7 infants and the other set had 8 infants. The assignment to each set was based on the balance of gender, gestational age, ethnic group, and the number of images. A model was first trained on the  $D_1$  set and validated on the  $D_2$  set. Subsequently, another model was trained on the  $D_2$  set and validated on the  $D_1$  set. Eventually, validation results from both models were combined to produce a predictive performance.

The predictive performance is reported using a Precision-Recall (PR) curve for each body part. A bounding box is regarded as positive if its overlap score is more than 0.50, where the score is defined as the intersection divided by the union of a detected bounding box and a ground-truth bounding box. Precision is the fraction of the correct detections to all detections. Recall is the proportion of correct detections to all ground truths. The area under the PR curve, called Average Precision (AP), provides a metric to describe the detection performance.

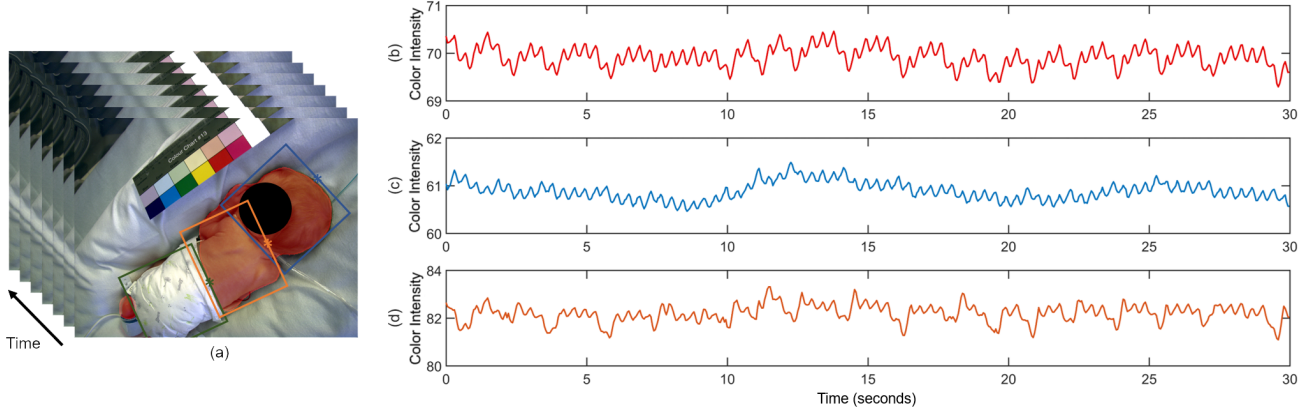


Figure 5. Extraction of raw PPGi signals from the entire skin, head and torso. (a) Video frames are overlaid with segmented skin area and detected body parts. 30-second time series signals extracted from the mean of the green channel for (b) the entire segmented skin area, (c) the segmented skin area inside the head bounding box and (d) the segmented skin area inside the torso bounding box respectively. The heart rate during this period was 158 beats/min.

## 4. HEART RATE ESTIMATION

### 4.1 Reference data

In the Neonatal HDU study, two heart rate measurements were provided by the patient monitor: the first computed from the ECG and the second computed from the PPG recorded using a pulse oximeter attached to the patient's foot. Similarly to [5], a reference heart rate was computed as the mean of the ECG and PPG heart rates during periods for which the two values do not differ by more than 5 beats/min (the maximum error allowed for medical devices according to the ANSI/AAMI EC13:2002 standard).

### 4.2 Extraction of PPGi signals

In the Neonatal HDU study, pre-term infants were nursed naked in front of the video camera. Typically, around 15–20% of a given video frame corresponds to skin pixels (see Fig. 1b). In the case that heart rate was estimated from the entire skin area, a raw PPGi signal was calculated by spatially averaging all pixels in the entire skin area from the green colour channel for each frame (see Fig. 5b). The green colour channel typically contains the strongest plethysmographic signal.<sup>18</sup> In the case of estimating heart rate from the head or torso, the raw PPGi signal was extracted from the skin area inside a detected bounding box for each frame in the video (see Fig. 5c–d). The raw PPGi signal was computed on a per-frame basis from raw uncompressed video data stored at the original resolution of  $1620 \times 1236$  pixels and a frame rate of 20 frames per second.

