Non-contact vital sign monitoring of pre-term infants



Sitthichok Chaichulee
St Catherine's College
University of Oxford

Supervised by
Professor Lionel Tarassenko
Dr Mauricio Villarroel

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2018

Declaration

I declare that this thesis is entirely my own work and, except where otherwise stated, describes my own research.

Sitthichok Chaichulee
St Catherine's College

Acknowledgements

This thesis would not have been completed without the help and support of many people around me. First and foremost, I would like to thank Professor Lionel Tarassenko for his guidance, freedom and encouragement throughout my studies at Oxford. It was a great privilege to work with him. I have learned so much through his leadership, expertise in the field and high-quality research. Thank you to Dr Mauricio Villarroel for his advice, enthusiasm for research and unending tolerance. His red pen helped shape my scientific thinking and writing. Thank you to Professor Andrew Zisserman for many inspiring meetings and insightful discussions. I am very grateful for his guidance and thoughtful feedback. Thank you to Dr Kenny McCormick for his clinical advice and assistance with the clinical study. It has been very fortunate to work with such exemplary people.

Over the past four years, it has been a great honour and pleasure to be a part of the camera research team and the Biomedical Signal Processing Group. Thank you to Dr João Jorge, Dr Alessandro Guazzi, Dr Jonathan Daly and Dr Hamish Tomlinson. Their useful discussions and intellectual support have made my research possible. Thank you to Dr Carlos Arteta for all our meetings and for providing guidance on deep learning. On the clinical side, I would like to thank Dr Gabrielle Green, Sara Davis and the staff of the John Radcliffe Hospital who involved in the clinical study for their dedication and excellent work. Special thanks are due to our participants and their parents who agreed to take part in the clinical study.

I would also thank my examiners Professor David Delpy and Professor Andrea Vedaldi for their detailed reading and helpful discussions.

Many people have made my DPhil study such an exciting and memorable journey. I have grown so much during the past four years. Thank you to my fellows in the CDT in Healthcare Innovation, my friends at St Catherine's College, St Cross College and the Institute of Biomedical Engineering and many others who contributed to my life. I am truly blessed to have met you all.

I would like to thank several funding agencies that supported my studies, including the Ministry of Science and Technology (Thailand), the UK Research and Innovation (formerly Research Councils UK) and the NIHR Oxford Biomedical Research Centre. I would also thank all my friends from far and near and, of course, my parents for their support and care.

Abstract

Non-contact vital sign monitoring using video cameras enables the measurement of heart rate and respiratory rate to be performed without sensors attached to the skin of the patient. This provides advantages in terms of patient comfort and the management of skin irritation and infection. However, the use of non-contact technologies in the clinic presents several challenges such as variations in human skin colour, changes in lighting conditions, the detection of the presence of a patient in the video frame and the selection of suitable regions of interests (ROIs) from which vital signs can be estimated.

This thesis proposes a framework for the accurate and continuous measurement of heart rate and respiratory rate in pre-term infants in a real-world hospital environment. The framework has been developed and validated on the data obtained from a clinical study of pre-term infants hospitalised in the high-dependency area of the neonatal intensive care unit (NICU) of the John Radcliffe Hospital in Oxford. The study involved the recording of videos and reference vital signs for 90 sessions during daytime from 30 pre-term infants, comprising a total recording time of 426.6 hours. The dataset provided a wide range of vital sign values for developing and validating the framework.

Multi-task deep learning algorithms were developed so that vital-sign estimation could be performed only when the infant was present in front of the video camera and no clinical interventions were undertaken. The algorithms are able to deal with different skin tones, lighting condition changes and variable body postures.

Heart rate is estimated from the photoplethysmographic imaging (PPGi) signal derived from subtle changes in the colour of the skin areas. Respiratory rate is estimated using data fusion from the PPGi signals along with the shape and morphological properties of the skin areas. Signal quality assessment algorithms are developed for both heart rate and respiratory rate to discriminate between clinically acceptable and noisy signals. In a test set, the mean absolute error between the reference and camera-derived heart rates is 2.3 beats/min for over 76% of the time for which the reference and camera data are valid. The mean absolute error between the reference and camera-derived respiratory rate is 3.6 breaths/min for over 78% of the time. Most of the time periods during which the video camera cannot provide estimates of heart rate and respiratory rate are lower than 30 seconds.

This thesis further investigated the feasibility of using non-contact algorithms to estimate heart rate and respiratory rate during a painful clinical procedure, a heel prick for withdrawing blood for blood gas analysis. Adaptation of the algorithms enables physiological indicators of pain to be derived from the analysis of the video camera data.

Publications

Journal paper

- M. Villarroel, S. Chaichulee, J. Jorge, S. Davis, G. Green, C. Arteta, A. Zisserman, K. McCormick, P. Watkinson, and L. Tarassenko, "Non-contact physiological monitoring of preterm infants in the Neonatal Intensive Care Unit," NPJ Digital Medicine, 2019. (In press)
- S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, K. McCormick, A. Zisserman and L. Tarassenko, "Cardio-respiratory signal extraction from video camera data for continuous non-contact vital sign monitoring using deep learning," Physiological Measurement 2019. (In press)
- J. Jorge, M. Villarroel, S. Chaichulee, G. Green, K. McCormick, and L. Tarassenko, "Assessment of direct and indirect methods of measuring the respiratory rate of preterm infants in the neonatal intensive care unit," IEEE Journal of Biomedical and Health Informatics, 2019. (In press)
- M. Harford, S. H. Clark, J. F. Smythe, S. Gerry, M. Villarroel, J. Jorge, S. Chaichulee, L. Tarassenko, D. Young, and P. Watkinson, "Non-invasive stroke volume estimation by transthoracic electrical bioimpedance versus Doppler echocardiography in healthy volunteers," Journal of Medical Engineering & Technology, vol. 43, pp. 33-37, 2019.

Conference papers

- S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, G. Green, K. McCormick, A. Zisserman and L. Tarassenko, "Localised photoplethysmography imaging for heart rate estimation of preterm infants in the clinic," in Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics, vol. 10501, 2018. (Oral presentation)
- J. Jorge, M. ViIllarroel, S. Chaichulee, K. McCormick, and L. Tarassenko, "Data fusion for improved camera-based detection of respiration in neonates," in Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics, vol. 10501, 2018. (Oral presentation)
- S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, G. Green, K. McCormick, A. Zisserman and L. Tarassenko, "Multi-task convolutional neural network for patient detection and skin segmentation in continuous non-contact vital sign monitoring," in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 266–272, 2017. (Oral presentation)
- J. Jorge, M. Villarroel, S. Chaichulee, K. McCormick, and L. Tarassenko, "Non-contact monitoring of respiration in the neonatal intensive care unit," in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 286–293, 2017. (Oral presentation)

Table of contents

List of figures xix

List of tables xxiii

List of abbreviations xxv

1 Introduction 1

1.1 Non-contact monitoring of pre-term infants . . . 1
1.2 Motivation . . . 2
1.3 Objectives . . . 3
1.4 Contributions . . . 3
1.5 Thesis outline . . . 4

2 Background 5

2.1 Introduction . . . 5
2.2 Physiological background . . . 6
2.2.1 Respiratory and circulatory systems . . . 6
2.2.2 Skin anatomy . . . 8
2.2.3 Vital signs . . . 9
2.2.4 Factors influencing vital signs . . . 10
2.3 Pulse oximetry . . . 11
2.3.1 Skin-light interactions . . . 11
2.3.2 Principles of pulse oximetry . . . 13
2.3.3 Photoplethysmography (PPG) . . . 14
2.3.4 Mode of operation . . . 15
2.3.5 Applications . . . 16
2.3.6 Limitations . . . 16
2.4 Vital sign monitoring using video cameras . . . 17
2.4.1 Photoplethysmographic imaging (PPGi) . . . 17
2.4.2 Imaging devices . . . 18
2.4.3 Vital sign estimation . . . 19
2.4.4 Limitations . . . 21
2.5 Conclusion . . . 22

3 Neonatal HDU Study 23

3.1 Introduction . . . 23
3.2 Current physiological monitoring of pre-term infants . . . 24
3.3 Clinical study design and protocol . . . 25
3.3.1 Research ethics . . . 26

3.3.2 Study participants . . . 26
3.3.3 Informed consent . . . 27
3.3.4 Participant confidentiality . . . 27
3.3.5 Data handling and record keeping . . . 27
3.4 Instrumentation . . . 28
3.4.1 Hardware for data acquisition . . . 28
3.4.2 Software for data acquisition . . . 32
3.5 Cohort description . . . 32
3.5.1 Participant demographics . . . 33
3.5.2 Recording statistics . . . 35
3.5.3 Reference vital signs . . . 35
3.6 Partitioning the dataset . . . 39
3.7 Conclusion . . . 40

4 Patient detection and skin segmentation **41**
4.1 Introduction . . . 41
4.2 Background . . . 42
4.2.1 Image classification . . . 43
4.2.2 Image segmentation . . . 44
4.3 Method overview . . . 45
4.4 Network architecture . . . 46
4.4.1 Shared core network . . . 46
4.4.2 Patient detection branch . . . 49
4.4.3 Skin segmentation branch . . . 50
4.5 Network training . . . 51
4.5.1 Dataset . . . 51
4.5.2 Data preprocessing . . . 55
4.5.3 Data augmentation . . . 55
4.5.4 Loss functions . . . 57
4.5.5 Model initialisation . . . 58
4.5.6 Training procedures . . . 58
4.6 Evaluation protocol . . . 59
4.6.1 Evaluation metrics . . . 59
4.7 Results . . . 60
4.7.1 Baseline results . . . 60
4.7.2 CNN results . . . 61
4.7.3 Analysis of design choices . . . 63
4.8 Discussion . . . 64
4.8.1 Qualitative assessment . . . 64
4.8.2 Network training . . . 64
4.8.3 Effects of multi-task learning . . . 65
4.8.4 Effects of data augmentation . . . 65

4.8.5 Runtime evaluation . . . 66
4.9 Conclusion . . . 66

5 Intervention detection **67**
5.1 Introduction . . . 67
5.2 Clinical interventions in the Neonatal HDU . . . 68
5.3 Method overview . . . 69
5.4 Optical flow network . . . 70
5.4.1 Sliding window operation . . . 70
5.4.2 Optical flow configuration . . . 71
5.4.3 Model architecture . . . 71
5.5 Fusion strategies . . . 72
5.5.1 Spatio-temporal fusion . . . 72
5.5.2 Multi-resolution temporal fusion . . . 74
5.5.3 Temporal context fusion . . . 76
5.6 Network training . . . 77
5.6.1 Dataset . . . 77
5.6.2 Data preprocessing . . . 79
5.6.3 Data augmentation . . . 81
5.6.4 Training procedures . . . 81
5.7 Evaluation protocol . . . 81
5.8 Results . . . 82
5.8.1 Baseline implementation . . . 82
5.8.2 Sliding window comparison . . . 83
5.8.3 Fusion strategy comparison . . . 83
5.9 Discussion . . . 85
5.9.1 Sliding window configurations . . . 85
5.9.2 Fusion strategies . . . 85
5.9.3 Failure modes . . . 86
5.9.4 Runtime evaluation . . . 87
5.10 Conclusion . . . 88

6 Reference heart rate **89**
6.1 Introduction . . . 89
6.2 Method overview . . . 89
6.3 Heart rate monitoring in the Neonatal HDU study . . . 90
6.4 Quality assessment of the ECG signal . . . 94
6.4.1 The ECG waveform . . . 95
6.4.2 Noise suppression . . . 95
6.4.3 QRS detection . . . 96
6.4.4 Signal quality assessment . . . 98
6.4.5 Heart rate estimation . . . 101

6.5 Computing the reference heart rate . . . 104
6.6 Results . . . 106
6.7 Discussion . . . 108
6.8 Conclusion . . . 110

7 Heart rate estimation from the video camera 111
7.1 Introduction . . . 111
7.2 Overview of the heart rate estimation process . . . 112
7.3 PPGi signal extraction . . . 113
7.3.1 Automatic PPGi acquisition . . . 113
7.3.2 Band-pass filtering . . . 116
7.3.3 Peak and onset detection . . . 116
7.4 Signal quality assessment . . . 118
7.4.1 Activity index . . . 119
7.4.2 Change point detection . . . 120
7.4.3 Frequency bounding . . . 121
7.4.4 Clipping detection . . . 122
7.4.5 Amplitude thresholding . . . 122
7.4.6 Multi-scale dynamic time warping . . . 123
7.4.7 Beat SQI . . . 126
7.5 Heart rate estimation methods . . . 128
7.5.1 Beat counting . . . 128
7.5.2 Fast Fourier Transform . . . 129
7.5.3 Dominant pole of an Autoregressive model . . . 130
7.5.4 Autoregressive modelling with the best model . . . 131
7.5.5 Kalman filtering . . . 131
7.6 Results . . . 133
7.6.1 Comparison of heart rate estimation methods . . . 134
7.6.2 Overall results . . . 135
7.6.3 Results by recording session . . . 135
7.6.4 Results by ethnicity . . . 141
7.7 Discussion . . . 142
7.7.1 Automated heart rate estimation process . . . 142
7.7.2 Differences between the training set and the test set . . . 142
7.7.3 Comparison with conventional monitoring . . . 142
7.7.4 Comparison to other clinical studies . . . 143
7.7.5 Continuous heart rate estimation . . . 143
7.7.6 Kalman filtering . . . 144
7.7.7 Effects of skin colour and body hair texture . . . 144
7.7.8 Challenges for heart rate estimation in the clinic . . . 145
7.8 Conclusion . . . 147

8 Reference respiratory rate 149

8.1 Introduction . . . 149
8.2 Overview of the estimation process . . . 150
8.3 Impedance Pneumography (IP) . . . 151
8.4 Respiratory signal extraction from ECG . . . 152
8.4.1 ECG-derived respiration (EDR) . . . 153
8.4.2 Respiratory sinus arrhythmia (RSA) . . . 153
8.4.3 R-peak amplitude (RPA) . . . 153
8.5 Signal filtering . . . 154
8.6 Peak and onset detection . . . 154
8.7 Signal quality assessment . . . 155
8.7.1 Physiological bounding . . . 155
8.7.2 Spectral density distribution . . . 155
8.7.3 Spectral Purity Index (SPI) . . . 156
8.7.4 Combining signal quality measures . . . 157
8.8 Estimation of reference respiratory rate . . . 157
8.8.1 Breath counting . . . 159
8.8.2 Autoregressive model . . . 159
8.8.3 Data fusion . . . 160
8.8.4 Data agreement . . . 161
8.8.5 Computing the reference respiratory rate . . . 163
8.9 Results . . . 163
8.10 Discussion . . . 164
8.10.1 Noise in respiratory signals . . . 164
8.10.2 Comparison to other relevant studies . . . 166
8.11 Conclusion . . . 167

9 Respiratory rate estimation from the video camera 169
9.1 Introduction . . . 169
9.2 Overview of the estimation process . . . 170
9.3 Respiratory signal extraction . . . 172
9.3.1 Photoplethysmography imaging (PPGi) . . . 173
9.3.2 Skin region properties . . . 173
9.3.3 Elliptical fitting . . . 177
9.3.4 Summary . . . 180
9.4 Signal filtering . . . 180
9.5 Peak and onset detection . . . 182
9.6 Signal quality assessment . . . 184
9.6.1 Peak agreement . . . 185
9.6.2 Multi-scale dynamic time warping . . . 186
9.6.3 Combining signal quality measures . . . 188
9.7 Respiratory rate estimation methods . . . 189
9.7.1 Estimating respiratory rate . . . 189

9.7.2 Data fusion . . . 190
9.8 Results . . . 190
9.8.1 Overall results . . . 191
9.8.2 Results by recording sessions . . . 198
9.9 Discussion . . . 198
9.9.1 Respiratory rate estimation process . . . 198
9.9.2 Comparison to conventional monitoring . . . 199
9.9.3 Comparison with other clinical studies . . . 199
9.9.4 Data fusion techniques . . . 199
9.9.5 Effects of different body postures . . . 200
9.9.6 Gaps in respiratory rate estimates . . . 200
9.9.7 Heart rate and respiratory rate estimation . . . 201
9.9.8 Challenges in respiratory rate estimation in the clinic . . . 201
9.10 Conclusion . . . 203

10 Towards pain assessment using video cameras **205**
10.1 Introduction . . . 205
10.2 Pain assessment in pre-term infants . . . 206
10.2.1 Pain perception in infants . . . 206
10.2.2 Heel prick procedure . . . 207
10.2.3 Physiological and behavioural indicators . . . 207
10.2.4 Pain assessment tools . . . 209
10.3 Dataset . . . 210
10.3.1 Dataset demographics . . . 210
10.3.2 Heel prick phases . . . 211
10.4 Vital sign estimation . . . 212
10.4.1 Reference vital signs . . . 212
10.4.2 Extraction of the PPGi signals . . . 212
10.4.3 Estimation of vital signs . . . 213
10.5 Results . . . 214
10.5.1 Vital sign estimation during the heel prick procedure . . . 214
10.5.2 Physiological responses across heel prick phases . . . 216
10.6 Discussion . . . 218
10.6.1 Non-contact vital sign estimation . . . 218
10.6.2 Physiological responses . . . 218
10.6.3 Integration with behavioural responses . . . 219
10.6.4 Study limitations . . . 220
10.6.5 Recommendations for future research . . . 220
10.7 Conclusion . . . 220

11 Conclusion **221**
11.1 Introduction . . . 221

11.2 Thesis review . . . 221
11.2.1 Clinical study . . . 221
11.2.2 Patient detection and skin segmentation . . . 222
11.2.3 Intervention detection . . . 222
11.2.4 Reference heart rate . . . 223
11.2.5 Heart rate estimation from a video camera . . . 223
11.2.6 Reference respiratory rate . . . 224
11.2.7 Respiratory rate estimation from a video camera . . . 224
11.2.8 Preliminary study of pain assessment . . . 224
11.3 Future work . . . 225
11.3.1 Improving the selection of ROIs using deep learning techniques . . . 225
11.3.2 Clinical activity detection using deep learning techniques . . . 225
11.3.3 Pain assessment . . . 225
11.3.4 Monitoring of neonatal resuscitation . . . 226
11.3.5 Monitoring of newborns in the first 72 hours of life . . . 226
11.4 Concluding remarks . . . 226

Appendices

A Convolutional neural network **227**
A.1 Building blocks . . . 227
A.2 Network training . . . 230

B Interactive image segmentation **231**
B.1 Graph cuts with geodesic star convexity . . . 231
B.2 Gaussian mixture models . . . 234

C Optical flow estimation **235**
C.1 Variational model . . . 235
C.2 Energy minimisation framework . . . 236

References **237**

List of figures

2.1 Diagram of the circulatory system . . . 7
2.2 Diagram of the skin structure . . . 8
2.3 Extinction spectra for materials presented in the skin . . . 12
2.4 Extinction spectra of oxyhaemoglobin and deoxyhaemoglobin . . . 13
2.5 The PPG signal component . . . 14
2.6 Two operational modes of pulse oximetry . . . 15
2.7 Diagram of skin-light interaction . . . 18
2.8 Amplification of subtle colour changes in the human skin . . . 18

3.1 Conventional vital sign monitoring in the Neonatal HDU . . . 24
3.2 Data acquisition setup . . . 26
3.3 The JAI AT-200CL 3-CCD camera . . . 29
3.4 Example images from the video camera . . . 30
3.5 Patient monitoring devices . . . 31
3.6 Distribution of gestational ages and infant weights for the entire dataset . . . 33
3.7 Timeline plot for recording times of each recording session in the dataset . . . 34
3.9 Distributions of vital signs for the entire dataset . . . 36
3.10 4-hour sample time series collected from a typical recording session . . . 37
3.11 Sample waveforms collected from a typical recording session . . . 38
3.12 Distribution of population demographics in the training and test sets . . . 39

4.1 Concepts of neural networks . . . 42
4.2 Example of a CNN model with two convolutional layers . . . 42
4.3 Hierarchical feature representations in CNNs . . . 43
4.4 The VGG-16 network . . . 44
4.5 Overview of the workflow for non-contact vital sign estimation . . . 45
4.6 Outline of the proposed multi-task CNN model . . . 46
4.7 Graphical representations of layers used in the VGG-16 network . . . 47
4.8 The proposed multi-task CNN model . . . 49
4.9 Downsampling of the input image . . . 49
4.10 Graphical representation of the convolutional transpose layer . . . 50
4.11 Pipeline for semi-automatic skin annotation . . . 52
4.12 Semi-automated image annotation tool . . . 53
4.13 Skin annotation results . . . 54
4.14 Example of lighting augmentation . . . 56
4.15 Patient detection results . . . 62
4.16 Examples of skin segmentation results . . . 62
4.17 Training loss for each task in the first and second training stages . . . 65

4.18 Example results from the networks trained with and without data augmentation . . 66

5.1 Examples of typical clinical interventions in the Neonatal HDU 69
5.2 Outline of the proposed CNN model 69
5.3 Example of optic flow calculation 70
5.4 Diagram of the processing of the input to the optical flow network 71
5.5 The optical flow network 72
5.6 Spatio-temporal fusion 73
5.7 Graphical representation of the concatenation operation 74
5.8 Outline of two different approaches for cropping video frames 75
5.9 Multi-resolution temporal fusion 75
5.10 Temporal context fusion 77
5.11 Intervention annotation tool 78
5.12 Intervention annotation results 80
5.13 ROC curves for the optical flow network with different sliding window sizes 83
5.14 ROC curves for the fusion networks with different fusion strategies 84
5.15 Examples of false negative results 87

6.1 Diagram for the computation of the reference heart rate 90
6.2 Examples of heart rate signals provided by the Philips patient monitor 91
6.3 Examples of discrepancies between the two heart rate values 92
6.4 Comparison between the ECG and PPG heart rates 93
6.5 Diagram for the estimation of heart rate from the ECG signal 94
6.6 Components of the ECG waveform 94
6.7 Noise suppression for an ECG signal 95
6.8 Diagram of the Pan-Tompkins QRS detection algorithm 96
6.9 Diagram of the length-transform-based QRS detection algorithm 98
6.10 Signal quality index computation for 30-second ECG waveforms 100
6.11 Heart rate estimation for a 5-minute ECG waveform 102
6.12 Comparison between the ECG heart rate from the Philips monitor and the heart rate estimates from the ECG signal 103
6.13 Agreement between the ECG heart rate from the Philips monitor and the heart rate estimates from the ECG signal 103
6.14 Time alignment between two heart rate signals 105
6.15 Results of the time alignment for the entire dataset 105
6.16 Agreement between the ECG heart rate and PPG heart rate from the Philips monitor 106
6.17 Relationship between the reference signals and image processing results 109

7.1 Diagram for the estimation of heart rate from the camera 112
7.2 Extraction of PPGi signals from the skin area 114
7.3 Raw PPGi signals extracted from the green channel with five different ROI sizes . . 115
7.4 Digital signal filtering of an PPGi waveform 116
7.5 Peak and onset detection for a PPGi waveform 117

7.6 Diagram for the signal quality assessment process . . . 118
7.7 Computation of the activity index for a typical 2-hour video segment . . . 119
7.8 Change point detection for a 60-second PPGi waveform . . . 120
7.9 Frequency bounding detection for a 60-second PPGi waveform . . . 121
7.10 Clipping detection for a 60-second PPGi waveform . . . 122
7.11 Amplitude thresholding for a 15-second PPGi waveform . . . 123
7.12 Template construction for multi-scale dynamic time warping . . . 124
7.13 Local cost matrix computation for dynamic time warping . . . 125
7.14 Computation of multi-scale dynamic time warping . . . 126
7.15 Signal quality assessment for noisy PPGi signals . . . 127
7.16 Signal quality assessment for a clean PPGi signal . . . 128
7.17 Heart rate estimation for a clean PPGi signal using the beat counting method . . . 129
7.18 Heart rate estimation for a clean PPGi signal using the FFT method . . . 129
7.19 Heart rate estimation for a clean PPGi signal by the autoregressive model method . . . 130
7.20 Example of heart rate estimates before and after applying Kalman filtering . . . 133
7.21 Heart rate estimation for a 60-minute video segment . . . 134
7.22 Comparison between the reference and camera heart rates for the training set . . . 136
7.23 Comparison between the reference and camera heart rates for the test set . . . 136
7.24 Examples of skin colour of non-white subjects . . . 141
7.25 Gaps in heart rate estimation . . . 144
7.26 Contribution of patient activities to errors in estimating heart rate . . . 145

8.1 Diagram for the computation of the reference respiratory rate . . . 150
8.2 Examples of IP waveforms extracted from typical recording sessions . . . 151
8.3 Respiratory signal extraction for a 60-second ECG waveform . . . 152
8.4 Respiratory signal extraction using the EDR method . . . 153
8.5 Digital signal filtering of an IP waveform . . . 154
8.6 Peak and onset detection for an IP waveform . . . 155
8.7 Computation of spectral density distribution for an IP waveform . . . 156
8.8 Signal quality assessment for a clean IP waveform . . . 158
8.9 Signal quality assessment for a noisy EDR waveform . . . 158
8.10 Respiratory rate estimation for a sample 5-minute IP signal using the breath counting and autoregressive model methods . . . 159
8.11 Diagram for data fusion using multiple Kalman filters . . . 160
8.12 Example of respiratory rate estimates using the data fusion method . . . 161
8.13 Comparison between the respiratory rates from the breath counting method and autoregressive model method for the entire dataset . . . 162
8.14 Agreement between the respiratory rates from the breath counting method and the autoregressive model method for the entire dataset . . . 164
8.15 Distribution of the Philips respiratory rate and the reference respiratory rate . . . 166
8.16 Distribution of respiratory rate estimates from different respiratory signals . . . 167

9.1 Diagram for the estimation of respiratory rate . . . 171

9.2 Overview of the respiratory signal extraction process . . . 172
9.3 Examples of PPGi signals extracted from a quiet video segment . . . 174
9.4 Examples of PPGi signals extracted during different body postures . . . 174
9.5 Examples of skin labels provided by the deep learning algorithm . . . 175
9.6 Extraction of the boundary of skin regions . . . 176
9.7 Respiratory signal extraction using the shape properties of skin regions . . . 176
9.8 Definition of an ellipse in Cartesian coordinates . . . 177
9.9 Examples of an ellipse fitted to the skin region . . . 179
9.10 Respiratory signal extraction using the ellipse fitting method . . . 180
9.11 Twelve respiratory signals extracted from a sample video segment . . . 181
9.12 Diginal signal filtering of all twelve respiratory waveforms . . . 182
9.13 Peak and onset detection for a 30-second clean respiratory waveform . . . 183
9.14 Peak and onset detection for a 30-second noisy respiratory waveform . . . 183
9.15 Examples of respiratory waveforms . . . 184
9.16 Peak agreement for a 30-second respiratory signal . . . 185
9.17 Template construction for multi-scale dynamic time warping . . . 186
9.18 Multi-scale dynamic time warping for a 30-second respiratory signal . . . 187
9.19 Signal quality assessment for a clean respiratory signal . . . 188
9.20 Respiratory rate estimation for a 60-second PPGi signal . . . 189
9.21 Data fusion of respiratory rate estimates for a 5-minute video segment . . . 190
9.22 Respiratory rate estimation for a 60-minute quiet video segment . . . 191
9.23 Comparison between the reference and camera respiratory rates for the training set 192
9.24 Comparison between the reference and camera respiratory rates for the test set . . . 192
9.25 Relationship between the mean absolute error and the estimated time . . . 193
9.26 Contribution of each respiratory rate estimate to the fusion respiratory rate . . . 200
9.27 Respiratory rate estimation for a 2-hour video segment during which the patient's body posture is changed from a prone to supine position . . . 200
9.28 Gaps in respiratory rate estimation . . . 201
9.29 Estimation of heart rate and respiratory rate for a 90-minute video segment . . . 202

10.1 Neurodevelopment of pain perception pathways . . . 206
10.2 Phases of the heel prick procedure . . . 211
10.3 Regions of interest for estimating heart rate and respiratory rate . . . 213
10.4 Estimation of heart rate and respiratory rate for the session HP1 . . . 214
10.5 Estimation of heart rate and respiratory rate for the session HP2 . . . 214
10.6 Estimation of heart rate and respiratory rate for the session HP3 . . . 215
10.7 Estimation of heart rate and respiratory rate for the session HP4 . . . 215
10.8 Maximum heart rate and respiratory rate for different heel prick phases . . . 217
10.9 Mean heart rate and respiratory rate for different heel prick phases . . . 217
10.10 Standard deviation of heart rate and respiratory rate for different heel prick phases . 217

B.1 Example of skin segmentation using graph cuts . . . 231
B.2 Geodesic star convexity . . . 232

List of tables

3.1 List of physiological parameters recorded during a typical study session 31
3.2 Summary of population demographics 33
3.3 Statistics for the video camera and reference recording times 35
3.4 Statistics for the reference vital-sign data recorded by the Philips monitor 36
3.5 Summary of population demographics in the training and test sets 39

4.1 Summary of population demographics in the D_1 and D_2 groups in the training set . 59
4.2 Patient detection performance of the baseline skin filters and the CNN models . . . 61
4.3 Skin segmentation performance of the baseline skin filters and the CNN models . . . 61
4.4 Skin segmentation performance on different network variants 63

5.1 Daily nursing activities for pre-term infants in the Neonatal HDU 68
5.2 Performance of the baseline two-stream architecture 82
5.3 Performance of the optical flow network with different sliding window sizes 83
5.4 Performance of the local optical flow network and context network 84
5.5 Performance of the fusion networks with different fusion approaches 84

6.1 Comparison between the ECG heart rate provided by the Philips monitor and the heart rate re-estimated from the ECG signal 104
6.2 Comparison between the ECG heart rate and PPG heart rate provided by the Philips monitor 106
6.3 Valid reference heart rate data for each recording session 107

7.1 Summary of heart rate estimation results for both training and test sets 135
7.2 Heart rate estimation results for each recording session in the training set (1/2) . . . 137
7.3 Heart rate estimation results for each recording session in the training set (2/2) . . . 138
7.4 Heart rate estimation results for each recording session in the test set (1/2) 139
7.5 Heart rate estimation results for each recording session in the test set (2/2) 140
7.6 Heart rate estimation results by ethnicity 141

8.1 Comparison of respiratory rates from the patient monitor, the breath counting method and the autoregressive model method for the entire dataset 162
8.2 Agreement between the respiratory rates from the breath counting method and the autoregressive model method for the entire dataset 163
8.3 Valid reference respiratory rate data for each recording session 165

9.1 Summary of respiratory rate estimation results for both training and test sets 193
9.2 Respiratory rate estimation results for each recording session in the training set (1/2) 194
9.3 Respiratory rate estimation results for each recording session in the training set (2/2) 195

9.4 Respiratory rate estimation results for each recording session in the test set (1/2) . . 196
9.5 Respiratory rate estimation results for each recording session in the test set (2/2) . . 197

10.1 Physiological and behavioural indicators of pain in pre-term infants 208
10.2 Premature Infant Pain Profile (PIPP) 210
10.3 Neonatal Infant Pain Scale (NIPS) 210
10.4 Patient demographics of each heel prick session 211
10.5 Physiological changes for the different phases of the heel prick procedure 216

List of abbreviations

AGC . . . . . Automatic gain control

AAMI . . . . . Association of the Advancement of Medical Instrumentation

ANSI . . . . . American National Standards Institute

API . . . . . Application programming interface

AR . . . . . Autoregressive model

AUC . . . . . The area under the receiver operating characteristic curve

BRC . . . . . Biomedical Research Centre

BSSF . . . . . Boxed slope sum function

CCD . . . . . Charged-coupled device

CRF . . . . . Case report form

CNN . . . . . Convolutional Neural Network

CMOS . . . . . Complementary metal-oxide semiconductor

DTW . . . . . Dynamic time warping

ECG . . . . . Electrocardiogram

FFT . . . . . Fast Fourier transform

FIR . . . . . Finite impulse response

FNR . . . . . False negative rate

FPGA . . . . . Field programmable gate array

FPR . . . . . False positive rate

GSC . . . . . Geodesic star convexity

HDU . . . . . High dependency unit

HR . . . . . Heart rate

HSL . . . . . Hue, saturation and lightness

ICU . . . . . Intensive care unit

IIR . . . . . Infinite impulse response

IP . . . . . Impedance pneumography

IQR . . . . . Interquartile range

LED . . . . . Light-emittng diode

LNCS . . . . . Low-noise cabled sensors

MAC . . . . . Mean average curve

MAD . . . . . Mean absolute deviation

MAE . . . . . Mean absolute error

MREC . . . . . Medical Research Ethics Committee

NDAU . . . . . Neonatal data analysis unit

NHS . . . . . National Health Service

NICU . . . . . Neonatal intensive care unit

NTP . . . . . Network time protocol

PPG . . . . . Photoplethysmography

PPGi . . . . . Photoplethysmographic imaging

ReLU . . . . . Rectified linear unit

RGB . . . . . Red, green and blue

ROC . . . . . Receiver operating characteristic

ROI . . . . . Region of interest

RR . . . . . Respiratory rate

SpO_2 . . . . . Peripheral oxygen saturation

SPI . . . . . Spectral purity index

SSF . . . . . Slope sum function

SQI . . . . . Signal quality index

TPR . . . . . True positive rate

TNR . . . . . True negative rate

Chapter 1
Introduction

1.1 Non-contact monitoring of pre-term infants

Vital signs such as heart rate, respiratory rate and peripheral blood oxygen saturation are essential parameters for assessing the health and well-being of pre-term infants, born before 37 weeks of gestation. Pre-term infants are usually nursed in the Neonatal Intensive Care Unit (NICU) or in the High Dependency Unit (HDU) depending on the severity of their condition. They are surrounded by complex medical equipment to monitor their physiological status. Pre-term infants are unstable and have fluctuating vital signs [1]. Conventional vital sign monitoring technologies, using pulse oximetry or electrocardiography (ECG), require the attachment of adhesive electrodes or transducers to the skin surface. The skin of pre-term infants is fragile and very sensitive, especially for those born before 27 weeks of gestation when the bond between the attached sensor and dermis could be stronger than that between the dermis and epidermis [2]. The attachment of sensors may damage the skin and increase the risk of developing infections. In addition, the movement of the infant may cause epidermal stripping [3].

The advances in digital camera sensors and computational capabilities in the last two decades have opened up research in non-contact vital sign monitoring using video cameras [4–14]. Although different methods for non-contact monitoring of vital signs exist, such as radar [15–17] and laser [18, 19], camera-based technologies are ubiquitous, low-cost and capable of high performance. They can be integrated into a patient ward or a telemedicine system. In addition, they could be expanded to provide other bedside assessments such pain scores through the analysis of facial expressions and the movement of the patient. The use of non-contact monitoring technologies for monitoring pre-term infants can provide advantages over or in addition to conventional vital-sign monitoring techniques.

This thesis investigates, in depth, the application of continuous camera-based non-contact monitoring of heart rate and respiratory rate in pre-term infants nursed in the HDU in the hospital. This thesis also explores the possibility of assessing the variability of vital signs during painful events, such as heel pricks for drawing blood from the infant.

1.2 Motivation

Heart rate is a measure of the number of heart contractions per minute, as it pumps blood carrying oxygen and other nutrients throughout the body. It can vary widely according to the body's physical demands and the health status of the patient. Heart rate measurement is important in clinical settings. In pre-term infants, a very low or high heart rate can indicate an underlying condition such as infection, pain or illness [20].

In pre-term infants, heart rate is continuously monitored during the course of the hospitalisation. Heart rate is usually computed from the Electrocardiogram (ECG) or the Photoplethysmogram (PPG) signals. The ECG measures the electrical activity of the heart during each cardiac cycle through electrodes attached to the skin surface. The PPG measures changes in the light absorption characteristics of blood. The PPG signal is typically recorded by a pulse oximeter attached to the ear, finger or toe.

Respiratory rate is a measure of the frequency of respiration cycles per minute. Respiratory rate is an important vital sign yet difficult to measure through conventional techniques such as impedance pneumography (IP) [21]. IP is widely used in the NICU for measuring changes in the electrical impedance of the thorax during breathing. This technique is, however, subject to a high degree of noise as a result of patient movement [22]. In a general ward, clinical staff often count the number of chest wall movements manually to ensure that a high-quality respiratory rate is obtained. Abnormal respiratory rate is often associated with hypoxaemia (low level of oxygen in the blood), hypercapnia (high level of carbon dioxide in the blood) or acidosis (high level of acidity in the blood) when the body attempts to maintain the oxygen supply [23]. Since the lungs of pre-term infants are underdeveloped, the early detection of these symptoms is very important to provide a prompt clinical intervention.

Continuous non-contact vital sign monitoring has a number of advantages over conventional contact monitoring methods and the intermittent manual measurements made by clinical staff. In non-contact monitoring, for example, heart rate can be measured by the analysis of subtle colour changes on the skin surface recorded by the video camera (see chapter 7). Respiratory rate can be measured by the analysis of the movement of the body or the torso (see chapter 9). Non-contact methods can make it possible to measure the physiological parameters in some scenarios for which it is difficult to employ conventional technologies such as patients with burn injuries, patients with fragile skin or patients in an isolation unit (a ward designated to isolate patients for the purpose of infection control). Non-contact methods also remove the risk of sensors falling off and of infection from traditional contact sensors. Non-contact technology is easy to deploy and could be used to monitor pre-term infants during hospitalisation in the NICU or those waiting to be discharged from the hospital.

Apart from the monitoring of vital signs, a video camera could also be used to monitor an infant's physical activity, distress and pain. Pain is an important measure in the clinic since it may raise

a concern about the infant's well-being and health status. Pain assessment is a routine medical procedure that is performed several times a day in pre-term infants. Current procedures for pain assessment are based on the subjective observations of changes in vital signs, behavioural indicators and the infant's state of arousal [24–27]. With the advance of non-contact vital sign monitoring and deep learning techniques, the video camera could be used to measure these parameters objectively and provide an assessment of the infant's physiology to the clinical staff.

It is crucial to continuously monitor the vital signs and maintain the health status of pre-term infants, especially for those infants with moderate or extreme prematurity. Comprehensive neonatal care can reduce the risk of complications, infections, mortality and long-term disabilities [1]. Non-contact monitoring using the video cameras has the potential to be expanded to detect other adverse clinical events such as apnoea (pauses in breathing) or bradycardia (low heart rate).

1.3 Objectives

The objectives of the thesis are:

- To create a video database for the development and calibration of non-contact algorithms for the neonatal population
- To investigate the use of deep learning techniques for non-contact vital sign estimation
- To create open-source software for labelling image and video data in order to generate ground truths for training deep learning algorithms
- To investigate the accuracy and the proportion of time that heart rate and respiratory rate can be estimated from pre-term infants using a video camera in an uncontrolled clinical environment
- To assess the feasibility of deriving physiological indicators of pain in pre-term infants.
- To investigate the variability of vital signs estimated from video camera data during painful medical procedures.

1.4 Contributions

Continuous non-contact estimation of vital signs in the neonatal intensive care environment poses several challenges such as: changes in the illumination characteristics during video recording; the determination of the presence or absence of a subject in the image frame; the detection of clinical interventions; the correct identification of the patient with respect to other people (parents, nurses, etc.); and the selection of suitable regions of interest (ROI) from which vital signs can be estimated.

The first contribution of this thesis is the development of multi-task deep learning techniques for patient detection, skin segmentation and clinical intervention detection. These three tasks are important as they provide the essential information which enables the automatic extraction of

vital signs from the video camera data in the clinic. This contribution, as a result, generates very large datasets for training the deep learning algorithms proposed in the thesis. All the datasets are annotated by three individuals in order to guarantee high-quality ground truth. These datasets could be used for extending and improving the algorithms in the future. Two open-source software codes are also developed: the first software code is for labelling the skin regions of the infant in images; and the second software code is for annotating the time periods during which clinical intervention is present in videos.

The second and main contribution of this thesis is the development of non-contact algorithms for estimating heart rate and respiratory rate. This involves the extraction of cardiac and respiratory signals from the video camera data via deep learning algorithms and the development of robust techniques for estimating heart rate and respiratory rate in challenging clinical scenarios. The vital-sign estimation pipelines allow the continuous estimation of heart rate and respiratory rate to be performed automatically, without human supervision or intervention. These algorithms are developed and validated on more than 420 hours of data recorded from 30 pre-term infants nursed in the Neonatal HDU.

The final contribution of this thesis is the investigation of the variability of vital signs during a pain-inducing clinical procedure, extracted from the Neonatal HDU dataset. This investigation further explores the feasibility of developing non-contact algorithms for deriving objective pain scores in pre-term infants using video camera data.

1.5 Thesis outline

This thesis is composed of 11 chapters. Chapter 1 introduces the objectives of the research. Chapter 2 reviews the literature regarding human physiology, non-contact vital sign monitoring and deep learning techniques. Chapter 3 provides an overview of the clinical study, the data from which are used in this thesis and describes the study population as well as the reference vital sign data. Chapter 4 describes the development of a deep learning algorithm for patient detection and skin segmentation. Chapter 5 presents a deep learning algorithm for identifying periods of clinical interventions in the video camera data. Chapter 6 analyses the validity of the heart rate obtained from conventional monitoring and computes the gold-standard reference heart rate to be compared with the camera-derived heart rate. Chapter 7 details an algorithm for estimating heart rate from the video camera data. Chapter 8 describes an approach to compute the gold-standard reference respiratory rate. Chapter 9 proposes an algorithm for estimating respiratory rate from video camera data. Chapter 10 presents exploratory work on the assessment of the variability of vital signs during a pain-inducing event. Chapter 11 summarises the research and outlines future work.

Chapter 2
Background

2.1 Introduction

When the heart pumps blood into a series of blood vessels, the volume of blood in cutaneous capillaries increases. On the contrary, when blood flows back to the heart through the veins, the amount of blood in cutaneous capillaries decreases [9, 28–30]. Photoplethysmography (PPG) is a non-invasive optical technique used to measure blood volume changes by measuring the proportion of light that cutaneous microvascular beds absorb [28]. Over the past decades, PPG has been extensively studied and used for monitoring various physiological parameters such as heart rate, peripheral oxygen saturation (SpO_2), respiratory rate, blood pressure and cardiac output [28].

Photoplethysmography imaging (PPGi) is an emerging technique to monitor the vasculature of tissue remotely using a digital video camera [4, 6]. PPGi removes the requirements of contact sensors and cables, thus the risk of infections can also be eliminated. However, most studies in PPGi have only been carried out either in controlled environments or with a small number of participants. Research by the computer vision community has recently shown that convolutional neural networks (CNNs) are capable of performing various computer vision tasks (such as image classification, image segmentation, video classification and body part detection) with high accuracy. The application of CNNs may enable non-contact physiological monitoring to be performed continuously and automatically in uncontrolled clinical settings such as a neonatal intensive care environment (NICU).

This chapter aims to provide the necessary background on human physiology, non-contact vital sign monitoring and the computer vision techniques used throughout the thesis. This chapter is organised as follows. Section 2.2 introduces basic physiological background and the vital signs. Section 2.3 describes relevant knowledge regarding PPGi including light-skin interactions, imaging devices and the extraction of PPGi signals. This section also reviews the current state of research in non-contact monitoring of heart rate and respiratory rate. Concluding remarks are presented in section 2.5.

2.2 Physiological background

This section summarises the relevant background of the respiratory and circulatory systems, the measurement of vital signs, and the anatomy of the skin, which leads to an understanding of the fundamental concepts behind PPG and PPGi technologies. A detailed description of the factors that influence vital signs in pre-term infants is also provided.

2.2.1 Respiratory and circulatory systems

Cells need a continuous supply of oxygen and nutrients to generate energy and sustain life. The respiratory system primarily deals with the movement of air into and out of the lungs and the gas exchange between oxygen and carbon dioxide at the lungs [31].

Inspiration (or inhalation) is a process of transporting air from outside the body into the lungs. When the diaphragm contracts and moves downwards, the thoracic cavity increases and atmospheric pressure pushes the air through respiratory passages into the lungs. Expiration (or exhalation), on the contrary, is a passive process by which the diaphragm relaxes following inspiration, thus it moves upwards and the lungs return to their original shape. The air is pushed out of the body through the airways with no muscle action. The movement of the chest walls during these processes can be tracked visually, especially for pre-term infants whose torso visibly moves upwards and downwards during inspiration and expiration respectively [31, 32].

In the lungs, oxygen diffuses into blood from the alveolus through the alveolar capillary network (see figure 2.1) and then binds to haemoglobin forming oxyhaemoglobin. Carbon dioxide diffuses, in the opposite direction, from the blood to the alveolus. The diffusion process occurs until the partial pressure between the alveolus and blood is at equilibrium [31].

The path of blood circulating through the human body is closed (see figure 2.1). Oxygenated blood travels from the lungs and enters the heart through the pulmonary veins to be distributed to the body through the arteries. Blood flows from the arteries into arterioles which branch into capillaries, where blood pressure and velocity decrease and the total cross-sectional area of blood vessels increases. Body cells receive oxygen and nutrient supply through the systemic capillaries close to them. Blood flowing through the capillaries retrieves carbon dioxide from the body cells and transports it back to the heart via the veins to be pumped to the lungs via the pulmonary artery [31, 32].

In the systemic capillaries, oxygenated blood releases oxygen to the tissue cells in exchange for carbon dioxide via a diffusion process through a thin layer of capillary walls sitting between the blood and surrounding tissues [31].

Pre-term infants are born before their completed development of approximately 37 weeks. Every body function may be immature and not capable of functioning at the same level as full-term infants or adults. The respiratory system is one of the last systems of the human body to develop. Pre-term

Figure 2.1: Simplified diagram of the circulatory system. The heart receives oxygenated blood (shown in red) from the lungs and distributes it throughout the body through the arteries. Oxygenated blood travels through a series of capillaries to tissue cells to which oxygen and other nutrients are supplied. Deoxygenated blood (shown in blue) leaves the tissue cells and flows through the veins back to the heart and lungs. The exchange of oxygen and carbon dioxide occurs through a diffusion process at the alveolus through a network of alveolar capillaries. (Image reproduced from [31], courtesy of McGraw-Hill Education)

infants delivered before 28 to 30 weeks' gestation may not have fully-functional alveoli in their lungs and may breathe through their terminal bronchioles and primitive alveoli [1]. They may have several problems affecting breathing, leading to respiratory complications such as surfactant deficiency, immature respiratory control centres, unstable chest wall and small respiratory passage [20]. Pre-term infants need to maintain a high level of oxygen supply to their circulatory system and other organs in order to make a successful transition to life outside the mother's uterus [20]. Some infants may have difficulty in transitioning from intra-uterine circulation in which gas exchange happens in the placenta, to extra-uterine circulation in which gas exchange occurs in their lungs [1].

Pre-term infants are at risk of various cardiovascular disorders ranging from congenital heart defects to dysfunctional autoregulation of blood flow. The ductus arteriosus (a blood vessel in the fetus connecting the pulmonary artery to the aorta, bypassing the lungs) may not close properly

after birth allowing oxygenated blood to flow back to the lungs, resulting in a patent (open) ductus arteriosus (PDA) [1]. This incidence affects approximately 30 percent of pre-term infants, according to a survey conducted in the United States in 1996 [33]. Apnoea (pauses in breathing) and bradycardia (low heart rate) are also commmon as a result of an immature cardio-respiratory system [1].

2.2.2 Skin anatomy

The skin is the largest human organ forming a barrier between the body and the outside. The skin has two distinct layers: the epidermis and dermis (see figure 2.2). The thickness of each layer varies depending on different regions of the body and their function. The epidermis is an outer layer composed entirely of stratified squamous epithelium (a dense stack of epithelial or lining cells). The corneum stratum is the most superficial layer of the epidermis providing protection against heat loss, dehydration, infections and injury [34]. Melanocytes are cells which lie at the deepest level of the epidermis. They produce melanin that gives the skin its distintive colour. There are two primary types of melanin: phaeomelanin, providing yellow and red pigmentation; and eumelanin, providing brown and black pigmentation. Differences in skin colour results from the amount, size and distribution of melanin. The ratio between phaeomelanin and eumelanin, which is mainly determined by genetics, affects skin colour [31].

The dermis is a layer below the epidermis composed of dense connective tissue, providing toughness and elasticity together with muscle tissue and nervous tissue. The dermis contains small blood vessels that provide oxygen and other nutrients to skin cells. These vessels can also control body temperature though the dilation and constriction of blood flow near the skin surface [31].

Figure 2.2: Diagram of the skin structure. The skin has two layers, the epidermis and dermis, on top of the subcutaneous layer in which large blood vessels as well as lobules of fat are located. The epidermis is a thin layer that lacks blood vessels. (Image reproduced from [31], courtesy of McGraw-Hill Education)

The subcutaneous (or hypodermis) layer is located beneath the dermis. It consists mostly of continuous connective tissue, along with adipose tissue, which help to insulate the body and conserve temperature [31]. The subcutaneous layer contains major blood vessels supplying the skin layers above (see Figure 2.2).

The development of the epidermis is complete within 34 weeks of gestation [35]. Thus, pre-term infants are usually born with deficiency and immaturity of the skin, resulting in increased transepidermal water loss, transcutaneous heat loss, difficulty in maintaining autoregulation and an increased risk of infection [36]. The development of the skin in pre-term infants not only causes changes in its thickness and opacity but also changes in melanin production and distribution which, in turn, result in changes in the reflection of light from the skin [37]. In Caucasian infants, the appearance of skin colour changes over time and is routinely used by medical staff to determine the gestational age of the infant [38].

2.2.3 Vital signs

Vital signs are a group of physiological measurements which are fundamental for assessing the health status of a patient, determining a treatment plan and monitoring the recovery progress during hospitalisation [39]. The measurement of vital signs is the most common intervention performed in the hospital. Vital signs generally include body temperature, blood pressure, heart rate, respiratory rate and peripheral blood oxygen saturation (SpO_2) [40]. The measurement of vital signs is an essential skill set for health practitioners. For hospitalised patients, vital signs are monitored continuously in critical care environments and periodically in general wards [39]. Abnormalities in vital signs provide useful information that may inform or support a diagnosis or a treatment decision [41]. This thesis is interested in two important vital signs: heart rate and respiratory rate.

Heart rate

Heart rate is the rate of heart beats or contractions, typically reported over a minute. Heart rate is the most frequent vital sign used in the clinic and varies according to individuals and the physical needs of the body. Heart rate can be used to assess the patient's health status and the heart's mechanism. Normal resting heart rate varies between 120 and 180 beats/min in pre-term infants at 28^{th} week of gestation, between 90 and 160 beats/min in infants at one month of age and between 60 and 100 beats/min in adults [42–44].

In the clinic, the electrocardiogram (ECG) is the gold standard technique to measure heart rate using electrodes placed on the patient's skin. A pulse oximeter, which can measure heart rate through an optical sensor placed on the fingers or ear lobes [45], is often used in general practice and at home. Manual measurement can be performed by using a stethoscope or counting the number of beats at the wrist or neck. Manual measurements can contain a high degree of

error. In a clinical competency study, only 25% of clinical staff could accurately perform the manual measurement of heart rate in a paediatric care unit [46].

Respiratory rate

Respiratory rate measures the number of respiratory cycles or breaths per minute. Each respiratory cycle is composed of inspiration followed by expiration. Respiratory rate is an important vital sign and is often used as a marker of potentially adverse clinical events such as respiratory tract infections and cardiac arrests [23]. Normal respiratory rate ranges between 20 to 80 breaths/min in pre-term infants, between 30 to 60 breaths/min in infants at one month of age and between 15 to 20 breaths/min in adults [42, 43, 47].

Various techniques can be used to monitor respiratory rate such as: using a strain-gauge transducer wrapped around the chest to measure changes in the diameter of the thoracic cavity; nasal cannula or mouth piece to collect the air flowing out of the lungs [48, 49]; and impedance pneumography (IP). Impedance pneumography is a common technique used to monitor respiratory rate in clinical settings. This technique measures electrical impedance changes of the thorax caused by changes in the conductivity of the lungs during respiration. The measurement of electrical impedance requires the injection of high-frequency current [50], which can be done through the same electrodes as for the ECG, thus the need of additional sensors to be attached the patient is eliminated.

The measurement of respiratory rate can be affected by breathing patterns, body position, equipment position and motion artefacts [51]. The IP signal not only represents the impedance of the lungs but also the impedance of other surrounding tissues, e.g. the muscle and skin. In addition, the impedance amplitude differs significantly according to body position and electrode placement [52]. IP is prone to motion artefacts which lead to false breath estimation and false detections of clinical adverse events such as central sleep apnoea and cardiogenic artefacts [53]. The measurement of respiratory rate is challenging in patients with abnormal breathing patterns and may be affected by sporadic events such as sneezing, coughing, eating, drinking and swallowing [54].

Manual measurement of respiratory rate can be performed by counting the number of chest wall movements. The manual measurement is, however, perceived to be a time-consuming process and can lead to an error up to 34% with a high tendency of overestimation [55, 56]. In a clinical competency assessment, only 67% of clinical staff could perform the manual measurement of respiratory rate with an error of up to 10 breaths/min [46].

2.2.4 Factors influencing vital signs

Vital signs are influenced by various long-term or temporary factors. In pre-term infants, vital signs can signify both the progression of growth and development as well as the effectiveness of the care

provided. Vital signs can vary according to multiple factors, including: age, oxygen requirements, physical conditions, body size, body temperature, medication and heart or lung problems.

The physiologic range of heart rate and respiratory rate in pre-term infants is normally higher than those in adults. Due to the immaturity of the heart's muscle fibres, the infant's heart needs to pump faster in order to maintain blood flow and pressure throughout the body [31, 57]. Any disorders that affect the respiratory, circulatory and renal systems may result in abnormal heart rate or respiratory rate. Pain is also an important factor to which the body responds by increasing heart rate and respiratory rate as a result of an increase in oxygen consumption of the tissues and muscles [31]. Medications, such as morphine, affect the actions of the heart and the lungs and have an influence on vital signs [58]. Infections, such as lower respiratory tract infections or sepsis, can also cause changes in vital signs as the body is fighting the diesease [58].

2.3 Pulse oximetry

This section introduces the relevant knowledge of skin-light interactions and the principles of pulse oximetry which forms the basis of the PPGi technology.

2.3.1 Skin-light interactions

Visible light (or the visible spectrum) is part of the electromagnetic spectrum in the wavelength range between 400 to 780 nm that is visible to the human eye (see figure 2.3a). The interaction between light and matter primarily includes: refraction, which is the bending of light rays after passing into a medium; reflection, which is the return of light rays to the originating medium occurring either at a smooth surface (specular reflection) or an inhomogeneous surface (diffuse reflection); scattering, which occurs when light penetrates into and scatters within a medium in multiple directions; and absorption, which occurs in the interior of a medium when the energy of photons is absorbed by the medium [59, 60].

The Beer-Lambert law describes the attenuation of light travelling into a light-absorbing medium. When a beam of light I_0 at a wavelength λ enters an absorbing medium, some part of the light is transmitted and the other part of the light is absorbed [61]. The light intensity I passing though the medium decreases with respect to its travelling distance and the medium's absorption characteristics, as defined by:

$$I(\lambda) = I_0(\lambda) \cdot \exp\{-\mu(\lambda)L\} \tag{2.1}$$

where $\mu(\lambda)$ is the absorption coefficient of the absorbing medium at the wavelength λ; L is the optical path length for the light travelling through the medium [62]. The extinction coefficient indicates the amount of light passing through the medium.

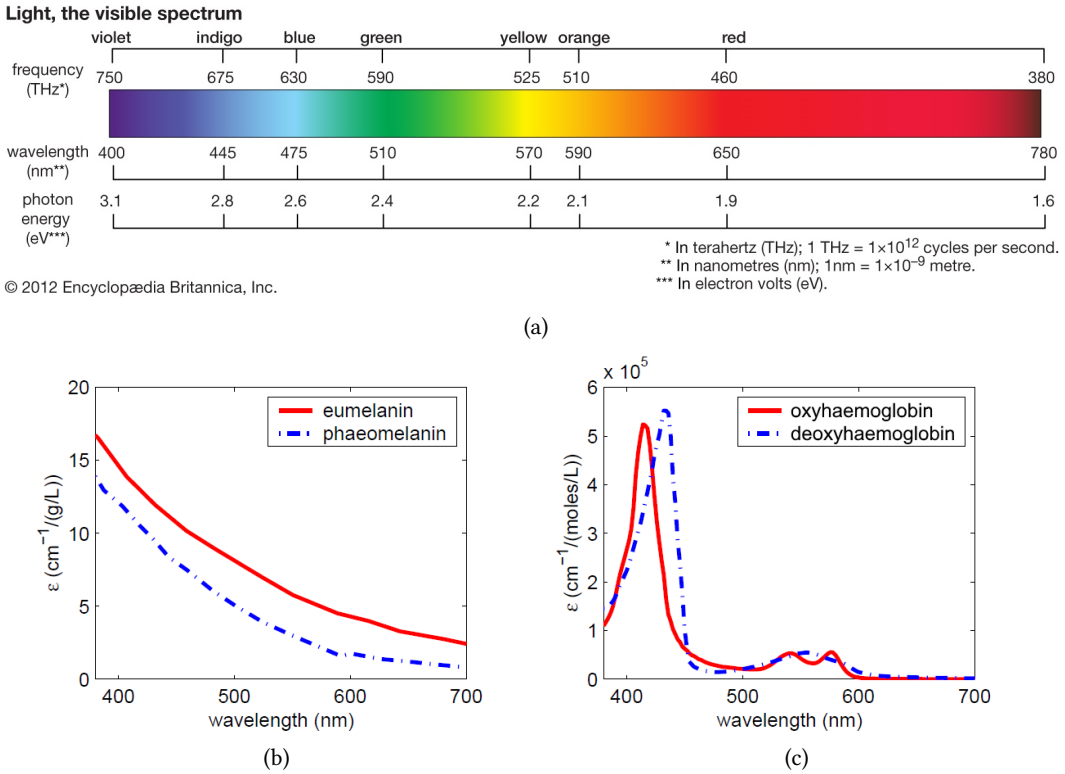


Figure 2.3: Extinction spectra for materials presented in the skin. (a) Visible spectrum. (Image courtesy of Encyclopædia Britannica, Inc.) (b) Extinction spectrum of eumelanin and phaeomelanin at different wavelengths. (c) Extinction spectrum of oxyhaemonglobin and deoxyhaemonglobin at different wavelengths. (Images reproduced from [63], courtesy of S. Prahl and the Oregon Medical Laser Center)

When light reaches the skin, its biological characteristics affect its propagation and absorption [63]. The epidermis reflects, propagates and absorbs light. About 5-7% of the incoming light is reflected back from the stratum corneum to the environment [63]. The absorbance property of the epidermis primarily depends on skin pigments or melanin. There are two types of melanin in human skin: the brownish black eumelanin and the reddish yellow phaeomelanin. The proportion of phaeomelanin and eumelanin is mostly associated with skin colour and skin types [64]. The absorption spectra of phaeomelanin and eumelanin are broad and higher at shorter wavelengths (400 nm) of the visible spectrum (see figure 2.3b). The absorbance level of the epidermis depends on the concentration of melanosomes, which are membranous particles that encapsulate melanin [63]. The epidermis' absorption property varies among individuals and even between different body parts.

The dermis also reflects, propagates and absorbs light. The dermis contains primarily connective tissues with blood vessels and nerves. Blood is distributed homogeneously in the dermis. The proportion of blood in the tissues varies from $0.2 - 7\%$, and the fluence rate of light (the total number of photons divided by the area) decreases exponentially as the depth of the dermis increases [65]. Oxygenated and deoxygenated haemoglobin have different absorption characteristics (see figure 2.3c). Approximately $90 - 95\%$ of haemoglobin in the arteries is oxygenated, while the amount of oxygenated haemoglobin is generally lower in the veins as a result of uptake by the cells [66].

The absorption properties of these haemoglobins are used for the measurement of the oxygen level in the blood by a pulse oximeter. The subcutaneous or hypodermis layer contains adipose tissues which predominantly prevent the absorption of light in the visible spectrum. Most of the incident light is reflected back to the layers above [63].

In pre-term infants, skin reflectance in visible light, measured using reflectance spectrometry, increases as gestational age and the skin matures [67]. Skin reflectance depends on the developmental stage and morphological pattern of the skin. It is different between male and female infants and also varies according to skin colour [34, 67]. Skin absorption increases in the first postnatal days, due to a poorly-developed epidermis and a thin stratum corneum, and declines as gestational age and skin maturity increase [68].

2.3.2 Principles of pulse oximetry

Pulse oximetry is an optical technique used to measure peripheral arterial oxygen saturation in blood by measuring the amount of light that the blood, particularly haemoglobin, absorbs [61]. Pulse oximetry is a standard technique widely used in general wards, operating rooms, intensive care and home care. The changes of blood volume in the arteries during systole and diastole modify the transmission and reflection of light passing through a body segment. A pulse oximeter uses a pair of wavelengths of light-emitting diodes (LEDs) to illuminate the tissue and measures the variations of light intensity at each wavelength with a photodetector [61]. The variations are related to the absorbances and pulsatile nature of the arterial blood, which can be used to determine the oxygen content in the blood. Venous blood has a low pressure and is non-pulsatile as a result of the veins that have thin vessel walls with thin layers of muscle and elastic fibres [61].

The main principle behind pulse oximetry is the difference between the absorbance of oxyhaemogloin and deoxyhaemoglobin in the red and infrared parts of the spectrum, typically measured

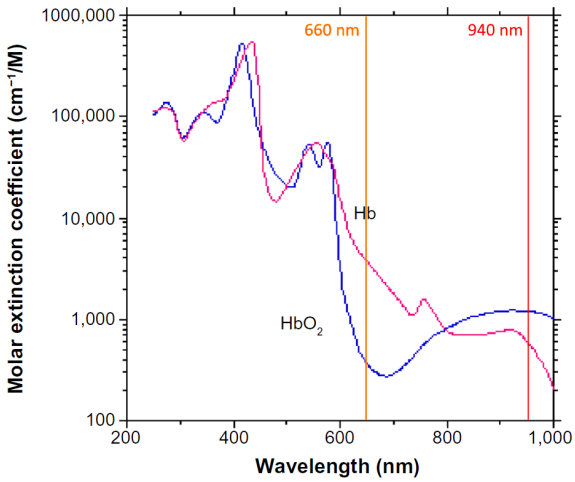


Figure 2.4: Extinction spectra of oxyhaemoglobin (HbO_2) and deoxyhaemoglobin (Hb). (Image reproduced from [69], courtesy of S. Prahl and the Oregon Medical Laser Center)

at approximately 660 nm and 940 nm respectively (see figure 2.4). Oxygenated blood is more absorbant at around 660 nm; while, deoxygenated blood is more absorbant at 940 nm. In addition to the difference in haemoglobin absorbance, light is greatly absorbed by melanin and water in the human tissue at the wavelengths shorter than 600 nm and higher than 980 nm respectively [70]. The penetration depth of light is wavelength-dependent, considering that higher wavelengths penetrate deeper than shorter wavelengths. For Caucasian skin, 600-nm red light can penetrate into living skin up to 1.5 mm [9]. Peripheral blood oxygen saturation (SpO_2) can be calculated from from the ratio of the absorption of the light at 660 nm and the light at 940 nm.

Each light source (at either 660 nm or 940 nm) passing through the tissue to be recorded by the photodetector results in a PPG waveform which varies according to the changes in blood volume during each cardiac cycle and the pulsatile nature of arterial blood.

2.3.3 Photoplethysmography (PPG)

Photoplethysmography (PPG) reflects the changes in blood volume over time in the underlying tissue beneath the skin. Figure 2.5 shows the PPG waveform containing an AC signal caused by the pulsatile component of arterial blood and a DC non-pulsatile quasi-static signal caused by venous blood and other tissues [71]). The DC signal may contain various low frequency components from respiration, thermoregulation or sympathetic nervous system activity [28].

The standard Beer-Lambert law, formulated in equation 2.1, defines the light absorption in a static non-scattering medium. The tissue is dynamic, thus the intensity passing through the tissue varies over time according to blood volume changes underneath the tissue [62]. Given an incident light I_0 at wavelength λ travelling into a dynamic non-scattering medium with an absorption coefficient $\mu_a(\lambda)$, the resulting intensity I measured by a photodetector can be modelled through

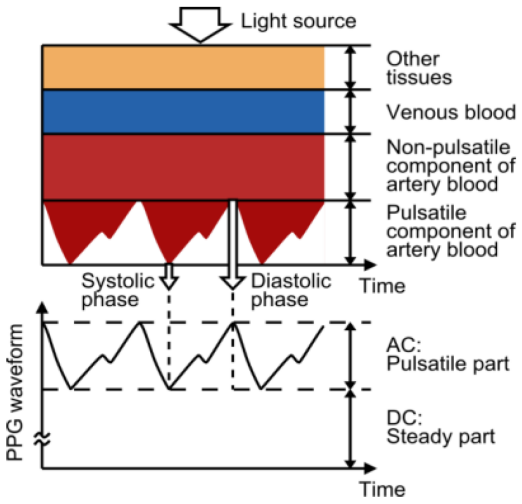


Figure 2.5: The PPG signal component. The PPG signal is composed of an AC pulsatile part and DC steady part. The AC signal changes according to the pulsatile component of arterial blood and is superimposed over the DC signal from venous blood, bone and other tissues. (Image reproduced from [71])

the inclusion of a time-varying optical path length $Lp(t)$ as:

$$I(\lambda) = I_0(\lambda) \cdot \exp\{-\mu_a(\lambda) Lp(t)\}. \tag{2.2}$$

The optical path length can be broken down into two components: static and dynamic, representing the optical path length of the DC (venous blood, non-pulsatile part of arterial blood and other tissues) and AC (pulsatile part of arterial blood) components respectively (see figure 2.5):

$$I(\lambda) = I_0(\lambda) \cdot \exp\{-[\mu_{a(\text{static})}(\lambda) L_{\text{static}} + \mu_{a(\text{arterial blood})}(\lambda) L_{\text{dynamic}} p(t)]\}. \tag{2.3}$$

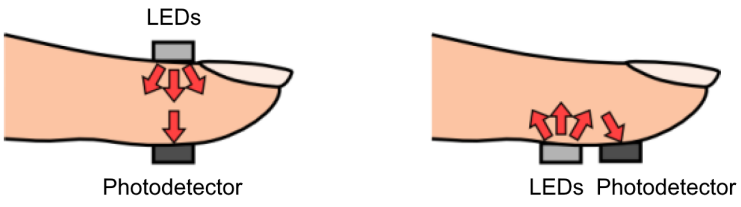
The amplitude of the AC component typically accounts for approximately 5% of the total intensity [61, 72]. The AC component has been shown to be useful in the assessment of cardiac performance, vascular disease and arterial compliance [9].

2.3.4 Mode of operation

Pulse oximetry requires two optical light sources at different wavelengths to illuminate the skin and a photodetector to record the light propagating through the skin. Two operational configurations of the light sources and photodetector are generally used in the clinic: transmission mode and reflectance mode (see figure 2.6a).

In transmission mode, the photodetector is placed at a known distance on the opposite side of the light sources. The light intensity recorded by the photodetector shows the changes of blood volume in the skin. During systole, the amount of red blood cells and haemoglobin increases, the optical density and path length also increase, thus the intensity of light recorded by the photodetector decreases. Transmission-mode pulse oximeters are suitable for a body segment with high vascularisation and less tissue such as fingers or earlobes, since the light is required to propagate to the opposite end of the body segment in order to reach the photodetector [73].

In reflectance mode, the photodetector and light sources are placed side by side over the skin, and the light from the light sources is scattered back to the photodetector (see figure 2.6b). Reflectance-mode pulse oximeters can be applied to a flat skin surface with high superficial vascularisation such as the forehead and chest or wrapped around the wrist [73]. Reflectance-mode pulse oximetry generally



(a) (b)

Figure 2.6: Two operational modes of pulse oximetry. (a) Tranmission mode, with the light sources and photodetector placed on opposite sides of the skin. (b) Reflectance mode, with the light sources and photodetector placed on the same side of the skin. (Images reproduced from [71])

produces a waveform with a lower signal-to-noise ratio than transmission-mode pulse oximetry. This is partly due to the smaller fraction of the emitted light reaching the photodetector [74].

Pre-term infants are generally equipped with an elastic adhesive emitter and detector strip wrapped around the foot [75]. Since their skin is moist and they are often shivering and crying, the attachment needs to be firmly secured to allow accurate measurements to be performed throughout the day. Any elastic strip allows clinical staff to adjust the sensor to fit to the foot. The pulse oximeter is normally designed to work in transmission mode. Several studies have suggested the use of reflectance-mode pulse oximetry in pre-term infants by wrapping a sensor around the head [76] or attaching it to the thigh [77].

2.3.5 Applications

Pulse oximetry is ubiquitous in the clinic thanks to its versatility and ability to provide a continuous measurement of various physiological parameters such as peripheral oxygen saturation, heart rate and respiratory rate.

As described before, oxygen saturation in blood flowing underneath the skin can be estimated by measuring the light absorbance at two wavelengths. Heart rate can be computed from variations in the AC component of the PPG signal which reflect changes in arterial blood volume. Breathing causes slow variations in the peripheral circulation and modulates the DC component of the PPG signal, allowing the respiratory rate to be estimated from the PPG signal [78, 79]. Several studies have been reported on the estimation of finger arterial blood pressure through a PPG sensor embeded inside an automatic inflatable finger cuff [80].

In pre-term infants, the continuous monitoring of oxygen saturation allows the clinical staff to maintain the correct oxygenation level in the blood by adjusting supplemental oxygen. This can prevent life-threatening cases related to reactive oxygen species such as retinopathy of prematurity and bronchopulmonary dysplasia [81]. The measurement of PPG waveforms from multiple sites allows the calculation of pulse phase difference which can be used to identify patent ductus arteriosus (the blood vessel bypassing the lungs not closing properly after birth) in pre-term infants [76].

2.3.6 Limitations

Despite the advantages and usefulness of pulse oximetry in the clinic, several factors can have negative effects on the derived measurements. Pulse oximeters are designed, calibrated and tested with healthy volunteers, thus discrepancies in the measurements may occur for patients. The presence of significant dysfunctional haemoglobins in sick patients, which are highly absorbant at the wavelengths used in pulse oximetry, may reduce the intensity of pulsatile components [82]. The effects of local hyperthermia (high body temperature) and hypothermia (low body temperature) may change the reflectance properties of the skin [69]. The presence of hypoxia

(oxygen deficiency) can reduce the oxygen supply, leading to a lower value measured by a pulse oximeter [83]. Large measurement errors of both heart rate and peripheral oxygen saturation derived from pulse oximeters have been reported in subjects with dark skin color or with a high melanin concentration in the skin [84–86].

The versatility of pulse oximetry is limited by the location where a measurement probe can be placed. Probes cannot be used on skin areas for which a dermal contact is undesirable, usually caused by burns, lacerations, dehydrations or poor hygiene [54]. Movement of a patient or the sensor often leads to motion artefacts and reduces the quality of the PPG signal [87]. Bright fluorescent light in the room may interfere with the light emitted by a pulse oximetry device [88]. Poor positioning or loosening of the sensor may create the penumbra effect (part of the light from the LEDs going directly to the photodetector) which also reduces the signal quality [69]. Pulse oximeters may cause complications such as infections, burns and pressure necrosis (skin sore or wound caused by pressure) [86, 89].

2.4 Vital sign monitoring using video cameras

2.4.1 Photoplethysmographic imaging (PPGi)

Photoplethysmographic imaging (PPGi) is an optical technique to measure pulse-induced subtle colour changes on the human skin surface or the epidermis using a video camera from a distance. The signals derived from the video have similar characteristics to the PPG signal [6].

There is a variety of terminology used in the literature to refer to the measurement of pulsatile signals from a video camera including rPPG (remote photoplethysmography), iPPG (imaging photoplethysmography), PPGi (photoplethysmographic imaging) and DistancePPG (distance photoplethysmography). Throughout this thesis, the term PPGi is used.

The principles of PPGi are similar to the reflectance-mode PPG in which the light source is ambient light or a dedicated light source (such as normal room lights, fluorescent light or halogen light) and the photodetector is a video camera, which can be perceived as an array of photodetectors. The light propagating into the skin is absorbed by both tissue and blood, thus the video camera can be used to measure the intensity of light reflected from the epidermis layer of the skin which corresponds to changes in blood volume underneath the skin during systole and diastole. However, instead of measuring the light intensity reflected from relatively deep tissue under the skin (such as arterioles) as in reflective-mode PPG, the non-contact PPGi measures light intensity reflected from more shallow vasculature (such as capillaries) [6, 30, 90, 91]. Figure 2.7 shows a diagrammatic representation of the PPGi technique overlaid with the interactions of light with the skin.

Different theories exist in the literature regarding the origin of the PPGi signal. Blazek *et al.* [4] proposed that the pulsatile variations in the signal are caused by volume changes in venous

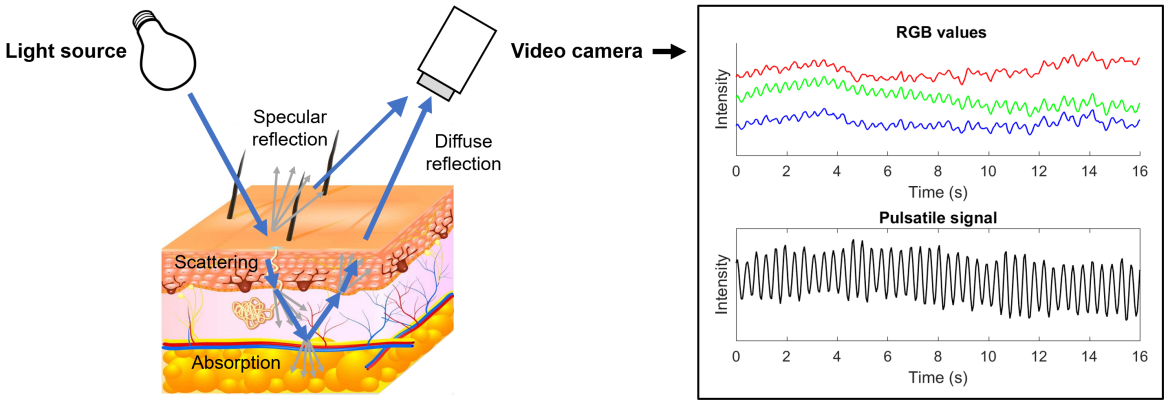


Figure 2.7: Diagram of the interaction between light and the tissue beneath the skin. The RGB values measured using the video camera contain useful pulsatile information related to heart beats.



Figure 2.8: Amplification of subtle colour changes in the human skin shows the periodic color variation of the human pulse. Images were produced using the Eulerian Video Magnification technique [11].

vascular beds which lead to changes in absorption characteristics during a cardiac cycle. Figure 2.8 shows the amplification of subtle colour changes in the human skin [11]. Balakrishnan *et al.* [92] showed that heart rate can be measured from changes in skin reflectance that are caused by musculoskeletal oscillations as a result of subtle head movements due to the influx of blood from the carotid arteries in each cardiac cycle. Thus a part or all of the PPGi signal might not originate from light absorption variations. Kamshilin *et al.* [93] suggested that pulsatile variations in the PPGi waveform may stem from the pulse oscillations caused by the elastic deformations of capillary beds in the upper dermal layer due to the expansion of the deeper arteries.

In the RGB (red, green and blue) colour recording, it has been found that the PPGi signal extracted from the green colour channel has the strongest pulsatile component, followed by those extracted from the blue colour channel and the red colour channel respectively [14, 94, 95]. This characteristic is thought to be because green light at the wavelength between 520 and 570 nm has large absorption peaks of oxyheamoglobin (see figure 2.4) [6].

2.4.2 Imaging devices

The characteristics of the video cameras play a significant role in the quality of the video and thus the PPGi signal. Most of the research in non-contact monitoring has employed charge-coupled device (CCD) cameras which have high sensitivity to visible light, low-noise performance and good

dynamic range [9, 54]. Complementary metal-oxide semiconductor (CMOS) cameras have also been employed in several research studies. Compared to CCD cameras, CMOS cameras have several advantages including higher frame rate and lower power consumption [54].

The imaging sensor should have sufficient sensitivity over the spectral range of the physiological parameters of interest; for example, sufficient sensitivity over the red and near-infrared wavelengths for measuring peripheral blood oxygen saturation. The colour filters, generally positioned over the image sensor, should provide a good separation between colour channels. The spectral range of each colour channel should have minimal overlap with nearby colour channels [30].

The video camera should allow the functions relating to automatic image adjustments (e.g. focus, shutter speed, white-balance, gain and exposure) and any other processing techniques (e.g. image filters and video compression) to be switched off. These settings often cause changes in pixel values and may diminish pulsatility from the video data [95]. The video camera should be able to capture and process video data in an uncompressed format and with a constant frame rate. Video compression may reduce the quality of video data and produce compression artefacts [96]. A variable frame rate may lead to an incorrect measurement of physiological parameters [95].

Finally, the video camera should be positioned to capture a sufficient amount of skin pixels such that the quantisation noise of the imaging sensor can be averaged out.

2.4.3 Vital sign estimation

Studies over the last decade have demonstrated that PPGi technology enables non-contact measurement of physiological parameters. PPGi technology has been shown to be able to estimate heart rate, respiratory rate, heart rate variability, and peripheral blood oxygen saturation [4–6, 14, 30, 91, 97]. The technology can be employed in the hospital, either in a general ward, an emergency room or a reception area. It can measure physiological parameters from multiple individuals at the same time [98]. PPGi technology can also be employed at home to provide a non-instrusive method for measuring and tracking the health condition of individuals [99, 100]. The system can also be integrated into existing video monitoring systems that are used for monitoring infants [101] or the elderly population [102], assessing gait function [103] and detecting risk of falls [104].

Blazek *et al.* [4] first introduced a near-infrared imaging system to analyse the blood volume changes in the top skin layers and visualise the skin vessels in 2000. Wieringa *et al.* [5] later showed that arterial oxygen saturation together with heart rate and respiratory rate could be derived from signals obtained from a video camera at different wavelengths. Tanako and Ohta [97] used a CCD camera to measure heart rate and respiratory rate remotely in a small group of healthy subjects using auto-regressive spectral analysis in 2007. Verkruysee *et al.* [6] showed that PPGi signals could be measured under ambient visible light from a consumer-level video camera in 2008.

Compared with conventional PPG which provides a single one-dimensional signal, the PPGi gives spatial physiological information which could be used, not only for monitoring vital signs, but also for detecting haemodynamics, visualising the location of blood vessels, computing skin perfusion maps and analysing local blood perfusion in tissue. A large sensing area enables physiological information to be derived from any skin regions. Although the signals are susceptible to motion artefacts, the cause of the artefacts can be identified and their effects can be attenuated by utilising signal processing and computer vision techniques [105, 106]. Computer vision algorithms could also be applied to the video to perform behavioural assessments and detect clinical adverse events. Although PPGi technology has been studied extensively by the research community in the last decade, there are many challenges and calibration issues that remain before the technology can be deployed into clinical practice.

Recent studies have demonstrated that vital signs, e.g. heart rate, respiratory rate and peripheral oxygen saturation, can be continuously derived from measurement of the ambient light reflected from the skin using a video camera in clinical environments [14, 107, 108]. Here we describe previous work in our research group, Tarassenko *et al.* [14] developed a method for non-contact cardio-respiratory vital sign monitoring which works under most illumination conditions. The light reflected from the subject's regions of interest was analysed using auto-regressive models. Heart rate and respiratory rate were estimated from the analysis of the frequency spectrum. The ability to derive heart rate estimates from multiple skin regions leads to the estimation of the strength of the heart rate and respiratory rate signals at various sites. Later, Guazzi *et al.* [108] presented a method to track changes in peripheral oxygen saturation using a colour camera under visible light. These studies showed significant progress towards the clinical application of non-contact vital sign monitoring. Since vital signs can be estimated remotely, the physical restrictions on cabling and measurement sites are reduced. This type of system is desirable in some challenging scenarios such as pre-term infants in the NICU, intensive care patients and patients with severe mental illness.

Several studies have investigated non-contact approaches for the vital-sign monitoring of pre-term infants in a clinical environment using a video camera. Scalise *et al.* [10] first investigated the non-contact monitoring of 7 pre-term infants for 30 seconds using a webcam. The authors reported that the infant's heart beat could be identified from colour changes in the skin measured under normal ambient light. Aarts *et al.* [12] further performed a study investigating 19 pre-term infants during short and stable periods of 1 to 5 minutes. Recently, Villarroel *et al.* [107] from our research group examined continuous monitoring of the vital signs during daytime under ambient light. They showed, in their preliminary results, that heart rate, respiratory rate and peripheral oxygen saturation could be derived remotely and continuously for several hours. They further

demonstrated that these parameters could be used to identify abnormal events such as bradycardia episodes in which oxygen saturation decreased due to apnoea.

2.4.4 Limitations

One of the major limitations of the camera-based technology, compared to convention technology such as a portable and wearable pulse oximeter, is that PPGi technology can only monitor a subject when he or she is in the field of view of the camera.

The characteristics of the light source is important for PPGi technology. The light source should provide sufficient energy covering the absorption spectrum of the physiological parameter of interest (see figure 2.3). Higher levels of light intensity could lead to higher strengths of the measured pulsatile signal as more photons penetrate the skin and are reflected back [95]. However, too much light intensity may cause signal clipping which could prevent the measurement of vital signs. The light source should also be stable in time [94, 95]; any flickering or illumination changes can be problematic as it can cause perturbations in the recorded PPGi signal.

The properties of the skin have an influence on the PPGi signal recorded from the video camera. The skin has different skin composition and thickness across the body. Some areas, such as the face and palms, have stronger pulsatile components than others [109]. Due to high melanin concentration, dark-colour skin absorbs more energy, therefore less energy is reflected back from the skin surface. This leads to a low signal-to-noise ratio of the PPGi signal recorded from an individual with dark skin colour [95]. Skin motion also presents challenges to PPGi technology. Movement of the patient can cause changes in the reflections of the light from the skin which lead to distortions to the PPGi signal [95]. The effects of skin motion can be partly mitigated by computing the PPGi signal by averaging over a sufficiently large area of pixels across the skin area [14].

Over the past decade, most of the research in camera-based vital sign monitoring has been conducted in controlled environments where healthy subjects remained relatively still during the video recording [5–8, 11–14, 106]. Several research teams have derived vital signs from regions of interest (ROIs) that were manually selected and fixed across the video sequence [5, 6, 11–13]; however, this is not practical and may lead to false measurements if the subjects move. Recent studies have investigated the automatic selection of measurement sites by incorporating techniques from the computer vision community [7, 8, 14, 106]. These studies employ well-established face detection and feature tracker methods to automatically derive regions of interest. Face detection is used to define the face region in the first frame and the face region is then automatically tracked in consecutive frames using a feature tracker. This approach, however, performs well only on frontal faces and a few profile faces. It is less reliable if the subject's face is partially occluded which is a

typical scenario in a hospital. In addition, the feature tracker can only tolerate small movements and has difficulties in the presence of large motion or occlusions.

A number of patent applications relating to the selection of a measurement site have been filed in the United States and Europe. In 1999, Toshiba patented a method based on the difference of two successive frames [110]. Since 2006, Philips has filed a family of patent applications regarding region of interest selection: the incorporation of the well-known Viola-Jones face detection algorithm and the block matching algorithm for region-of-interest tracking [111, 112], the use of body markers on the face and chest of the subject to define regions of interest [113] and the use of spectral frequency analysis of movement patterns estimated on a grid-by-grid basis from the whole image [114–116]. Similarly, Mediatek have filed a method based on skin colour detection to obtain skin regions in the face area [117]. Collectively, these aforementioned approaches are ideal when the subject is facing the camera directly. The variation in lighting conditions could potentially cause problems for the Toshiba frame difference and Mediatek skin detection techniques.

To summarise, conventional methods based on shape analysis of the face or chest are less reliable as they are limited by the posture and position of the subject. A feature tracker usually fails as a result of natural movements such as face tilting, hand motion and turning over. Illumination changes negatively affect the process of determining skin regions and motion vectors. In practice, the method for the selection of a measurement site has to be fast, robust and accurate especially under challenging scenarios such as in a NICU where the illumination changes throughout the day and the subjects are actively moving and typically are not facing the camera.

2.5 Conclusion

Pulse oximetry is a ubiquitous optical technique that allows the measurement of physiological parameters to be performed continuously and non-invasively. Most pulse oximetry systems on the market can only measure vital signs from one location, e.g. ears, fingers or toes. In pre-term infants, a pulse oximeter is typically placed around the foot.

Camera-based vital-sign monitoring enables the non-contact monitoring of physiological parameters. The majority of the literature has reported studies carried out with healthy volunteers in well-controlled environments. A limited number of studies have been conducted in a real hospital environment, but their population size and recording length were not sufficiently large to demonstrate the robustness of algorithms. It is also challenging to develop non-contact vital-sign algorithms to perform in a real-world hospital environment.

Chapter 3
Neonatal HDU Study

3.1 Introduction

This chapter describes the Neonatal High-Dependency Unit (Neonatal HDU) study which was designed to investigate the use of video cameras for monitoring the vital signs of 30 pre-term infants. The study was carried out in the high-dependency area of the Neonatal Intensive Care Unit (NICU) at the John Radcliffe Hospital in Oxford. Each pre-term infant was recorded under regular ambient light during daytime for up to four consecutive days.

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 [118]. Babies born at the same gestational age may have very different health outcomes. In the NICU, they are surrounded by complex medical equipment depending on their unique needs. Constant nursing and medical supervision are provided to the infants until they are strong enough and ready to be discharged [1]. In the UK, the median length of stay in the hospital is 93 days for pre-term infants of less than 27 weeks of gestation, 44 days between 28 to 31 weeks of gestation and 13 days for pre-term infants between 32 to 36 weeks of gestation [119]. The median postmenstrual age (the sum of gestational and postnatal ages) at discharge from the hospital is 38.6 weeks [119].

According to the Neonatal Data Analysis Unit (NDAU) [119], over 110,977 infants were admitted to specialist neonatal care in England, Scotland and Wales in 2015. In Northern Ireland, over 2,100 infants were admitted to neonatal care in 2014 [120]. In the past decade, the number of admissions has continued to rise by approximately 13% each year in England, Scotland and Wales [121]. Pre-term infants, especially those who are born very early, are often associated with disabilities (e.g. motor and learning disabilities and visual and hearing impairment) and account for half of disabilities in children and young adults [122].

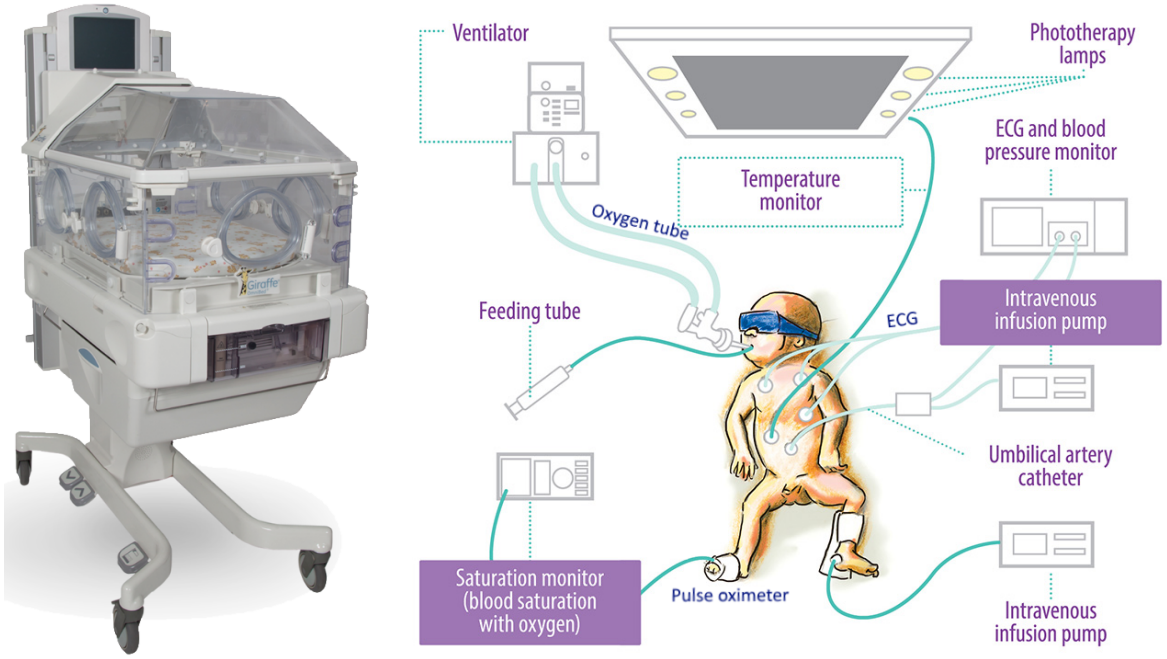
Since pre-term infants are physiologically unstable, their vital sign values can vary substantially in a short time period [1]. The wide spectrum of these values benefits the development and validation of non-contact vital-sign estimation algorithms.

3.2 Current physiological monitoring of pre-term infants

Pre-term infants admitted to specialist neonatal care are required to be nursed in incubators and to be regularly monitored by clinical staff (see figure 3.1a). Vital signs, such as heart rate, respiratory rate, oxygen saturation and blood pressure, are essential for assessing the well-being and body functions of the infants during hospitalisation.