Since the raw PPGi signal contains pulsatile components correlated with the cardiac frequency as well as motion artefacts and often other sources of noise, the signal was first detrended in order to remove any DC offset and then filtered using a cascade of a 40th-order low-pass Finite Impulse Response (FIR) filter with a cut-off frequency of 4.5 Hz (270 beats/min) and a 60th-order high-pass FIR filter with a cut-off frequency of 1.5 Hz (90 beats/min). The filters attenuate the frequency components outside the physiological range of interest. Once the signal is filtered, peak and onset detection<sup>19</sup> was carried out to identify peaks and onsets corresponding to cardiac cycles in the PPGi signal.

### 4.3 Signal quality assessment

The assessment of the quality of the PPGi signal is of high importance as data corruption by subject movement and changes in the lighting conditions present considerable challenges during video analysis. The signal quality assessment in this work was extended from the previous work presented in [5].

Firstly, an activity index was computed on a per-frame basis based on changes in the centroid of the segmented skin in the case of estimating heart rate from the entire skin or changes in the centroid of the detected bounding

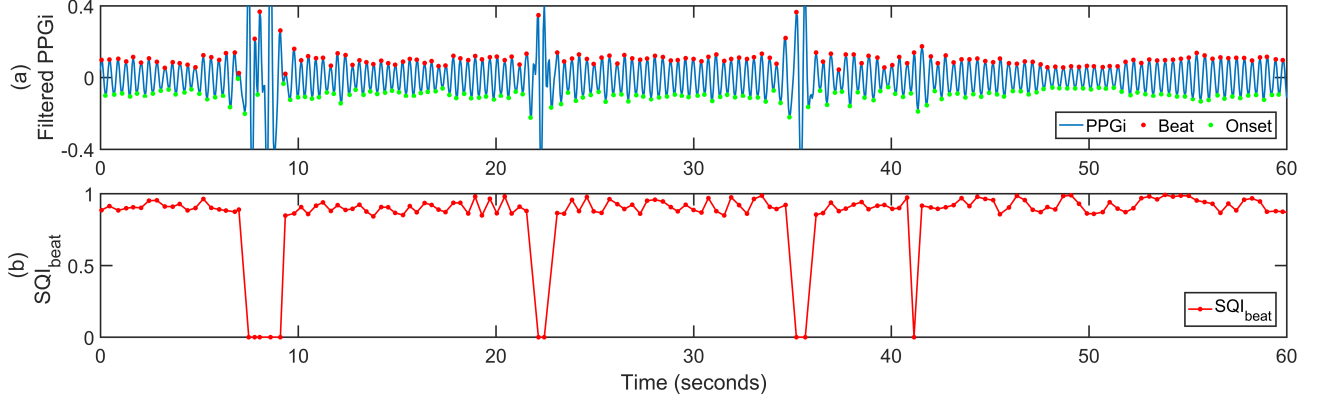


Figure 6. Signal quality assessment for a 60-second PPGi signal extracted from the torso during which the patient was awake. (a) Filtered PPGi signal with the peaks and onsets of each beat detected. (b) Signal Quality Index (SQI), noisy beats are flagged as invalid ( $SQI_{\text{beat}} = 0$ ).

boxes in the case of estimating heart rate from the head or torso. The beats occurring during time periods of high motion, judged as the centroid shifted more than 1 cm (measured using the ruler as part of the colour chart), were flagged as invalid.

Bayesian change point detection<sup>20</sup> was carried out on the raw PPGi signal to identify step changes, often caused by motion artefacts or changes in the lighting conditions (overhead light turned on or off; window blinds opened or closed; or clinical staff walking pass by the incubator). Given a time series data  $x$  of length  $N$  is from two piece-wise constant inputs with Gaussian noise added, the probability of a step change at location  $m$  is defined as in equation Eq. 1.<sup>5,20</sup> The beats occurring during a step change were flagged as invalid.

$$P(m|x) \propto \frac{1}{\sqrt{m(N-m)}} \left[ \sum_{i=1}^N x_i^2 - \frac{1}{m} \left( \sum_{i=1}^m x_i \right)^2 - \frac{1}{N-m} \left( \sum_{i=m+1}^N x_i \right)^2 \right]^{-\frac{N-2}{2}} \quad (1)$$

Finally, a beat-by-beat quality assessment was performed over a 15-second running window with a step size of 5 seconds. For each window, the beats were flagged as invalid if they were outside a physiological range between 90 and 270 beats/min, the amplitude of the beats were outside the range of the mean plus three standard deviations of the window or the signal was clipped. A template was computed as the mean of all valid beats in the 15-second window. A multi-scale Dynamic Time Warping (DTW) algorithm was employed to measure the similarity of each valid beat in the window by calculating the minimum distance between each beat to the window template. The multi-scale DTW algorithm can cope with nonlinear and non-stationary changes in the beat morphology occurring due to changes in heart rate or cardiac output.<sup>5,21</sup> The signal quality index  $SQI_{\text{beat}}$  was defined as

$$SQI_{\text{beat}}(k) = 1 - \frac{DTW(k)}{100} \quad (2)$$

where  $DTW(k)$  is the minimum distance between the  $k$ th beat and the window template. Since the calculation of the SQI was performed using a running window, the highest SQI value was taken when the windows overlap. Fig. 6 shows an example of signal quality assessment for a 60-second PPGi signal from the torso.

#### 4.4 Estimation of heart rate

Heart rate estimation was performed using autoregressive modelling by extending the technique presented in [3]. The autoregressive model has the benefit of having no frequency resolution limitations on short time-series data compared to the Fast Fourier Transform (FFT).

Heart rate was estimated using a window length of 8 seconds with a step size of 1 second. There were 160 samples or approximately 18 cardiac cycles for each time window. An autoregressive model was fit to the time-series data such that the coefficients of the model are estimated. The number of coefficients used to describe the

model is called the model order. The choice of model order is a compromise between a high model order, which can provide a better approximation of the signal but can also overfit to the noise in the signal; and a low model order, which may not be sufficient to represent the signal.<sup>3,5</sup> Although many techniques were presented for the selection of the model order, there is no definite method that can guarantee an optimal model order.

In this work, the time-series data were fitted to a set of autoregressive models ranging from 4th to 12th model order computed using the Burg method.<sup>22</sup> The best model was the model for which the difference between the frequency of the dominant pole and the highest peak of the frequency response in the frequency band of 1.5 to 4.5 Hz is less than 1 beat/min. If more than one model falls in this criterion, the model with the highest amplitude of the dominant pole was chosen. If no model falls in this criterion, heart rate for that time window was not estimated. Heart rate was estimated by finding the highest peak of the frequency response from the best model. The SQI of the heart rate  $SQI_{hr}$  was defined as

$$SQI_{hr} = \frac{1}{N} \sum_{k=1}^N SQI_{beat}(k) \quad (3)$$

where  $k = 1, 2, \dots, N$  is the list of beats in the corresponding time window.

Once heart rate was computed, a Kalman filter was applied to the heart rate estimates to track and adjust the heart rate based on their signal quality and reduce the effects of transient changes, noise and artefacts.<sup>5,23</sup>

In the case of the fusion of multiple heart rate estimates, the overall heart rate estimate of each time window  $HR_{fusion}$  was combined by weighting each Kalman-filtered heart rate estimate  $HR_k$  as:<sup>5,23</sup>

$$HR_{fusion} = \sum_{k=1}^N \left( \frac{\prod_{i=1, i \neq k}^N \sigma_i^2}{\sum_{i=1}^N (\prod_{j=1, j \neq i}^N \sigma_j^2)} \cdot HR_k \right) \quad (4)$$

where  $k = 1, 2, 3$  corresponds to the list of heart rate estimates from the head, torso and entire skin respectively. Given the SQI of heart rate  $SQI_{hr}$  and Kalman residual error  $r$ , the weighted innovation  $\sigma^2$  is defined as<sup>23</sup>

$$\sigma^2 = \left( \frac{r}{SQI_{hr}} \right)^2. \quad (5)$$

This procedure weights the heart rate estimates from cleaner data (high  $SQI_{hr}$  and low  $r$ ) higher than noisy data (low  $SQI_{hr}$  and high  $r$ ).<sup>23</sup>