Patient monitors are used to obtain the physiological parameters through sensors and electrodes directly attached to the infant's body. The vital signs of the infant are gathered continuously and clinical staff document them in patient records on a regular basis. Some parameters such as blood pressure are recorded and documented at variable rates depending on the state of the infant. In addition, patient monitors are equipped with alarm systems which are activated when vital signs deviate from their normal ranges.

Figure 3.1b shows the medical equipment that is commonly attached to an infant who is nursed in an incubator in the NICU. Three electrodes are typically placed on the infant's chest for recording the electrocardiogram (ECG) and impedance pneumography (IP) signals which are used to assess the cardiovascular physiology and estimate the values of the heart rate and respiratory rate. A pulse oximeter is often attached to the foot to monitor the oxygen supply. An umbilical artery



(a) (b)

Figure 3.1: Monitoring of vital signs in the Neonatal HDU (a) Front view of a typical incubator. (Image courtesy of GE Healthcare) (b) Top view of an infant nursed in the incubator. An oxygen tube is attached to the ventilator to provide supplemental respiratory support. Electrodes are placed on the body to record the electrocardiogram (ECG) for assessing cardiovascular physiology. A pulse oximeter is placed around the foot for monitoring the oxygen supply. Intravenous (IV) pump delivers nutrients or other fluids to the infant. (Image courtesy of "Our Premature Children" Foundation)

catheter is inserted into the umbilical artery to provide access to arterial blood supply and allow the measurement of blood pressure; however, this apparatus is rarely used after the first seven days of life [118]. These standard-of-care sensors may cause skin irritation or infection in early pre-term infants since their skin is very delicate and sensitive. Neonatal monitoring technology has not changed since the introduction of pulse oximetry to the standard of care in the early of 1980s [28].

3.3 Clinical study design and protocol

Recent work in our group described in Tarassenko *et al.* [14] has demonstrated that heart rate, respiratory rate and peripheral oxygen saturation can be continuously monitored using a video camera under ambient lighting by measuring the light reflected from the skin. This work has led to several studies in both healthy subjects and patients under hospitalisation [107, 108, 123]. These studies aimed to evaluate the feasibility of estimating the vital signs from a digital video camera in different population groups and different clinical environments.

The main aim of the Neonatal HDU study was to develop and validate non-contact vital-sign monitoring methods in pre-term infants nursed in hospital. The clinical study was designed to run without affecting regular patient care. Pre-term infants who participated in the study were required to be nursed in a designated study incubator in the high-dependency area of the NICU at the John Radcliffe Hospital in Oxford. The participants were double-monitored with a digital video camera and the standard patient monitoring devices. The study incubator was similar to a standard incubator currently used in the hospital (see figure 3.2). The video camera was mounted over the incubator on a camera stand which was clamped to a custom-built trolley. Both the video camera and camera stand were CE-marked and compliant to European health, safety and environmental protection standards. Detailed information on the study was given to the infant's parents before the start of video recording.

All the standard patient monitoring and ongoing care were continued throughout the study session. The set-up of all the research equipment (video recording and storage) was designed to minimise the inconvenience to clinical staff during the study. No additional sensors were attached to the infants. Access to the incubator was not in anyway restricted by the position and location of the video camera and the associated equipment.

The clinical study was performed during daytime under regular ambient light conditions. As part of routine care, there was a designated quiet period of six hours each day during which the infant was allowed to rest without any disturbance by clinical staff or parents except for emergencies.

Video recording could be temporarily paused or the video camera could be temporarily covered at the discretion of clinical staff during some clinical procedures (e.g. phototherapy and cannula insertion) or when the infants were taken out of the incubator for cuddling by their parents (kangaroo



Figure 3.2: Data acquisition setup on a typical recording session in the Neonatal HDU study. The camera was positioned over a specifically-drilled hole on top of the study incubator (red circle). In the picture, a mannequin was placed inside the incubator.

care). If the infant was to be transferred to another unit, video recording was terminated and data were recorded until that point.

Our Neonatal HDU study is different from other studies of non-contact monitoring in a neonatal intensive care [12, 124–126] in the following ways: (1) it was performed on a bigger cohort of 30 pre-term infants, (2) the infants were continuously monitored for several hours a day in parallel with regular patient care and (3) no restrictions were imposed regarding gestational age, infant weight, ethnic group or skin tone.

3.3.1 Research ethics

The Neonatal HDU study was part of a research programme in the Oxford University Hospitals NHS (National Health Service) Foundation Trust and the Oxford Biomedical Research Centre (BRC). The research was compliant with the relevant government and regulations of the Oxford University NHS Foundation Trust. The study was approved by the South Central – Oxford Research Ethics Committee under the reference number 13/SC/0597.

3.3.2 Study participants

This study aims to monitor 30 pre-term infants for up to seven consecutive days. Each infant was nursed in a designated incubator in the high-dependency area of the NICU. The clinical team recruited the infants based on the British Association of Perinatal Medicine's Categories of Care 2011 [127].

Inclusion criteria

Participants needed to satisfy all of the following criteria:

- < 37 weeks postmenstrual age (gestational age plus chronological age)
- Requiring high-dependency care
- Requiring continuous monitoring of heart rate, respiratory rate and oxygen saturation
- Requiring to be nursed naked

Exclusion criteria

The study excluded any infants who presented life-threatening conditions that prevented the continuous monitoring in the high-dependency area of the NICU.

3.3.3 Informed consent

Consent was required to be given by the parents of each baby prior to any recording. Parents whose infants fitted the inclusion criteria (see section 3.3.2) were approached by the study personnel (NICU clinicians) and given full verbal and written information about the study. They were later offered a follow-up meeting with the research team to discuss any questions that could have arisen. Parents were informed that video camera images and vital sign reference data would be used in an anonymised data format by the research team solely for research purposes. They were made aware that participation was voluntary and that they were able to withdraw their baby from the study at any time without giving any explanation and without affecting patient care. Once they agreed for their infant to participate in the study, an informed consent form was given to them to read and sign before recording could start.

3.3.4 Participant confidentiality

The study complies with the UK's Data Protection Act. The data were anonymised as soon as it was practical to do so and stored in an encrypted storage cluster only accessible by members of the research team. Participants were only identifiable by their participant number, the letters 'PN' followed by a two-digit number (e.g. 'PN01' for the first participant). Researchers were not allowed to access identifiable personal information.

3.3.5 Data handling and record keeping

Patient demographics, physiological records, clinical events and patient case notes were anonymously recorded in electronic case report forms (CRFs). This study used the OpenClinica platform [1] to store clinical research data and capture CRFs.

The following participant demographic information was recorded: ethnicity, gender, gestational age at birth (weeks), corrected gestational age (weeks), weight at birth (grams) and current weight

[1] OpenClinica is a clinical trial software package for storing patient records and capturing clinical study data

(grams). Clinical staff also recorded, where appropriate, the type of respiratory support (e.g. none, low-flow nasal cannula, high-flow therapy), medications administered, rate of enteral feeding (ml/kg/day) and rate of continuous infusions (ml/kg/day).

Any clinical procedures or interventions performed during the recording were also recorded in the CRFs. Examples of procedures or events were: changing sensing probes, measuring blood pressure, collecting a blood sample or visits by parents. Other episodes of clinical deterioration were also noted in the CRFs.

3.4 Instrumentation

The continuous recording of video images and vital-sign data in a real hospital environment presents many challenges. For example, the monitoring equipment is required to be compact, easy to transport and safe for hospital use. The video recording software must be able to handle in a reliable manner the high-throughput data produced by the camera.

Our study used a custom-built recording system and in-house developed software. The hardware passed a rigorous health and safety test carried out by the Department of Medical Physics and Clinical Engineering at the John Radcliffe Hospital. The software was specifically developed to satisfy the requirements of the study, with video images stored in raw format without any alterations to the data recorded by the video camera sensor.

3.4.1 Hardware for data acquisition

The research team developed a data acquisition system (see figure 3.2) which consists of a trolley carrying the video camera and two recording workstations. One workstation was used to record video data from the camera, and the other workstation was connected to the NICU's patient monitors in order to record the reference physiological signals.

Trolley

A trolley was specifically assembled for this study by a technician team in the mechanical workshop of the Institute of Biomedical Engineering at the University of Oxford. The trolley was designed to facilitate the research team moving the entire monitoring equipment quickly inside the hospital.

Incubator

Pre-term infants were cared for in a designated Giraffe OmniBed® Carestation incubator (General Electric, Connecticut, USA). A modification was made to the incubator's canopy by drilling a small hole in the top plastic panel of the incubator's canopy to allow a video camera to be positioned inside the incubator's chamber in order to film the infants without reflection and attenuation from the perplex layer (see figure 3.2). This modification to the incubator was approved by the Medical

Research Ethics Committee (MREC). After the modification was carried out on the study incubator, a series of humidity and temperature tests were performed over a period of two weeks to ensure that the incubator had the same level of environmental control within the chamber as a standard incubator. These tests were carried out by the health and safety team at the Department of Medical Physics and Clinical Engineering at the John Radcliffe Hospital. This was done to ensure that the modification made to the incubator was safe and did not pose any risk to the patients.

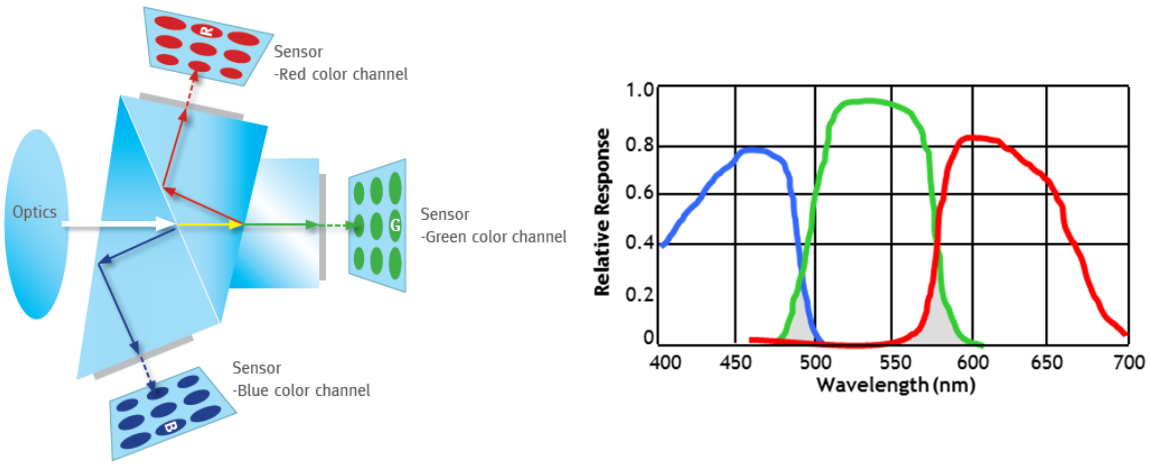
Workstations

Data acquisition was performed using two workstations: one for video recording (Workstation A) and the other for medical data recording from the patient monitor (Workstation B). Workstation A was a custom-built Dell PowerEdge® server (Dell, Texas, USA) equipped with 24-core Intel® Xeon® processor (Intel, California, USA), 32 GB memory, 40 TB hard disk, Nvidia® Quaddro graphics card (Nvidia, California, USA) and the Fedora™ 20 operating system. Workstation B was a Dell OptiPlex® workstation equipped with an Intel® Core™ i7 processor (Intel, California, USA), 8 GB memory, 4 TB hard disk and Microsoft Windows® 7 operating system.

A medical-grade Accumed® keyboard and mouse (Accuratus, UK) were used with both workstations. These devices comply with the IP67 standard for dust and water protection and the JIS Z 2801 test for antimicrobial activity of plastics.

Video camera system

Video recordings were acquired using a 3-CCD JAI® AT-200CL digital video camera (JAI A/S, Denmark). The camera employs three Sony® ICX274AL 31/1.8" image sensors (Sony, Japan) to measure



(a) (b)

Figure 3.3: The JAI AT-200CL 3-CCD camera used in the study. (a) The prism used for splitting the incoming light into three different colour channels: red, green and blue. (b) Spectral response curve for the three imaging sensors. (Images courtesy of JAI Inc.)



Figure 3.4: Typical images from the video camera. The whole infant body was filmed in each session. A X-RITE BST-13 colour checker chart (X-RITE, Michigan, USA) was also placed inside the incubator for calibration purposes.

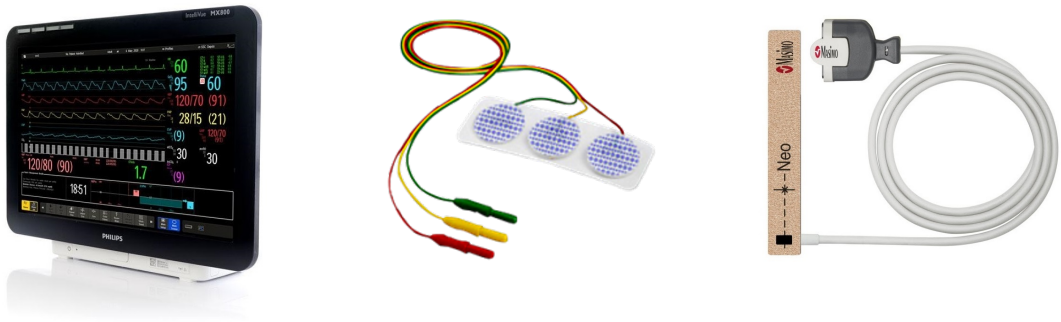
the light intensity of each colour channel (red, green and blue) separately. Figure 3.3a shows a block diagram for a 3-CCD camera, and Figure 3.3b shows the spectral responses of the camera sensors.

The video camera was equipped with a VS Technology SV-0614H lens kit (VS Technology, Japan) which allowed full control over focal length and lens aperture. Before, and occasionally during a recording session, attending clinical staff were required to adjust these parameters to ensure that the infant's skin was in the focal plane and that the brightness level was adequate (see figure 3.4). The camera was securely clamped to the trolley via a Manfrotto® 224RC camera arm (Manfrotto, Italy). The arm could be extended horizontally over the incubator.

The video camera system acquired 24-bit lossless colour images (8-bit per colour) at a resolution of 1620×1236 pixels. The video recordings were performed at a rate of 20 frames per second. The video was recorded on Workstation A via a microEnable IV AD4-CLR framegrabber card (Silicon Software, Germany). Since continuous transmission of large data streams may cause image corruption or data loss when the workstation is busy and unable to request an additional video image from the camera, Workstation A had a video frame buffer implemented using a Xilinx® Spartan Field Programmable Gate Array (FPGA) board (Xilinx, California, USA). A typical 1-hour recording produced 408 GB of data.

Patient monitor

Reference vital signs (heart rate, respiratory rate and oxygen saturation) were monitored using a Philips® IntelliVue MX800 patient monitor (Philips, Netherlands). The Philips® patient monitor were installed with a Philips® IntelliVue MMS measurement module and a Masimo® SET SpO2 Vuelink IntelliVue measurement module (Masimo, California, USA). All participants in this study were equipped with a set of neonatal electrodes for recording ECG and IP signals and a Masimo®



(a) (b) (c)

Figure 3.5: Patient monitoring devices: (a) Philips® IntelliVue MX800 patient monitor (Image courtesy of Philips), (b) a set of neonatal electrodes (Image courtesy of Philips) and (c) Masimo® LNCS Neo SpO2 adhesive sensor (Image courtesy of Masimo).

Table 3.1: Physiological parameters recorded during a typical study session

Name	Description	Sampling rate
ECG	1-lead electrocardiography signal	500 Hz
IP	Impedance pneumography signal	64 Hz
PPG	Photoplethysmography signal	125 Hz
ECG-HR	Heart rate derived from ECG by Philips® IntelliVue MMS module	1 Hz
PPG-HR	Heart rate derived from PPG by Masimo® SET SpO2 module	1 Hz
RR	Respiratory rate derived from IP by Philips® IntelliVue MMS module	1 Hz
SpO_2	Oxygen saturation provided by Masimo® SET SpO2 module	1 Hz

LNCS (Low Noise Cabled Sensors) Neo SpO2 adhesive sensor for recording oxygen saturation. Figure 3.5 shows the patient monitoring equipment used in the Neonatal HDU study.

The patient monitor was connected to Workstation B via a RS232 serial interface. The ixTrend® software (Ixellence GmbH, Germany) was used to record the data streams generated by the Philips patient monitor. The following waveforms were recorded: 1-lead ECG signal (at 500 Hz), IP signal (at 64 Hz) and the PPG signal (at 125 Hz). The following physiological parameters were recorded at a rate of 1 Hz: heart rate from ECG, heart rate from PPG, respiratory rate from IP and oxygen saturation. Blood pressure was manually measured by the neonatal nurses and the values were recorded in the Case Report Form (CRF). Table 3.1 gives a summary of the physiological parameters that were recorded in the Neonatal HDU study.

Audio recorder

The level of noise inside the incubator's chamber was measured using a TASCAM DR-07 MKII digital voice recorder (TASCAM, California, USA) through a Drift Ghost HD microphone (Drift Innovation, UK). The audio was recorded at 44.1 Hz in a 16-bit WAV format. The microphone was disposed of after each recording session.

Storage cluster

A storage cluster was installed in the data centre of the Institute of Biomedical Engineering. The cluster consisted of a Dell PowerVault MD3200® head node (Dell, Texas, USA) and Dell PowerVault MD1200® (Dell, Texas, USA) storage nodes. Each storage node has 12×4 TB hard disks configured in RAID (Redundant Array of Independent Disks) 60 which provides fault tolerance for up to 3 disks per storage node.

After each recording session at the John Radcliffe Hospital, video and reference data were exported into an encrypted LaCie 2big Quadra 8TB external storage device (LaCie, France). The disk was then transported to the Institute of Biomedical Engineering where the data were uploaded to the storage cluster.

3.4.2 Software for data acquisition

Apart from the ixTrend software for recording medical data using Workstation B, the research team developed custom software specifically designed for data acquisition and other routine tasks.

Video collector

Software for video recording, running in Workstation A, was developed to interface directly with the camera manufacturer's API (Application Programming Interface) such that it could have full control of the camera during recording. The software recorded videos at a full camera resolution of 1620×1236 pixels, with a continuous trigger mode of 20 frames per second. Automatic gain control (AGC) was disabled in all colour channels. Video data were stored in a 24-bit raw format in blocks of one minute (1200 video frames) without any encoding or compression.

Time synchronisation

Since the recording of the video and reference data was carried out by two separate workstations, time synchronisation was required at the beginning of each recording session in order to ensure the accuracy of the timestamps across workstations. Any inconsistency of timestamps between the video and reference data will cause difficulties for the retrospective analysis of the recorded data.

Synchronisation was performed using the network time protocol (NTP). A time server was set up on Workstation A. At the start of every recording, Workstation B synchronised its clock with that of Workstation A. This process was performed automatically.

3.5 Cohort description

This section describes the demographics of the research participants and presents the statistics of the study sessions and the reference vital-sign data.

3.5.1 Participant demographics

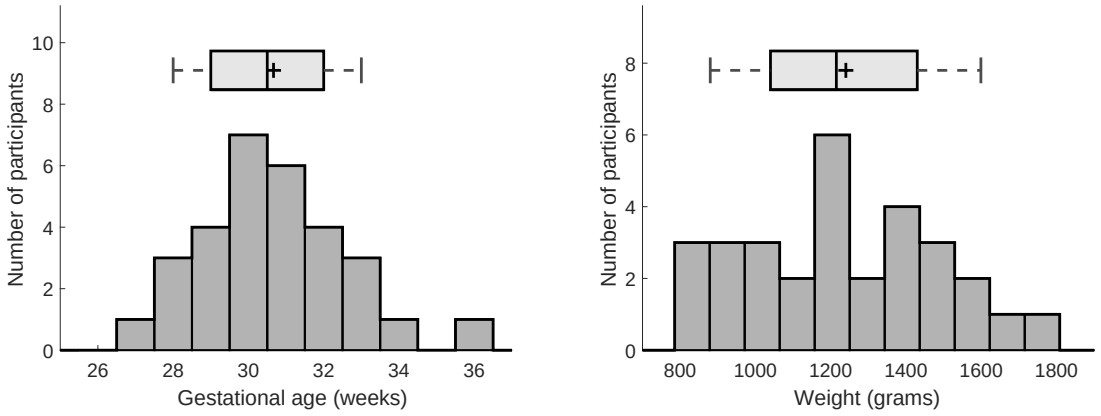
The Neonatal HDU study ran for 15 months from February 2014 to May 2015. A total of 30 preterm infants were enrolled in the clinical study, and a total of 90 study sessions were recorded. 18 out of 30 participants were male (60%). The study gained interest from a wide spectrum of ethnicity backgrounds. 18 out of 30 infants were White British (60%). The mean of gestational age was 30.7 weeks, and the mean infant weight was 1,240 grams. Table 3.2 provides a summary of participant demographics.

Figure 3.6 shows the distribution of gestational ages and infant weights. These values were collected from each participant during the first day of study. The range of gestational age was 27.6 – 36.4 weeks; and the range of infant weights was 830 – 1746 grams.

Table 3.2: Summary of population demographics

Total participants	30
Total recording sessions	90
Gestational age[1,2] (weeks)	30.7 (±2.0)
Weight[1,2] (grams)	1, 240 (±252)
Gender (number of participants)	
Male	18
Female	12
Ethnicity (number of participants)	
White British	18
White – Any Other Background	2
Asian or Asian British – Pakistani	2
Black British or Black African	1
Mixed – White and Asian	2
Mixed – White and Black Caribbean	2
Mixed – White British and Japanese	1
Mixed – Any Other Mixed	2

[1] Value shown in Mean (±SD)
[2] ON the first day of recording



(a) (b)

Figure 3.6: Distribution of a) gestational age and (b) infant weight collected on the first day of recording. Above the histograms, a box plot bounds the 25% and 75% quartiles along with whiskers marking 9% and 91% quantiles. The middle lines indicate the median whereas a '+' mark signifies the mean.

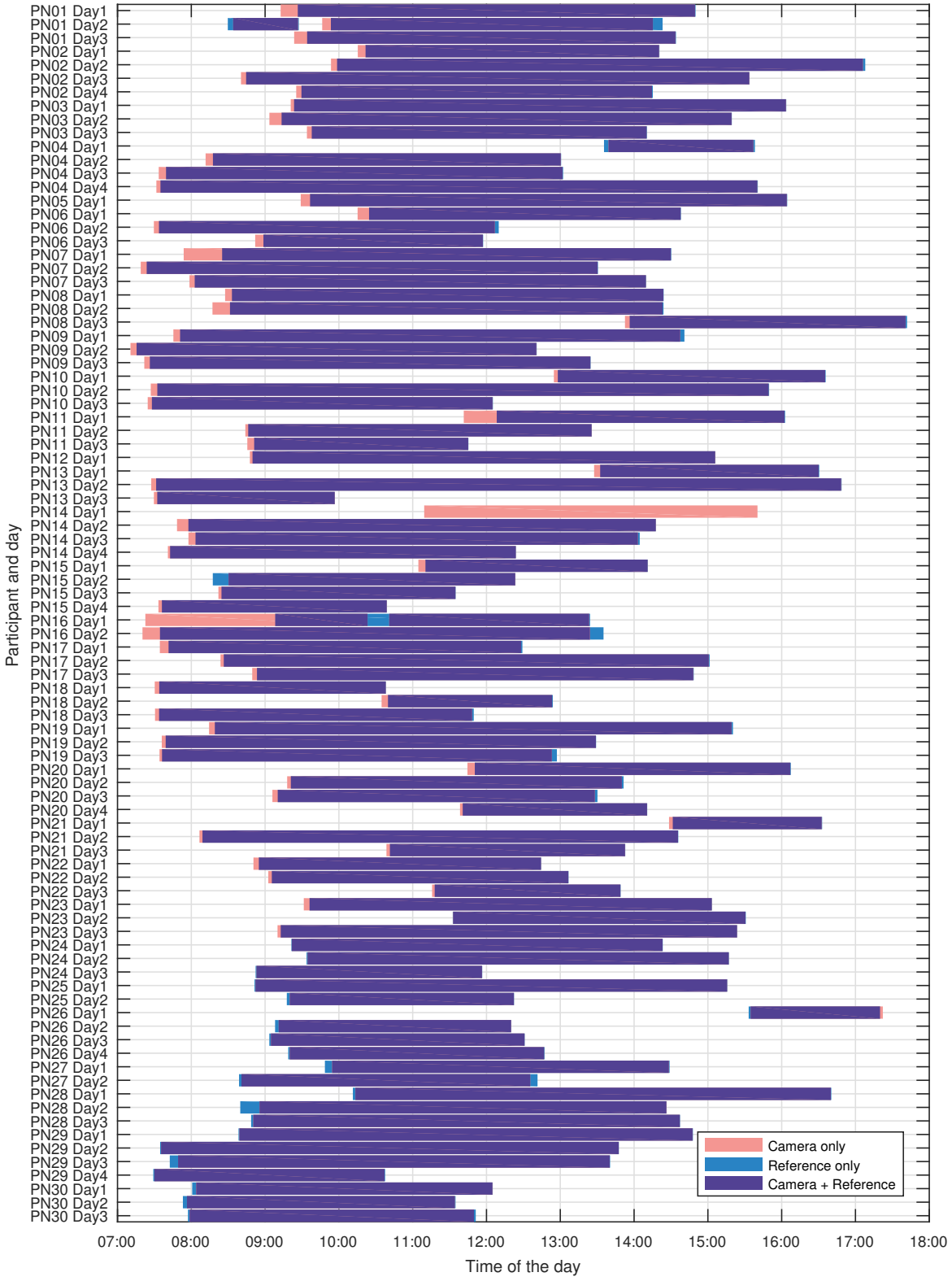


Figure 3.7: Recording times for each session in the Neonatal HDU study. The recording of reference data usually started a few minutes after the video camera.

3.5.2 Recording statistics

Figure 3.7 shows a timeline plot of the recording times for all recording sessions. Studies usually started in the morning and ended in the early afternoon. Since equipment malfunction occurred during the recording of session PN14 Day 1, the reference data for this session were not recorded. The total time that all participants spent in the study was 442.0 hours (video camera or reference data were being recorded). Of 442.0 hours, 426.6 hours (96.5%) were the periods during which the recording times for both the video camera and the reference monitoring equipment overlapped (both video camera data and reference data were being recorded at the same time).

Table 3.3 summarises the statistics for the video recording and vital-sign data acquisition. The total video recording time was slightly higher than the reference equipment because the recording of video camera data usually started before the recording of the reference vital signs. Moreover, there was one session without the recording of reference data (PN14 Day1).

Table 3.3: Statistics for the video camera and reference recording times

Camera recording time	
Total recording time (hours)	439.1
Average recording time per session[1] (hours)	4.9 (±1.6)
Average recording time per subject[1] (hours)	14.9 (±5.7)
Reference vital sign recording time	
Total recording time (hours)	429.5
Average recording time per session[1] (hours)	4.8 (±1.6)
Average recording time per subject[1] (hours)	14.3 (±5.2)

[1] Value shown in Mean (±SD)

3.5.3 Reference vital signs

Table 3.4 gives the descriptive statistics for the reference vital-sign data recorded from the Philips monitor. The descriptive statistics of the heart rate provided by the ECG are consistent with those provided by the PPG. The mean of heart rate values was 156.5 beats/min for the heart rate provided by the ECG and 156.3 beats/min for the heart rate provided by the ECG. The mean of respiratory rate was 51.8 breaths/min. The mean of oxygen saturation was 94.6%.

Figure 3.9 shows the distributions of vital sign values (heart rate, respiratory rate and peripheral oxygen saturation (SpO_2)) recorded from the Philips patient monitor at 1 Hz over the entire Neonatal HDU study. The Neonatal HDU dataset has a wide range of physiological values.

Figure 3.10 shows an example of the time-series recorded from a typical 4-hour study session. The values of vital signs vary greatly during the day. When the infant was undergoing clinical interventions (e.g. changing a diaper and taking blood samplings), clinical staff may temporarily remove physiological sensors or pause the vital sign monitoring. Figure 3.11 shows short sections of 30-second and 3-second time-series waveforms.

Table 3.4: Statistics for the reference vital-sign data recorded by the Philips monitor

Vital sign	Mean (±SD)	Median (IQR)
Heart rate derived from ECG (beats/min)	156.5 (±14.0)	157 (17)
Heart rate derived from PPG (beats/min)	156.3 (±13.9)	157 (17)
Respiratory rate (breaths/min)	51.8 (±18.5)	50 (27)
Oxygen saturation (%)	94.6 (±5.1)	96 (5)

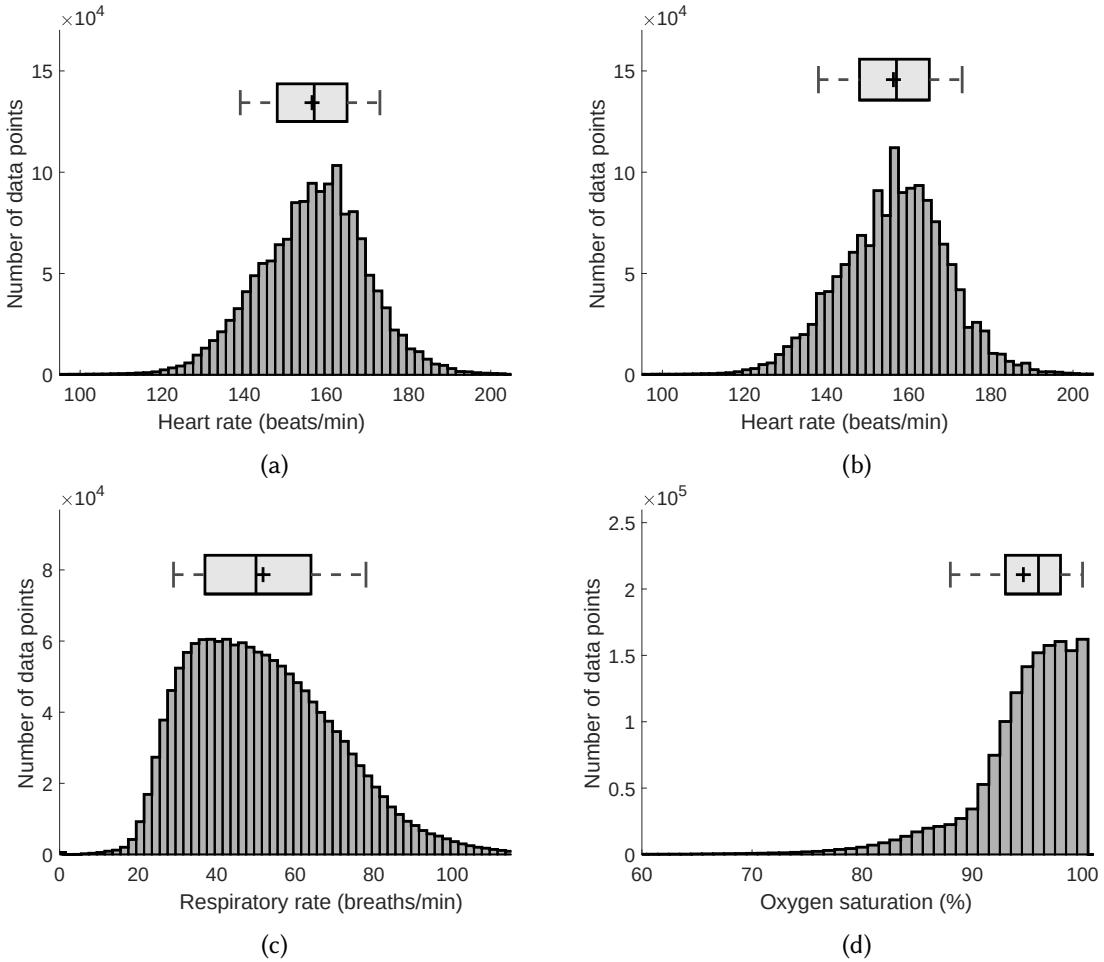


Figure 3.9: Distributions of vital signs for the entire Neonatal HDU study (a) heart rate from ECG, (b) heart rate from PPG, (c) respiratory rate from IP and (d) oxygen saturation from the pulse oximeter.

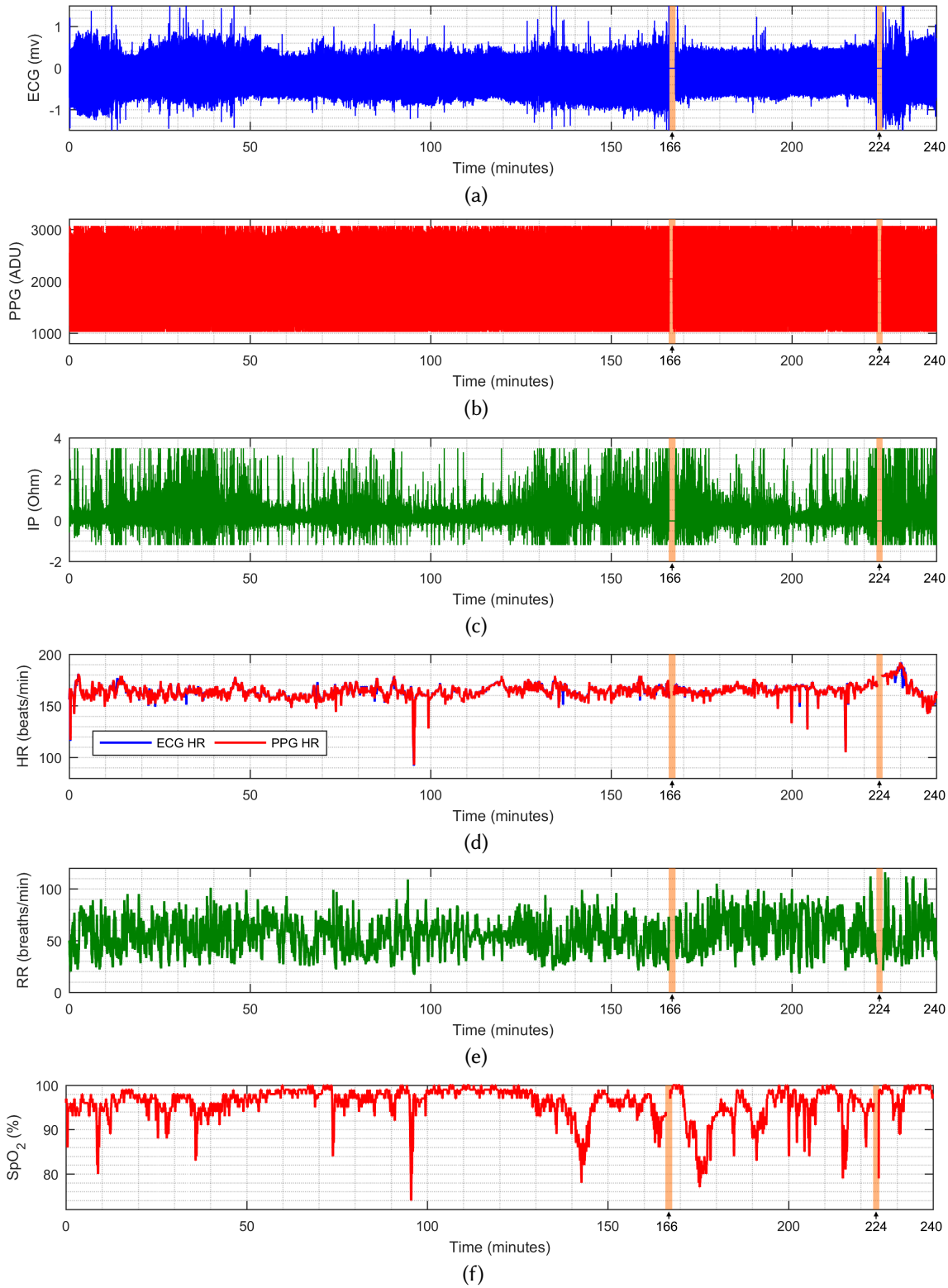


Figure 3.10: 4-hour sample from a recording session: (a) ECG waveform, (b) PPG waveform, (c) IP wavefrom, (d) Heart rate derived from ECG (blue) and PPG (red), (e) respiratory rate derived from IP signal and (f) oxygen saturation from the pulse oximeter. Note that the infant was undergoing clinical interventions at minutes 166 and 224, during which the monitoring was temporarily paused.

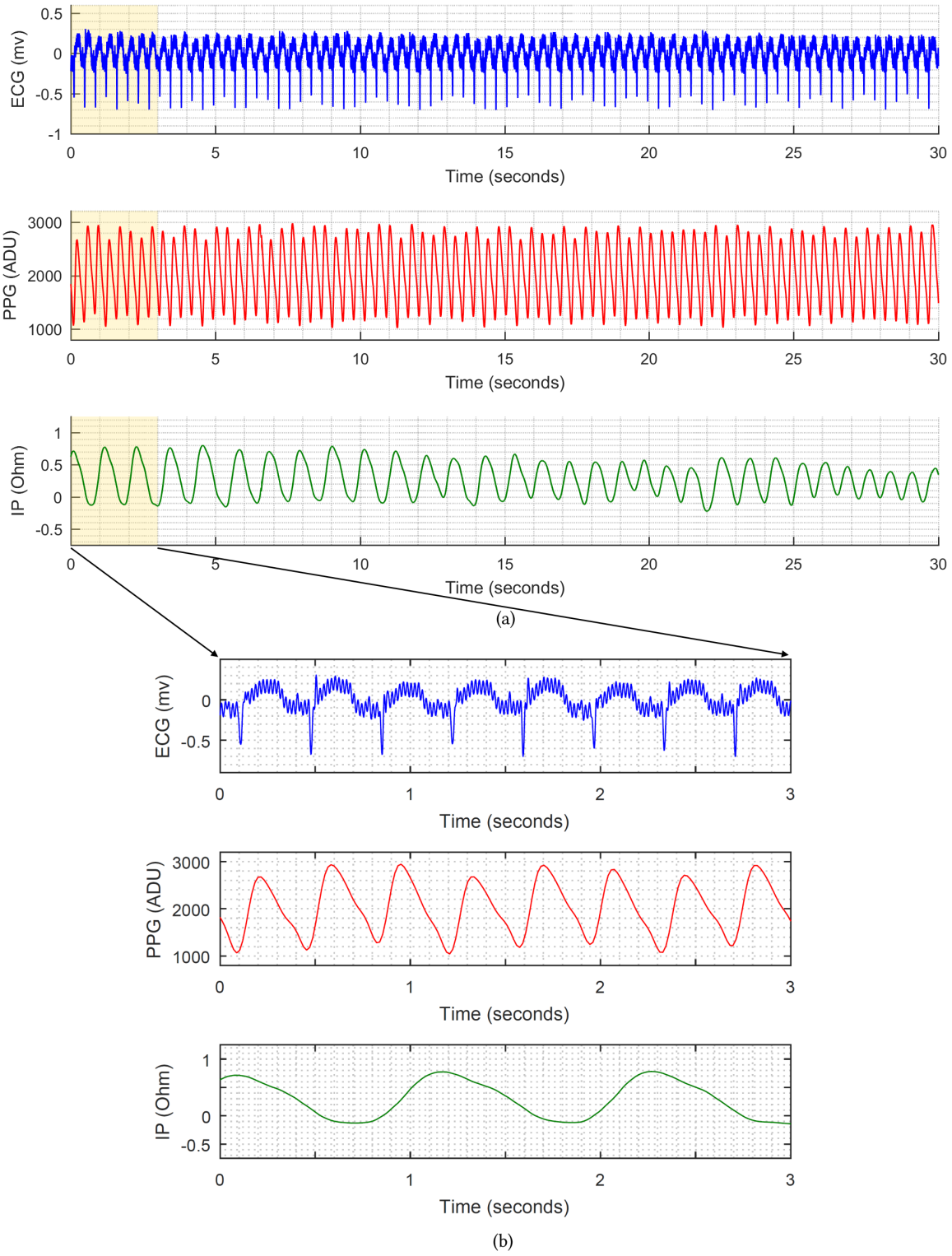


Figure 3.11: Sample waveforms from a typical recording session: (a) 30-second recording and (b) the first three seconds for the recording session in (a).

3.6 Partitioning the dataset

The clinical study protocol submitted to the ethics committee required the algorithms to be developed and validated on half the study participants and then tested on the other half. In compliance with the protocol, the participants were divided into two sets of 15 infants each: the *training* set and the *test* set. The video recording information (date, time and duration) and patient demographics (ethnicity, gestational age and weight) were chosen to be as balanced as possible in each set.

The training set was used to develop algorithms for pre-processing the video data and to optimise the global parameters of the vital-sign estimation algorithms. Once the vital-sign algorithms produced estimates that correlated well with the corresponding values from the patient monitor, the algorithms were tested on the test set which contains the data from the other 15 infants.

Table 3.5: Summary of population demographics in the training and test sets

Set	Number of subjects	Number of sessions	Total time[1] (hours)	Gender		Ethnicity[2]					
				Male	Female	W	B	A	WB	WA	O
Training	15	43	216.6	8	7	10	1	1	1	1	1
Test	15	47	210.0	10	5	10	–	1	1	2	1
Total	30	90	426.6	18	12	20	1	2	2	3	2

[1] Period during which both reference data and camera data were being recorded simultaneously.
[2] W = White, B = Black, A = Asian, WB = Mixed White & Black, WA = Mixed White & Asian and O = Other.

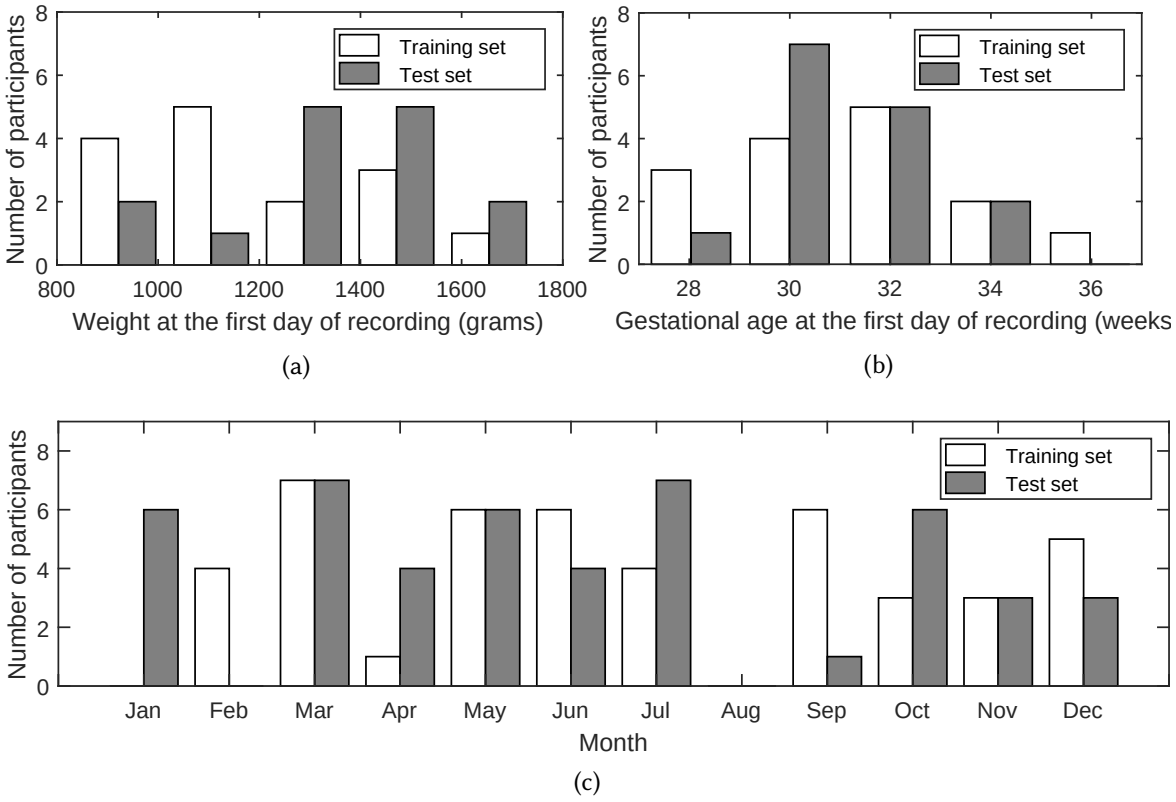


Figure 3.12: The distribution of (a) infant weights, (b) gestational ages and (c) study dates in the training and test sets of the Neonatal HDU study.

Table 3.5 gives a summary of the participant demographics for the training set and the test set. The distributions of gestational ages and weights are shown in Figure 3.12a–b. Also, the distribution of the months during which the studies took place is shown in Figure 3.12c.