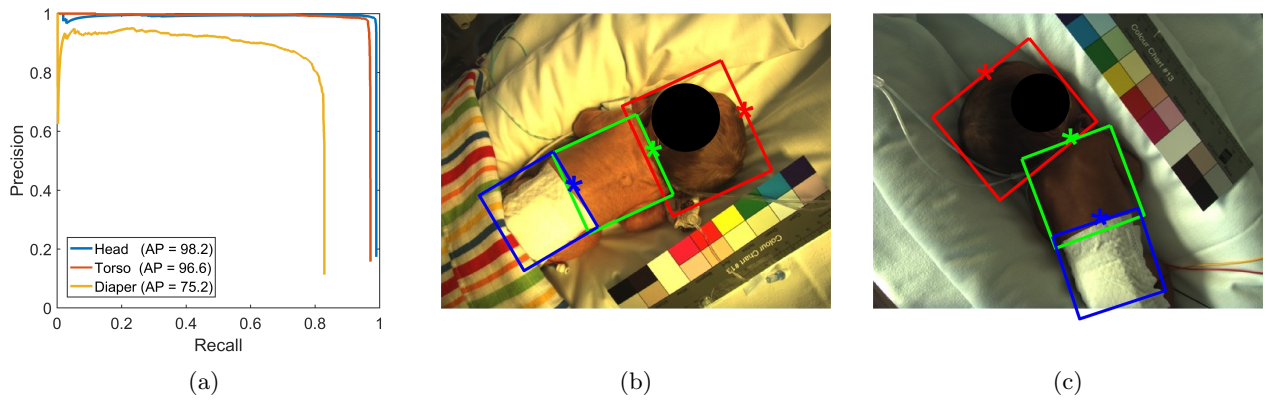


Figure 7. Body part detection (a) Precision-recall curves reporting detection performance on three categories: head, torso and diaper. (b–c) Examples of body part detection on typical video frames taken from videos that were not part of the training data. The CNN model was able to predict the coordinates and rotation of each body part correctly.

## 5. RESULTS

### 5.1 Body part detection results

Fig. 7a shows the precision-recall (PR) curves describing the performance of the body part detection task. Each PR curve denotes different body parts (head, torso and diaper). The average precisions (AP) obtained for the head, torso and diaper classes are 98.2, 96.6 and 75.2 respectively. Fig. 7b-c shows examples of the results of the body part detection algorithm.

### 5.2 Heart rate estimation results

Fig. 8 shows continuous heart rate estimates for a 1-hour segment during a quiet period. The training set was used to compute the estimation parameters such as filter coefficients and the appropriate thresholds for signal quality assessment, while the test set used these parameters to assess the performance of the algorithm. The comparison was performed on the valid data which were defined as the periods during which the reference heart rate was available and the infant was present. The valid data were available for 50.7 hours (87.7%) of the total time of 57.8 hours for the training set and 38.5 hours (66.6%) of the total time of 53.5 hours for the test set.

Tab. 2 summarises the heart rate estimation results for all chosen sessions in both training and test sets. The mean absolute errors (MAE) and mean absolute deviations (MAD) of the heart rate computed by taking the entire skin regions (baseline) were lower than the errors of that computed by taking only the head or torso as a region of interest. The heart rate derived from the data fusion of the head and torso regions yielded similar results to the heart rate derived from the entire skin. By combining the heart rate estimates from all regions