3.7 Conclusion

Participants in the Neonatal HDU study were double monitored using a video camera in parallel with a standard patient monitor, without disturbing regular patient care. This study produced a large dataset of video data and reference signals collected from participants from diverse backgrounds. It was used to evaluate the performance of non-contact vital-sign estimation algorithms. The dataset can be used not only to estimate physiological parameters but also for the development of algorithms for detecting important clinical events.

Chapter 4
Patient detection and skin segmentation

4.1 Introduction

In this chapter, we aim to answer two important questions arising during continuous non-contact monitoring of vital signs: is a patient in the scene? and if so, which pixels belong to skin? Skin pixels are used to extract the pulsatile information required for the estimation of vital signs such as heart rate, respiratory rate and oxygen saturation.

In a neonatal intensive care environment, infants are usually active and clinical staff constantly interact with them. Infant postures and positions are spontaneous. Many routine clinical activities are undertaken several times a day. The infants are regularly taken out of the incubator by clinical staff for a clinical procedure, or cuddling by their parents (kangaroo care). Lighting conditions change not only throughout the day but they also depend on the season of the year, with long and bright days during summer and short and dark days during winter. Artificial light sources, such as room and overhead lights, often cause reflections on the skin that change how the video cameras record colour [14, 128]. Shadows are cast over the baby when clinical staff walk by the incubator in which the infant is nursed. These scenarios pose challenges for the development of algorithms for estimating vital signs from video cameras. New methods are required to be less dependent on skin tones, body postures, patient position and lighting conditions.

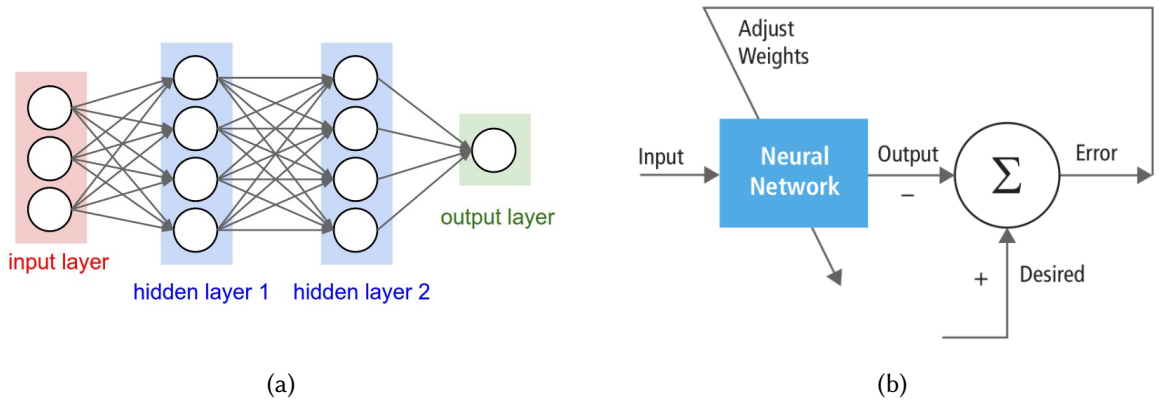
This chapter begins with background and previous work on convolutional neural networks (CNNs) or deep learning in section 4.2. An overview of the pre-processing pipeline prior to vital sign estimation is presented. in section 4.3. The chapter then presents a multi-task CNN model for the detection of the presence or absence of the pre-term infant in the video frame and the segmentation of the skin regions in section 4.4. In section 4.5, a dataset and its annotation protocol are described, followed by the procedures used to train the model on this dataset. The protocol used to evaluate the model is explained in detail in section 4.6. In section 4.7, experimental results

describing the performance of the model are reported, followed by discussion in section 4.8. A conclusion is finally presented in section 4.9.

4.2 Background

A neural network is a system consisting of neurons that are connected together. Each neuron has a weight vector. The neuron that receives information from other neurons can process it and pass on to successor neurons. Figure 4.1a shows an example of a 3-layer neural network [129]. The numerical weight vector associated with each neuron is determined during the training process. The training process is performed on a dataset with labelled examples. The dataset generally contains a wide variety of input patterns. Each input is tagged with a desired response. The training is performed by iteratively changing the weight vectors for each layer and each neuron in the network. Figure 4.1b shows a simplified diagram of the training process of neural networks [130]. A properly trained neural network can respond correctly to its inputs.

A convolutional neural network or CNN is a special variation of a neural network specially designed for handing data with spatial information, such as images. The CNN employs a convolution



(a) (b)

Figure 4.1: Concepts of neural networks. (a) Example arrangement of a neural network with three layers. Neurons in each layer are connected with every neuron in the previous layer. Each neuron is associated with a numerical weight. It can process information and pass it on to other neurons. (b) Training process of neural networks. Numeric weights are iteratively determined during the training process. Image courtesy of Andrej Karpathy [129] and Hijazi *et al.* [130].

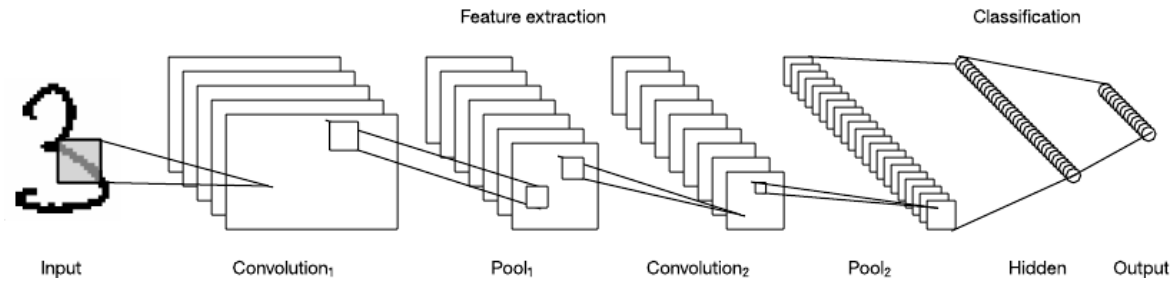


Figure 4.2: Example of a CNN architecture with two convolutional layers (LeNet2 [131]). This network is designed for classifying hand-written digits. Image courtesy of Tim Jones [132].

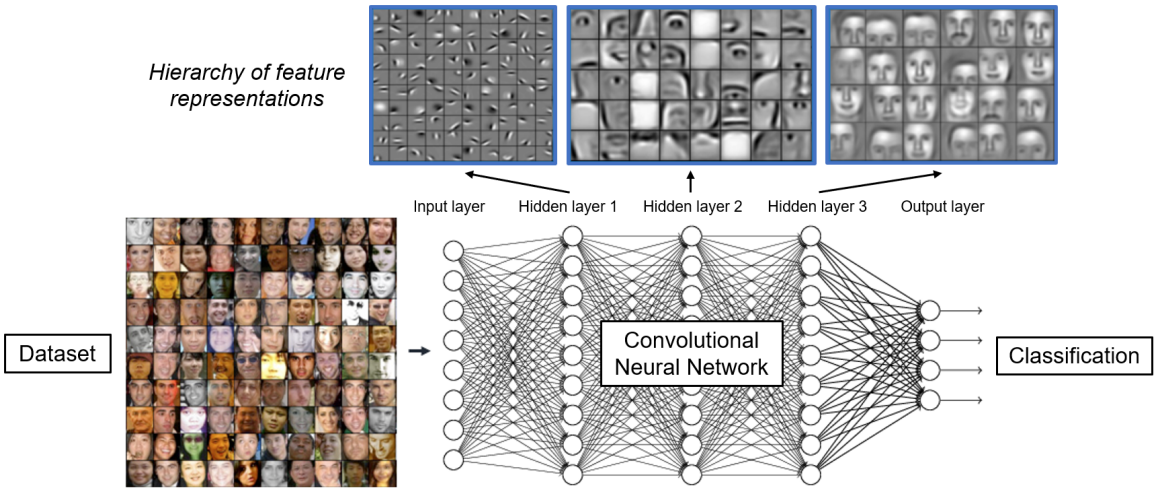


Figure 4.3: Hierarchical feature representations in CNNs. The network typically comprises a series of convolution operations which allow it to extract features at multiple levels of abstraction. The first few layers extract low-level local features (edges and corners), while deeper layers handle progressively more complex features (objects). (Image courtesy of Michael Nielsen and Andrew Ng [147])

operation to compute the response of the neuron's weights to the multi-dimensional input data. The design of a CNN was inspired by the connectivity pattern of the animal visual cortex [133]. The visual cortex contains cells (cortical neurons) that response only to small and overlapping regions of the reception field. The convolution operation emulates the function performed by these cells. A typical CNN model contains several convolutional layers stacked together, and each layer consists of a set of learnable filters (or neurons). A pooling layer is often placed between convolutional layers to reduce the spatial resolution of the features (see figure 4.2). This arrangement allows the extraction of image features at multiple levels of abstraction: the first few layers extract low-level local features (edges and corners), while deeper layers handle progressively more complex features (objects) [134]. With the convolution operation, the model is able to process high-dimensional data such as images, audio and videos. CNN has brought great success in various computer vision tasks such as image classification [135–139], image segmentation [140–142] and video classification [143–145]. The network is able to learn how to discover complex structures in the datasets by adjusting the weights of each hidden layer through the backpropagation algorithm [146]. Figure 4.2 shows an example of hierarchical feature representation in CNNs.

4.2.1 Image classification

Image classification refers to the task of identifying visual contents, e.g. objects, within an image. In the past decades, it has commonly relied on handcrafted features extracted from local image patches. CNNs have recently led to major advances in large-scale image classification [135–139] thanks to their ability to learn rich feature representations and combine feature and classifier training together.

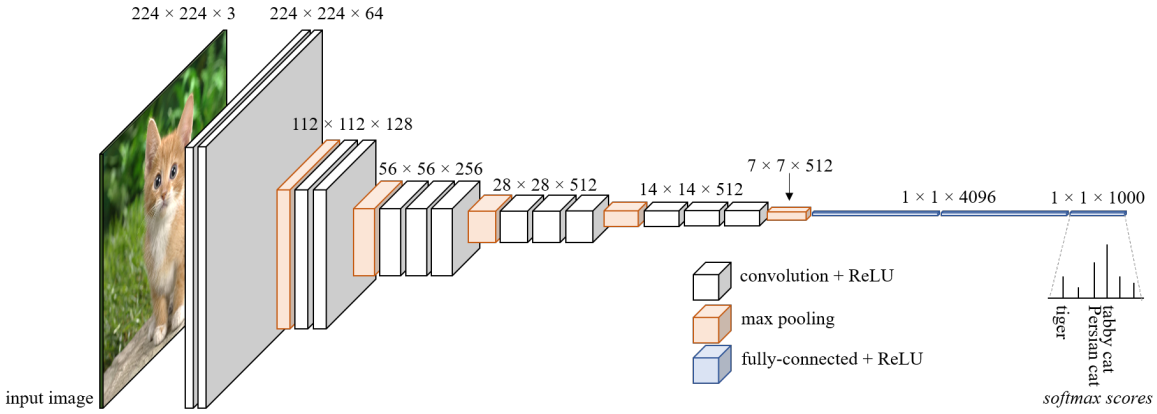


Figure 4.4: The VGG-16 network is made up of a stack of convolutional layers with non-linearity layers (ReLU), and periodically followed by max-pooling layers (for downsampling). The numbers over each layer indicate the dimension Height × Width × Channels of its output. The structure repeats until the output has a small spatial size and a decision is made on that output by fully-connected layers and a softmax layer (for calculating class scores). The VGG-16 has 13 convolutional and 3 fully-connected layers.

The success of CNNs began with AlexNet [135] which won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, a challenge for large-scale image classification, using a network of five convolutional layers and two fully-connected layers. The fully-connected layers made use of all outputs in the preceding layer to provide high-level reasoning. Later, the visual geometry group (VGG) networks by Simonyan and Zisserman [137] won the ILSVRC competition in 2014. The authors demonstrated that network depth was an important factor for achieving high levels of performance [137]. Their famous network, VGG-16 (see figure 4.4), has thirteen small convolutional layers and three fully-connected layers in a homogeneous arrangement. This network gained popularity through its good generalisation on different datasets or even in different domains.

Lin *et al.* [148] proposed global average pooling to replace fully-connected layers for classification. The global average pooling takes a spatial average of the input to the layer. It forces the network to generate one output for each class in the last convolutional layer. This technique performs competitively with the traditional fully-connected layers but significantly fewer parameters are required. Szegedy *et al.* [138] employed this idea in their GoogLeNet which was designed to reduce the computing resources used by the network while keeping its performance high. They demonstrated that both depth and width of the network are important factors for achieving a high performance. He *et al.* [139] investigated the depth of the network up to 152 layers. They used short connections to connect non-successive intermediate layers together in order to help training very deep networks, and global average pooling as a classifier. Their networks, ResNets, won the first place in the ILSVRC competition in 2015.

4.2.2 Image segmentation

Image segmentation is a method for assigning a label to each pixel of an image. Many deep architectures are designed specifically to tackle image segmentation. Most networks maintain the VGG16

network as a core network and the segmentation is performed by decoding the last convolutional layer which contains high-level features of a class label of interest but in a very low resolution.

A fully convolutional network (FCN) [140] converts the VGG's fully-connected layers to convolutional layers and adds deconvolutional layers to upsample feature maps across the network in order to produce meaningful segmentation results. A deconvolution network [141] implements a decoding architecture as a mirror version of the entire VGG-16 network. This multiplies the number of parameters by a factor of two and makes the network very difficult to train. The performance of such a network was reported to be slightly better than that of the FCN network but with significantly higher inference time. SegNet [142] leveraged the problem by dropping the VGG's fully-connected layers and implementing a decoding branch using a mirror version of the encoder with connections that link the encoder's pooling layers with the decoder's pooling layers. Its performance was comparable to that of the FCN network with less memory consumption but higher inference time due to a large stack of convolutional layers.

Several proposed architectures have employed a conditional random field (CRF) to improve the segmentation performance of deep networks. The CRF exploits boundary information, e.g. edges, to refine segmentation outputs. DeepLab by Chen *et al.* [149] used the CRF as a post-processing step after the final fully-connected layer. The authors reported a performance boost due to refined segment boundaries. Lin *et al.* [150] explored a multi-scale architecture which extracts local and global features at different scales and then feeds to a trainable CRF layer. However, the network was reported to be difficult to train due to the large parameter size and the requirement for a special training strategy. Zheng *et al.* [151] augmented the FCN network with a CRF-based graphical models whose parameters were jointly trained with the rest of the network. However, large improvements in performance were observed only when extra training data were used.

4.3 Method overview

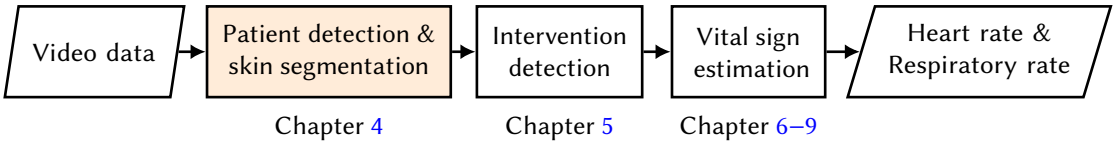


Figure 4.5: Overview of the workflow for non-contact vital sign estimation.

There are three major pre-processing steps prior to the non-contact estimation of vital signs in the Neonatal HDU study: patient detection, skin segmentation and intervention detection (see figure 4.5). This chapter focuses on patient detection and skin segmentation, as highlighted in figure 4.5. Intervention detection is later discussed in chapter 5. The clinical interventions, e.g. changing nappies or sensors, can cause motion artefacts preventing the estimation of vital signs.

Therefore, vital sign estimation is performed from the infant's skin regions only when the infant is present in the scene and no clinical interventions are undertaken.

In this chapter, a multi-task CNN model is developed for detecting the presence of the infant and segmenting the infant's skin regions. The multi-task model consists of a shared core network with two branches which handle two different tasks: one for patient detection and the other one for skin segmentation (see figure 4.6). The patient detection task is performed first, followed by skin segmentation only if a patient is found in the video frame. The video recording is not processed further when the infant is absent from the scene.

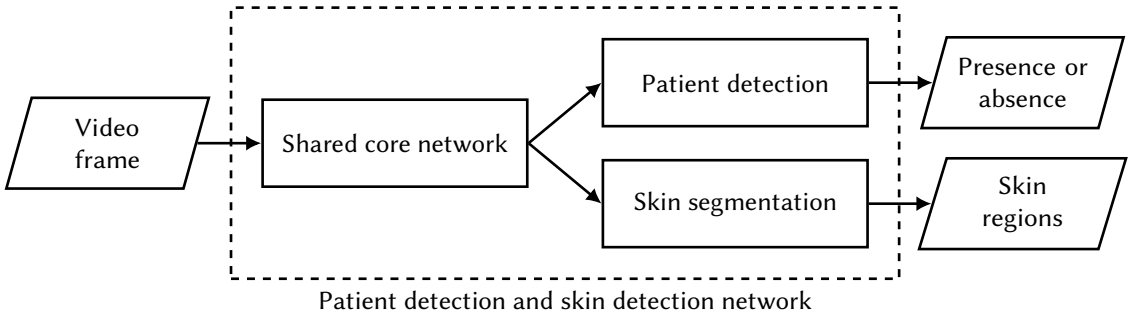


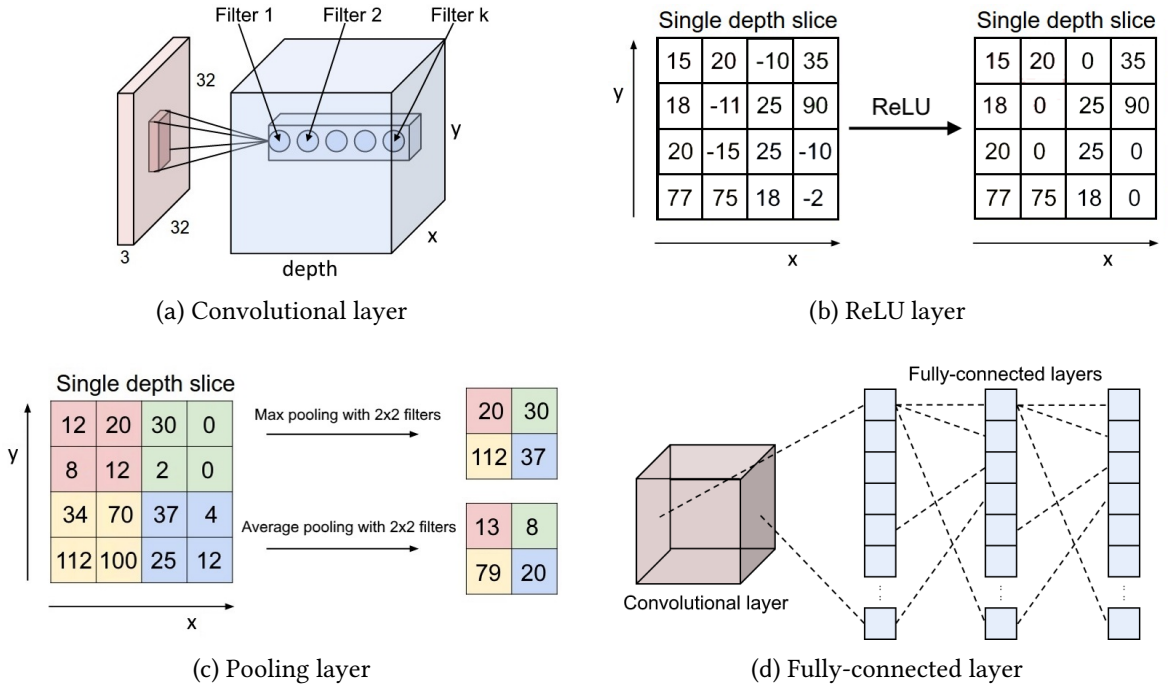
Figure 4.6: Outline of a multi-task CNN model for patient detection and skin segmentation.

4.4 Network architecture

The CNN method is used to tackle the problem of patient detection and skin segmentation. The problem is formulated as a joint task of image classification and image segmentation built into a single CNN model. The network takes a single video frame and provides a decision on whether the infant is in the scene together with the segmented skin regions if the infant is found. Our multi-task network has a shared core network, implemented using the VGG-16 architecture [137], with two output branches: the patient detection branch, implemented using global average pooling; and the skin segmentation branch, implemented using hierarchy upsampling of image features across the shared core network. All the modules in the entire network are trained simultaneously.

4.4.1 Shared core network

The shared core network was used to extract image features that are shared between the two tasks (patient detection and skin segmentation). The core network was implemented based on the VGG-16 network, originally proposed by Simonyan and Zisserman [137]. The VGG-16 network was orignally developed for image classification and was previously trained on 1.3 million images of the ImageNet dataset [137]. It has been recognised as a generic feature extractor and has demonstrated good generalisation in transfer learning towards other datasets [137].



(a) Convolutional layer (b) ReLU layer

(c) Pooling layer (d) Fully-connected layer

Figure 4.7: Graphical representations of layers employed in the VGG-16 architecture. (a) Convolutional layer. (b) Rectified linear unit (ReLU) layer. (c) Pooling layer. (d) Fully-connected layer. (Images adapted from Andrej Karpathy [129])

The VGG-16 network is a complex architecture. The network consists of five types of layers: convolutional layers, non-linear activation layers, pooling layers, fully-connected layers and a softmax layer. Appendix A contains more details on the mathematical formulation for each layer.

- **Convolutional layers** perform the convolution operation to extract features from the input data. This layer is made of a set of filters. Each filter, typically 3×3 and 2×2, is slid over every spatial location of the input data and performs a dot product between the filter and each small region of the input data, from left to right and top to bottom. The resulting output features are stacked together along the depth axis. Figure 4.7a shows a graphical representation of the convolution operation.
- **Non-linear activation layers** is an important building block that is usually placed immediately after each convolutional layer in the network. The role is to remove negative values so the important and distinct output features from the preceding layer are emphasised. One of the most common functions used in non-linear layers is the rectified linear unit (ReLU) which is defined as:

$$y = \max(x, 0) \tag{4.1}$$

where x is the input to the layer. The output size of the non-linear activation layer is the same as the input size. Figure 4.7b shows a graphical representation of a ReLU layer. This layer ignores the negative values produced by the preceding convolutional layer, thereby increasing the non-linearity property of the network.

- **Pooling layers** reduce the resolution of the output of the preceding layer. The two commonly used pooling operations are: max pooling and average pooling. For max pooling, the maximum values over small regions of the preceding layer are taken. For average pooling, in the same way, the average values over small regions are taken. Figure 4.7c shows a graphical representation of the max pooling layer. This layer allows the succeeding convolutinal layer to operate on the larger regions of the input data. They make the succeeding convolution operations more invariant towards small transformations, such as translations and rotations.
- **Fully-connected layers** are often placed later in the network, after the convolutional layers. These layers perform a weighting sum of the output of the preceding layer. Each neuron in the fully-connected layer is connected to all neurons in the preceding layer, and each neuron has its own weighting factor. These layers are often followed by a softmax layer for determining a target output. These layers, hence, act as a classifier. Figure 4.7d shows a graphical representation of the fully-connected layer. The fully-connected layer can be seen as a convolutional layer with a filter size equal to the input. Unlike the convolutional layer, there is no sliding operation of the filter across the input.
- **Softmax layers** are usually placed as a final layer of the network, after the fully-connected layers, to compute the score for multi-class classification. This layer employs a softmax function which is typically defined as:

$$y_j = \frac{e^{x_j}}{\sum_{k=1}^{N} e^{x_k}} \tag{4.2}$$

  where x is the input to the layer. This function transforms an input into values between 0 and 1. This layer returns a classification score for each class.

The VGG-16 architecture is made up of a stack of convolution layers (to extract image features) with ReLU layers (to highlight distinct features) and periodically followed by max pooling layers, as in figure 4.8. The max pooling layer reduces the spatial resolution of the feature maps by half. The structure repeats until the output has a small spatial size and a decision is made upon that output by a set of three fully-connected layers for high-level reasoning and a softmax layer for calculating class scores.

Our extension to the VGG-16 network follows that of the fully convolutional network [140]. The original VGG-16 network produces class scoring estimates. Several modifications were needed to enable the skin segmentation branch to perform pixel-level segmentation on the output of the shared core network. All fully-connected layers in the VGG-16 network were converted into convolution layers by having them perform convolution operations by sliding their filters across the input data. These layers were then able to produce a spatial output map with spatial coordinates

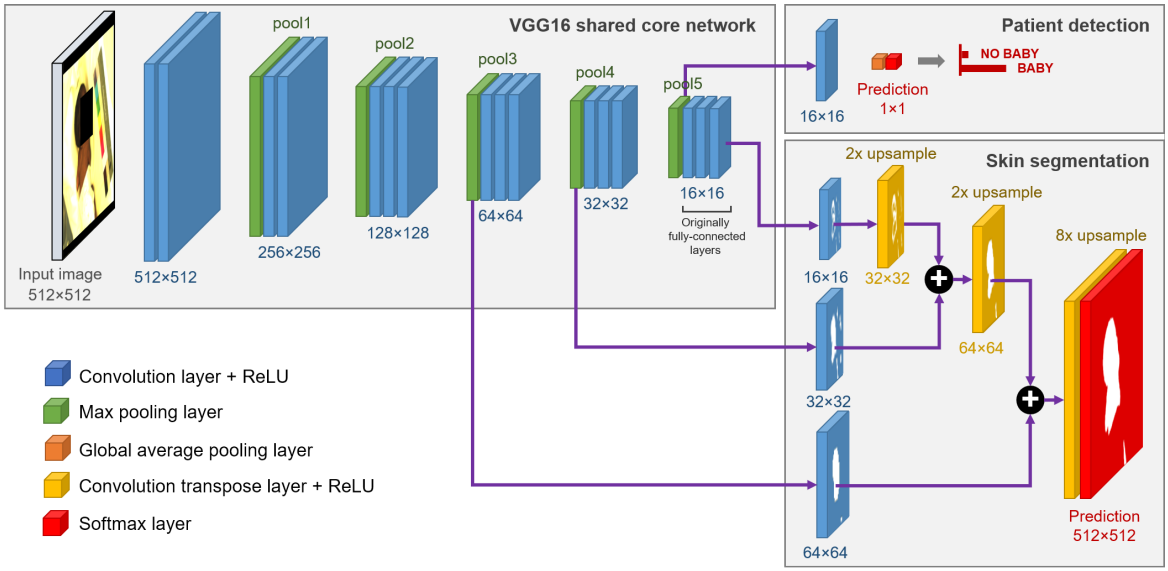


Figure 4.8: The proposed CNN model extended the VGG-16 network [137] with two branches: skin segmentation branch, implemented using a fully convolutional network as in [140]; and a classification branch, implemented using global average pooling over feature maps similar to [148]. The network was modified to evaluate the segmentation branch only if the classification branch found a subject in the image.

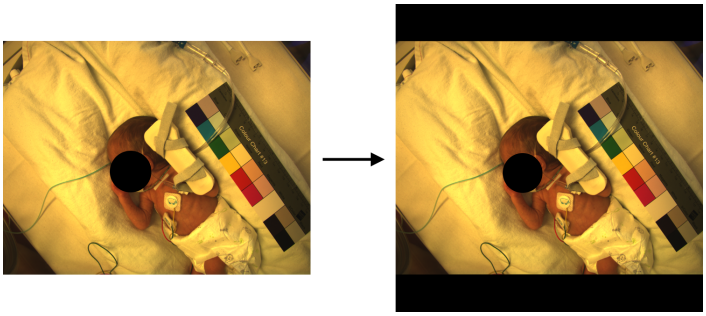


Figure 4.9: Input image was downsampled from the original size of 1620×1236 pixels to 512×512 pixels with black bars added at the top and bottom of the image to maintain the original aspect ratio of the content.

preserved (see figure 4.8). The last convolution layer was modified to produce 2-class scoring outputs for the non-skin and skin classes.

A sufficiently large input image is required to achieve accurate segmentation results [140]. However, the spatial size was limited by the amount of memory and computational time required during training. Hence, our proposed architecture was designed to take a 512×512 RGB image. The original aspect ratio of the image was maintained by adding black pixels at the top and bottom of the image (see figure 4.9).

4.4.2 Patient detection branch

The patient detection branch was implemented using global average pooling for classification, similarly to Lin *et al.* [148] and Szegedy *et al.* [138]. Global average pooling was used in the original papers to replace the use of fully-connected layers for classification. Global average pooling was implemented using an average pooling layer whose filter size was equal to the input size. It

takes the spatial average of each slide of the input data, resulting in one output for each slide. The resulting output vector is then fed to a softmax layer for calculating class scoring estimates. A major advantage of using global average pooling for classification is that there was no weight parameter to optimize [148].

In our implementation, a 1×1 convolution layer with 2 outputs was added on top of the `pool5` layer (the layer before the original fully-connected layers in the shared core network). The 1×1 convolution layer performs a linear combination across feature maps (no visual pattern being learned in this layer) in order to reduce the size of the feature dimension. The 1×1 convolution layer was followed by a global average pooling layer which averages out the spatial information resulting in an output vector fed to the softmax layer. The patient detection branch, hence, produces a two-output vector containing class scoring estimates for infant absent and present classes.

4.4.3 Skin segmentation branch

The skin segmentation branch was implemented using a fully convolutional network for image segmentation which performs a series of spatial learnable upsampling from cross-network feature maps [140]. It employs convolutional transpose layers to project feature maps onto a larger-dimensional space and produce pixel-level labelling of skin regions. The convolutional transpose layer can be seen as the convolutional layer that goes in an opposite direction. Figure 4.10 shows a graphical representation of the convolutional transpose layer.

Our implementation followed that of Long *et al.* [140]. Given that the size of the input to the network is 512×512 pixels, the feature maps of the last convolutional layer in the shared core network have a spatial size of 16×16 pixels (a factor of 32 reduction of the input size). A 1×1 convolution layer with 2 outputs was first added on top of this layer to produce a coarse prediction of non-skin and skin classes at 16×16 pixels reduction (see figure 4.8).

Next, as the feature maps of the `pool4` and `pool3` layers in the shared core network have a spatial size of 32×32 and 64×64 pixels respectively, a 1×1 convolution layer with 2 outputs was added on top of each of these layers to produce two additional predictions of non-skin and skin classes at finer resolutions, at 32×32 and 64×64 pixels respectively (see figure 4.8).

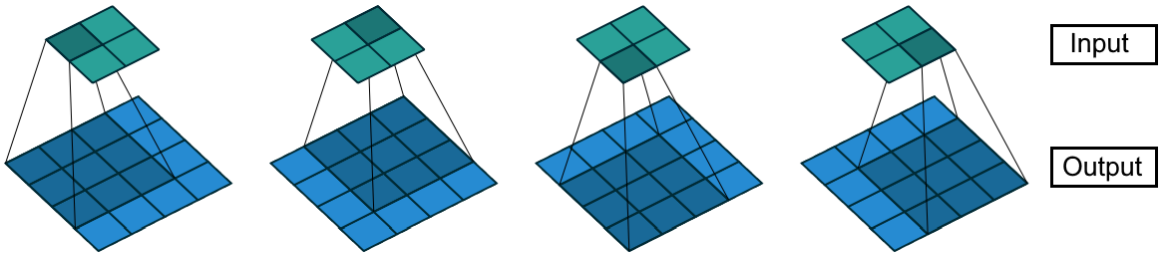


Figure 4.10: Graphical representation of the convolutional transpose layer. A 2×2 input feature map on the top is projected to a 4×4 output feature map on the bottom. This convolutional transpose layer performs 2 times upsampling. (Image adapted from Dumoulin *et al.* [152]).

Once the predictions at 16×16, 32×32 and 64×64 pixels have been processed, the next step is to combine these prediction together. The coarse prediction at 16×16 pixels is spatially upsampled through a convolutional transpose layer with a factor of 2, producing a finer prediction at 32×32 pixels. The resulting prediction is later fused with the prediction of the `pool4` layer at 32×32 pixels. Subsequently, in the same manner, the prediction fused from these two layers, at 32×32 pixels, is spatially upsampled by a factor of 2, producing a finer prediction at 64×64 pixels. The resulting prediction is then fused with the prediction of the `pool3` layer. This results in a prediction at a factor of 8 of the original resolution (64×64 pixels).

Finally, a convolutional transpose layer with a factor of 8 is added in order to obtain a final prediction at the same spatial size as the input image (512×512 pixels). The network is completed by a softmax layer which produces per-pixel class scoring estimates. The skin segmentation branch is executed only if the presence of the infant is identified by the patient detection branch.

4.5 Network training

This section firstly introduces the dataset and explains how it was constructed in section 4.5.1, followed by the pre-processing steps applied to training images in section 4.5.2. Subsequently, a multi-objective loss function used to train the model for two different tasks is described in section 4.5.3. Finally, the initialization applied prior to training is explained in section 4.5.4, followed by a description of the training procedures in section 4.5.5.

4.5.1 Dataset

The development of the patient detection and skin segmentation network required ground truth data. The first step towards assembling such a dataset was to annotate skin regions on video images obtained from the Neonatal HDU study. The images in this dataset were extracted only from the *training* set of 15 babies (see chapter 3) and annotated by three individuals. To this end, a dataset was created, consisting of positive images in which an infant is present with pixel-level skin labels provided and negative images in which the infant is absent.

Semi-automatic skin annotation

Skin annotation is a process of labelling skin regions in images. It is not efficient to annotate every frame in the dataset since multiple consecutive frames mostly contain similar information. Although changes in lighting conditions throughout the day can affect skin colours, sudden changes rarely happen (e.g. an overhead light turned on or a window blind opened). Therefore, the annotators were asked to label skin regions in a representative number of frames from the whole dataset.

The image annotation was performed using a semi-automatic approach to reduce the human effort required for labelling skin regions. A custom software package was developed for the

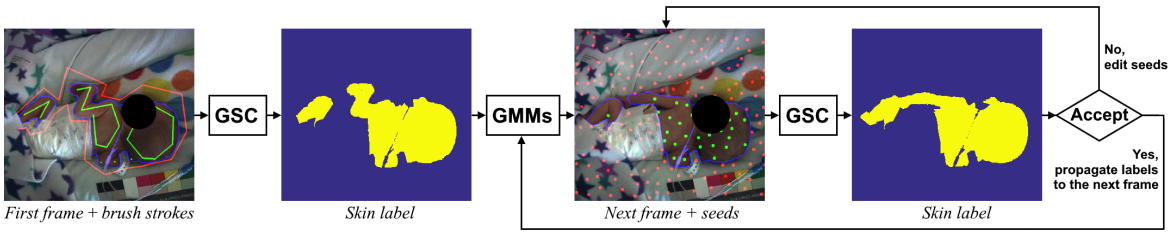


Figure 4.11: Flowchart of semi-automatic skin annotation. An annotator was asked to label skin areas in the first image of each session. The label was propagated to the next frame using GMMs. The annotator can interact with *seeds* (green and red circles) to modify the skin label.

annotation process. An annotator was required to annotate the first frame, then the annotation was propagated to successive frames. The propagation was performed using an algorithm for annotating skin regions, implemented on top of graph-cut segmentation using geodesic star convexity (GSC), as proposed by Gulshan *et al.* [153]. A Geodesic star is a star shape prior which exploits both colour and image gradient information for representing the shape of an object of interest. The GSC algorithm generates a partition on the skin and non-skin regions in the image from brush strokes provided by the annotator. Appendix B provides more details of the description and mathematical formulations of the GSC algorithm. The flow chart describing the algorithm is shown in figure 4.11.

The graph cut segmentation requires *brush strokes* of skin and non-skin pixels in order to represent hard constraints for the segmentation procedure. As discussed in [153], these brush strokes were used for creating colour models and for exploiting image structure through Geodesic stars. The colour models employed 5-component Gaussian Mixture Models (GMMs) as in the original work. Final annotation was obtained using energy minimisation subject to colour, boundary and geodesic star constraints (see appendix B for more details of mathematical formulations).

The annotation was propagated to the next image using GMMs which learned the colour models of skin and non-skin from previously segmented images. For the subsequent image, the GMMs produced the probability of each pixel being skin. These probabilities were used to generate *seeds*, which are simulated brush strokes. Skin annotation labels were computed as a result of the segmentation using those seeds. The new skin labels could be altered by the annotators modifying the seeds if they were not satisfied with the result.

The generation of segmentation seeds was as follows. First, evenly distributed N circular seeds with a radius of S pixels were randomly generated using Mitchell's Best Candidate algorithm [154] on a zero matrix that had the same size as the image to be annotated. Next, the seed matrix was spatially aligned to the probability map generated from the GMMs learned from the previous G images. For each seed in the matrix, the seed which had an average probability greater than U was regarded as a skin seed and the seed which had an average probability lower than L was regarded as a non-skin seed. All unassigned seeds were discarded. Finally, segmentation seeds

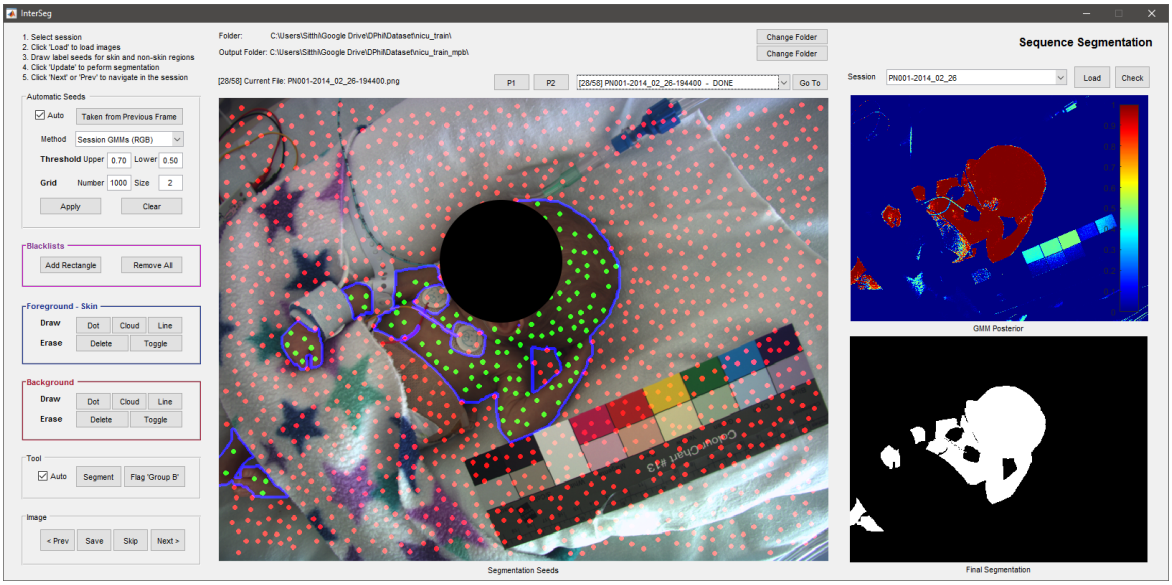


Figure 4.12: Semi-automated image annotation tool. This specifically-developed software was implemented on top of the graph cut algorithm with geodesic star convexity (GSC) [153].

were obtained. These seeds provide an effective way for the user to interact with automatically-generated labels for the subsequent images.

The annotators were allowed to adjust: the number of previous images used for training the colour models G (initially set to all previous images), the probability thresholds U and L (initially set to 0.50 and 0.70 respectively), the number of seeds N (initially set to 1,000 seeds) and the radius of circular seeds S (initially set to 2 pixels). These parameters control the sensitivity of the generation of propagated labels. For example, for a session with highly-fluctuating illumination, the annotator may want to generate a propagated label from the GMMs that learn from just one or two previous video frames in order to limit false positives. The annotation tool is shown in Figure 4.12.

Annotation protocol

For medical purposes, high-quality ground truth data are required. Using the training set of the Neonatal HDU study, images were evenly sampled in order to cover various lighting conditions. These images were obtained from the training set containing video recordings of the 15 infants (see section 3.6). Considering the trade-off between annotation effort and lighting variations, images were taken every 6 minutes. So, 2,269 images were obtained covering a total video recording time of 226 hours. The annotation was performed on the original images at 1628×1236 pixels.

Three annotators were asked to label the same set of video images. Even though the annotation was done using the semi-automated tool, all images were viewed and all labels were confirmed by the three annotators.

Construction of the dataset

The dataset was composed of *positive images* for which the infant is present and *negative images*

for which the infant is absent. The positive images had pixel-level skin annotations provided by the three annotators. The negative images were collected from periods, indicated in the electronic data records, during which the infants were not inside the incubator.

Positive Images. The annotations of skin regions from the three annotators were combined to form the positive images. Images were regarded as *positive* if more than two annotators provided skin labels without skipping the images, so each positive image has at least two different skin labels. With this criterion, 1,718 out of the 2,269 images (76%) were labelled as positive. For each image, a pixel was regarded as skin if at least two annotators agreed, otherwise the pixel was marked as non-skin. The inter-annotator agreement for skin labelling in the positive images was 96.7%. Figure 4.13a shows the histogram of inter-annotator agreements and a sample image with annotated skin regions.

Negative Images. During the studies, infants were often taken from the incubator for kangaroo care (skin-to-skin contact with their mother) or for other clinical interventions. There was a designated quiet period of about 6 hours for every 24 hours when the infants were not disturbed except in case of an emergency. During this period the lights were dimmed and the infants were allowed to rest. However, from reading the records created by the neonatal nurses, it is not clear when a subject was absent. For example, during the quiet period, the incubator's canopy was covered in order to limit the light entering the canopy; however, the camera could also be covered or left to film the subject during these time periods of low ambient light.

The focus for obtaining negative images was on the following clinical activities: kangaroo care, quiet periods, infant taken to another clinical study, camera covered, and during cannulation. For the 15 infants of the training set, these periods accounted for 23.5 hours of the 226 hours of recordings. Images were taken every 20 seconds during these periods. This generated 4,227 images for the annotator to decide. The same annotation strategy, the agreement of two out of three annotators, was utilised for the negative set. All images were presented to the three annotators. An annotator needed to identify each image as infant or non-infant. The images with two or more agreements were regarded as *negative*. With this scheme, 2,885 negative images were obtained with 99.47% inter-annotator agreement (only 12 images were marked as non-infant by just one annotator).



Figure 4.13: (a) Histogram of inter-annotator agreements based on a two-out-of-three approach on the positive images (mean agreement = 96.7%), (b) example image and (c) binary ground truth processed from three annotators.

4.5.2 Data preprocessing

All training images and ground truth labels were resized to 512×512 pixels with the aspect ratio maintained by allowing black spaces at the top and bottom of each image (see figure 4.9). This was done to ensure that the network and intermediate data fitted into the available memory of the computer used for training of the network. Mean subtraction was performed to normalise the images. The mean of each colour feature was computed across the training set and then subtracted from the three colour components across all pixels of the image. Therefore, the data was centred around the origin along each dimension. This process tends to improve the numerical condition of the network during training [146].

4.5.3 Data augmentation

Data augmentation is a method utilised to avoid overfitting and achieve the generalisation of the deep learning network [135, 155]. Multiple variations of each training image were generated. The orientation of the infant in the incubator changes at the clinical staff's discretion. In addition, infant skin colours vary according not only to the infant's phenotype but also the external lighting conditions. We employed three augmentation techniques during training: rotational, mirroring and lighting augmentations.

Rotational augmentation

CNNs have several degrees of translation and rotation invariance as a result of the convolution and pooling processes which progressively increase the level of abstraction of the image [156]. In order to encourage the network to learn rotational invariance, seven additional images were generated for each original image by rotating the image at 45-degree increments between 0 and 360 degrees without resizing.

Mirroring augmentation

In order to encourage the network to learn the symmetry of the human body, two additional images were generated by mirroring each original image with respect to the centre of the image on the x-axis and y-axis.

Lighting augmentation

Lighting conditions fluctuated a lot during the daytime as a result of changes in both natural and artificial light sources. By varying the lighting characteristics in each image, the network could be made invariant to illumination changes. Eventually, the ability to generalise for different skin colours under different lighting levels should be improved as a result of this form of data augmentation.