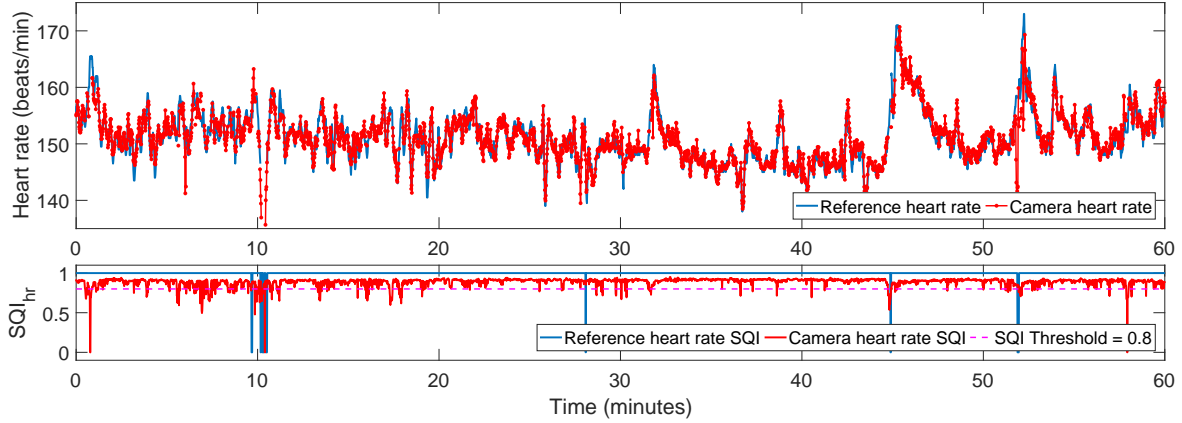


Figure 8. Heart rate estimates, shown in red, combined from the head, torso and entire skin for a 60-minute segment during which the infant was quietly sleeping. The reference heart rate is shown in blue. The SQL threshold of 0.80 defines the boundary between good-quality and poor-quality heart rate estimates.

Table 2. Summary of heart rate estimation results for both training and test sets

Region of interest	Training set			Test set		
	MAE <sup>†</sup>	MAD <sup>†</sup>	Estimated time <sup>‡</sup>	MAE <sup>†</sup>	MAD <sup>†</sup>	Estimated time <sup>‡</sup>
Entire skin (baseline)	2.8	4.5	30.4 h, 60.0 %	2.2	3.3	27.4 h, 71.1 %
Head	3.8	5.7	25.4 h, 50.2 %	3.3	4.4	22.7 h, 58.7 %
Torso	3.2	4.7	19.9 h, 39.4 %	2.8	4.3	17.2 h, 44.5%
Combined head and torso	3.1	5.0	33.0 h, 65.2 %	2.8	4.3	27.0 h, 70.1%
Combined skin, head and torso	3.0	5.1	37.9 h, 75.0 %	2.4	4.1	31.1 h, 80.6%

<sup>†</sup> Heart rate in beats/min

<sup>‡</sup> Percentage of valid data during which the reference data were available and the infant was present.

of interests (the entire skin, head and torso), the errors were reduced and the proportion of estimated time are improved by approximately 10% compared to the baseline.

## 6. DISCUSSION

### 6.1 Body part detection

Our proposed network can estimate the location and orientation of the head and torso with high accuracy even under low ambient lighting. However, the network had some difficulties in detecting the diaper when it was mostly covered by a blanket. Even though the location of the diaper does not contain cardiac information for the estimation of heart rate, it can be used to derive the motion of the lower torso which could be used as a clue for the estimation of respiratory rate.

### 6.2 Network training and run-time evaluation

The network converged after 150k mini-batches and the training took approximately 36 hours on a Nvidia GTX Titan 6GB GPU. The network was able to process  $512 \times 512$  video frames at a rate of 3.5 frames per second (including a non-maxima suppression process in the body part detection task). There is room for improvements towards real-time performance; for example, changing the shared core network to a model with fewer parameters and switching the object detection branch to an architecture that does not require a proposal generation.

### 6.3 Comparison to other clinical studies

The errors obtained from the estimation of 20 recording sessions are comparable to those presented in the initial analysis of the first two pre-term infants in 2014<sup>4</sup> (MAE of 2.8 beats/minute for 80% of the time). For an adult population, Villarroel *et al.*<sup>5</sup> reported the results of non-contact vital sign estimation for over 370 hours for patients undergoing haemodialysis treatment, the MAE between the reference heart rate and camera-derived heart rate was 2.8 beats/minute for 65% of the time. These errors were similar to those obtained in this study of pre-term infants. Compared to adults, neonates generally have a higher range of heart rate, depending on age and clinical condition. Episodes of short-term fluctuations can be found even during sleeping (see Fig. 8).

### 6.4 Continuous heart rate estimation

In a hospital environment, it is beneficial to monitor the status of the patient in a continuous manner. Even though the non-contact monitoring of heart rate can be challenging in the clinic due to the active nature of the infants, heart rate can be estimated with a reasonable accuracy during quiet and stable periods (see Fig. 8). Fig. 9 shows the histogram of the periods for which heart rate cannot be estimated. The gap durations are mostly concentrated below 30 seconds. The interruption of heart rate estimates for several seconds would not be considered a major problem in a practical clinical environment.<sup>24</sup>

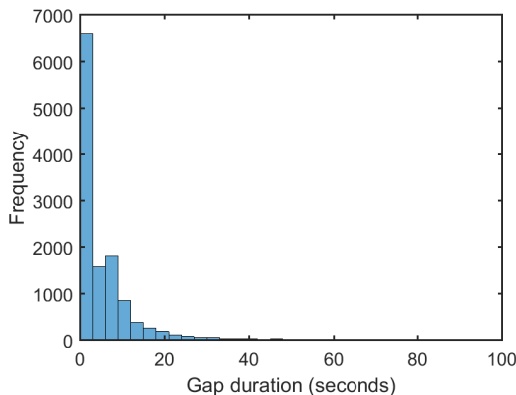


Figure 9. Gaps in camera-derived heart rate estimates for all 20 infants, most of them are under 30 seconds.