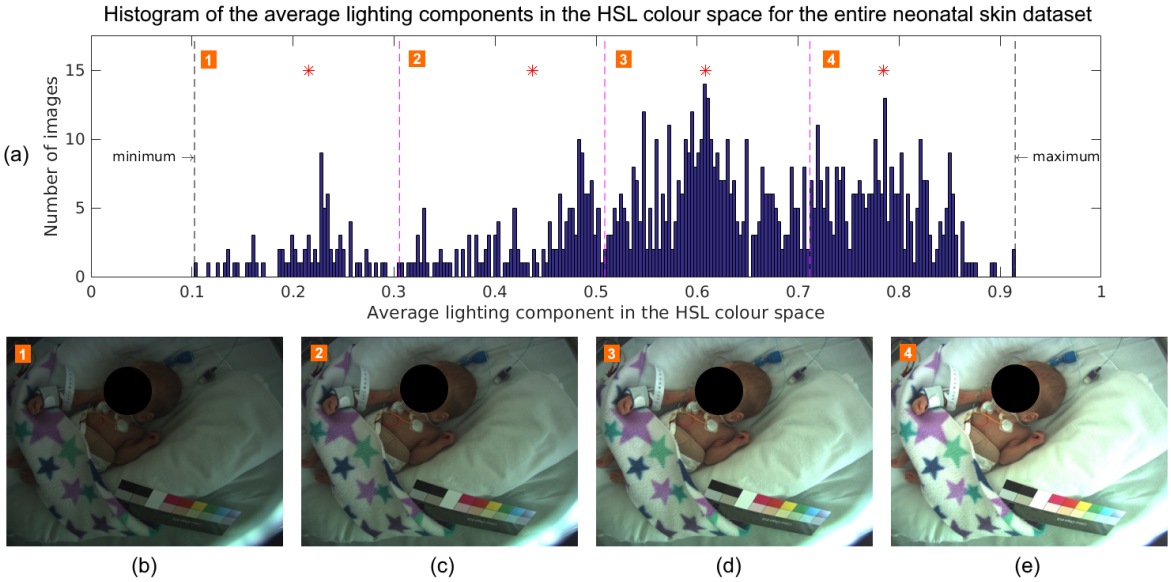


Figure 4.14: Lighting augmentation was applied to generate more training images with different lighting conditions. (a) The histogram of the average lighting components of all training images was divided into four uniform intervals. The mean of each interval is marked with a red asterisk. (b) Original image with an average lighting component corresponding to the first interval above. (c–e) Three more images were generated by scaling the lighting component of the original image to the mean of the interval (c) 2, (d) 3, and (e) 4 respectively.

The augmentation was performed by converting the original image into the HSL (Hue-Saturation-Lighting) colour space, scaling the lightness component and then converting the image back into the RGB colour space. The HSL colour space defines colour in a natural way in which the hue component H represents the shade of colour, the saturation component S specifies the intensity of colour and the lightness component L defines the brightness or closeness of colour to white [157]. The transformation between RGB and HSL is described in [157]. Our implementation used an algorithm for the conversion of the RGB colour space to the HSL colour space (and vice versa) built in the Image Processing Toolbox of MATLAB.

Figure 4.14 shows how the images were augmented. The average lightness component was calculated for each training image. A lightness range was defined by the minimum and maximum of the averages of the illuminant component computed over all the images in the training set. The range was divided into four uniform intervals and the mean of each interval was calculated. For each image, if the average lightness fell in one interval, three additional images were generated by scaling the lightness component in the HSL space using the values calculated from the three other intervals. For the interval i, the scale was defined as the ratio of the interval mean to the average lightness

$$s_i = \frac{m_i}{a} \tag{4.3}$$

where m_i is the mean of the interval i and a is the average lightness of the current image.

4.5.4 Loss functions

The whole CNN network was trained jointly using a unified multi-objective loss function as in [158]. This section first introduces two loss functions, one for the patient detection task and one for the skin segmentation task, and then describes the unified multi-objective loss function.

Patient detection

Patient detection is the task of detecting the presence or absence of the infant in the image. For this task, a loss function is introduced, which measures the discrepancy between the softmax probability estimates produced by the patient detection branch and the associated ground truth labels.

Given an input image x, the output $y_{\text{det}} = \{d_0,\ d_1\}$ of the patient detection branch is a two-class softmax probability vector. Suppose that $B = \{b_0,\ b_1\}$ is a ground truth label where $B = \{1,\ 0\}$ indicates the absence of the infant in the image and $B = \{0,\ 1\}$ indicates the presence of the infant in the image. The loss function is defined as the multinomial logistic loss of the softmax output [159]:

$$\text{Loss}_{\text{det}} = -b_0 \log(d_0) - b_1 \log(d_1). \tag{4.4}$$

Skin segmentation

Skin segmentation is the task of partitioning the image into two regions: skin and non-skin. For this task, the loss function is also defined as the multinomial logistic loss. Since the output of the skin segmentation branch is a skin label whose spatial size is equal to the input size, the loss is summed across all pixels. As the number of non-skin pixels tends to be larger than that of skin pixels, the contribution to the loss of the skin class is then weighted according to the ratio of the number of ground truth non-skin pixels, $N_{\text{non-skin}}$, to that of ground truth skin pixels, N_{skin}.

Given that $\mathcal{P}$ is a set of pixels in an input image x, the output $y_{\text{seg}} = \{s_0,\ s_1\}$ of the skin segmentation branch is a two-class softmax probability vector for each pixel, where the subscripts 0 and 1 denote the non-skin and skin classes respectively. $L = \{l_0,\ l_1\}$ is the ground-truth skin annotation where $l_0,\ l_1 \in \{0, 1\}$. The loss function of the skin segmentation branch is defined as:

$$\text{Loss}_{\text{seg}} = -\sum_{i=1}^{\mathcal{P}} l_0(i) \log(s_0(i)) - \lambda \sum_{i=1}^{\mathcal{P}} l_1(i) \log(s_1(i)) \tag{4.5}$$

where the weighting factor λ is defined as:

$$\lambda = \frac{N_{\text{non-skin}}}{N_{\text{skin}}}. \tag{4.6}$$

Total loss function

Given that x is an input image with ground truth labels $G_x = \{B_x,\ L_x\}$ for patient detection and skin segmentation labels respectively, and $f(x) = \{y_{\text{det}},\ y_{\text{seg}}\}$ is the output of the network , a

multi-objective loss function is defined as the weighted sum of the two loss functions:

$$\text{Loss}\,(f(x), G_x) = \alpha_{\text{det}}\text{Loss}_{\text{det}}(y_{\text{det}},\ B_x) + \alpha_{\text{seg}}\text{Loss}_{\text{seg}}(y_{\text{seg}},\ L_x) \tag{4.7}$$

where α_{det} and α_{seg} are weighting parameters which are defined based on the relative importance of the patient detection and skin segmentation tasks respectively in the unified loss function.

4.5.5 Model initialisation

The training of the CNN model is an iterative process usually performed using a small set of images (batch training). Prior to training, there was a need to include extra processing layers to facilitate the training process and to initialise the parameters of each layer.

Firstly, batch normalisation layers were added between the convolutional layer and the ReLU layer in the shared core network. The batch normalisation operation was employed to normalise the output of the convolutional layer to have a zero mean and unit variance by using the statistics computed from the entire batch. The batch normalisation layer can improve the generalisation of the network and reduce training time [160]. In the skin segmentation branch, dropout layers with a probability of 0.50 were added after each ReLU layer. The dropout operation was used to randomly discard activation units (or neurons) and prevent them from being used in the next layer of the forward pass of the neural network. A dropout rate of 0.50 is commonly used in CNNs. It results in an equal probability of dropping out activation units. The dropout layer can help to avoid overfitting [161].

The model was initialised with the original VGG-16's weights which hold accumulated knowledge on edges, patterns and shapes learned from the 1.3-million images in the ImageNet dataset [137]. All the new weight layers, except for the convolutional transpose layers, were initialised using the Xavier algorithm [162] with zero bias, as in [137]. The Xavier initialization process created a reasonable range of weight values that were uniformly distributed across the layers. Such an initialisation can lead to faster convergence during training [162]. The learnable convolutional transpose layers were initialised using bilinear interpolation filters, as suggested by Long *et al.* [140]. Bilinear interpolation computes each output from a weighted average of the nearest four points. The entire network was then trained from the output layers to the input layer using the backpropagation algorithm [146].

4.5.6 Training procedures

The CNN network was implemented within the MatConvNet framework presented by Vedaldi and Lenc [163]. The training was performed using standard Stochastic Gradient Descent (SGD) optimization in two stages. Two-stage training tends to reduce the amount of training time [140, 142].

The network was first trained for the skin segmentation task using only the images containing the infant with annotated skin regions. Pre-training was done using the unified loss function (see equation 4.7) with the parameters $\alpha_{\text{det}} = 0$ and $\alpha_{\text{seg}} = 1$. The learning rates were scheduled

Table 4.1: Summary of population demographics in the D_1 and D_2 groups

Set	Subjects	Sessions	Hours	Gender		Ethnicity[1]					
				Male	Female	W	B	A	WB	WA	O
D_1	8	23	118.7	5	3	5	1	1	–	–	1
D_2	7	22	107.7	3	4	5	–	–	1	1	–
Total	15	45	226.4	8	7	10	1	1	1	1	1

[1] W = White, B = Black, A = Asian, WB = Mixed White and Black, WA = Mixed White and Asian and O = Other.

to start at 10^{-2} and reduced by a factor of 10 for every two epochs until convergence, with a momentum of 0.90 and a batch size of 20. These parameters allowed the training and validation losses to gradually reduce and eventually converge to steady values. The choice of the batch size was limited by the memory of the graphics card.

The network was subsequently trained jointly for the patient detection and skin segmentation tasks using the whole dataset. The individual loss functions for each task were weighted equally: $\alpha_{\text{det}} = 1$ and $\alpha_{\text{seg}} = 1$. The learning rate started at 10^{-4} and was decreased by a factor of 10 for every two epochs until convergence, with a momentum of 0.90 and a batch size of 20.

4.6 Evaluation protocol

The Neonatal HDU study protocol allowed algorithms to be developed on only 15 infants in the *training* set (see chapter 3). Although the dataset has the necessary number of images suitable for the development of the algorithms, the images were collected from very few subjects.

An approach to obtaining predictive performance is cross-validation using two independent folds. The training set was firstly divided into two groups, D_1 and D_2, such that one group had 8 subjects and the other group had 7 subjects. The assignment to different sets was based on a balance of choice between skin phenotype, gestational age and the number of positive images. For each set, positive images were taken directly from the positive pool, and negative images were randomly sampled without replacement from the negative pool so that the number of positive and negative images were equal.

Table 4.1 presents a summary of the statistics for the two sets. A model was first trained on D_1 and validated on D_2. Then, another model was trained on D_2 and validated on D_1. The validation results from both models were combined to produce the overall predictive performance. The main advantage of this approach is that all images are used for both training and validation.

4.6.1 Evaluation metrics

For the patient detection task, the classifier's performance is described using the Receiver Operating Characteristics (ROC) curve, accuracy, precision, true positive rate (TPR or recall) and true negative

rate (TNR or specificity), defined by:

$$\text{Accuracy} = \frac{\sum \text{TP} + \sum \text{TN}}{\sum \text{Total population}} \tag{4.8}$$

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{Predicted positives}} \tag{4.9}$$

$$\text{True positive rate (TPR)} = \frac{\sum \text{TP}}{\sum \text{Actual positives}} \tag{4.10}$$

$$\text{True negative rate (TNR)} = \frac{\sum \text{TN}}{\sum \text{Actual negatives}} \tag{4.11}$$

$$\text{False positive rate (FPR)} = \frac{\sum \text{FP}}{\sum \text{Actual negatives}} \tag{4.12}$$

where a true positive (TP) is a correctly-identified positive sample and a false positive (FP) is an incorrectly-identified positive sample.

The ROC curve is generated by plotting TPR versus FPR for different patient detection scores. The area under the ROC curve or AUC provides a single measure to describe overall performance.

For the skin segmentation task, a pixel-wise intersection over union (IOU), which is the standard metric for evaluating a segmentation algorithm, is used to describe the segmentation performance. The IOU metric is defined as:

$$\text{IOU} = \frac{y_p \cap y_g}{y_p \cup y_g} \tag{4.13}$$

where y_p denotes a predicted segmentation result and y_g denotes a ground truth label. Since we are only concerned with the segmentation of skin, only the IOU metric of skin labels is calculated and the IOU metric of non-skin (background) labels is ignored.

4.7 Results

In this section, we present the results of colour-based skin filters commonly used in the literature for comparison, followed by the results of our CNN model for patient detection and skin segmentation.

4.7.1 Baseline results

We implemented colour-based skin colour filters, which were previously adopted in several non-contact estimation algorithms [112, 164], as the baseline for skin segmentation using three classifiers: Naive Bayes [165], Random Forests [166] and GMMs [159]. The skin filters classify each pixel as a skin pixel based solely on skin colours and provide a skin probability map, which can be thresholded to a binary label. The skin models were trained on the training images that were converted to the HSL colour space [167] with white balance correction applied [168, 169]. The white balance correction

can reduce the effect of variations in skin colours [167]. Data augmentation was not applied to the training images. Patient detection was performed using the ratio of skin to non-skin pixels and the average probability of predicted skin pixels to make a decision, as in the method described in [165].

Table 4.2 shows that the Naive Bayes classifier achieved the highest accuracy in the patient detection task. It achieved 1.0% and 1.6% higher accuracy than Random Forests and GMMs respectively. In term of AUCs, GMMs obtained the highest score (98.8) followed by Naive Bayes (98.1) and Random Forests (97.2) respectively. For the skin segmentation task (see table 4.3), Random Forests achieved the best performance for skin segmentation, with 4.7% and 14.5% improvements in IOUs with respect to GMMs and Naive Bayes respectively.

4.7.2 CNN results

The multi-task CNN model trained with data augmentation outperformed other models for majority of metrics in both patient detection and segmentation tasks (see tables 4.2 and 4.3). For patient

Table 4.2: Patient detection performance of the baseline skin filters and the CNN models.

Model	AUC	Accuracy	Precision	Recall	Specificity
Baseline skin filters			*All values are expressed as a percentage*		
Naive Bayes	98.1	98.6	**97.8**	99.4	**97.8**
Random Forests	97.2	97.7	97.5	97.9	97.4
GMMs	98.8	97.1	97.8	96.4	**97.8**
CNN without data augmentation					
CNN patient detection only	97.7	98.0	96.1	**100.0**	96.0
Multi-task CNN	**99.7**	98.2	96.6	**100.0**	96.5
CNN with data augmentation					
CNN patient detection only	97.9	97.1	96.0	98.3	95.4
Multi-task CNN	98.2	**98.8**	97.6	**100.0**	96.8

Table 4.3: Skin segmentation performance of the baseline skin filters and the CNN models

Model	Pixel Accuracy			Intersection over Union		
	Mean (SD)	Min	Max	Mean (SD)	Min	Max
Baseline skin filters			*All values are expressed as a percentage*			
Naive Bayes	89.5 (8.3)	32.7	98.9	61.3 (17.4)	4.3	92.9
Random Forests	95.0 (4.6)	57.7	99.3	75.9 (16.1)	6.8	95.4
GMMs	93.4 (5.2)	47.5	99.1	71.2 (14.2)	16.8	94.7
CNN without data augmentation						
CNN skin segmentation only	92.2 (3.4)	71.1	74.4	57.4 (15.2)	0.00	84.5
Multi-task CNN	96.2 (2.0)	**75.9**	98.9	77.2 (9.9)	4.8	92.9
CNN with data augmentation						
CNN skin segmentation only	97.9 (1.2)	88.7	99.5	87.8 (6.0)	**49.4**	96.5
Multi-task CNN	**98.1 (1.9)**	75.6	**99.6**	**88.6 (7.5)**	39.0	**97.0**

Performance evaluated only on positive images with the presence of a subject.

detection, the model achieved an accuracy of 98.8% and an AUC score of 98.2%. For skin segmentation, the network yielded an IOU score of 88.6% and a pixel accuracy of 98.1%.

Compared to the best of the colour-based algorithms (Naive Bayes), the multi-task model trained with data augmentation achieved a 3.1% higher pixel accuracy and a 12.7% higher IOU score. Figure 4.15 shows the confusion matrix describing the classification performance of the colour-based algorithms and the CNNs. Figure 4.16 shows the skin segmentation results for several images in the dataset using each of the different approaches.

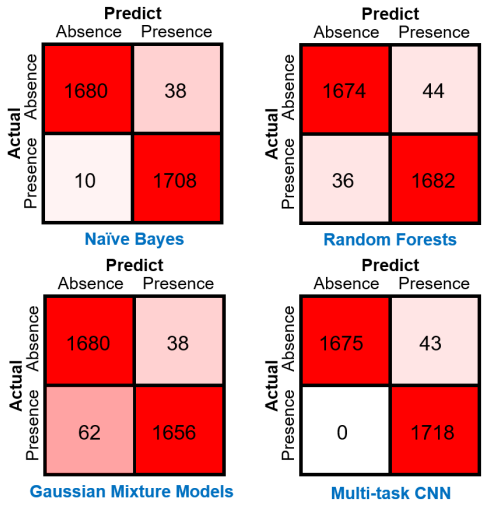


Figure 4.15: Confusion matrix of each approach to the patient detection task. GMMs has the highest classification errors.

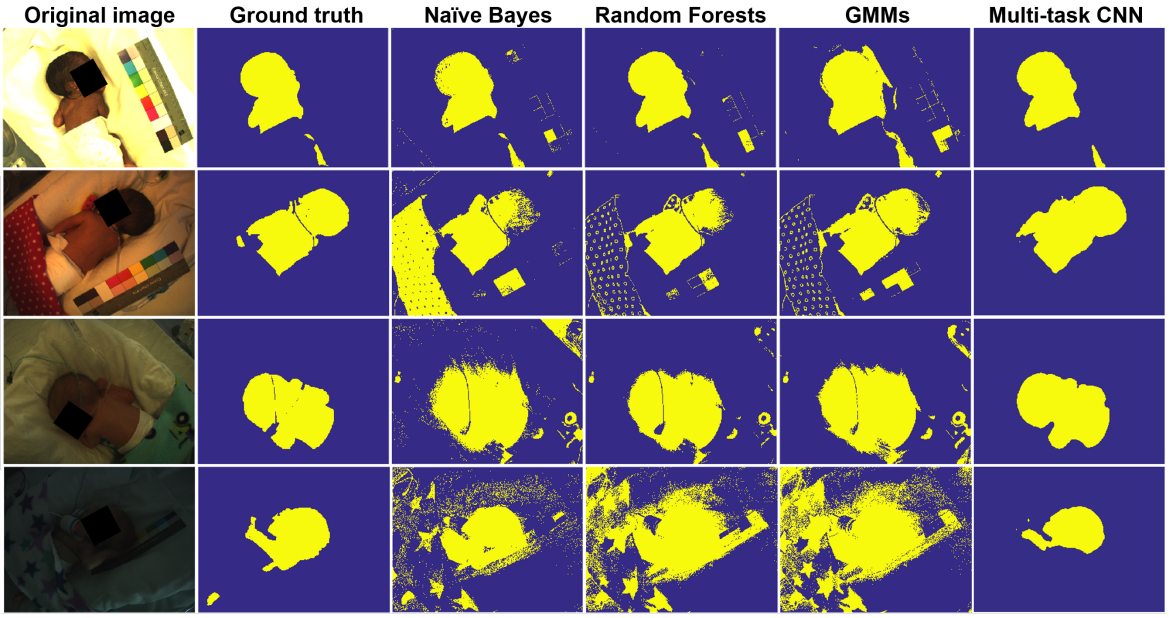


Figure 4.16: Example results for skin segmentation using both baseline skin filters and the proposed CNN. Among baseline skin filters, random forests outperformed other classifiers in terms of quantitative assessment; however, they did not perform well in low light scenarios. All baseline skin filters over-segmented the skin and generated false positives whose colours were similar to skin colours. The multi-task CNN model produced more accurate skin labels.

4.7.3 Analysis of design choices

Experimental results for several network variants are detailed below. These results have an influence on the design of the proposed network.

Patient detection

Two network variants for the patient detection branch were examined: traditional two fully-connected (FC) layers and global average pooling.

For the first variant, the first FC layer with 512 outputs was connected to `pool5` of the shared core network (see figure 4.8). The second FC layer with 2 outputs was connected to the first FC layer. This variant achieved an accuracy of 99.0% and an AUC of 99.5% for classification. Its segmentation scores were 98.7% for pixel accuracy and 85.2% for the IOU metric.

For the second variant, global average pooling [148], which enforces the correlation between feature maps and each class, was investigated. This variant achieved an accuracy of 98.8% and an AUC of 98.2% for classification, and a pixel accuracy of 98.1% and an IOU score of 88.6% for segmentation respectively. This variant achieved marginally lower scores in classification metrics than the first variant but for segmentation there was a 3.4% increase in the IOU score. Hence, the second variant (global average pooling) was chosen in the proposed multi-task network. Although this variant may reduce the complexity of the classification branch, the network still has enough complexity to represent the two-class classification problem.

Skin segmentation

Table 4.4: Skin segmentation performance on different network variants

	SegNet	Fully convolutional network			
		VGG-M	VGG-16	ResNet-50	ResNet-101
Mean Pixel Accuracy	93.3	94.9	**97.9**	97.2	97.1
Mean IOU	70.9	73.3	**87.8**	85.9	85.6
Speed (fps)	4.5	17.6	**9.9**	21.3	11.8

All values are expressed as a percentage.

Two network variants were examined for the skin segmentation task: fully convolutional network [140] and SegNet [142]. The SegNet model was found to achieve a lower IOU score of 70.9%, compared to the score of 87.8% for the fully convolutional network. In addition to the lower performance, the SegNet model produced noisy and grainy segmentation results, especially in low-light video images.

The implementation of the fully convolutional network was investigated on four pre-trained networks, VGG-M, VGG-16, ResNet-50 and ResNet-101, in order to observe the trade-off between speed and performance among different network variations. The results are shown in Table 4.4. The fully convolutional network implemented on the VGG-16 network had the highest performance.

4.8 Discussion

4.8.1 Qualitative assessment

Although the random forests algorithm yielded the highest performance among the baseline algorithms, it performed poorly in low-light conditions (see figure 4.16). As the illumination decreased, colours were distorted and the difference between skin and non-skin colours became less distinguishable [170]. The baseline skin filters were applied to images on a pixel-by-pixel basis and, since they only take colour information into consideration, noise was to be expected due to the analogue-to-digital conversion process used by video cameras.

The multi-task CNN model correctly segmented skin regions under normal ambient light (see figure 4.16). The model had some difficulties with low-light images; however, it performed better than the colour-based skin filters. In most of the negative images, the models did not produce skin labels when an infant was not present. The CNN network did not produce noisy and grainy skin labels as it processed the whole image at once. As expected, the CNN model cannot identify small skin regions as the CNN architecture periodically downsamples feature maps in the network.

The CNN model had some difficulties in very low-light conditions as it under-segmented skin regions; however, its results were better than those of the baseline algorithms.

One may argue that the CNN model without a Conditional Random Fields (CRF) layer as in [149, 151] may not be able to produce fine-grained segmentation; however, our results showed that the algorithm can produce accurate skin labels. This was due to a large image input (512×512 pixels) and a large skin area in the image ($15-20\%$ of the image pixels on average corresponding to the skin area).

4.8.2 Network training

In the first training stage, during which the network parameters were learned for the skin segmentation task only, the network converged after epoch 6. Training took 12 hours with a training speed of two images per second. In the second training stage, during which the network was further trained jointly for both patient detection and segmentation tasks, the patient detection branch was converged by epoch 5. This was likely caused by small variations in the negative images. The decrease in training and validation loss is shown in figure 4.17. The second training stage took 16 hours. Therefore, the whole network took around 28 hours to train on a single Nvidia GTX Titan 6GB GPU. The network trained with data augmentation was found to have lower training and validation losses than the network trained without data augmentation.

One of the common problems occurring during the training of neural networks is overfitting. It occurs when the error on the training set becomes very small. The network has learned to memory examples in the training set, but the network has not learned to generalise towards new examples. When the network starts to overfit the training examples, the error on the validation set starts to

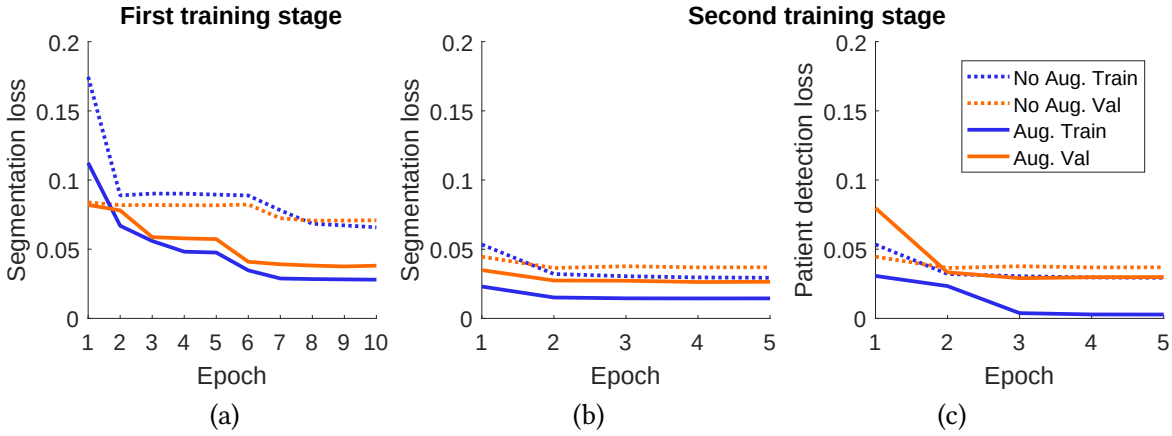


Figure 4.17: The decrease in training and validation loss over epochs in (a) the first training stage and (b–c) second training stage. The models trained with data augmentation had lower training and validation loss.

increase. The chance of overfitting is high when the network is trained with a limited number of training examples. We did not find overfitting to be a big problem in this study. Both training and validation losses were converged after several epochs of training.

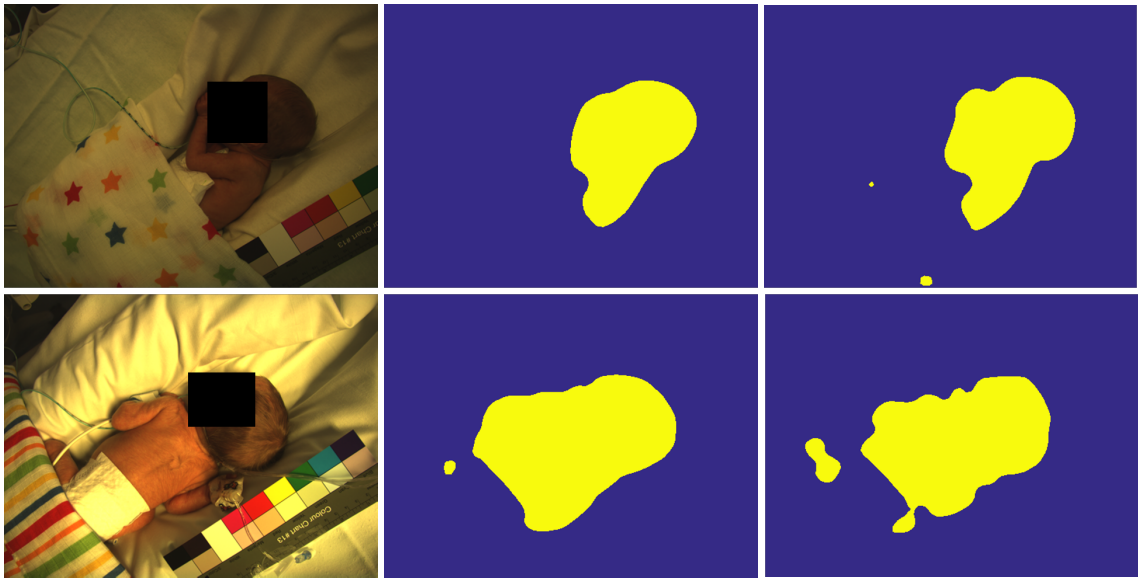
4.8.3 Effects of multi-task learning

The multi-task CNN model exhibited similar performance to the CNN models trained individually for a single task. The joint network achieved an improvement of 1.7% in accuracy and 0.3% in AUC score for patient detection and an improvement of 0.9% in IOU score for skin segmentation compared to that of the single-task networks. The increase in performance may be due to the fact that the network was fine-tuned with lower training rates for several epochs (see figure 4.17 for the decay in training and validation loss). Similar results were observed in the multi-task network of Gkioxari *et al.* [158]. As expected, the joint network did not show a bias towards one individual task. The multi-task network can perform both tasks two times faster than a cascade of two single-task networks.

4.8.4 Effects of data augmentation

Data augmentation was found to substantially improve the performance of the network (see table 4.3) as well as the quality of segmentation results (see figure 4.18). With data augmentation applied, 30.5% and 11.4% improvements in IOU were observed for the single-task model and the multi-task model respectively. Without data augmentation (see figure 4.18b), the CNN produced coarser segmentation results. This might be due to the small number of patients in our dataset, making it difficult for the network to learn a generic structure of the objects of interest.

Patient detection performance in the models trained with and without data augmentation was found to be similar. This is likely to be because the patient detection task is much simpler than the skin segmentation task.



(a) Original image (b) No data augmentation (c) Data augmentation

Figure 4.18: Image segmentation of (a) the original image on the CNN models trained (b) without and (c) with data augmentation applied on training the images. The model trained with data augmentation produced skin labels with better segmentation detail.

4.8.5 Runtime evaluation

The network can process a 512×512 image at a rate of 9.9 images per second if a subject is found in the image, and at a rate of 13.4 images per second if the subject is absent from the image. In practice, one could process the video recordings in the Neonatal HDU dataset at a reduced spatial size of 512×384 pixels with the original aspect ratio maintained by cropping a small portion (4 pixels) at the top and bottom of the image. A 512×384 video frame can be processed at a rate of 10.9 frames per second if the subject is found and at a rate of 16.4 frames per second if the subject is absent. Note that images were read directly from a magnetic hard drive. Real-time performance could be achieved for video processing with dual GPUs, a lower sampling rate or a smaller image size.

4.9 Conclusion

This chapter has proposed a multi-task CNN model for patient detection and skin segmentation. The model consists of the VGG-16 convolution layers with two processing branches for the patient detection and skin segmentation tasks. The whole network was trained using images labelled by three annotators. The proposed network achieved high performance on both tasks and demonstrated robustness against pose variations and illumination changes.

Overall, this approach is capable of determining the regions of interest for continuous non-contact vital sign monitoring. The accurate segmentation of skin regions is essential for the successful estimation of vital signs.

Chapter 5
Intervention detection

5.1 Introduction

Chapter 4 presented the convolutional neural network (CNN) model for patient detection and skin segmentation. However, there is another question that needs to be answered before estimating vital signs in an incubator: is the infant undergoing a clinical intervention? Interventions can cause severe motion artefacts which may prevent the estimation of vital signs from video camera data.

This work was inspired by similar work in video classification using deep learning techniques. In previous studies, Karpathy *et al.* [143] investigated different strategies for extending the connectivity of a CNN architecture for the task of video classification by fusing the information of multiple colour video frames over time. The authors proposed a multi-resolution fusion architecture that has two separate processing streams: the first stream operates on entire video frames and the other stream operates on the middle portion of the video frame. Simonyan et al. [144] introduced a two-stream CNN architecture which consists of spatial and temporal processing streams. The spatial stream operates on an individual video frame, whereas the temporal network operates on a stack of dense optical flows computed between successive frames. Feichtenhofer *et al.* [145] conducted an extensive investigation on different approaches for the fusion of spatial and temporal processing streams and proposed a two-stream spatiotemporal fusion network which uses convolutional operations to combine feature maps from the last layers of the spatial and temporal streams.

In this chapter, the proposed patient detection and skin segmentation network in chapter 4 was extended to identify time periods during which a clinical intervention occurred. The chapter begins with a description of typical interventions that pre-term infants experience during their hospital stay in section 5.2. An overview of the proposed methods for detecting such interventions is presented in section 5.3. A CNN model for analysing temporal information in videos is introduced in section 5.4. Section 5.5 discusses different fusion approaches for combining information from the patient detection and skin segmentation network with temporal information processed by this new network. Section 5.6 describes a training dataset and the procedures used to train the CNN models

Table 5.1: Daily nursing activities for pre-term infants in the NICU
(Data provided by research nurses in the Neonatal HDU at the John Radcliffe Hospital)

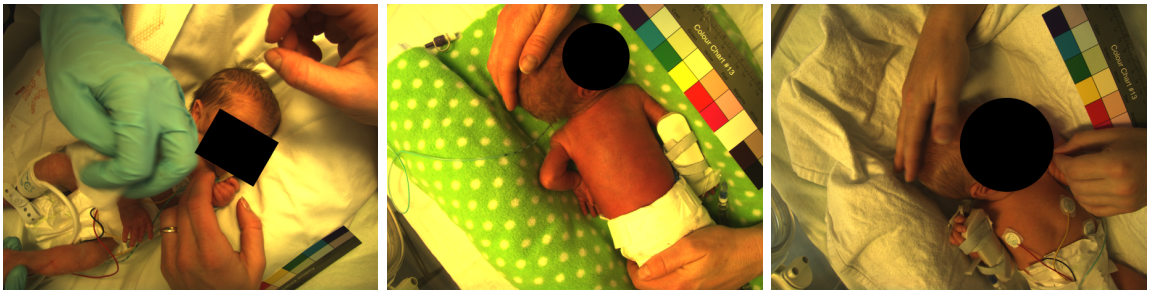
Time	Event
At nurse shift handover (typically every 8–12 hours)	• Lift incubator cover to look at the baby • Examine nasogastric tube (NGT) placement • Examine central venous line (CVL)
As required	• Check emergency equipment • Check ventilation equipment • Check infusion pumps • Replace ECG leads and nasal probes
After bradycardia, O_2 desaturation or apnoea	• Stimulate by touching • Change infant position
Every hour	• Take aspirate (fluid from a body cavity) • Give nasogastric tube (NGT) feed • Record vital sign parameters
Every 6 hours	• Take temperature and blood pressure • Change skin probe sites • Change infant position
Every 12 hours	• Take infant out of incubator for cuddles
Every 6–8 hours if under phototherapy Every 2–6 hours if hypoglycaemic	• Heel prick for blood test
Every 6–12 hours	• Give oral medication
Every 4–12 hours	• Give intravenous line (IV) medication

on this dataset. Section 5.7 explains the protocol used to evaluate the networks. In section 5.8, the performance of each fusion strategy is reported and compared. A discussion on fusion strategies and their failure modes follows in section 5.9. Finally, section 5.10 summarises the chapter.

5.2 Clinical interventions in the Neonatal HDU

In a high-dependency care environment, pre-term infants experience routine clinical interventions several times a day, for example: checking the normal functionality of medical equipment, changing a nappy, taking temperature readings, administering medications or withdrawing blood from the heel for a blood gas test. Table 5.1 summarises the activities carried out by the nurses in the care of the pre-term infants in the Neonatal HDU. Parents also visit their newborn baby regularly. Figures 5.1a–b show some sample images during these events.

Intervention detection in the Neonatal HDU is a challenging problem. During these events, clinical staff or parents actively interact with the infant causing motion artefacts that pose challenges to the estimation of vital signs from video camera data. The appearance and motion of clinical staff or parents in the video during an intervention are spontaneous and change from time to time,



(a) (b) (c)

Figure 5.1: Examples of typical interventions. (a) Clinical intervention performed by nurses. (b) Parental touch by mother during a visit. (c) Checking ventilation equipment and other sensors.

depending mainly on the clinical status of the infant. In addition, the infants in our study were filmed in an unconstrained environment in which the position and orientation of the camera were not fixed. The lighting conditions also changed throughout the day and with the season of the year.

Since the incubator was made of a transparent material, any events happening around the incubator were also captured in the video. The intervention detection algorithm needs to identify these time periods and discard them when estimating vital signs. An ideal algorithm should be able to detect intervention periods without being confused by other periods with substantial movement (e.g. when the infant is crying) or periods when there are changes due to external factors (e.g. when clinical staff are walking around the incubator).

5.3 Method overview

This chapter presents the extension of the patient detection and skin segmentation network developed in chapter 4 for the intervention detection task (see figure 5.2). The intervention detection extension uses the information processed in the patient detection and skin segmentation network together with temporal information extracted from multiple-frame optical flow to identify the occurrence of interventions.

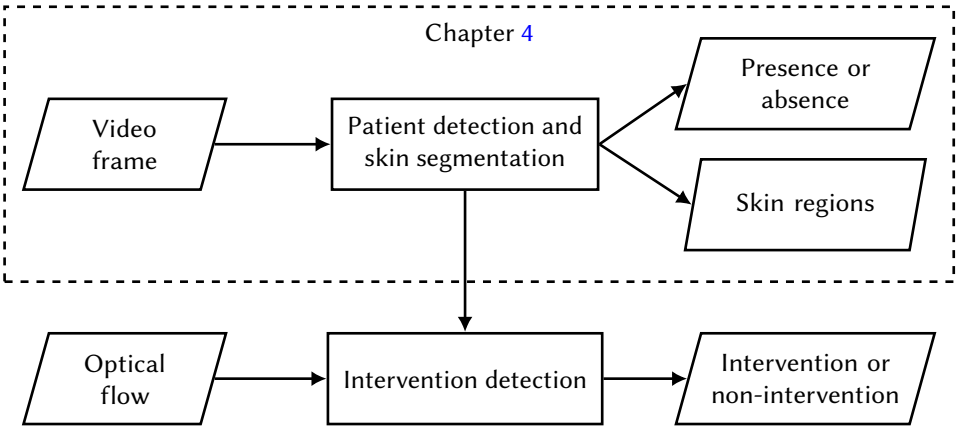


Figure 5.2: The outline of the CNN models used in this thesis. This chapter presents several variants of network architectures inside the intervention detection block.

In this chapter, we first introduce the optical flow network and describe how multiple video frames were processed through this network in section 5.4. In section 5.5, we then discuss three fusion strategies for combining the temporal information from the optical flow network with the information from the patient detection and skin segmentation network for detecting a clinical intervention event: (1) *spatio-temporal fusion*, which directly merges spatial information extracted from single or multiple RGB video frames with temporal information from optical flow; (2) *multi-resolution temporal fusion*, which combines the information from the optical flow computed from the entire frame with the information from the optical flow computed from the image area from which the patient was segmented; and (3) *temporal context fusion*, which combines the information from multiple-frame skin confidence maps with multiple-frame dense optical flow.

5.4 Optical flow network

While the patient detection and skin segmentation network operates on a single RGB image input, the optical flow network operates on multiple dense optical flows. Optical flow is the pattern of motion of objects in a visual scene calculated between two image frames. It comprises a 2-dimensional vector field containing the displacements of points from the first image to the second image (see appendix C for more details on the mathematical formulation). Figure 5.3 shows the optical flow computed from two sample video frames. The optical flow consists of the displacements in the horizontal and vertical directions.

5.4.1 Sliding window operation

In order to identify intervention periods during long video recordings (6 – 8 hours per video), a sliding-window approach was used to process the video sequence with a fixed window length T

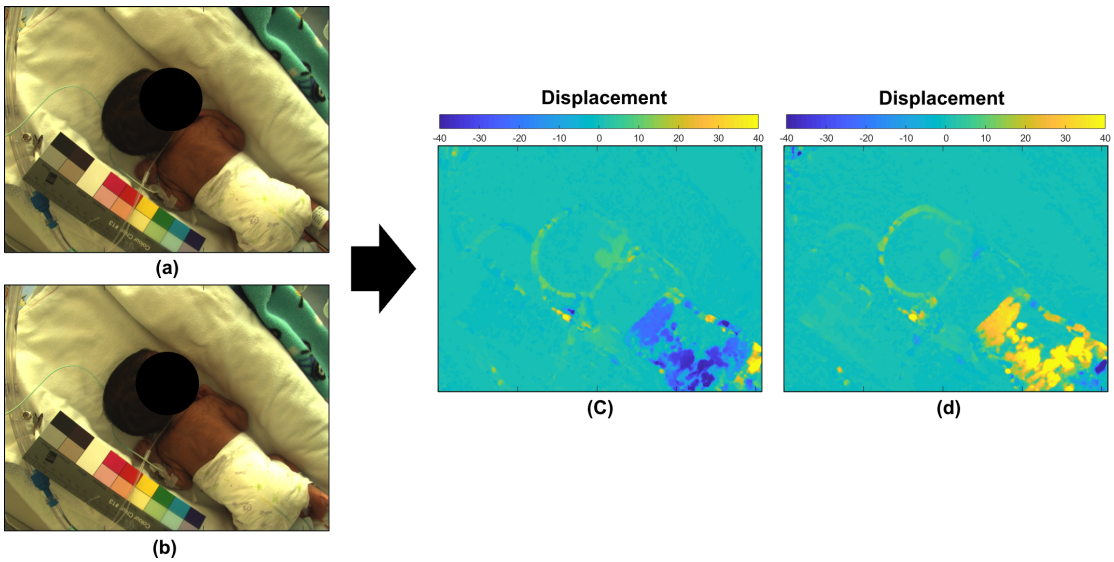


Figure 5.3: Optic flow calculation. (a) First image frame. (b) Second image frame. Optical flow in (c) horizontal direction and (d) vertical direction.

and a step size τ. Given that F_s is the number of frames in each second, each sliding window v holds a total of $T \times F_s$ consecutive video frames. Note that $\tau < T$ results in overlapping sliding windows. We evaluated different window lengths and different step sizes to find an appropriate configuration for the implementation of the optical flow network.

5.4.2 Optical flow configuration

Unlike the original implementation of the two-stream network in [144, 145] that used the optical flow between frame I_i and the next frame I_{i+1}, we instead used the optical flow between frame I_i and frame I_{i+F_s} where F_s is the number of frames in each second. So, the optical flow was computed every second. This was done to ensure that the implementation of the network was possible for the video recordings of the Neonatal HDU study that contains over 32.5 million video frames.

Given a T-second sliding window v, $L + 1$ video frames were first extracted, one image every second (see figure 5.4). The optical flow computes the motion of pixels between two consecutive extracted frames. Therefore, L optical flows were computed from $L + 1$ video frames. Each optical flow consists of displacement vectors in the horizontal and vertical directions at the same spatial size of the video frame. The horizontal and vertical components of each optical flow vector were stacked together across input channels, as suggested by Simonyan and Zisserman [144]. Figure 5.4 illustrates the calculation of optical flows using the sliding window scheme.

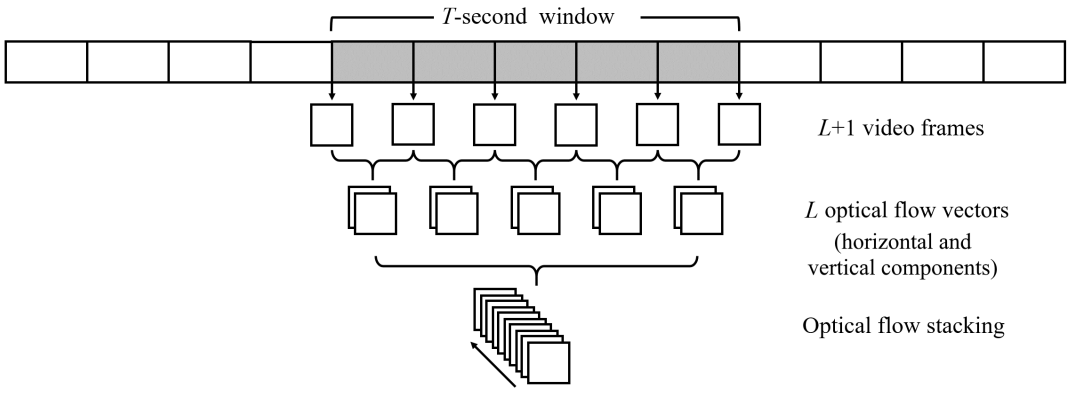


Figure 5.4: The processing of the input to the optical flow network. For each T-second window, $L + 1$ video frames were taken, one image every second. L optical flow vectors were calculated, and their components (horizontal and vertical) were then stacked together.

5.4.3 Model architecture

The optical flow network was implemented upon the ResNet50 network with 50 weighted layers, as proposed in [139] (see figure 5.5). Even though the number of weighted layers in the ResNet50 network is higher than the VGG16 network used in chapter 4, each layer in the ResNet50 is smaller and has fewer number of parameters. The implementation of the optical flow network using the ResNet50 network allows the task to be performed with fewer parameters and a lower amount of computational resources compared to the VGG networks used in [144, 145].

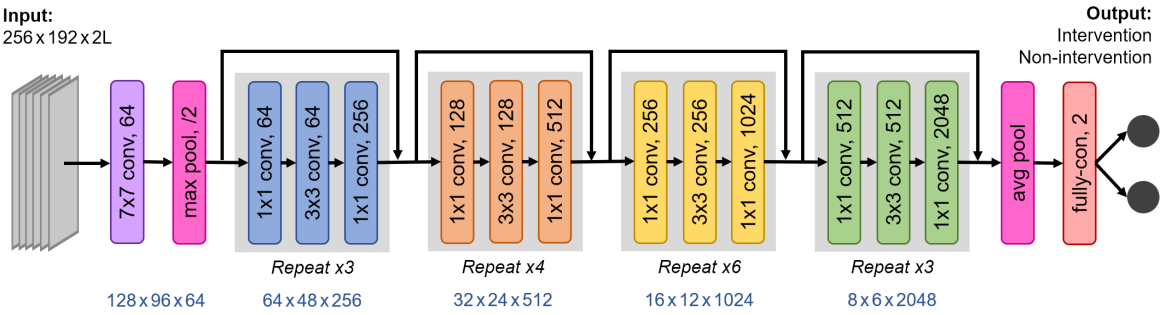


Figure 5.5: The optical flow network is built on the ResNet50 architecture [139] with 50 weighted layers. There are four sets of building blocks as shown in different colours with the numbers of blocks repeated below. Each building block consists of three small convolutional layers. The size of the feature maps (Width × Height × Channels) at each stage is shown at the bottom. The ResNet50 architecture has a shortcut connection from the previous block to the next block. The shortcut connection helps in training a deep network.

Instead of accepting a single RGB image, the architecture was modified to accept a stack of $2L$ dense optical flow components. The first convolutional layers were extended to $2L$ channels by stacking the spatial average of the first convolutional filters across the channels. In order to maintain the aspect ratio of the Neonatal HDU videos, the network was designed to take a $256 \times 192 \times 2L$ volume as an input. The input size was limited by the computation time and the memory requirements for the workstation used to develop the algorithms. The last fully-connected layer was modified to produce two outputs for intervention and non-intervention labels, and its filters were re-initialised with the Xavier algorithm with zero bias [162] (see section 4.4.5). The dropout ratio was changed to 0.85 as suggested in [145]. The output of this network was the classification score of intervention and non-intervention events.

5.5 Fusion strategies

This section describes the three strategies used for combining information from the patient detection and skin segmentation network with temporal information from the optical flow network for intervention detection: (1) spatio-temporal fusion; (2) multi-resolution temporal fusion; and (3) temporal context fusion.

One modification was made to the optical flow network prior to the fusion stage. When combining the feature maps from the optical flow network with those from another network, the classification part of the original optical flow network was no longer required. Therefore, the average pooling and fully-connected layers (the last two layers in figure 5.5) were removed from the optical flow network. The optical flow network outputs $8 \times 6 \times 2048$ feature maps.

5.5.1 Spatio-temporal fusion

Spatio-temporal fusion combines appearance information extracted from RGB video frame(s) with temporal information extracted from the multiple-frame optical flow network described above to

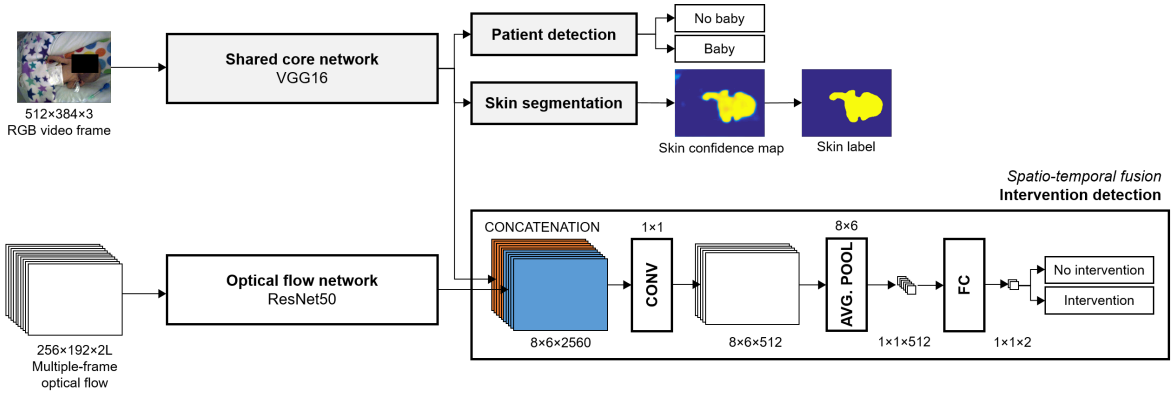


Figure 5.6: Spatio-temporal fusion combines feature maps from the patient detection and skin segmentation network with the optical flow network using a series of concatenation and convolution layers.

predict the presence of an intervention during a specific time window. Figure 5.6 shows the overall network architecture. Note that, in this implementation, the patient detection and skin segmentation network operates on a 512×384 input RGB video frame. The original aspect ratio of the video recordings was maintained by cropping out a small portion at the top and bottom of the video frame (see figure 4.9). This allows the network to operate faster and reduces computational resources. The appearance features were extracted through the shared VGG16 convolution layers in the patient detection and skin segmentation network, whereas the temporal features were extracted through the optical flow network. Both types of features were combined in the intervention detection branch. In this fusion strategy, the VGG16 convolution layers were used as a feature extractor without re-training to reduce the training complexity.

Model architecture

Several modifications were made to integrate the two networks together. In the patient detection and skin segmentation network, an additional 4×4 max-pooling layer was added after the last convolution layer of the shared core network (see figure 4.6) in order to produce an output with a 8×6 spatial size to match the output of the optical flow network.

In the intervention detection branch (see figure 5.6), the outputs from the patient detection and skin segmentation networks and the output from the optical flow network were fused together through a series of concatenation and convolution layers as in [145]. Concatenation is a process of stacking two or more feature maps with the same spatial size together across channels (see figure 5.7). Both output feature maps were stacked together and then convolved with a 1×1 convolution layer with 512 feature channels, producing $8 \times 6 \times 512$ feature maps (see figure 5.6). The convolution layer performed weighted combinations of spatial and temporal feature maps and reduced the dimension of the combined feature channels. This layer could learn information corresponding to a decision-making process from both networks [145]. The network ended with a

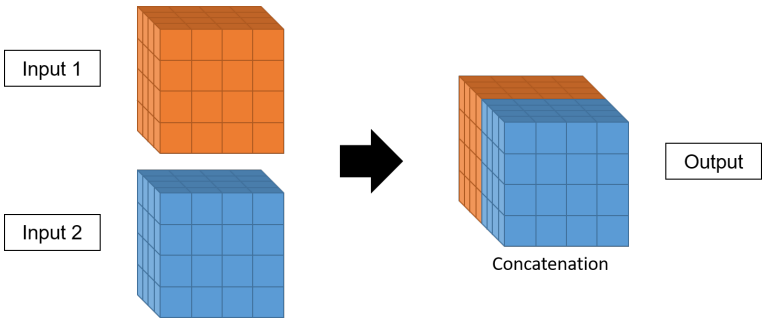


Figure 5.7: Graphical representation of the concatenation operation

global average pooling layer, 2-way fully-connected and softmax layers. The use of global average pooling and fully-connected layers for classification enforces the correspondence between feature maps and categorical outputs [139]. The last layer provides classification scores to distinguish between non-intervention and intervention events.

In the spatio-temporal fusion approach, we investigated two different fusion configurations from the patient detection and skin segmentation networks. In the first configuration, the fusion network was designed to take the feature maps of a single video frame obtained from the middle of a given time window. In the second configuration, the fusion network was designed to take the feature maps of three video frames. Each video frame was evenly sampled from a given time window. In this configuration, the shared core network was re-used for different images.

5.5.2 Multi-resolution temporal fusion

During the data acquisition in the Neonatal HDU study, the camera was occasionally positioned slightly off-centre, so the filming also included the area outside the incubator. False detections may arise when clinical staff move near the incubator without performing any intervention. This motivated an approach to use a multi-resolution network with two input streams: one input stream for the optical flows computed from the entire images (full-frame optical flow); and another input stream for the optical flows computed from cropped images containing only the patient area (patient-cropped optical flow), similar to what was proposed by [143]. A major advantage is that the multi-resolution fusion can then learn temporal information from both global and local contexts.

Two approaches for cropping video frames prior to the calculation of optical flows were investigated: fit cropping and centre cropping (see figure 5.8). For each video segment, the skin labels computed over the segment were cached and used to calculate the crop regions.

Fit cropping. For a given sliding window, the union of all skin labels over the video segment was calculated, and the video frames used for computing optical flow were cropped according to the minimum and maximum horizontal and vertical positions of the skin label union. The video frames were then resized to a fixed size of 256×192 pixels (the input size of the optical flow network). If the video segment did not contain skin regions, the original images were used

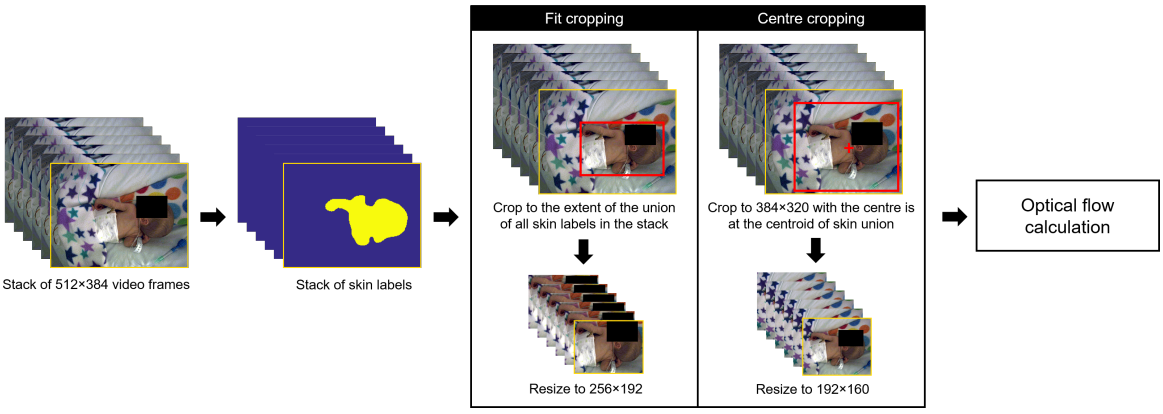


Figure 5.8: Two approaches for cropping video frames prior to optical flow calculation: fit cropping and centre cropping. The extent of the cropping is calculated from the union of skin labels over a video segment.

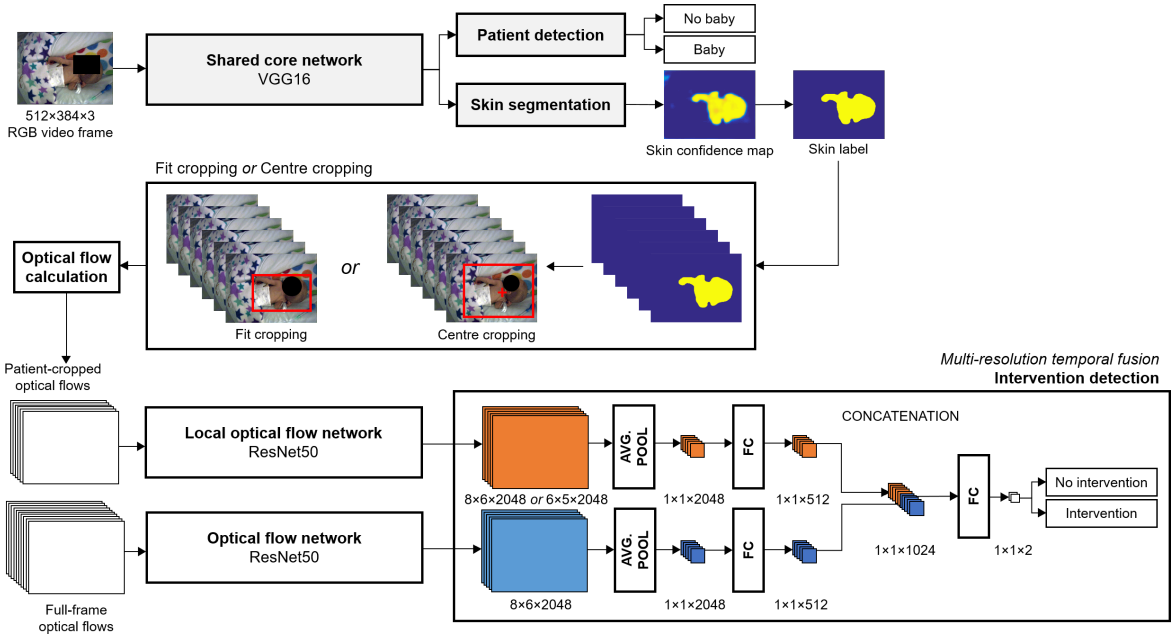


Figure 5.9: Multi-resolution temporal fusion combines temporal information from patient-cropped optical flow and full-frame optical flow using *concatenation fusion* [145] to predict the presence of interventions.

instead. The stack of fit-cropped optical flow data mainly contained the object of interest, so the effect of any camera offset was reduced. In this approach, the local optical flow network had the same architecture as the original optical flow network.

Centre cropping. With centre cropping, video frames were cropped to a fixed size. The union of all skin labels in a video frame was computed, and the centroid of the union was then calculated. When there were no skin regions, the centre of the image was taken to be the centroid. The video frames used for computing optical flow were therefore cropped to a 384×320 image with its centre situated at the centroid of the skin union. The aim of this was to remove the effect of motion that may occur around the incubator. The video frames were then resized by half to 192×160 pixels, and optical flow was computed from the cropped video frames.

Model architecture

The multi-resolution architecture consists of: the local optical flow network which takes the patient-cropped optical flow as its input; and the main optical flow network which takes the full-frame optical flow as its input. The local optical flow network was implemented and trained in the same way as the main optical flow network, except that with the centre-cropping approach, the size of the average pooling layer was changed from 8×6 to 5×6, according to the size of of the input. Neither network had shared parameters.

Since the main optical flow network processes a stack of full-frame optical flows and the local optical flow network processes a stack of patient-cropped optical flows, the feature maps from the last convolution layers from both networks were not spatially aligned. Fusion was performed through the concatenation operation, similar to the spatio-temporal fusion, with additional steps to combine two feature maps with different spatial sizes. In the intervention detection branch (see figure 5.9), the output from each network was first passed through a global average pooling layer to reduce its spatial dimension. The output from the average pooling layer was later passed through a fully-connected layer with 512 outputs. Consequently, the outputs from the fully-connected layer of both networks were concatenated across feature channels and then processed through another fully-connected layer with 2 outputs. Finally, the last softmax layer produced classification scores for intervention and non-intervention events.

5.5.3 Temporal context fusion

Unlike the two-stream spatio-temporal network presented in section 5.5.1, where fusion occurred in the early layers, the temporal context fusion approach used another network, a context network, to process multiple-frame skin confidence maps provided by the patient detection and skin segmentation network before fusion with the outputs of the optical flow network. The skin confidence map was defined as the softmax output of the skin segmentation branch of the patient detection and skin segmentation network before applying a threshold to compute the skin and background labels. The skin confidence maps over a video segment contain the information related to the motion of the infant as well as that of the clinical staff, if they were present in the image because intervention is being performed.

Model architecture

The architecture of the context network was similar to that of the optical flow network, but it instead accepts multi-frame skin confidence maps with the same spatial size as the optical flow data (256×192 pixels). We used the skin confidence maps of the same $L + 1$ video frames that were used to compute the L optical flow components. Hence, the context network processed a

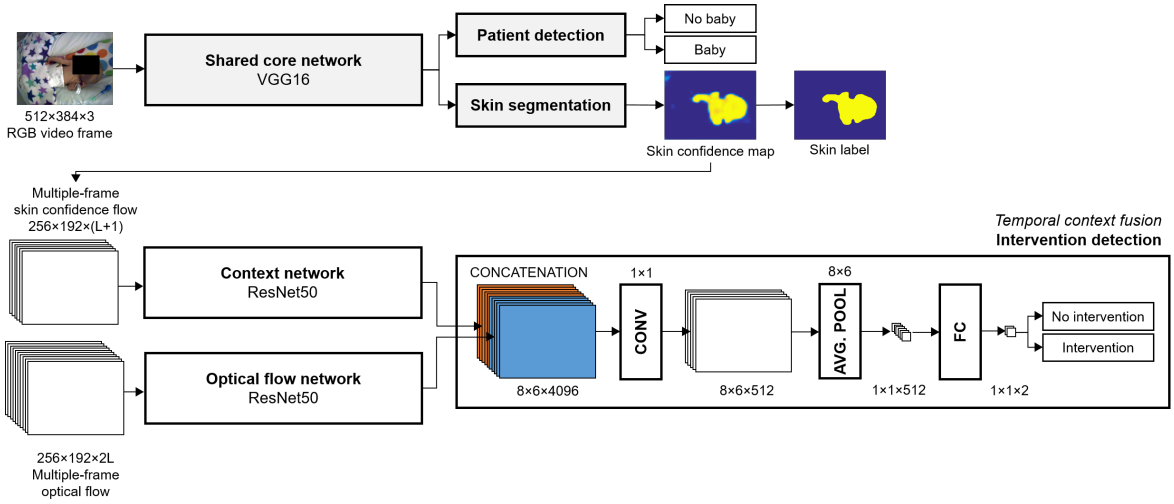


Figure 5.10: Temporal context fusion combines information from multiple-frame skin confidence maps and multiple-frame optical flows using a series of concatenation and convolution layers.

stack of $L + 1$ confidence maps, whereas the optical flow network processed a stack of $2L$ optical flow components. A final decision was made in the intervention detection branch based on the combination of information from both networks.

The intervention detection branch (see Figure 5.10) was implemented using the convolution fusion approach [145] similar to that of the two-stream spatio-temporal fusion. The outputs before the global average pooling layer of the context and optical flow networks were fused together through a series of concatenation and convolution layers. The feature maps of both networks were concatenated together along feature channels resulting in $8 \times 6 \times 4096$ feature maps, followed by a 1×1 convolution layer with 512 outputs. The branch outputs a decision on intervention and non-intervention events through a series of global average pooling, 2-way fully-connected and softmax layers.

5.6 Network training

This section describes the dataset and approaches used to train the network, including data pre-processing, data augmentation and network training procedures.

5.6.1 Dataset

The methods for intervention detection required a dataset of annotated intervention periods. Such information was not provided in full by the clinical staff during the clinical study. Therefore, we first needed to annotate intervention periods in the videos and then construct the dataset.

Intervention annotation

Even though information on clinical activities was recorded in the OpenClinica clinical study management platform (see chapter 3), the start and finish times for each clinical intervention

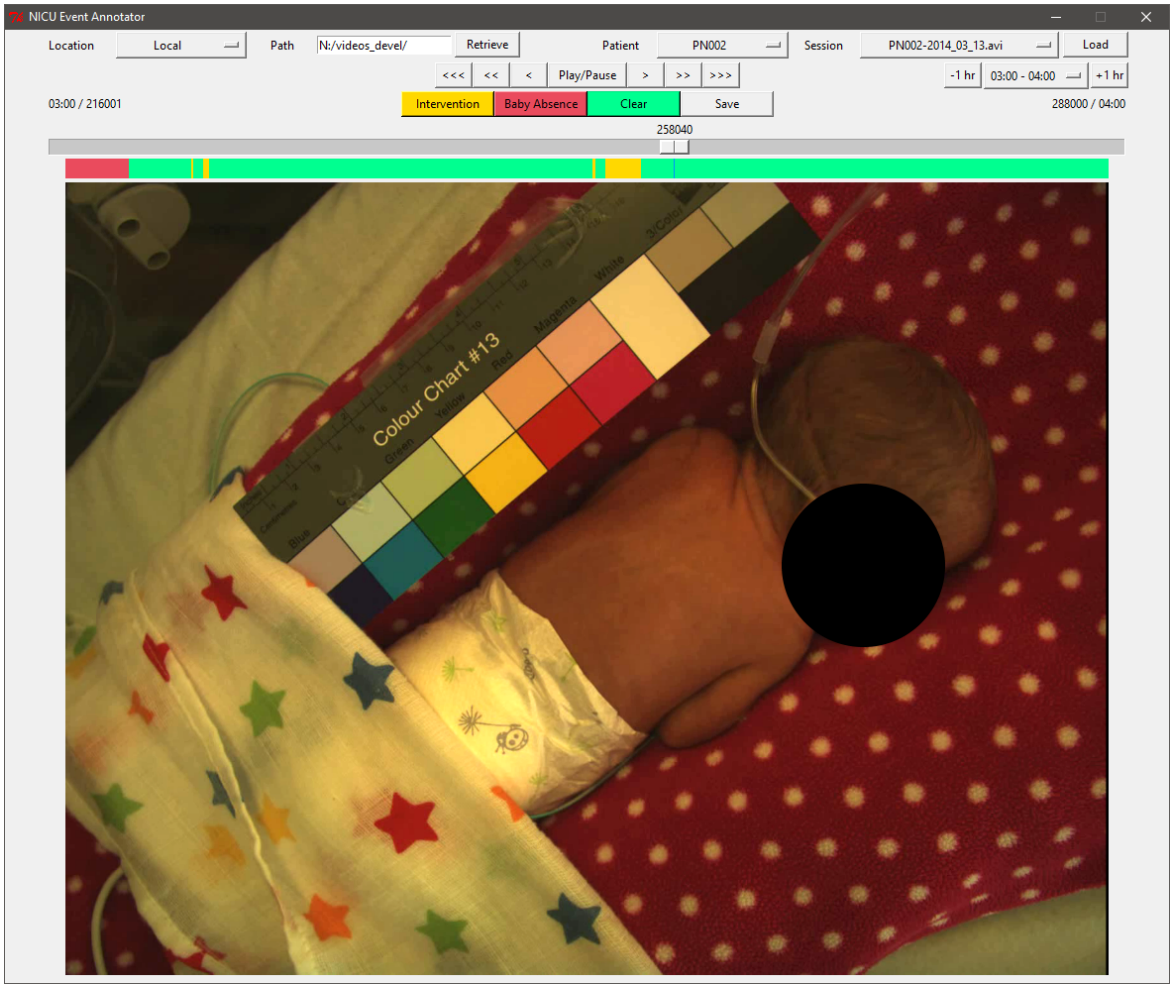


Figure 5.11: The annotation tool to label intervention periods in video recordings. The annotators were asked to watch each video and mark the start and end frames for intervention and infant-absence periods.

were not accurately noted down. In addition, several activities were not recorded, for example changing the infant's position or parental visits. Such intervention periods needed to be labelled by human annotators.

In addition to the intervention events, there was no clear indication in the clinical management system of the periods during which the infant was absent. Even though the annotation of the infant absence was labelled in the previous chapter as part of developing the training set, the annotation was performed on images extracted every 6 minutes from the video recordings. Hence, the annotation for this chapter focused on determining the start and end time points of two events: intervention and infant absence. Non-intervention is a period without any intervention or infant absence flags.

Annotation tool

The annotation of intervention periods was performed using a tool developed specifically to handle this task. Figure 5.11 shows the graphical user interface of the annotation tool. The software reads and shows the video files from the Neonatal HDU study. The videos were displayed to the three annotators in a block of one hour. The slider bar indicates the current position of the video that is

being played. The timeline bar under the slider bar indicates the annotation given by the annotators. In the timeline bar, intervention periods were shown in yellow and infant absence periods were shown in red. Fast forwarding was not allowed without watching the video first.

Annotation protocol

To ensure the quality of the training data, three human annotators were employed. All the annotators were postgraduate students. This also eliminated the chance of missing annotations which might have occurred when the annotator was fatigued and overlooked the occurrence of an event. The annotation involved all video sessions in the 15-infant training set. The periods during which the infant was under blue-light phototherapy (11.91 hours) were excluded from the annotation. Thus, the annotators were required to label the intervention and the baby absence periods in 214.0 hours of videos in total.

Annotators were asked to watch videos playing at 30× speed. This was to ensure that the entire videos were seen by the annotators. They could navigate forward and backward in time; however, forward navigation was not allowed unless the video section had been previously watched. The annotators were requested to mark a beginning frame number and an end frame number in the video for sections during which (1) medical staff or parent(s) were in the video frame (*intervention*) and (2) the baby was not in the frame (*baby absence*) by clicking on the 'Intervention' and 'Baby Absence' buttons respectively.

Construction of dataset

The intervention labels provided by three annotators were combined based on consensus among the annotators. For the intervention label, each frame was regarded as intervention only if at least two annotators agreed. This approach was also adopted for the baby absence label. Subsequently, the annotations, which were performed at the frame level, were converted into one-second labels using consensus among the labelled video frames for each second. For example, if the majority of the video frames in second t were labelled as intervention, that second t would be given the *intervention* label.

Dataset statistics

The Fleiss' kappa inter-rater reliability of agreement [171] between three annotators was 96.1. The intervention dataset contains 214.0 hours of annotated videos. Of that, 178.9 hours were marked as non-intervention, 16.7 hours were marked as intervention and 18.4 hours were marked as baby absence periods. Figure 5.12 shows the timeline plots of annotated video data in the training set of 15 babies.

5.6.2 Data preprocessing

Dense optical flows were pre-computed for each second in the videos using the optical flow algorithm of Brox *et al.* [172]. Optical flows were stored in compressed JPEG files as in [144] in order to reduce

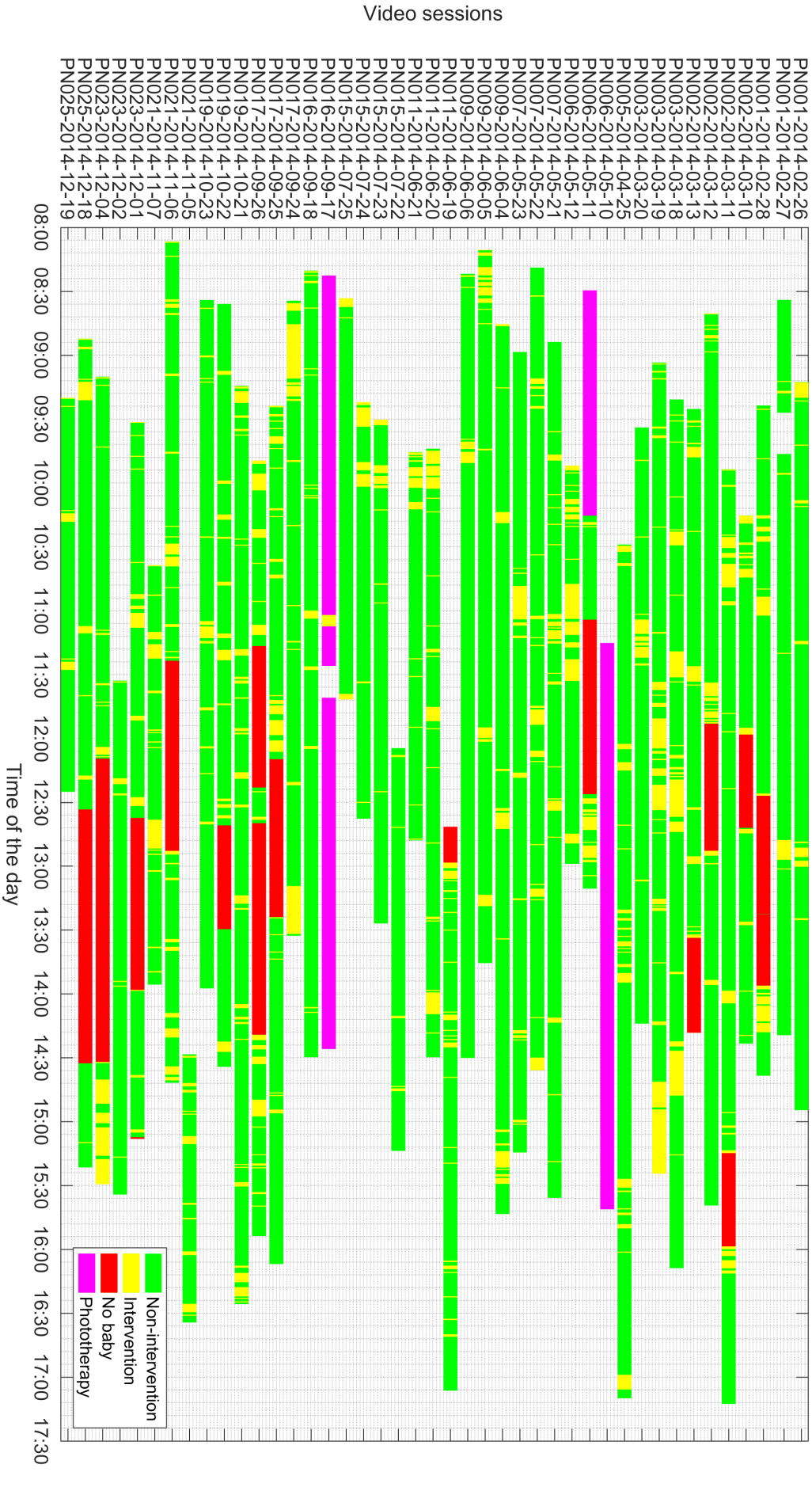


Figure 5.12: The agreement on intervention annotation provided by three annotators.

storage space and allow the dataset to be stored in memory. Since the optical flows were stored in an 8-bit integer format, the values can only vary from 0 to 255. For each optical flow component, displacement values bigger than 40 pixels were clipped to 40 pixels and these values were scaled to the range between 0 and 255, similarly to [145].

5.6.3 Data augmentation

The following transformations were applied to each image while training: random spatial resizing between scales of 1 and 1.25; random translation between -10% and 10% of the training image size in any direction; random mirroring in the horizontal and vertical directions; and random rotation between -45 and 45 degrees. These random parameters were applied together. For a stack of optical flows or skin confidence maps, the same augmentation was applied to all frames in the stack.

5.6.4 Training procedures

The CNN models were implemented on the MatConvNet framework [163]. Mean subtraction was applied to zero-centre the network input as in [144]. The sliding window approach resulted in a considerable amount of short video segments (from 1-second duration to 10-second duration). Each sliding window is regarded as *intervention* if more than half of the time is labelled as intervention, otherwise it is regarded as *non-intervention*. Each sliding window segment was treated as an independent and separate sample. For each training iteration, video segments were sampled uniformly across intervention and non-intervention classes to create a balanced training set.

The training was performed using standard Stochastic Gradient Descent in two stages in order to reduce training time and avoid overfitting [173]. In the first stage, the main optical flow network was trained with a momentum of 0.90 and a batch size of 24 samples. The learning rates were scheduled to start at 10^{-3} and reduced by a factor of 10 for every 12,000 iterations until convergence. The same configuration was used for the local optical flow network and the context network. In the second stage, the patient detection and skin segmentation network was integrated with the other network(s), and the intervention detection branch was added to form a fusion network. For each fusion network model, fine-tuning was performed with a momentum of 0.90, a batch size of 12 and a learning rate of 10^{-5}. The learning rate was decreased by a factor of 10 for every 6,000 iterations until convergence. Fine-tuning was performed on only the layers added after fusion, as suggested in [143, 145].

5.7 Evaluation protocol

The performance of the proposed networks was evaluated using cross-validation on two independent folds, as described in the previous chapter. Since the current work is an extension of the work described in chapter 4, the same dataset splits were also used for training and validating the

models investigated in this chapter. The validation results from both splits were combined to produce an overall classification performance.

As in section 4.6, the results are reported using accuracy, precision, recall or true positive rate (TPR), specificity or true negative rate (TNR), the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC). To tackle the class imbalance in the intervention dataset, the intervention class was oversampled with replacement to have the same number of training examples as the non-intervention class to create a balanced validation set.

5.8 Results

This section presents the results of the methods described above. Firstly, results for the baseline experiment are presented. Next, the performance of the optical flow networks on different sliding window configurations is reviewed. Finally, the results of the networks for different fusion strategies are reported and compared.

5.8.1 Baseline implementation

A baseline experiment was conducted on the two-stream convolutional architecture for action recognition proposed by Simonyan and Zisserman [144] implemented using the VGG-M-2048 model [136]. The VGG-M-2048 model has similar architecture and connectivity to the VGG-16 model but the number of weighted layers is lower (8 layers compared to 16 layers in the VGG-16 model). As suggested in [144], the number of optical flow stacks L is set to 10. To compare with the original implementation, we used the sliding window method, as described in 5.4.1, with a window length of 10 seconds and a step size of 1 second ($T = 10$, $\tau = 1$).

Table 5.2 shows the performance of the baseline implementation. The results are consistent with those of the study reported in [144]. The temporal network (87.8%) yielded higher accuracy than the spatial network (76.4%). The fusion of both networks using either averaging or a linear Support Vector Machine (SVM) led to further improvement in accuracy since they provide complementary information to support the classification task. The fusion network with SVM resulted in the highest accuracy of 92.4%, a 4.7% improvement over the temporal network.

Table 5.2: Baseline performance of the two-stream architecture proposed by [144]

Model	AUC	Accuracy	Precision	Recall	Specificity
Spatial	84.1	76.4	77.0	75.2	77.5
Temporal, $L = 10$	95.3	87.8	84.1	93.3	82.4
Fusion network (Average)	95.4	89.3	86.3	93.4	85.2
Fusion network (SVM)	**98.1**	**92.4**	**90.8**	**94.4**	**90.5**

Values expressed as percentage.

5.8.2 Sliding window comparison

Table 5.3 summarises the effects of using different sliding window configurations on the optical flow network. Window length T of 1, 5 and 10 seconds with overlapping ($\tau = 1$) and non-overlapping ($\tau = T$) step sizes were assessed. Since video frames were sampled for optical flow calculations at 1 Hz, the numbers of optical flow stacks L were 1, 5 and 10 respectively for each window length. The configuration of a 5-second sliding window with 1-second step size ($T = 5,\ \tau = 1$) led to 92.2% accuracy and outperformed the other configurations. Increasing the size of the window length from 5 to 10 decreased performance. The use of an overlapping step size was beneficial to the task.

Figure 5.13 shows the ROC curves characterising the classification performance of each sliding window configuration. The configuration of $T = 5,\ \tau = 1$ clearly outperforms the others.

Table 5.3: Performance of the optical flow network with different sliding window configurations (varying sliding window length T and step size τ)

Configuration	AUC	Accuracy	Precision	Recall	Specificity
$T = 1$ sec., $\tau = 1$ sec.	95.2	88.5	89.3	87.5	89.5
$T = 5$ sec., $\tau = 1$ sec.	**97.4**	**92.2**	**91.7**	**92.8**	**91.6**
$T = 5$ sec., $\tau = 5$ sec.	95.9	89.7	90.6	88.5	90.8
$T = 10$ sec., $\tau = 1$ sec.	96.2	88.2	88.2	91.5	87.8
$T = 10$ sec., $\tau = 10$ sec.	95.5	88.5	87.2	90.4	86.7

Values expressed as percentage.

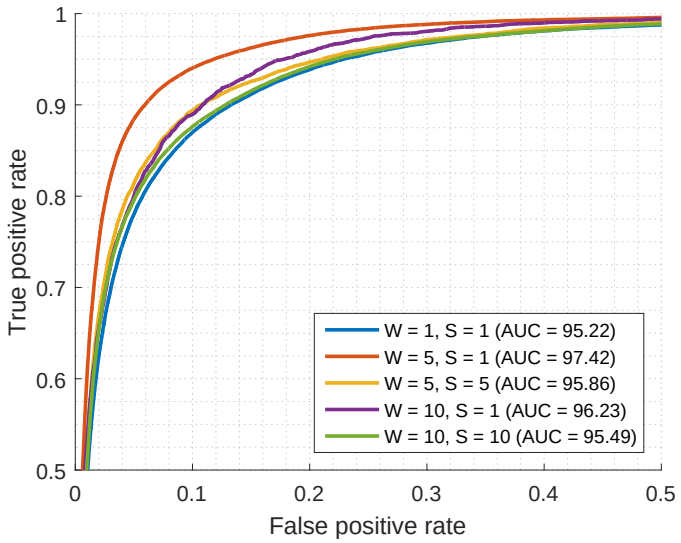


Figure 5.13: ROC curves show the classification performance of the optical flow network with varying sliding window configurations. The configuration of 5-second sliding window with 1-second step size outperformed the others.

5.8.3 Fusion strategy comparison

Table 5.4 reports the performance of the local temporal network and the context network that were trained individually and used for constructing the multi-resolution and temporal context fusion

networks. These networks were implemented using 5-second sliding windows with 1-second step size ($T = 5,\ \tau = 1$), which was the best performing configuration for the optical flow network.

Table 5.5 reports the classification performance for different fusion strategies. The temporal context fusion method yielded the highest performance with an accuracy of 94.5%, i.e. a 2.3% improvement with respect to the optical flow network alone. Multi-resolution temporal fusion, with both fit or centre cropping, gave marginal performance improvements. In contrast, the spatio-

Table 5.4: Performance of the local optical flow network and context network

Model	AUC	Accuracy	Precision	Recall	Specificity
Local optical flow network					
Fit cropping	96.6	90.5	89.8	91.4	89.6
Centre cropping	96.8	91.1	92.9	89.0	93.2
Context network					
Skin heatmap stacking	95.9	89.7	89.6	89.9	89.5

Values expressed as a percentage.

Table 5.5: Performance of the fusion networks on different fusion approaches

Model	AUC	Accuracy	Precision	Recall	Specificity
Spatio-temporal fusion					
Single frame, $M = 1$	97.0	91.7	89.8	94.1	89.3
Multiple frames, $M = 3$	96.0	90.5	90.4	90.8	90.3
Multi-resolution temporal fusion					
Fit cropping	97.8	93.7	92.8	93.7	92.7
Centre cropping	97.7	92.9	94.5	91.1	94.7
Temporal context fusion					
Skin heatmap stacking	**98.2**	**94.5**	**94.4**	**94.7**	**94.4**

Values expressed as a percentage.

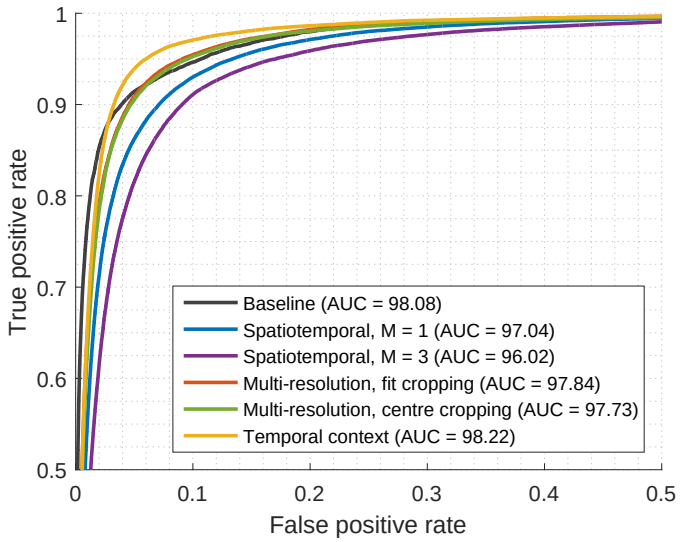


Figure 5.14: ROC curves show the classification performance of the fusion networks with different fusion strategies. The multi-stream temporal context network with skin confidence map stacking outperformed the others with an AUC of 98.2.

temporal fusion method was unable to make effective use of spatial information extracted through the patient detection and skin segmentation network.

Figure 5.14 shows the ROC curves for the networks implemented with different fusion strategies. The multi-stream temporal context network achieved the highest AUC of 98.2%.

5.9 Discussion

This section discusses the results presented in the previous section and provides further analysis of the implementation of the networks.

5.9.1 Sliding window configurations

From the different sliding window configurations of the optical flow network in table 5.3, a window length of $T = 5$ and a step size of $\tau = 1$ achieved the highest accuracy (3.7% over $T = 1$ and $\tau = 1$ as well as 4.0% over $T = 10$ and $\tau = 1$). A sufficiently large sliding window length is required to capture motion information over video frames. However a very large window length can cause deterioration in performance as it may lead to overestimation of the start and end times of an intervention interval. The overlapping sliding window scheme resulted in further improvements in performance. It significantly increased the size of training data by τ/T times. As a result, overfitting was reduced and the generalisation of the network was improved. In the $T = 5$ setting, the overlapping sliding window ($\tau = 1$) improved the accuracy by 2.5% compared to the non-overlapping sliding window ($\tau = T$).

Unlike the baseline implementation presented in section 5.8.1, the current work used the ResNet50 model. With the same settings (see table 5.2 for the VGG-M optical flow network and table 5.3 for the ResNet50 optical flow network with $T = 10$ and $\tau = 1$), the ResNet50 network resulted in a marginal improvement in accuracy over the VGG-M model (88.2% compared to 87.8% respectively).

5.9.2 Fusion strategies

This study explored several strategies for fusing spatial and temporal information from the video records. In the baseline implementation, the fusion of both networks using either softmax averaging or linear SVM led to further improvement in accuracy as the networks provided complementary information to support the classification task. However, the original implementation of [144] used just one image as an input to the spatial network. It was sensible for the implementation described in [144] as their datasets contained small videos with only one action per video. For our intervention dataset, in contrast, the nurses' hands can go in and out of the video frame at any time during a clinical intervention. One image is not enough to describe what is happening within a sliding window.

With spatio-temporal fusion, we expected the convolution layers of the patient detection and skin segmentation network, trained using our skin dataset, to work as a generic feature extractor. The

learned convolution layers were therefore used without fine-tuning. Surprisingly, the performance of the spatio-temporal network was found to be lower than that of the optical flow network before fusion. In addition, fusing information from multiple frames performed worse than using just a single frame. It is possible that the high-level convolutional features were too specific to the original patient detection and skin segmentation tasks. They may not carry meaningful information for the detection of intervention periods. By stacking the feature maps of multiple video frames together, the network found it difficult to learn, possibly because of the large numbers of free parameters.

Multi-resolution temporal fusion achieved an accuracy of 93.7% for fit cropping (images containing only the area of patient) and of 92.9% for centre cropping. Fit cropping ignores completely all motion information outside the extent of the subject. Manual inspection of the videos revealed that the fusion network produced fewer false positives, as expected, when the camera is positioned off-centre compared to the optical flow network. However, the overall improvement is very marginal.

Temporal context fusion yielded 94.5% accuracy which resulted in a 2.3% and 4.8% improvement over the optical flow network and the context network respectively. Apart from temporal information from optical flow, the stack of skin confidence maps also carries contextual information on both the location and motion of the infant. The improvement is likely to be due to the ability of the network to fuse correspondence information from each network. The improvement was found to be primarily due to the reduction in the number of false positives arising when: (1) a shadow is cast on the infant as a result of clinical staff walking past the incubator and (2) the camera is positioned off-centre and partially filming clinical staff walking in the room. Skin confidence maps have several degrees of tolerance with respect to skin colour variations and implicitly provide the location of the infant to the network.

The best performing fusion method was the temporal context fusion which resulted in a 2.5% improvement in accuracy over the fusion network in the baseline implementation.

5.9.3 Failure modes

We investigated on the failure modes of the network with the best-performing temporal context fusion network. Since intervention detection was performed using sliding windows, classification errors were to be expected at the start and end times of an intervention period. This was also made worse by mistakes from the human annotators since it was difficult to provide the exact start and end times of the annotation in the videos. Errors are also expected during interventions for which clinical staff constantly put their hands in and out of the incubator. Most false positives were found among the following scenarios:

1. The infant was very active or crying.
2. The position of the camera was adjusted by clinical staff.



(a) (b) (c)

Figure 5.15: Examples of false negative results. (a) Nurse touched the baby for a second when he or she was experiencing an apnoea. (b) Nurse's hand was off from the infant. (c) Nurse manually counted a respiratory rate and quietly held a timer in front of the camera.

3. Lighting conditions changed abruptly when fluorescent lights were switched on or off, or window blinds were opened or closed.
4. Daylight illumination changed quickly as the sky went from clear to cloudy, and vice versa.
5. Clinical staff were near the incubator causing shadows to be cast on the infant.
6. The incubator was touched and then shaken when clinical staff came to check equipment or to manually record vital sign values.

The changes in lighting conditions and the movement of the camera or incubator caused abrupt changes in optical flow. Even though mean subtraction was applied to reduce the effect of global motion as in [144], this only gave a marginal improvement.

False negatives occurred in the following scenarios:

1. Parents calmly held the infant in their hands during their visits to the Neonatal HDU (see figure 5.1b).
2. Clinical staff provided stimulation by touch to the infant, so their hands stayed still inside the incubator for a short time (see Figure 5.15a).
3. Clinical staff's hands were off from the baby during the intervention (see figure 5.15b).
4. Clinical staff quietly held a timer during manual respiratory counting (see figure 5.15d).

The errors were likely to have been caused by small changes in optical flow during the above scenarios, such that the network misinterpreted an intervention event as non-intervention.

In most cases, the network made reasonable mistakes. Even though we collected on very large dataset of more than two hundred hours of videos, the data came from just 15 different infants. It was difficult for the network to learn all possible clinical scenarios.

5.9.4 Runtime evaluation

The extension of the patient detection and skin segmentation network to perform intervention detection formed a multi-task network, which can perform all the necessary processing prior to the estimation of vital signs.

For every video frame, the network checked for the presence of a patient in front of the camera and segmented skin regions (see section 4.8.5 for computation time). For every second, the network identified the presence of an intervention event within the video segment. The computation of one set of optical flow vectors (in the horizontal and vertical directions) from a pair of video frames took approximately 0.07 seconds, so, for a given 5-second sliding window segment, the computation took approximately 0.35 seconds. The intervention detection network took 0.03 seconds to process the input. Hence, the whole processing time for each 5-second segment was approximately 0.38 seconds.