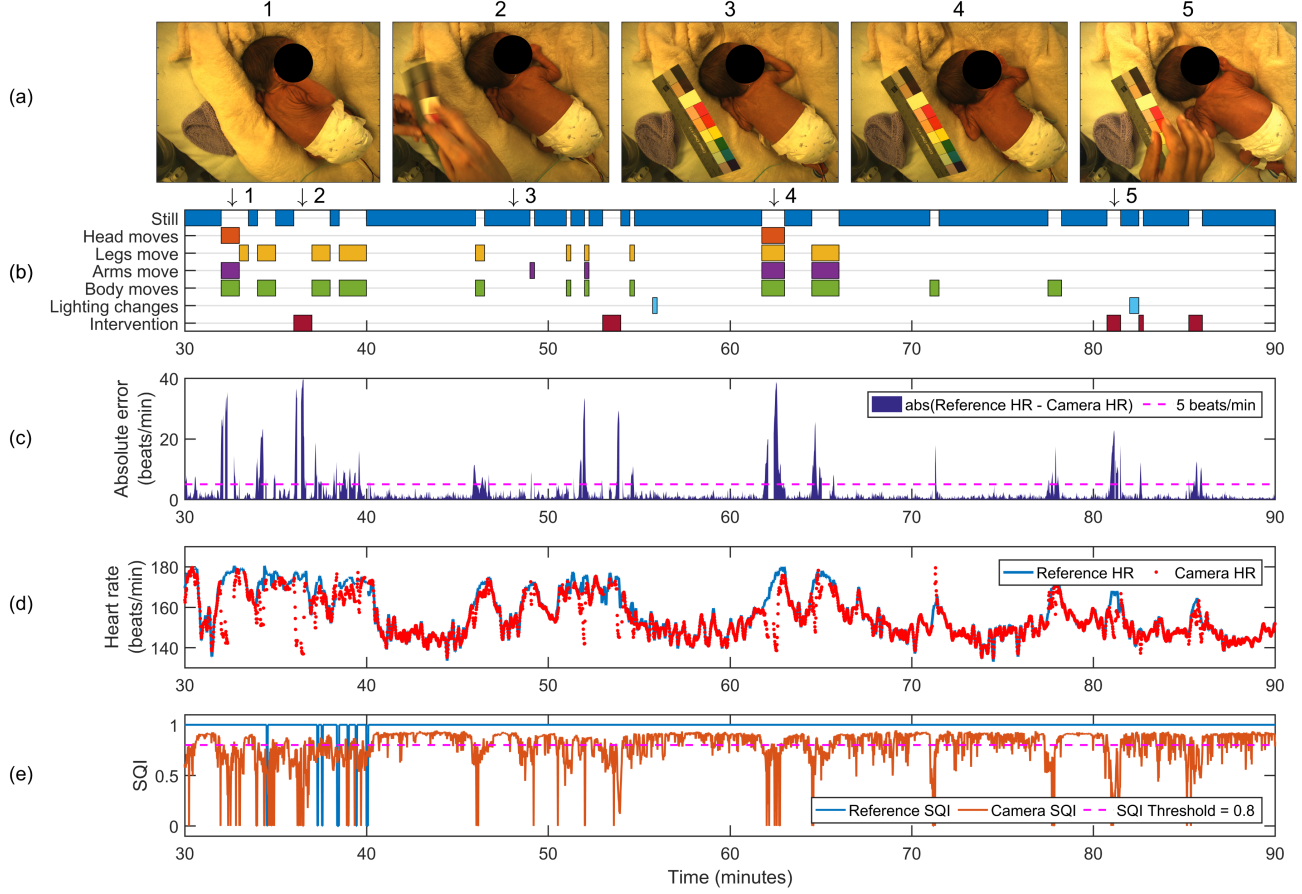


Figure 10. Typical activities of the patient and clinical staff can contribute to errors in estimating vital signs. (a) Video frames corresponding to the time in the timeline plot below. (b) The timeline of patient activities (manually annotated) over a 60-minute segment for a typical recording session. (c) Absolute difference between the reference and camera heart rate estimates. Almost all the large errors are associated with periods of patient movement. (d) Comparison of reference heart rate and camera-derived heart rate. (e) Heart rate SQIs.

## 6.5 Patient activities

Fig. 10 shows heart rate estimates for a 60-minute sample period with patient movement and clinical interventions annotated. During periods of patient motion, heart rate estimates become unreliable with low signal quality, the absolute error between the reference values and the camera-derived heart rate estimates increase substantially. Computer vision techniques could be used to identify and classify patient motion patterns and improve the assessment of signal quality.

## 7. CONCLUSION

We presented the comparison of heart rate estimates from different regions of interest (entire skin, head and torso) as well as heart rate estimates combined from multiple regions of interest. Our results showed that the heart rate estimates combined from all regions of interest achieved the lowest MAE of 2.4 beats/min for over 80% of the time. The estimation errors mostly occurred during short time periods of patient motion, typically under 30 seconds.

Future work includes the improvement to the CNN model to reduce the running time, the incorporation of computer vision techniques to improve the assessment of signal quality, and the use of Gaussian process regression to predict heart rate estimates during short-time periods where the data were incomplete.

## ACKNOWLEDGMENTS

SC and JJ acknowledge the RCUK Digital Economy Programme under grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation). SC was supported by the National Science and Technology Development Agency, Thailand. MV was supported by the Oxford Centre of Excellence in Medical Engineering funded by the Wellcome Trust and EPSRC under grant number WT88877/Z/09/Z. JJ was supported by the Fundação para a Ciência e Tecnologia, Portugal, under grant number SFRH/BD/85158/2012. GG and KM were supported by the NIHR Biomedical Research Centre Programme, Oxford. This work was supported by the EPSRC Programme Grant Seebibyte EP/M013774/1. We would like to thank our participants and their parents who agreed to take part in the clinical study and Sara Davis who carried out the study.