5.10 Conclusion

This chapter has presented the extension of the patient detection and skin segmentation network to detect interventions, resulting in a multi-task multi-stage network that can efficiently process a video recording to detect a patient, segment skin regions and identify time periods during which a clinical intervention occurs. The best performing fusion strategy uses skin confidence maps together with dense optical flows to identify the occurrence of an intervention. By providing contextual information on the location of the patient through the use of skin confidence maps, false positives are reduced. The fusion network performs better than the optical flow network alone. It does not require further processing to derive skin confidence maps since there have already been computed for skin segmentation. The proposed fusion network can be used to remove intervention periods prior to the estimation of vital signs.

Chapter 6
Reference heart rate

6.1 Introduction

The Neonatal HDU study is a clinical study established to investigate the use of video cameras for monitoring the vital signs of pre-term infants. The data provided by this study will be used to develop algorithms for estimating heart rate and respiratory rate from video cameras. The neonatal population presents different characteristics in comparison to the adult population. For instance, the resting heart rate and respiratory rate of a neonate are nearly double those of an adult [174]. This chapter describes the analysis of heart rate measurements provided by the Philips IntelliVue patient monitor and presents the computation of a new reference heart rate for later comparison with the heart rate estimate derived from the video camera.

This chapter is organised as follows. Section 6.2 outlines the overall methodology for computing the reference heart rate. Section 6.3 presents the analysis of the two heart rate measurements provided by the patient monitor. Section 6.4 describes the quality assessment of the ECG signal. Section 6.4.4 explains the computation of the reference heart rate. Section 6.4.5 reports the results of the reference heart rate estimation, followed by discussion in section 6.6. Finally, section 6.7 concludes the chapter.

6.2 Method overview

This section gives an overview of the methodology used for computing the reference heart rate. In the Neonatal HDU study, the Philips IntelliVue patient monitoring system provided two heart rate measurements: one derived from the electrocardiogram (ECG) and the other derived from the photoplethysmogram (PPG). The ECG sensor measures the electrical signals generated by the activity of the heart, whereas the PPG measures changes in blood volume underneath the skin. The heart rate estimates from these two sensors are not synchronised with each other as a result of different device clocks used by the ECG and PPG monitoring equipment. The ECG is considered to be the standard reference for measuring heart rate in pre-term infants in the hospital [175]. Since

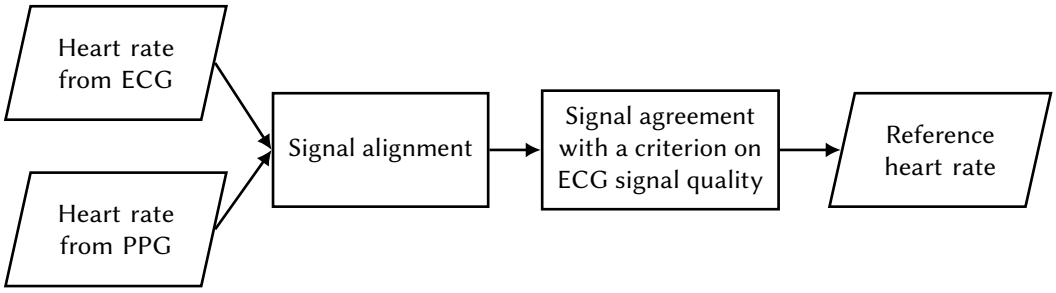


Figure 6.1: Overview of the computation of the reference heart rate.

the true value of the heart rate is not known, there is a need to analyse the heart rate measurements provided by the patient monitor and compute improved reference heart rate estimates with which to compare the estimates from the video camera.

An overview of the computation of the reference heart rate is shown in Figure 6.1. The process starts with the quality assessment of the ECG signal, as the ECG is taken to be the gold-standard for heart rate measurement [175]. The computation of the reference heart rate is performed by considering both the agreement of the two heart rate estimates and the quality of the ECG signal. The reference heart rate is computed as the mean of the two heart rate measurements when the two values are in close agreement and the ECG signal is deemed to be of good quality.

6.3 Heart rate monitoring in the Neonatal HDU study

The Philips IntelliVue patient monitor is a modular system that integrates the measurements from multiple devices that are fitted to the patient according to his or her needs. In the Neonatal HDU study, the Philips IntelliVue monitor provided two sets of heart rate measurements: one derived from the ECG, recorded by the Philips measurement module (ECG heart rate); and the other derived from the PPG, recorded by the Masimo pulse oximetry module (PPG heart rate). These heart rate estimates along with other physiological signals were acquired from the Philips patient monitor by a separate computer through a serial interface using the ixTrend software (see chapter 3).

In an ideal scenario, the monitor would report out any true heart rate values from the two sources which were the same; however, these values often varied. There are several reasons for these discrepancies between the two measurements, including: the different device clocks used by the Philips and Masimo measurement modules; different proprietary algorithms used to estimate heart rate; the intrinsic differences between ECG and PPG signals; different averaging or smoothing techniques; physiological events such as bradycardia or apnoea; clinical interventions such as changing measurement sensors and nappies; and motion artefacts.

Figure 6.2 shows the heart rate measurements provided by the Philips IntelliVue patient monitor for a typical recording session. The difference between the ECG heart rate and PPG heart rate is potentially due to the device manufacturers' proprietary algorithms, which are usually not

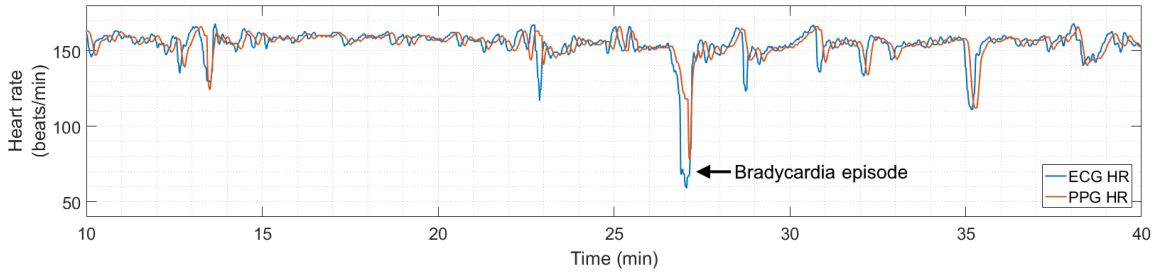


Figure 6.2: Heart rate signals provided by the Philips IntelliVue patient monitor for a typical recording session. ECG HR denotes the heart rate provided by the Philips neonatal ECG module, whereas PPG HR denotes the heart rate provided by the Masimo pulse oximetry module.

disclosed. The heart rate values from the pulse oximetry module are often delayed with respect to the ECG module (see figure 6.2). Nevertheless, both manufacturers (Philips and Masimo) claim compliance with the ANSI/AAMI EC13:2002 "Cardiac monitors, heart rate meters, and alarms" standard [176] recommended by the Association of the Advancement of Medical Instrumentation (AAMI) which specifies the maximum heart rate measurement error to be 1% or 5 beats/min, whichever is greater [176]. Even though the heart rate values were ultimately recorded on the same computer, these values came from two different hardware modules which had their own device clocks. The difference in these clocks, therefore, could have contributed to the time difference between the two heart rate measurements.

For the periods during which a bradycardia occurred (see figures 6.2 and 6.3a), the values of the ECG heart rate usually decrease faster and are lower than that of the PPG heart rate. By observing the video recordings for the time periods corresponding to these events, it is clear that the two heart rate values are slightly different, even when the infant is stable and has minimal motion. Other sources of differences are motion artefacts occurring when the baby moves the body segment (leg, arm or upper body) to which the sensor is attached (see figure 6.3b) or during a clinical intervention (see figures 6.3c–d).

Figure 6.4 shows the comparison between the ECG and PPG heart rate measurements for all recordings in the Neonatal HDU dataset, comprising a total recording time of over 426 hours. The comparison between the two measurements is performed using the Bland-Altman plot [177], a correlation analysis [159], and the mean absolute error (MAE) and the mean absolute deviation (MAD).

The Bland-Altman plot is designed specifically to assess the agreement between two clinical measurements [177]. It is constructed by plotting the mean of the two measurements on the x-axis and the differences between the two measurements on the y-axis. Figure 6.4a shows the Bland-Altman plot comparing the ECG and PPG heart rate values. The mean difference between the two measurements was 0.3 beats/min and 95% of the differences between the two heart rate sources lie within [-8.5, 9.1] beats/min. The histogram of the differences between the two measurements, in

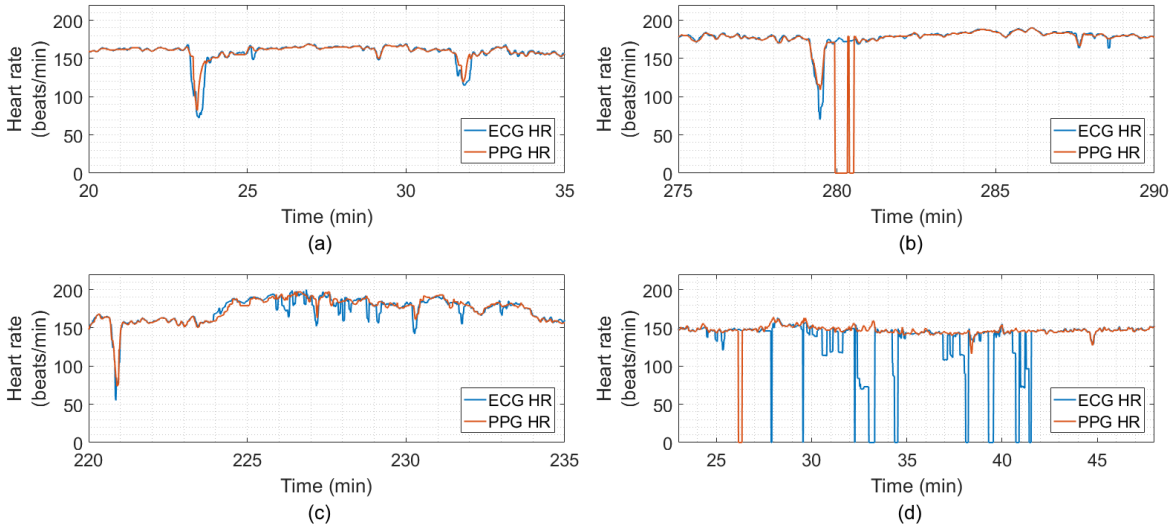


Figure 6.3: Examples of discrepancies between the two heart rate values recorded by the Philips patient monitor. (a) ECG heart rate usually decreased faster than PPG heart rate during bradycardia. (b) Time period during which a pulse oximeter sensor site was changed, causing the PPG heart rate to drop to zero. (c) Baby crying caused body movement and motion artefacts on the ECG signal. (d) ECG heart rate and PPG heart rate during a clinical intervention for which the ECG electrodes were replaced.

figure 6.4b, shows that the majority of errors were low, close to 0 beats/min. Figure 6.4c shows that the mean values were distributed across the expected physiological range for the neonatal population.

The correlation plot is used to assess the degree of relationship between two variables. It is obtained by plotting one variable against the other. In correlation analysis, the Pearson correlation coefficient r is used to measure the linear correlation between two time series x and y of length N [159]. The Pearson correlation coefficient is defined as:

$$r = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{N}(x_i - \mu_x) \cdot (y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 \cdot \sum_{i=1}^{N}(y_i - \mu_y)^2}} \tag{6.1}$$

where μ_x and μ_y denote the means of x and y respectively, and σ_x and σ_y denote the standard deviations of x and y respectively. A high value indicates that both variables are highly correlated, but it does not imply a good agreement. The two methods, the Bland-Altman plot and correlation analysis, are complementary as the Bland-Altman plot assesses the differences between the two measurements, whereas correlation analysis quantifies the relationship between the two measurements [178]. Figure 6.4d shows the correlation plot indicating that data points were located mostly around the identity line with a high correlation coefficient r of 0.95. The coefficient of determination r^2, which describes the goodness of fit of the data to the regression line, was 0.90.

The mean absolute error (MAE) is used to measure the error between two variables, and the mean absolute deviation (MAD) is used to measure the absolute deviation from the mean. Given two time series x and y of length N, the MAE is defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - x_i|. \tag{6.2}$$

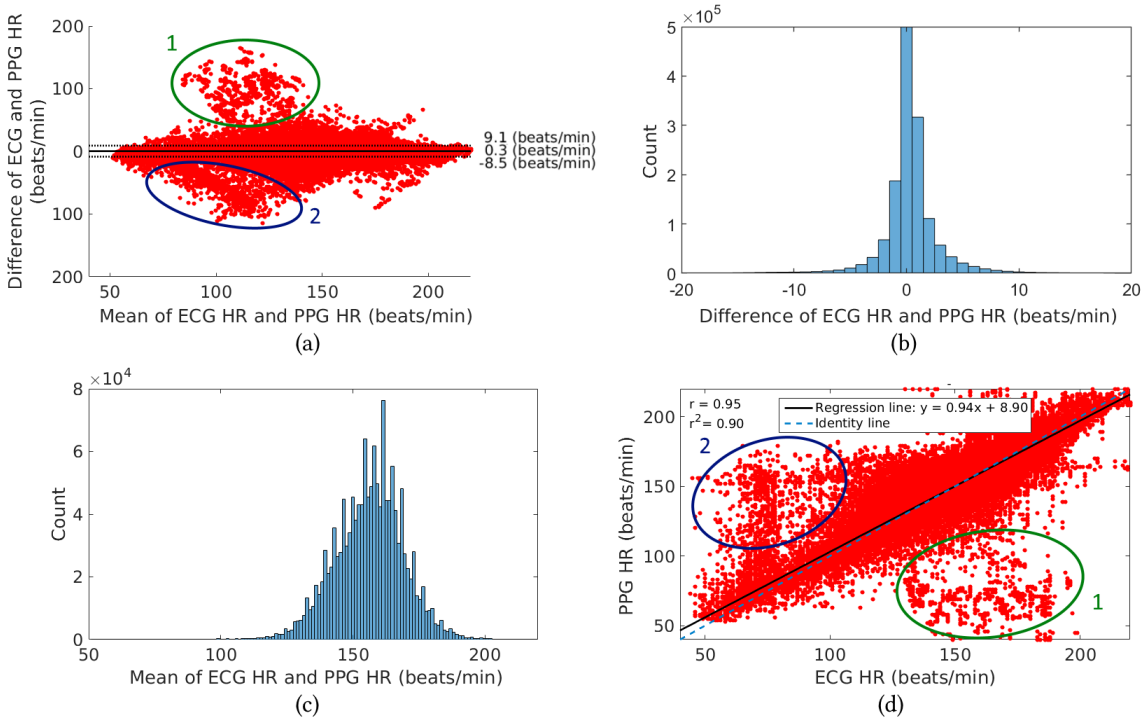


Figure 6.4: Comparison between the ECG heart rate (ECG HR) and PPG heart rate (ECG HR) values provided by the Philips monitor. (a) Bland-Altman plot reveals large errors between the two values. (b) Histogram of the difference between the ECG heart rate and PPG heart rate values. (c) Histogram of the mean values. (d) Correlation plot with a correlation coefficient r of 0.95. Ellipse "1" denotes data points for which the ECG heart rate is higher than the PPG heart rate. Ellipse "2" refers to data points for which the PPG heart rate is higher than the ECG heart rate.

Given z_i is the difference between the two time series x and y for each sample i, and μ_z is the mean of z, the MAD is defined as:

$$MAD = \frac{1}{N}\sum_{i=1}^{N}|z_i - \mu_z|. \tag{6.3}$$

The MAD measures the average distance between the data and its mean. In the Neonatal HDU study, the MAE between both heart rate measurements from the Philips patient monitor was 4.1 beats/min with a MAD of 4.0 beats/min.

In figures 6.4a and 6.4d, the green ellipse encloses sample data points for which the ECG heart rate is higher than the PPG heart rate. The blue ellipse encloses sample data points for which the PPG heart rate is higher than the ECG heart rate. The latter differences were largely due to periods of bradycardia during which heart rate decreased for several seconds and then recovered, as shown in figure 6.2.

The above analysis revealed discrepancies between the two heart rate measurements. A direct comparison between the camera-derived heart rate and either the ECG heart rate or PPG heart rate may lead to incorrect performance results. Even though the benefit of having two measurement devices is that the agreement between the two can imply that the heart rate measurements are of

good quality, a measure of the agreement alone is not sufficient to reflect the quality of the reference signals. As the signal quality of the reference data was not provided by the manufacturers, there is a need to compute a signal quality measure from the ECG or PPG time series. Signal quality can reveal periods during which the signal was severely corrupted by noise or motion artefacts, which may lead to errors in heart rate measurements (see figure 6.3).

6.4 Quality assessment of the ECG signal

In this section, the signal quality of the ECG waveform is analysed and the heart rate is re-estimated from the ECG using established algorithms validated using clinical databases in the public domain. The re-estimated heart rate is later compared to the heart rate provided by the Philips patient monitor to validate the signal quality. The derived signal quality is used to identify periods during which the reference signals are of poor quality.

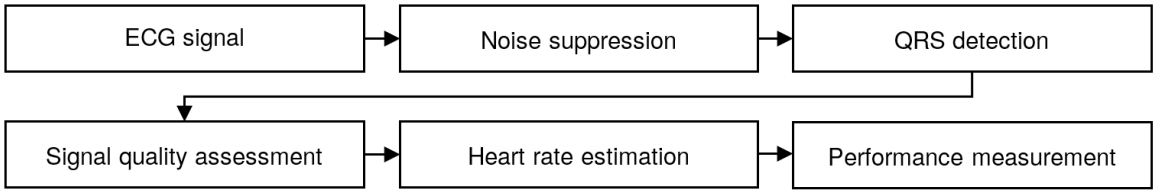


Figure 6.5: Block diagram of the estimation of heart rate from ECG waveform.

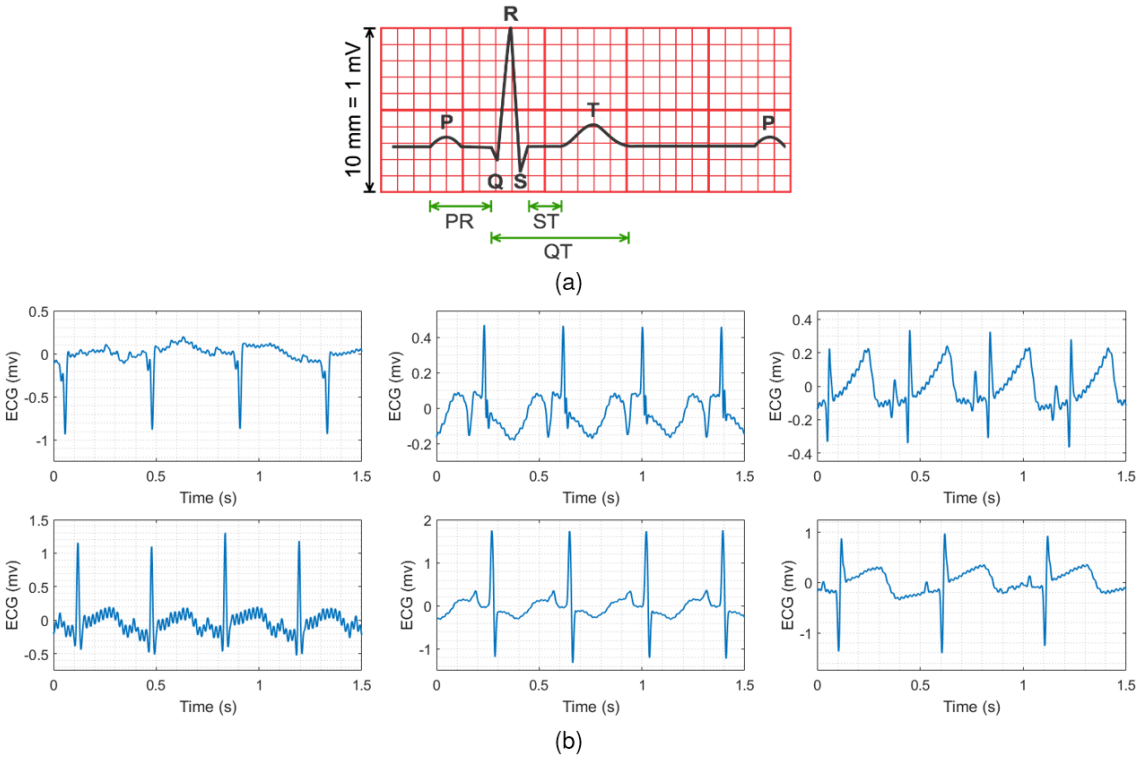


Figure 6.6: The ECG waveform (a) A typical ECG waveform of an adult contains a P wave, followed by a QRS complex and a T wave (Image adapted from [179]). (b) Sample ECG waveforms from pre-term infants in the Neonatal HDU dataset.

An overview of the quality assessment of the ECG signal is shown in figure 6.5. The ECG signal is first passed through a pre-processing step where noise components are attenuated. The components of the signal corresponding to each heart beat are then detected from the pre-processed signal. Subsequently, the quality of the signal is estimated for each detected heart beat. Finally, heart rate estimation is performed and compared with the Philips' ECG heart rate in order to validate the results.

6.4.1 The ECG waveform

A typical ECG consists of five components, as shown in figure 6.6a. A normal cardiac cycle starts with the P wave, followed by the QRS complex and the T wave [180]. The ECG in pre-term infants is different from that in adults because major changes in body and cardiac physiology occur at birth and during the first year of life [175, 181]. Figure 6.6b shows samples of ECG signals from pre-term infants in the Neonatal HDU dataset.

6.4.2 Noise suppression

ECG signals are usually contaminated by various sources of noise [182] (1) Power line interference can contaminate the signal at 50 Hz (the mains frequency in the United Kingdom) and its harmonics; (2) Electrode noise caused by the loss of contact between the electrode and the skin can introduce rapid transient interference [183]; (3) Motion artefacts caused by electrode motion having a rapid transient disturbance in the baseline signal with variable peak amplitudes and durations; (4) Muscle contraction causes transient interference of zero-mean band-limited Gaussian noise with a frequency content around 10 kHz; and (5) Baseline wander is a low-frequency interference caused by respiration or patient movement [180]. Noise filtering techniques can be applied to the ECG signal to reduce these types of noise.

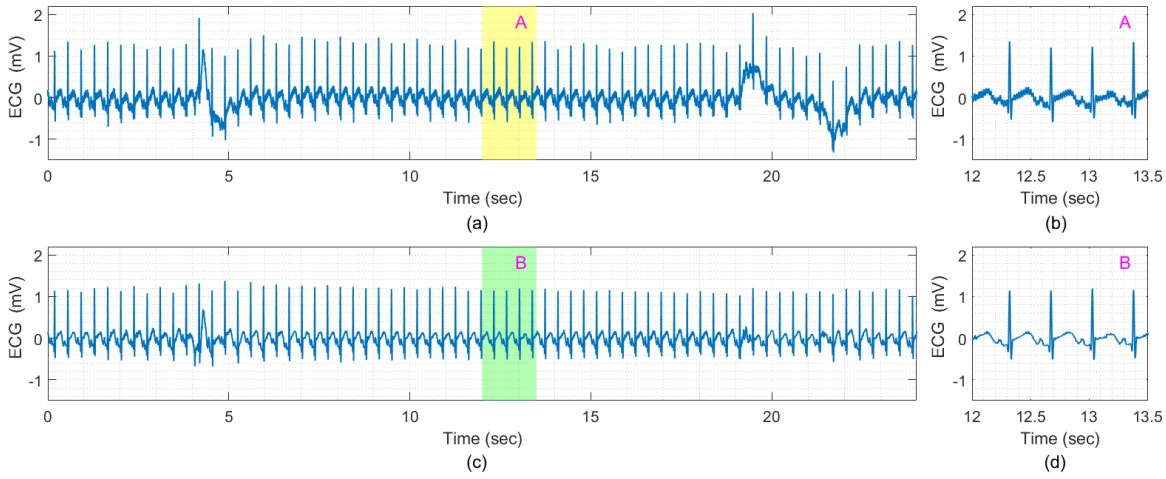


Figure 6.7: Noise-suppression for a ECG sample. (a) The original ECG signal is contaminated with baseline wander and power-line interference. (b) Close-up plot of the area 'A', highlighted in yellow in (a). (c) Processed ECG signal after noise suppression. (d) Close-up plot of the area 'B', highlighted in green in (c).

In this work, the ECG signal was processed through a cascade of a high-pass filter to suppress low-frequency baseline wander and a notch filter to eliminate power-line interference. The characteristics of these filters are similar to those described by Alste and Schilder [184]. The high-pass filter was designed as an infinite impulse response (IIR) filter of order 8 with a cut-off frequency of 0.7 Hz for the -3 dB point of the pass band. A notch filter was designed as a second-order IIR notch filter with a centre frequency of 50 Hz. Some high frequency noise still exists in the signal but will be suppressed during the detection of the QRS complex. This approach retains the structure of the signal from the distortion which may pose a problem in the analysis of signal quality in later stages. Figure 6.7 shows the ECG signal before and after pre-processing.

6.4.3 QRS detection

QRS detection generally implies a process of determining the components (Q, R, and S feature points) of each QRS complex in the ECG signal. This work employed two QRS detection algorithms: the Pan-Tompskin QRS detector [185, 186] and the length-transform based QRS detector proposed by Zong *et al.* [187]. The agreement between the QRS complexes detected by both algorithms was later used to assess the signal quality of the ECG. Accurate QRS detection is essential for the quality assessment of the ECG waveform and, therefore, an important step for the estimation of heart rate.

Pan–Tompkins QRS detector

The Pan–Tompkins QRS detection algorithm is based on digital filtering and integration techniques [185, 186]. The process is outlined in figure 6.8.

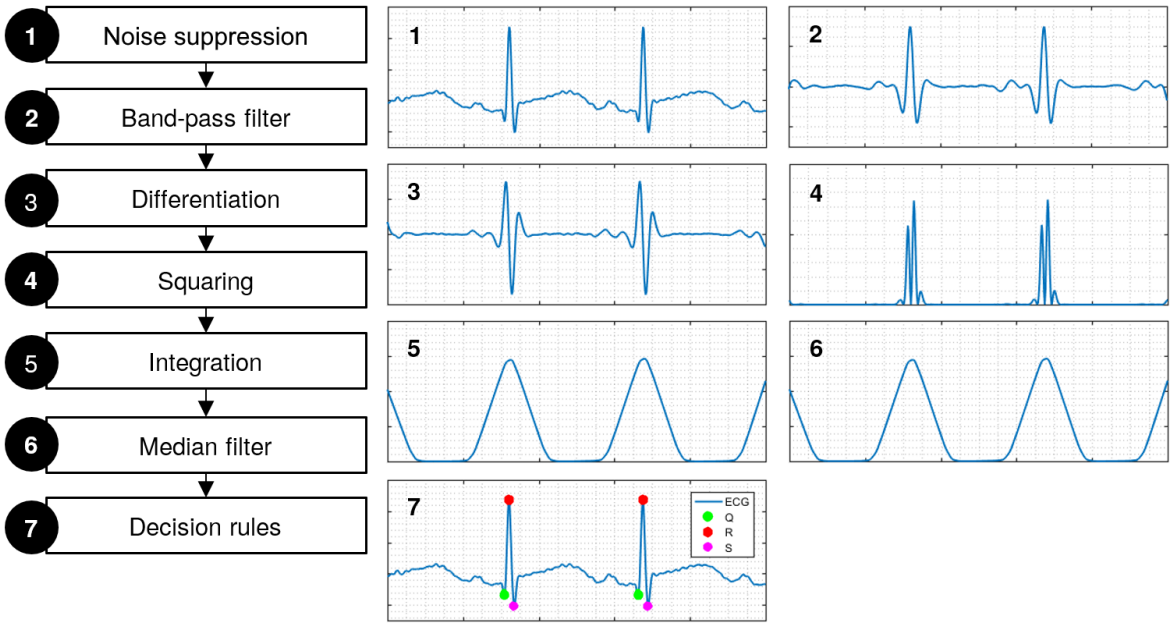


Figure 6.8: (Left column) Block diagram of the Pan-Tompkins QRS detection process. (Middle and right column) The resulting sample signals from each stage.

The ECG signal is first passed through a band-pass filter with cut-off frequencies of 15 and 25 Hz, as suggested by Schiavenato *et al.* for the neonatal population [188]. These cut-off frequencies are higher than those used in the original algorithm which was designed for the adult population since pre-term infants have nearly double the heart rate of adults [188]. The band-pass filter was implemented using a second-order Butterworth filter. Signal filtering is performed to maximize the frequency content of the QRS complex as well as to attenuate the P and T wave interference and noise (see figure 6.8).

The filtered signal x was then differentiated to emphasise the QRS complex using a seven-point central difference defined by:

$$y_i = \frac{-x_{i-3} + 9x_{i-2} - 45x_{i-1} + 45x_{i+1} - x9_{i+2} + x_{i+3}}{60}. \tag{6.4}$$

The differentiated signal y is then squared to produce a positive time series and integrated to further amplify the QRS complex (see figure 6.8). Subsequently, a five-point median filter is applied to reduce the effect of artefacts.

Eventually, R peaks are detected using an adaptive thresholding method [185]. For each recording session, the signal is first divided into ten learning periods of 5 minutes, evenly distributed over the whole recording for that session for that baby. The R peak is identified as the maximum value (or minimum value if the ECG waveform is inverted – see figure 6.6) in the ECG when the signal exceeds a given threshold, set as the mean of the processed signal over the learning periods. To avoid missing beats, the threshold is reduced by half if the inter-beat (RR) interval between successive R peaks is larger than twice the median of the RR intervals of R peaks already detected. For each detected R peak, Q and S points are defined in the ECG as the minimum value (or maximum value if the ECG waveform is inverted) on the left and right of the R peak respectively.

Length–transform–based QRS detector

Zong et al. [187] proposed an algorithm to find the QRS complex based on a curve-length transform (LT). Figure 6.9 shows a block diagram of the algorithm. Initially, the ECG signal is passed through a second-order Butterworth low-pass filter with a cut-off frequency of 25 Hz. The filtered signal is then differentiated and squared to enhance the features of the QRS complex, similar to the Pan-Tompkins QRS detector.

The algorithm assumes that the QRS complex generally has a larger amplitude than that of the other ECG components. Given y is the filtered ECG signal, the curve length transform (LT) of y over an interval $[i - w, i]$ is defined as:

$$LT(w, i) = \sum_{k=i-w}^{i} \sqrt{1 + (\frac{y_k - y_{k-1}}{\Delta t})^2 \Delta t}. \tag{6.5}$$

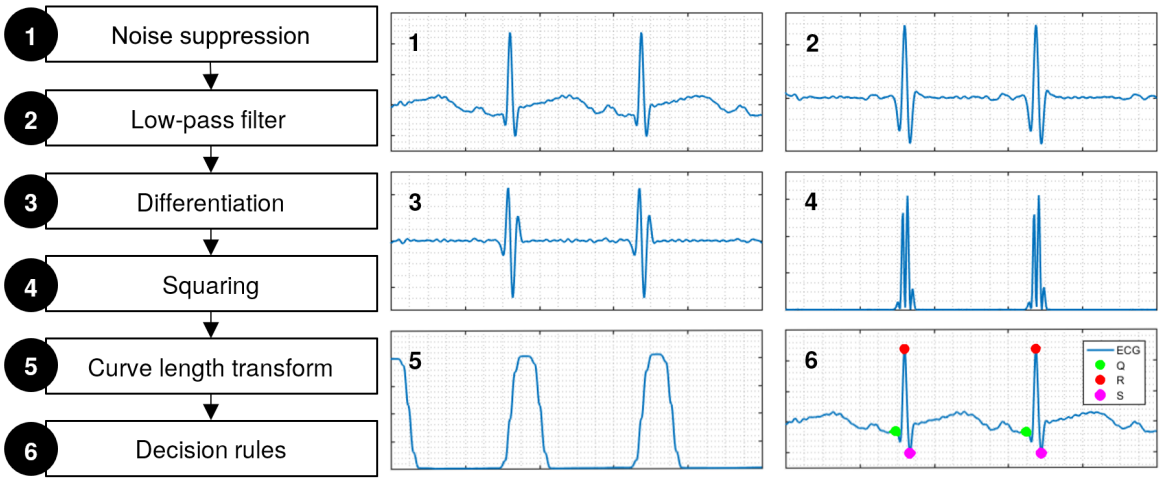


Figure 6.9: (Left column) Block diagram of the wQRS algorithm. (Middle and right columns) The resulting sample signals from each stage.

where Δt is the sampling interval and w is chosen to be the maximum duration of the neonatal QRS complex, taken to be 80 milliseconds for pre-term infants [175]. This process has the effects of not only enhancing the QRS complex but also of suppressing low-frequency noise [187].

Eventually, an adaptive thresholding technique and decision rules are applied to the LT signal to detect the QRS onset. Similar to the Pan-Tompkins detector, the LT signal is first divided into ten small learning periods of 5 minutes as before. The threshold is initially set to be the mean of the LT signal over the learning periods and adaptively adjusted to 1/3 of the median of the maximum LT values for the detected QRS complexes in a 15-second window. The adaptive thresholding allows QRS detection to be performed accurately even if the ECG amplitude changes. It also prevents miss-detection of the QRS complex when the ECG amplitude momentarily changes due to body movement during an apnoea event or a hiccup. For each threshold-crossing point, the algorithm searches backward to find a Q point and forward to find an S point. The R peak is defined as the maximum (or minimum if the ECG waveform is inverted) of the ECG signal between each pair of the Q and S points.

6.4.4 Signal quality assessment

The assessment of the signal quality of the ECG is important as data corruption by noise or motion artefacts is difficult to avoid during the continuous recording of physiological signals. The assessment was based on the algorithm proposed by Li *et al.* [189], in which the signal quality index of the ECG (SQI_{ECG}) depends on the combination of four signal quality indices ($SQIs$): (1) the instantaneous heart rate calculated using the interval between adjacent R peaks (SQI_{phys}); (2) the agreement of R peaks detected between two beat detectors (SQI_{agree}); (3) the calculation of kurtosis (SQI_{kur}); and (4) the analysis of the signal-to-noise ratio of the ECG (SQI_{snr}). Our implementation is different from the original algorithm since the Neonatal HDU study recorded a single-lead ECG waveform. Therefore, the calculation of the SQI for multiple-lead ECGs was ignored.

Physiological bounding

SQI_{phys} measures whether the instantaneous heart rate falls within the physiological range typical for pre-term infants, taken to be from 40 to 230 beats/min. Given HR_{i} as the instantaneous heart rate of the kth beat, SQI_{phys} is defined as:

$$SQI_{\text{phys}}(k) = \begin{cases} 1 & \text{if } 40 \leq HR_{\text{i}}(k) \leq 230 \\ 0 & \text{otherwise} \end{cases} . \tag{6.6}$$

The R-peaks are detected using the Pan-Tompkins QRS detector described in the previous section. The physiological range is defined according to the clinical guidelines in [176] and the range of heart rates in the Neonatal HDU dataset (see chapter 3).

Beat detection agreement

SQI_{agree} employs two beat detectors: the Pan-Tompkins QRS detector and the length-transformed-based detector. The accuracy of the two detectors can be used to estimate the level of noise in the ECG signal. SQI_{agree} for the kth beat centred around a window w of 10 seconds is defined as:

$$SQI_{\text{agree}}(k) = \frac{N_{\text{match}(k,w)}}{N_{\text{all}}(k,w)} \tag{6.7}$$

where N_{matched} is the number of beats for which both methods agree, within 80 milliseconds (the duration of a typical QRS complex for the neonatal population [175]), and N_{all} is the number of all beats that both algorithms detected minus N_{matched}. SQI_{agree} is assigned a value of 1 if all beats detected agree, otherwise a value of 0 is assigned.

Kurtosis

SQI_{kur} is based on the estimation of kurtosis, with an assumption that noise tends to have a Gaussian-like distribution. Kurtosis measures the sharpness of the peak of a distribution. The kurtosis of the ECG data x in a 10-second window w centred around the kth beat is defined as:

$$\hat{K}(k) = \frac{1}{M_x} \sum_{i=1}^{M_x} \left[\frac{x_i - \hat{\mu}_x}{\hat{\sigma}_x} \right]^4 \tag{6.8}$$

where M_x, $\hat{\mu}_x$ and $\hat{\sigma}_x$ are the length, mean and standard deviation of the data x respectively. The kurtosis of a Gaussian-like distribution is approximately 3, and ECG contaminated with noise and artefacts usually has a kurtosis lower than 5 [190]. Therefore, SQI_{kur} is defined as:

$$SQI_{\text{kur}}(k) = \begin{cases} 1 & \text{if } \hat{K}(k) > 5 \\ 0 & \text{if } \hat{K}(k) \leq 5 \end{cases} . \tag{6.9}$$

Signal-to-noise ratio

SQI_{snr} measures the strength of the signal carrying cardiac information. According to Rigg *et al.* [191], the spectrum of the neonatal ECG is concentrated below 40 Hz. This value is significantly

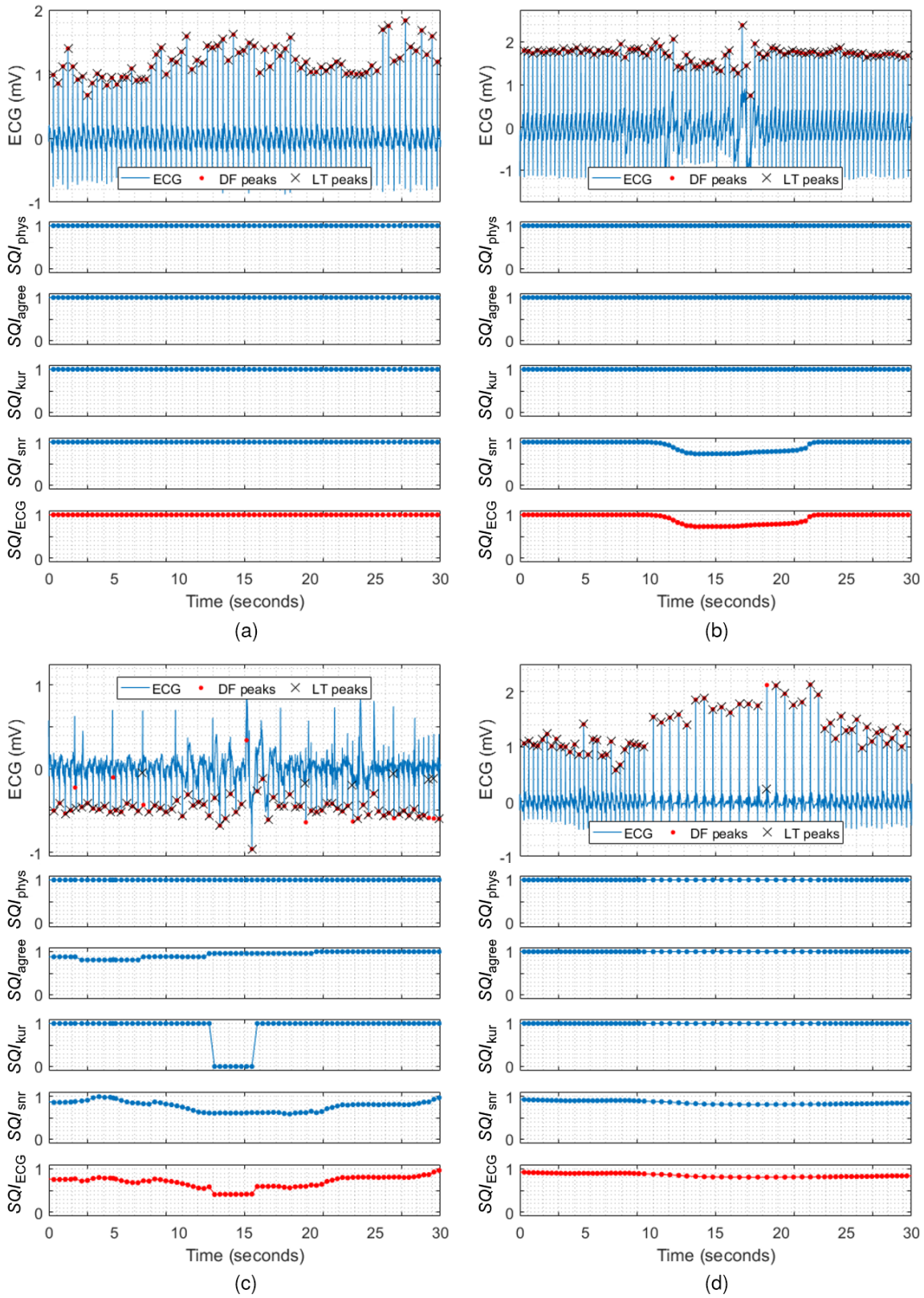


Figure 6.10: Signal quality indices for four 30-second ECG signals. (a) ECG signal scores high on all SQIs for a period of good quality. (b) ECG signal extracted from a period of moderate infant motion, resulting in a lower SQI_{snr}. (c) ECG signal extracted from a period during which the infant had hiccups (shown as high-amplitude spikes), yielding a low SQI_{ECG}. (d) ECG signal extracted from a bradycardic episode, resulting in a slight decrease in the SQI_{ECG}. DF peaks indicate the peaks detected by the Pan-Tompkins QRS detector, whereas LT peaks indicate the peaks detected by the length-transform-based detector.

different from that for the adult ECG as a result of the shorter duration of the cardiac cycle in the neonatal patient [191]. Given x is an ECG signal in a 10-second window w centred around the kth beat and x_{filt} is the ECG signal x filtered with a band-pass filter with cut-off frequencies of 1 and 40 Hz, the signal-to-noise ratio is defined as:

$$SNR(k) = 10\log_{10}\frac{\text{Var}(x_{\text{filt}})}{\text{Var}(x - x_{\text{filt}})} \tag{6.10}$$

where $\text{Var}(\cdot)$ denotes the variance function. A high $SNR(k)$ implies a good quality for the signal. Hence, SQI_{snr} is defined as

$$SQI_{\text{snr}}(k) = \begin{cases} 1 & \text{if } SNR(k) > 10 \\ \dfrac{SNR(k)}{10} & \text{if } 0 \leq SNR(k) \leq 10 \\ 0 & \text{otherwise} \end{cases} . \tag{6.11}$$

Combined signal quality index

Finally, SQI_{ECG} combines the results from the four SQIs described above:

$$SQI_{\text{ECG}}(k) = \begin{cases} SQI_{\text{phys}}(k) \cdot SQI_{\text{agree}}(k) \cdot SQI_{\text{snr}}(k) & \text{if } SQI_{\text{kur}}(k) = 1 \\ SQI_{\text{phys}}(k) \cdot SQI_{\text{agree}}(k) \cdot SQI_{\text{snr}}(k) \cdot \eta & \text{if } SQI_{\text{kur}}(k) = 0 \end{cases} \tag{6.12}$$

where η is a coefficient denoting the presence of statistical noise. The value of η was chosen to be 0.7 as in the original paper [189] which implies that SQI_{ECG} is reduced by 30% if statistical noise is present. SQI_{phys} penalises SQI_{ECG} when an R peak is undetected. SQI_{snr} reduces the SQI_{ECG} if the signal-to-noise ratio falls below a threshold of 10, which often occurs when there noise is present outside the frequency range related to the cardiac cycle. Figure 6.10 shows examples of the computation of the SQIs for various ECG signals extracted from typical recording sessions.

6.4.5 Heart rate estimation

Heart rate was computed using sliding 10-second windows with a step size of 1 second by counting the number of detected beats in the time window (as is the clinical practice for pre-term infants [189]). Once the heart rate was estimated, the SQI of the heart rate (SQI_{HR}) was calculated as the median of the SQI_{ECG} of the underlying detected beats for the given time window. Figure 6.11a shows the heart rate estimated from the ECG using the beat counting technique.

Signal alignment

In order to compare two time series, e.g. the heart rate measured by two different devices, the signals need to be aligned in time to reduce the effects from the sources of discrepancies (see section 6.3).