## REFERENCES

- [1] Scalise, L., Bernacchia, N., Ercoli, I., and Marchionni, P., “Heart rate measurement in neonatal patients using a webcam,” in [*Proceedings of the IEEE Int. Symp. on Medical Measurements and Applications*], 6–9 (2012).
- [2] Aarts, L. A. M., Jeanne, V., Cleary, J. P., Lieber, C., Nelson, J. S., Bambang Oetomo, S., and Verkrusysse, W., “Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit,” *Early Human Development* **89**(12), 943–48 (2013).
- [3] Tarassenko, L., Villarroel, M., Guazzi, A., Jorge, J., Clifton, D. A., and Pugh, C., “Non-contact video-based vital sign monitoring using ambient light and auto-regressive models,” *Physiological Measurement* **35**(5), 807–31 (2014).
- [4] Villarroel, M., Guazzi, A., Jorge, J., Davis, S., Green, G., Shenvi, A., Watkinson, P., McCormick, K., and Tarassenko, L., “Continuous non-contact vital sign monitoring in neonatal intensive care unit,” *Healthcare Technology Letters* **1**(3), 87–91 (2014).
- [5] Villarroel, M., Jorge, J., Pugh, C., and Tarassenko, L., “Non-contact vital sign monitoring in the clinic,” in [*Proceedings of the IEEE Conf. on Automatic Face and Gesture Recognition*], 278–285 (2017).
- [6] Jorge, J., Villarroel, M., Chaichulee, S., Guazzi, A., Davis, S., Green, G., McCormick, K., and Tarassenko, L., “Non-Contact Monitoring of Respiration in the Neonatal Intensive Care Unit,” in [*Proceedings of the IEEE Conf. on Automatic Face and Gesture Recognition*], 286–293 (2017).
- [7] Behrman, R. E. and Butler, A. S., “Mortality and acute complications in preterm infants,” in [*Preterm Birth: Causes, Consequences, and Prevention*], Behrman, R. E. and Butler, A. S., eds., 263, National Academies Press, Washington (DC), USA (2006).
- [8] Lloyd, R., Goulding, R., Filan, P., and Boylan, G., “Overcoming the practical challenges of electroencephalography for very preterm infants in the neonatal intensive care unit,” *Acta Paediatrica* **104**(2), 152–157 (2015).
- [9] Chaichulee, S., Villarroel, M., Jorge, J., Arteta, C., Green, G., McCormick, K., Zisserman, A., and Tarassenko, L., “Multi-Task Convolutional Neural Network for Patient Detection and Skin Segmentation in Continuous Non-Contact Vital Sign Monitoring,” in [*Proceedings of the IEEE Conf. on Automatic Face and Gesture Recognition*], 266–272 (2017).
- [10] Kenner, C., Wright, J., and Lott, J. W., [*Comprehensive neonatal care: an interdisciplinary approach*], vol. 6, Saunders Elsevier, St. Louis, MO (2007).
- [11] Dutta, A., Gupta, A., and Zissermann, A., “VGG Image Annotator (VIA).” Accessed: 3 September 2017 <http://www.robots.ox.ac.uk/~vgg/software/via/> (2016).
- [12] Simonyan, K. and Zisserman, A., “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in [*Proceedings of the Int. Conf. on Learning Representations*], (2015).
- [13] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., “Going Deeper with Convolutions,” in [*Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*], 1–9 (2015).
- [14] Long, J., Shelhamer, E., and Darrell, T., “Fully Convolutional Networks for Semantic Segmentation,” in [*Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*], 3431–40 (2015).

- [15] Ren, S., He, K., Girshick, R., and Sun, J., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(6), 1137–1149 (2017).
- [16] Girshick, R., “Fast R-CNN,” in [*Proceedings of the Int. Conf. on Computer Vision*], 1440–1448 (2015).
- [17] Vedaldi, A. and Lenc, K., “MatConvNet – Convolutional Neural Networks for MATLAB,” in [*Proceedings of the ACM Int. Conf. on Multimedia*], 689–92 (2015).
- [18] Verkruysse, W., Svaasand, L. O., and Nelson, J. S., “Remote plethysmographic imaging using ambient light,” *Optics Express* **16**(26), 21434–45 (2008).
- [19] Zong, W., Heldt, T., Moody, G., and Mark, R., “An open-source algorithm to detect onset of arterial blood pressure pulses,” in [*Computers in Cardiology*], 259–262 (2003).
- [20] Ó Ruanaidh, J. J. K. and Fitzgerald, W. J., “Retrospective Changepoint Detection,” in [*Numerical Bayesian Methods Applied to Signal Processing*], 96–121 (1996).
- [21] Li, Q. and Clifford, G. D., “Dynamic time warping and machine learning for signal quality assessment of pulsatile signals,” *Physiological Measurement* **33**(9), 1491–1501 (2012).
- [22] Burg, J. P., “Maximum Entropy Spectral Analysis,” in [*Proceedings of the 37th Ann. Int. SEG Meeting*], **6** (1975).
- [23] Li, Q., Mark, R. G., and Clifford, G. D., “Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter,” *Physiological Measurement* **29**(1), 15–32 (2008).
- [24] Evans, D., Hodgkinson, B., and Berry, J., “Vital signs in hospital patients: A systematic review,” *International Journal of Nursing Studies* **38**(6), 643–650 (2001).