The time delay between two signals can be estimated using cross-correlation [192]. Given that f and g are vectors containing a time series signal, the cross-correlation measures the similarity

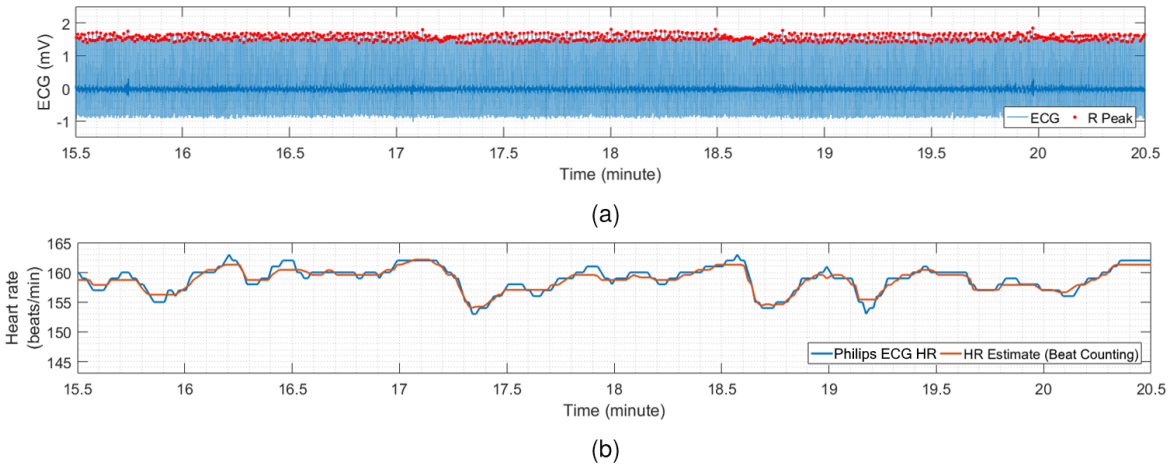


Figure 6.11: Heart rate estimation from a sample 5-minute ECG signal. (a) Pre-processed ECG Signal. (b) Heart rate estimated using the beat counting method compared to heart rate reported by the Philips monitor.

between f and a shifted version of g. The cross-correlation at a time lag t is defined as:

$$R_{xy}(t) = \frac{\sum_{i-1} N(f_i - \mu_f)(g_{i-t} - \mu_g)}{(N-1)\sigma_f \sigma_g} \tag{6.13}$$

where μ_f and μ_g are the means of f and g respectively and σ_f and σ_g are the variances of f and g respectively. The maximum of the cross-correlation indicates the time delay in which the two signals are best aligned. The time delay is defined as:

$$T = \underset{t}{\operatorname{argmin}}\, R_{xy}(t). \tag{6.14}$$

The time delay indicates how much g is shifted, along the x-axis (time axis), with respect to f.

Since physiological signals such as heart rate are nonstationary in nature, the presence of sudden spikes can have a negative impact on the cross-correlation computation. A median filter is usually applied to the time series signal to reduce the effect of these occurrences and smooth the signal prior to the computation of the cross-correlation. The median filter should not be too long as it may then cause significant distortion of the signal. In this work, a median filter with a 10-second time constant was applied to align these signals.

Results

Figure 6.12 shows the comparison between the ECG heart rate estimates provided by the Philips monitor and the heart rate re-estimated from the ECG by the beat counting method using sliding windows of 10 seconds for which the SQI_{HR} is higher than 0.5, a threshold implying a moderate quality of the signal. This criterion covers approximately 392.9 of 426.6 hours (92.1%) of the total recording time. According to the Bland-Altman plot, the mean bias is -0.1 beats/min and 95% of the differences between the two data sources fall within a range of [-5.6, 5.3] beats/min. The correlation coefficient r is 0.90. The mean absolute error is 1.1 beats/min with a mean absolute deviation of 1.1 beats/min.

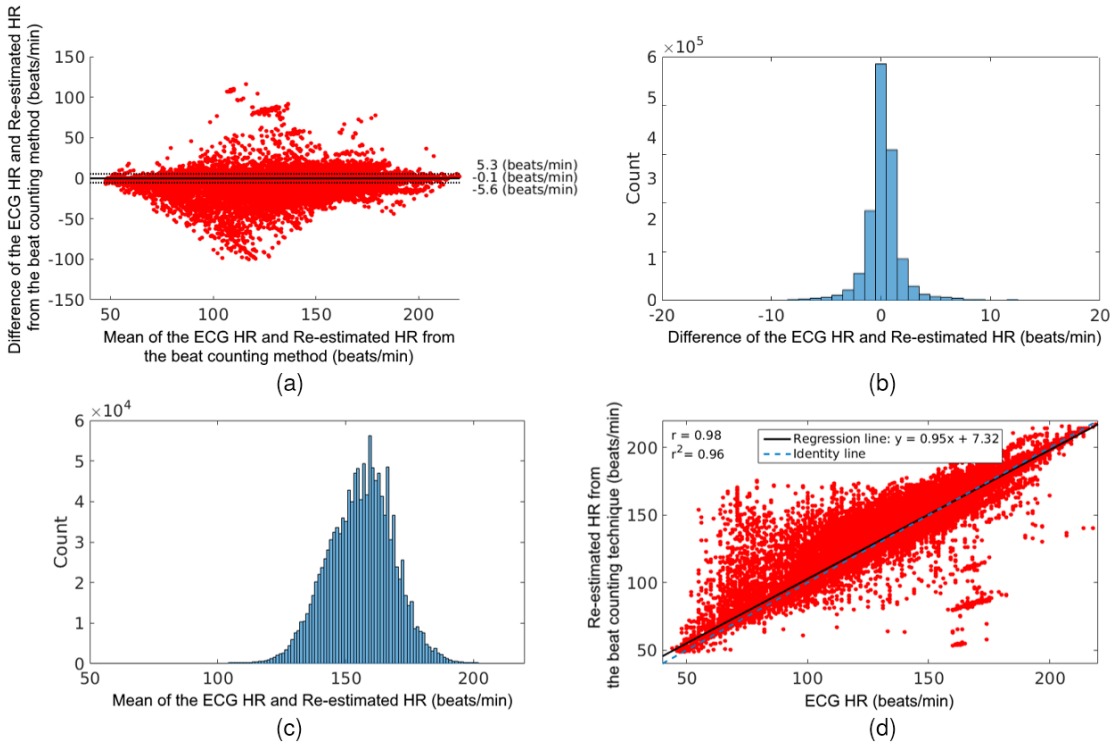


Figure 6.12: Comparison between the ECG heart rate from the Philips monitor and the heart rate re-estimated from the ECG using the beat counting method for all sessions. (a) Bland-Altman plot. (b) Histogram of difference values. (c) Histogram of mean values. (d) Correlation plot of ECG heart rate against the re-estimated heart rate.

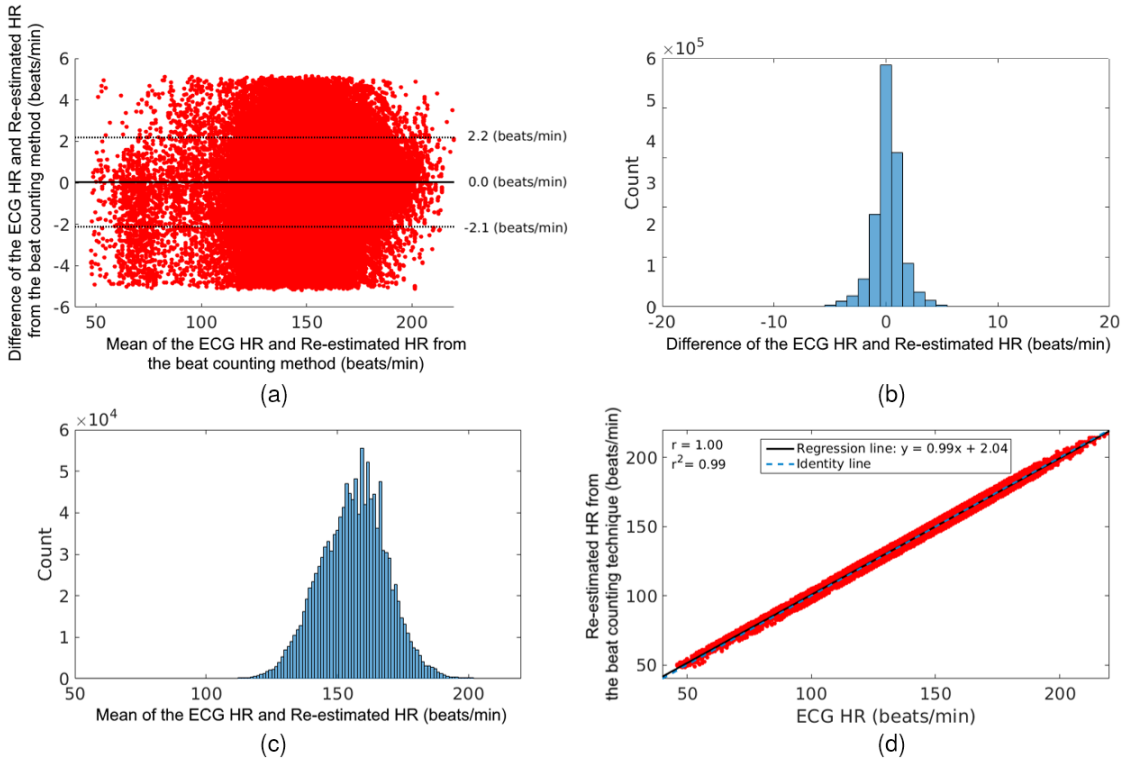


Figure 6.13: Comparison between the Philips' ECG heart rate and the heart rate re-estimated from the ECG signal using the beat-counting method for all sessions. The absolute difference between the two values was limited to 5 beats/min. (a) Bland-Altman plot. (b) Histogram of difference values. (c) Histogram of mean values. (d) Correlation plot of ECG heart rate against the re-estimated heart rate.

Table 6.1: Comparison between the ECG heart rate and re-estimated heart rate from the beat counting technique for all sessions in the Neonatal HDU dataset

Values compared	MAE (beats/min)	MAD (beats/min)	Estimated time (%, hours)
ECG heart rate versus re-estimated heart rate			
Values with $SQI_{\text{HR}} \geq 0.5$	1.1	1.1	92.1% (392.9 hours)
Values with $SQI_{\text{HR}} \geq 0.5$ and difference ≤ 5 beats/min	0.8	0.8	90.0% (384.1 hours)

Figure 6.13 shows the comparison between the re-estimated heart rate and the ECG heart rate for the Philips monitor using the beat counting method, as with the previous figures. However, the comparison was performed during these times for which the SQI_{HR} is higher than 0.5 and the absolute difference is less than 5 beats/min, the maximum heart rate estimation error as recommended by the ANSI/AAMI EC13:2002 standard [176]. Approximately 384.1 of 426.6 hours (90.0%) of heart rate values are within these criteria. The Bland-Altman plot shows a mean bias close to zero and 95% of the differences fall with a range of [-2.1, 2.2] beats/min. The correlation coefficient r is 1.00. The mean absolute error of the two signals is 0.8 beats/min with a mean absolute deviation of 0.8 beats/min. Table 6.1 summarises the results of the heart rate estimation presented in figures 6.12 and 6.13 above. The comparison between the ECG heart rate and the re-estimated heart rate showed a good agreement.

Summary

Without applying the limit on the absolute difference between the ECG heart rate and the re-estimated heart rate, the heart rate estimation results in low errors, in both MAE and MAD metrics, and a small bias. There are a few extreme values away from identity line of the correlation plot (see figure 6.12d). By applying the limit of 5 beats/min on the absolute difference, the heart rate estimation results in lower errors, with both time series being highly correlated.

This implies that the SQI_{HR} could be used as a measure to identify periods during which the camera-derived heart rate cannot be compared with the reference heart rate because of the low quality of the reference signals.

6.5 Computing the reference heart rate

The comparison between the ECG heart rate and PPG heart rate provided by the Philip monitor in section 6.3 shows discrepancies in both time and physiological values. Since the true value of the heart rate is not known, there is a need to compute a new reference heart rate to be compared against the estimates from the video camera.

The computation of the reference heart rate was briefly described in section 6.2 and figure 6.1. It is performed by considering the agreement of the two heart rate measurements computed

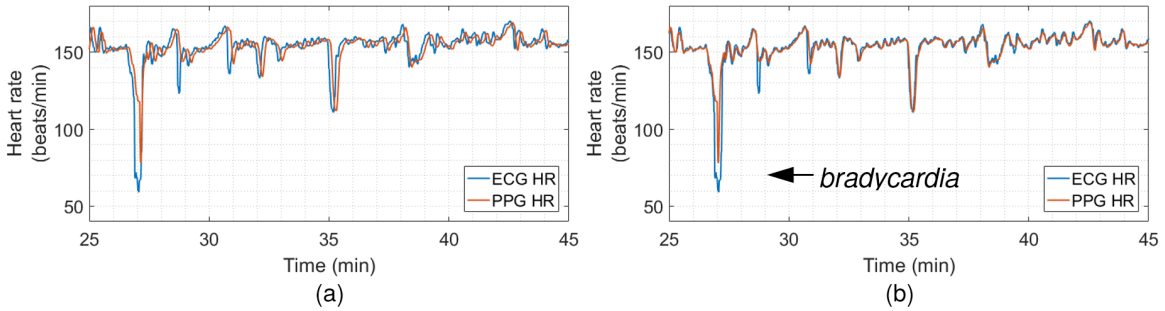


Figure 6.14: Heart rate estimates provided by the Philips monitor for a typical session (a) before signal alignment and (b) after signal alignment. The time delay between the two signals is estimated as 6 seconds.

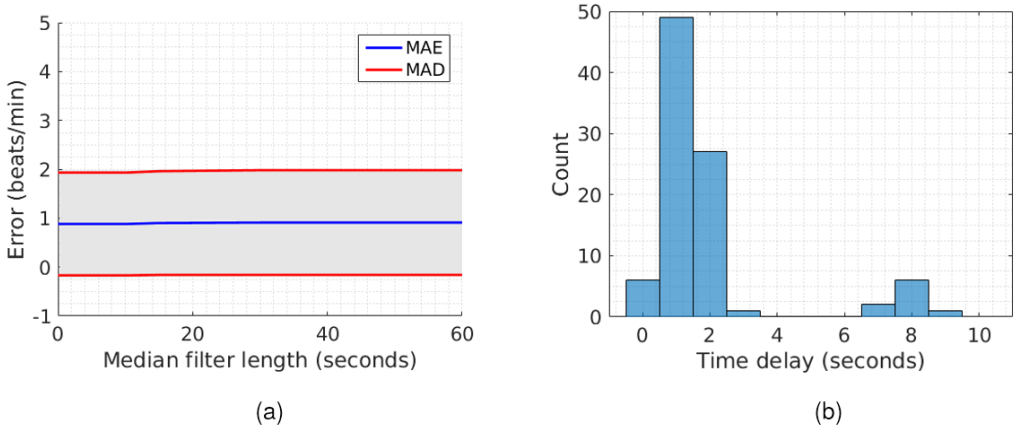


Figure 6.15: Results of time alignment between the ECG heart rate and PPG heart rate for all sessions in the Neonatal HDU dataset. (a) Errors while varying the length of median filter applied to each measurement before time alignment. (b) Histogram of time delays used to align the measurements.

by the Philips IntelliVue patient monitor and the quality of the ECG signal. Even though both measurements are available at a frequency of 1 Hz, the two measurements need to be time-aligned. This is performed by calculating the time delay of the PPG heart rate with respect to the ECG heart rate using the cross-correlation method as described in section 6.4.5. Figure 6.14 shows samples of the ECG heart rate and PPG heart rate time series before and after signal alignment. The new reference heart rate is calculated, on a second-by-second basis, as the mean of the two heart rate measurements for which both values do not differ by more than 5 beats/min, as recommended by the ANSI/AAMI EC13:2002 standard [176], and the SQI_{HR} is greater than 0.5, which implies a moderate signal quality as suggested by [189].

Figure 6.15 shows the results of the time alignment between the ECG heart rate and PPG heart rate for all sessions in the Neonatal HDU dataset. The evaluation of the median filter used in the time alignment process shows that the length of the median filter has a small effect on signal alignment (see figure 6.15a). A 10-second median filter, therefore, is applied to reduce sudden spikes or decreases in signal value prior to the calculation of the time delay (as discussed in section 6.4.5). Figure 6.15b shows the time delays calculated across all the sessions. Most sessions had time

delays of between 1 and 3 seconds. Several sessions have time delays between 8 and 10 seconds as a result of errors with the neonatal module for monitoring PPG during the first few sessions. This module was subsequently replaced for the later sessions.

6.6 Results

Table 6.2 summarises the errors calculated when comparing the ECG heart rate, PPG heart rate and the new reference heart rate. The comparison between the original ECG heart rate and PPG heart rate, as shown in figure 6.3, has a mean absolute error of 4.1 beats/min and a mean absolute

Table 6.2: Comparison between the ECG heart rate and PPG heart rate provided by the Philips monitor for all sessions in the Neonatal HDU dataset, corresponding to a total recording of 426.6 hours

Values compared	MAE (beats/min)	MAD (beats/min)	Estimated time (%, hours)
Original values, no time alignment	4.1	4.0	98.6%, 420.6 hours
Reference heart rate with the difference $\leq$ 5 beats/min	1.0	1.1	94.5%, 403.1 hours
Reference heart rate with the difference $\leq$ 5 beats/min and $SQI_{\mathrm{HR}} \geq 0.5$	0.9	1.0	91.2%, 388.9 hours

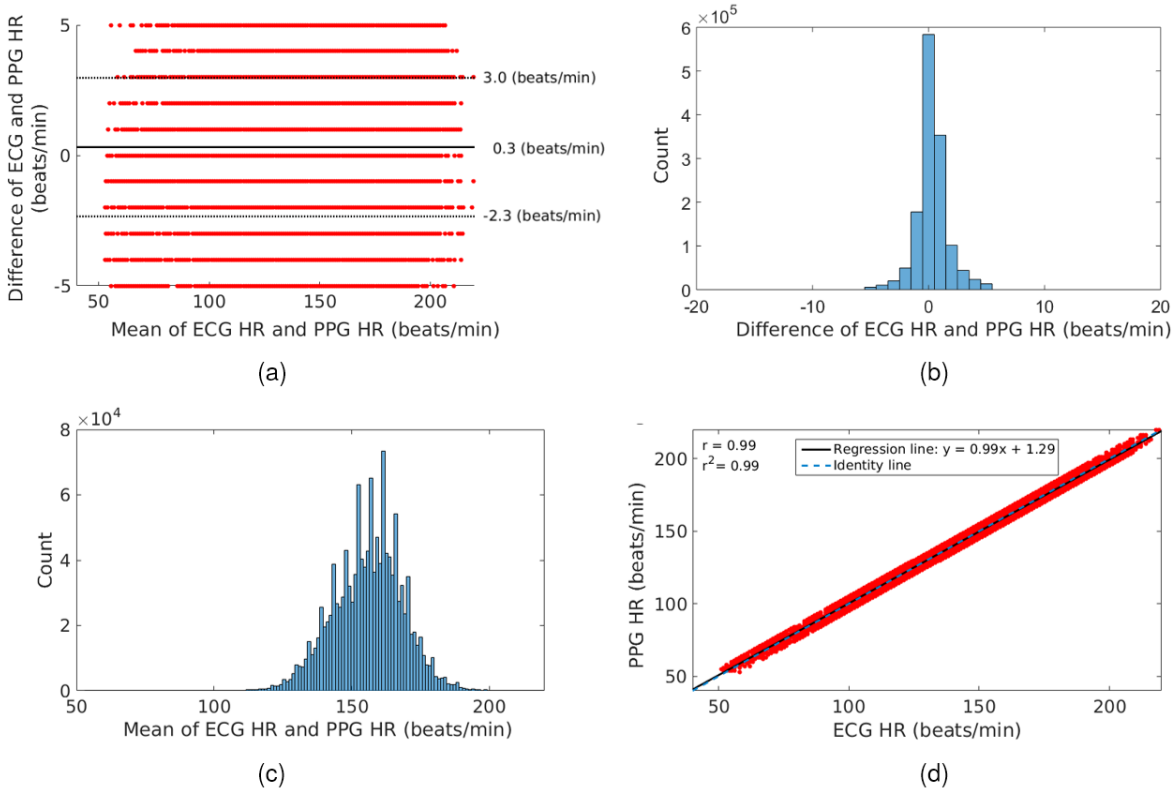


Figure 6.16: Comparison between ECG heart rate and PPG heart rate measurements where the absolute difference between the two measurements was limited to 5 beats/min and the SQI_{HR} was higher than 0.5. (a) Bland-Altman plot. (b) Histogram of difference values. (c) Histogram of mean values. (d) Correlation plot of the ECG heart rate against the PPG heart rate.

Table 6.3: The amount of valid heart rate reference data organised by recording sessions

(a) Training set

Session	Total reference recording time	Valid time	
		hh:mm	%
PN01 Day 1	05:23	05:01	93.3
PN01 Day 2	05:15	05:04	96.2
PN01 Day 3	04:59	04:41	94.0
PN02 Day 1	03:59	03:43	93.7
PN02 Day 2	07:07	06:52	96.4
PN02 Day 3	06:49	05:08	75.3
PN02 Day 4	04:45	04:33	95.9
PN03 Day 1	06:40	06:27	96.8
PN03 Day 2	06:06	05:46	94.7
PN03 Day 3	04:32	04:22	96.3
PN05 Day 1	06:28	05:54	91.3
PN06 Day 1	04:13	04:03	96.1
PN06 Day 2	04:33	04:09	91.0
PN06 Day 3	02:58	02:41	90.4
PN07 Day 1	06:05	05:46	94.9
PN07 Day 2	06:07	05:55	96.8
PN07 Day 3	06:07	05:01	82.1
PN09 Day 1	06:47	06:02	88.9
PN09 Day 2	05:25	04:44	87.5
PN09 Day 3	05:58	05:28	91.7
PN11 Day 1	03:54	03:33	91.0
PN11 Day 2	04:39	04:31	97.1
PN11 Day 3	02:54	02:50	97.4
PN15 Day 1	03:00	02:55	96.9
PN15 Day 2	03:53	03:47	97.5
PN15 Day 3	03:10	03:08	98.8
PN15 Day 4	03:03	02:51	93.8
PN16 Day 1	03:57	03:29	88.2
PN16 Day 2	05:49	05:15	90.2
PN17 Day 1	04:47	03:27	72.2
PN17 Day 2	06:34	05:42	86.8
PN17 Day 3	05:55	05:32	93.7
PN19 Day 1	07:00	06:48	97.2
PN19 Day 2	05:50	05:22	92.2
PN19 Day 3	05:17	05:07	96.9
PN21 Day 1	02:01	01:56	95.9
PN21 Day 2	06:27	05:28	84.7
PN21 Day 3	03:11	03:02	95.2
PN23 Day 1	05:27	05:13	95.7
PN23 Day 2	03:58	03:52	97.5
PN23 Day 3	06:11	05:59	96.7
PN25 Day 1	06:23	06:13	97.4
PN25 Day 2	03:02	02:58	97.7
Total	216:36	200:19	92.5

(b) Test set

Session	Total reference recording time	Valid time	
		hh:mm	%
PN04 Day 1	01:58	01:55	97.5
PN04 Day 2	04:43	04:39	98.5
PN04 Day 3	05:22	05:04	94.5
PN04 Day 4	08:05	07:43	95.3
PN08 Day 1	05:51	05:30	94.0
PN08 Day 2	05:52	05:37	95.8
PN08 Day 3	03:44	03:33	94.8
PN10 Day 1	03:37	03:34	98.4
PN10 Day 2	08:17	08:01	96.9
PN10 Day 3	04:37	04:29	97.2
PN12 Day 1	06:16	05:24	86.0
PN13 Day 1	02:58	02:48	94.7
PN13 Day 2	09:17	08:12	88.4
PN13 Day 3	02:24	02:20	96.9
PN14 Day 1	*Reference data not recorded*		
PN14 Day 2	06:20	06:06	96.3
PN14 Day 3	05:60	05:37	93.8
PN14 Day 4	04:41	04:18	91.9
PN18 Day 1	03:04	02:47	90.9
PN18 Day 2	02:13	01:58	88.7
PN18 Day 3	04:14	02:24	56.6
PN20 Day 1	04:16	03:46	88.1
PN20 Day 2	04:29	04:20	96.5
PN20 Day 3	04:18	04:01	93.5
PN20 Day 4	02:30	02:18	92.2
PN22 Day 1	03:49	03:40	95.8
PN22 Day 2	04:01	03:55	97.6
PN22 Day 3	02:31	02:14	88.5
PN24 Day 1	05:01	04:55	97.9
PN24 Day 2	05:42	04:43	82.7
PN24 Day 3	03:03	02:36	85.4
PN26 Day 1	01:45	01:43	97.4
PN26 Day 2	03:09	00:59	31.1
PN26 Day 3	03:26	02:60	87.2
PN26 Day 4	03:27	02:22	68.7
PN27 Day 1	04:34	04:03	88.8
PN27 Day 2	03:55	02:59	76.3
PN28 Day 1	06:27	06:16	97.2
PN28 Day 2	05:31	05:20	96.7
PN28 Day 3	05:47	05:02	87.0
PN29 Day 1	06:08	05:31	89.8
PN29 Day 2	06:12	04:49	77.7
PN29 Day 3	05:51	05:08	87.9
PN29 Day 4	03:07	02:42	86.6
PN30 Day 1	04:01	03:31	87.5
PN30 Day 2	03:38	03:03	83.9
PN30 Day 3	03:51	03:45	97.4
Total	210:02	188:38	89.8

deviation of 4.0 beats/min for 98.6% of the total recording time (420.6 hours). For about 1.4% of the time (6 hours), the data from at least one of the devices is not available (the Philips monitor reported a value of 0 or undefined – NaN). For around 94.5% of the total recording time (403.1 hours) the two measurements do not differ by more than 5 beats/min, with a mean absolute error of 1.0 beats/min and a mean absolute deviation of 1.0 beats/min. When computing the reference heart rate, it is found that, for approximately 91.2% of the total recording time (388.9 hours), the two measurements do not differ by more than 5 beats/min and the SQI_{HR} is greater than 0.5, with a mean absolute error of 0.9 beats/min and a mean absolute deviation of 1.0 beats/min.

Figure 6.16 shows the detailed comparison between the ECG and PPG heart rate estimates for which the mean absolute difference between the two measurements is less than 5 beats/min and the SQI_{HR} is higher than 0.5. According to the Bland-Altman plot (see figure 6.16a), the bias is 0.3 beats/min, and 95% of the difference between the two measurements falls within [-2.3, 3.0] beats/min. The mean values of the two measurements are distributed within the expected range of the neonatal population (taken to be between 60 and 230 beats/min). The correlation plot reveals a high correlation with a coefficient r of 0.99. These results imply a good agreement between the ECG heart rate and PPG heart rate estimates under these conditions.

Table 6.3 shows the amount of valid heart rate reference data for each recording session, listed separately for the training set and the test set. In the training set, over 200.1 hours of 216.6 hours total reference recording time (92.5%) are found to be valid. Similarly, in the test set, over 188.6 hours of 210.0 hours (89.8%) are valid.

6.7 Discussion

The Bland-Altman analysis between the ECG and PPG heart rate estimates in Figure 6.16a reveals a small bias of close to zero and a low variance. The correlation between the two measurements is very high, demonstrating a good agreement between the two heart rate estimates.

The differences between the two values are mainly caused by the manufacturer's signal processing algorithms, motion artefacts and interactions between the infant and the clinical staff. The difference in physiological values is mostly due to infant activities in the incubator. As mentioned before, the disagreement between the two heart rate estimates can provide a measure of the quality of the reference data. By analysing the ECG signal, periods during which the ECG is contaminated with noise or artefacts could also be detected and excluded from the comparison with the video camera estimates (see section 6.3). The new reference heart rate, therefore, can be said to provide a trustworthy reference signal.

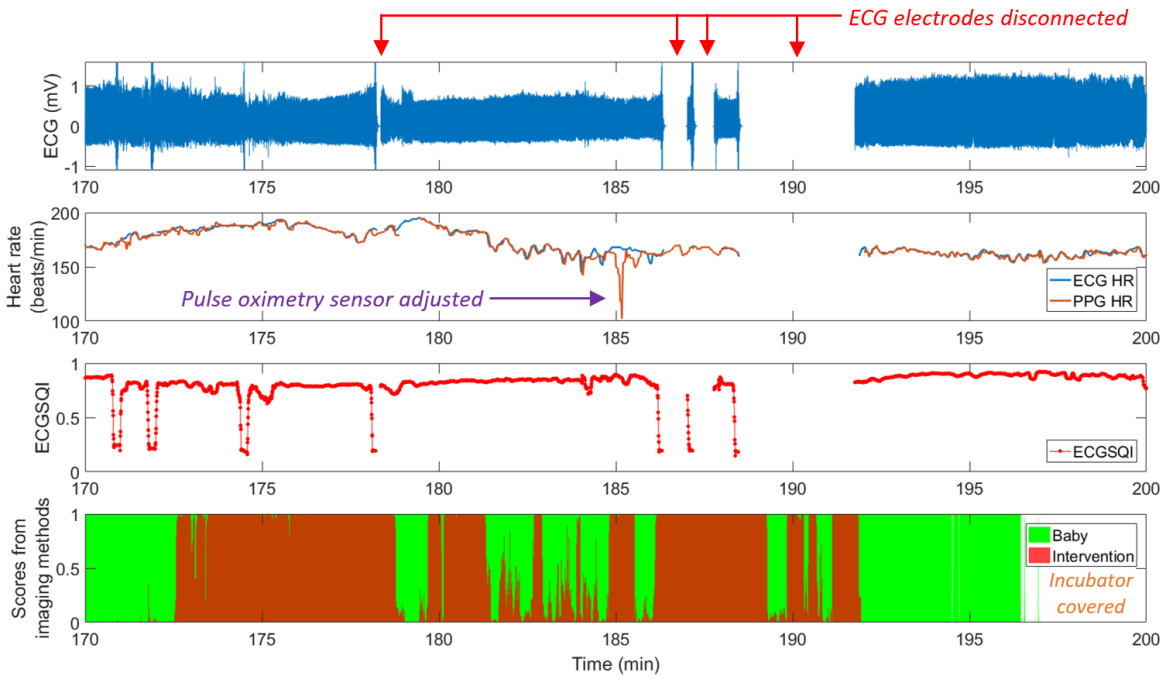


Figure 6.17: The relationship between the reference signals and the results from the image processing algorithms described in chapters 4 and 5. The plot shows a session during which the ECG electrodes and the pulse oximetry probe were replaced as part of routine care. The imaging algorithms can correctly identify periods during which the infant is not present in the video recording or is undergoing a clinical intervention. During these periods, the reference signals are noisy. The replacement of the sensor caused large errors between the ECG and PPG heart rate estimates. As expected, the presence of noise in the ECG has a negative impact on the SQI_{HR}.

Analysis of reference signals and image processing results

The image processing algorithms presented in chapters 4 and 5 can also be used to explain the quality of the data. They can provide an insight into the patient and nursing activities in the video recording. For example, when the nurse covers the incubator with a blanket during a resting period and the patient detection algorithm cannot detect the presence of the infant, the reference signals generally agree well as the baby falls asleep and move less. During the time periods that clinical interventions (e.g. changing sensors or a diaper) are undertaken, the reference signals are often noisy or not being recorded at all.

Figure 6.17 shows the plot of the reference signals together with the SQI_{ECG} and the results of the image processing algorithms, for a 30-minute section during which the ECG electrodes and the pulse oximetry sensor were replaced as part of routine clinical care. The ECG heart rate and PPG heart rate are not provided by the Philips monitor whilst the sensors are disconnected. The imaging algorithms identify when the baby is present in the video frame (labelled as baby) and whether clinical staff are in the video frame (labelled as intervention). These results match with the periods during which the reference signals are of poor quality

6.8 Conclusion

The analysis described in this chapter reveals inner discrepancies in the ECG and PPG heart rate estimates provided by the Philips monitor when they are compared against each other. A new reference heart rate is computed by first aligning the two heart rate measurements and then calculating the mean of the two measurements that do not differ by more than 5 beats/min and SQI_{HR} is greater than 0.5. The new reference heart rate is suitable for comparison with the camera-derived estimates.

Chapter 7
Heart rate estimation from the video camera

7.1 Introduction

Several techniques have been proposed for the estimation of heart rate from a video camera (see chapter 2). The majority of them were not designed to perform continuous heart-rate estimation in the clinic. This chapter describes an algorithm for doing so using the video camera data obtained from pre-term infants in the Neonatal HDU study (see chapter 3). The estimates of heart rate are then compared against the reference heart rate computed in the previous chapter.

The development of an algorithm that can automatically perform continuous non-contact vital-sign monitoring in an unconstrained clinical setting is very challenging. The robust framework presented in chapter 4 provides accurate per-frame skin segmentation which can be used to select a region of interest (ROI) for estimating physiological parameters. The aim of this chapter is to investigate the feasibility of using the segmented skin area to extract a photoplethysmographic imaging (PPGi) signal from the video data and to compute heart rate estimates from this signal.

The chapter is structured as follows: Section 7.2 outlines the overall methodology for heart rate estimation. Section 7.3 describes the process of extracting the PPGi signal from the video camera data. Section 7.4 explains the techniques used to assess the quality of the PPGi signal. Section 7.5 describes the methods, in both the time and frequency domains, that are used to compute heart rate estimates from the PPGi signal. Section 7.6 presents the comparison of the reference heart rate and the camera-derived estimates as well as detailed results for each recording session. Section 7.7 discusses the results and provides further insight into the use of the whole skin area for extracting the PPGi signal. Finally, section 7.8 summarises the chapter and presents possible directions to improve the estimation of heart rate in a clinical setting.

7.2 Overview of the heart rate estimation process

This section outlines the process of estimating heart rate from the video camera data. Figure 7.1 shows an overview of the estimation process. The patient detection and skin detection algorithms described in chapter 4 are applied to the video camera data to detect the presence of the patient and simultaneously segment the skin area. The intervention detection algorithm described in chapter 5 is applied to detect the periods during which no clinical intervention is undertaken. The raw PPGi signal is extracted from the video camera data by taking the mean of the segmented skin area over time. Since the extracted signal is likely to be contaminated by noise (e.g. motion artefacts and changes in the ambient light), a digital filter is used to suppress frequency components below 1.5 Hz and above 4.5 Hz, corresponding to a heart rate range of 90 to 270 beats/min. Peak and onset detection algorithms are then applied to the signal to identify salient points for each heart beat. Subsequently, the quality of the PPGi signal is analysed to identify good-quality beats. A change-point detection algorithm is applied to detect step changes in the raw PPGi signal. Infant motion is computed based on changes in the segmented skin area over time. Beats are flagged as invalid if they occur during

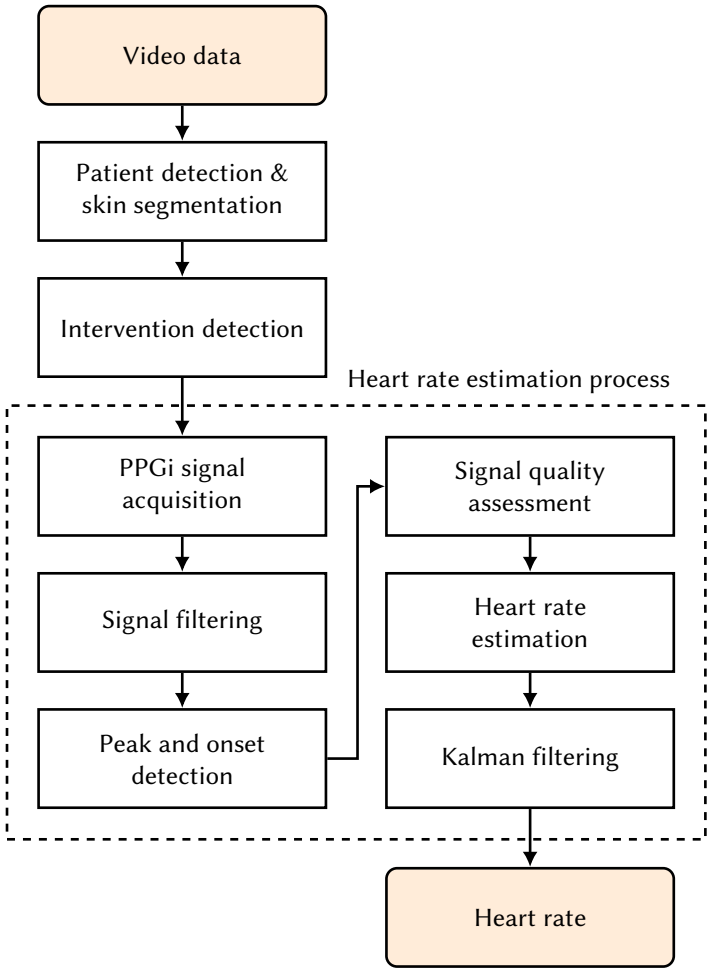


Figure 7.1: Block diagram for the heart rate estimation process from the video camera data.

a step change or during a period of high infant motion, or their amplitude is clipped, or the beat interval corresponds to a frequency outside the valid heart rate range, or the beat morphology is different from that of an average beat estimated using a multi-scale dynamic time-warping algorithm.

Heart rate estimation is performed for periods during which the PPGi signal is of good quality. A Kalman filter is applied to reduce the effects of transient artefacts and to improve the accuracy of heart rate estimation. The resulting heart rate estimates are compared against the gold-standard reference heart rate computed in chapter 6.

7.3 PPGi signal extraction

The extraction of the PPGi signal from the video camera data involves computing the mean colour intensity from pixels across the skin area (section 7.3.1), filtering the resulting signal to remove non-physiological frequencies (section 7.3.2) and finally detecting peaks and onsets in the resulting waveform (section 7.3.3).

7.3.1 Automatic PPGi acquisition

The PPGi signal can be remotely measured from the skin pixels in the image during periods of normal ambient light conditions [6]. The signal is extracted from the analysis of the subtle skin colour changes caused by blood flow in the body. One major challenge for the adoption of a non-contact vital-sign estimation system in the clinic is its ability to automatically acquire a continuous pulse signal from the video, and hence to estimate physiological signals with a clinically-relevant accuracy [5–8, 12, 13, 106]. The deep learning algorithms for patient detection and skin segmentation presented in chapter 4 compute accurate skin labels even during challenging scenarios, and this enables the automatic acquisition of the PPGi signal (see figure 7.2).

In the Neonatal HDU study, pre-term infants were nursed naked before the video camera mounted on top of the incubator. Typically, around 15% – 20% of the video frame corresponds to skin pixels (see figure 7.2a–b). Verkruysse *et al.* [6] showed that a consumer-level digital camera could be used to acquire PPGi signals. The authors reported that the green colour channel contained the strongest plethysmographic signal. This corresponds to the peak of the absorption spectrum of oxyheamoglobin in pulse oximetry [193]. Maeda *et al.* [194] compared the signal from green-light reflected PPG with the signal from infrared transmitted PPG and found a strong correlation between the green-light PPG signal and the R-peaks of the ECG signal recorded in the same study. Since most consumer-level cameras use a Bayer filter with a larger number of green-colour filters than other colour wavelengths to produce the final image (see chapter 3), the use of consumer cameras in [6] also contributed to the strength of the plethysmographic signal obtained from the green

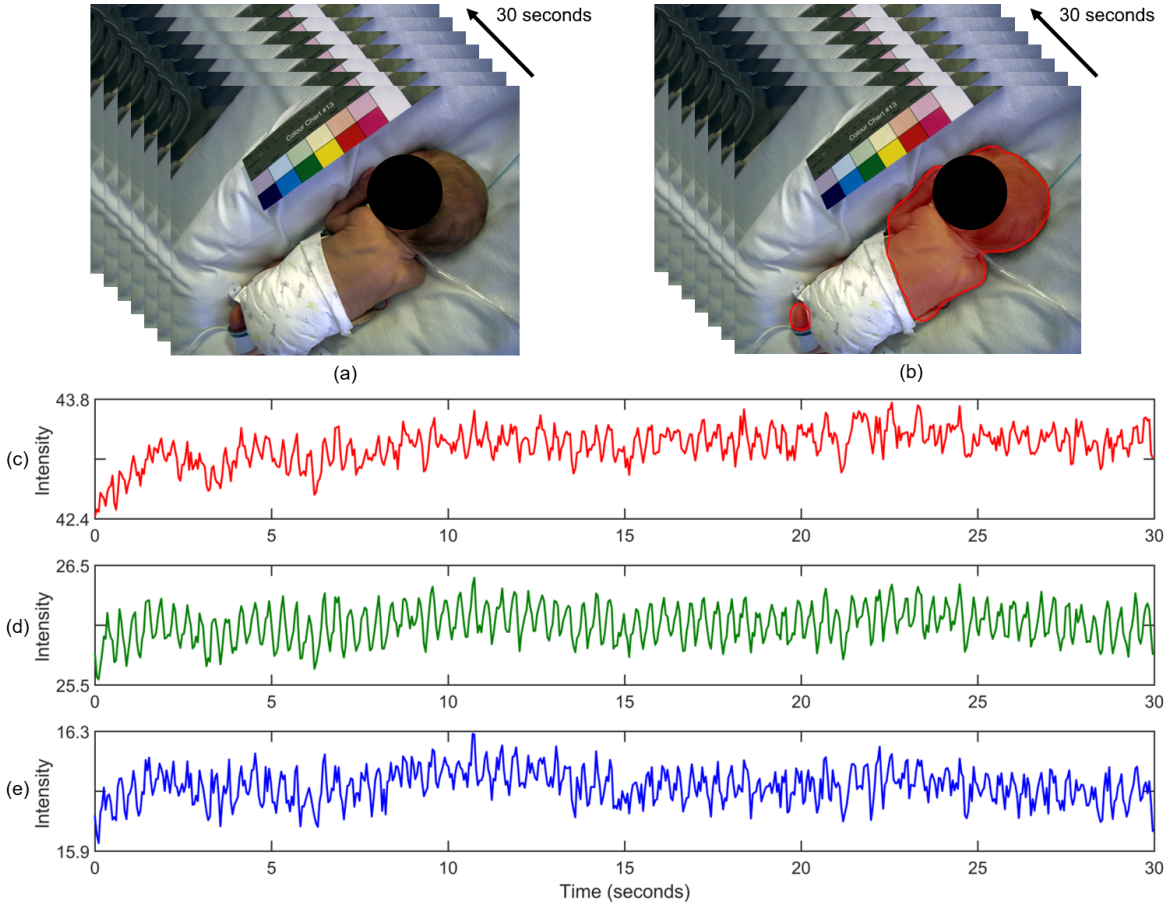


Figure 7.2: Extraction of raw PPGi signals from the skin area. (a) Video frames obtained from a typical recording session. (b) Segmented skin area provided by the deep learning algorithms presented in chapter 4. (c–e) 30-second time series extracted from the mean of the entire segmented skin area over time in the red, green and blue channels respectively. The raw PPGi signal from the green channel contains the strongest plethysmographic component [6]. The heart rate during this period is 152 beats/min.

colour channel. Tarassenko *et al.* [14] showed that a skin ROI needs to be large enough such that a PPGi signal with a strong cardiac component can be obtained.

In this work, the output from the deep learning model (originally 512×512 pixels, with the aspect ratio maintained) is cropped and resized to match the size of the image frame in the original videos (1620×1236). The raw PPGi signal is derived by spatially averaging all pixels in the whole skin area for the green colour channel for each frame in the video. The raw PPGi signal is extracted from the raw uncompressed video data stored at a resolution of 1620×1236 pixels and a frame rate of 20 frames per second. Figure 7.2 shows an example of the acquisition of the raw PPGi signal from the skin.

Let G_t be the green channel of a video frame I_t at time t and $S_t \in \{0, 1\}$ be a skin label of I_t computed by the algorithm presented in chapter 4, where the subscripts 0 and 1 denote non-skin and skin classes respectively. The raw PPGi signal is defined as

$$\text{PPGi}_{\text{raw}} = \frac{1}{N_{\text{skin}}} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} G_t(i,j) S_t(i,j) \tag{7.1}$$

where i and j are spatial coordinates and N_{skin} is the number of segmented skin pixels in the image. The raw PPGi signal is computed on a per-frame basis such that the signal's sampling rate is equal to the video frame rate of 20 Hz. The signal contains the pulsatile AC variations superimposed on a non-pulsatile DC component, as shown in figure 7.2.

Pre-term infants are particularly active. Their activity patterns, especially from the hands and the legs, may distort the cardiac signal. Some skin regions, such as the nose or the head, are covered with hair and can introduce measurement errors that reduced the ability of the camera to obtain a strong pulsatile component. This effect is compensated for by the larger skin areas found in the face (if visible) and upper torso. In addition, pre-term infants, compared to full-term infants or adults, have an immature epithelial layer of skin, limited subcutaneous fat, and increased skin vascularity [195]. These physiological characteristics can increase the strength of the cardiac signal and benefit the estimation of physiological signals from the pre-term infant population.

Figure 7.3 shows four comparisons of raw PPGi signals obtained from square ROIs of size 25×25, 50×50, 100×100 and 150×150 pixels compared with the signal obtained from the whole segmented skin area during a resting period of 15 seconds for which the average heart rate of the patient was 148 beats/min. Heart beats are more prominent in larger ROIs (100×100 and 150×150 pixels); the smaller ROIs (25×25 and 50×50 pixels) are corrupted by noise and motion likely to be caused

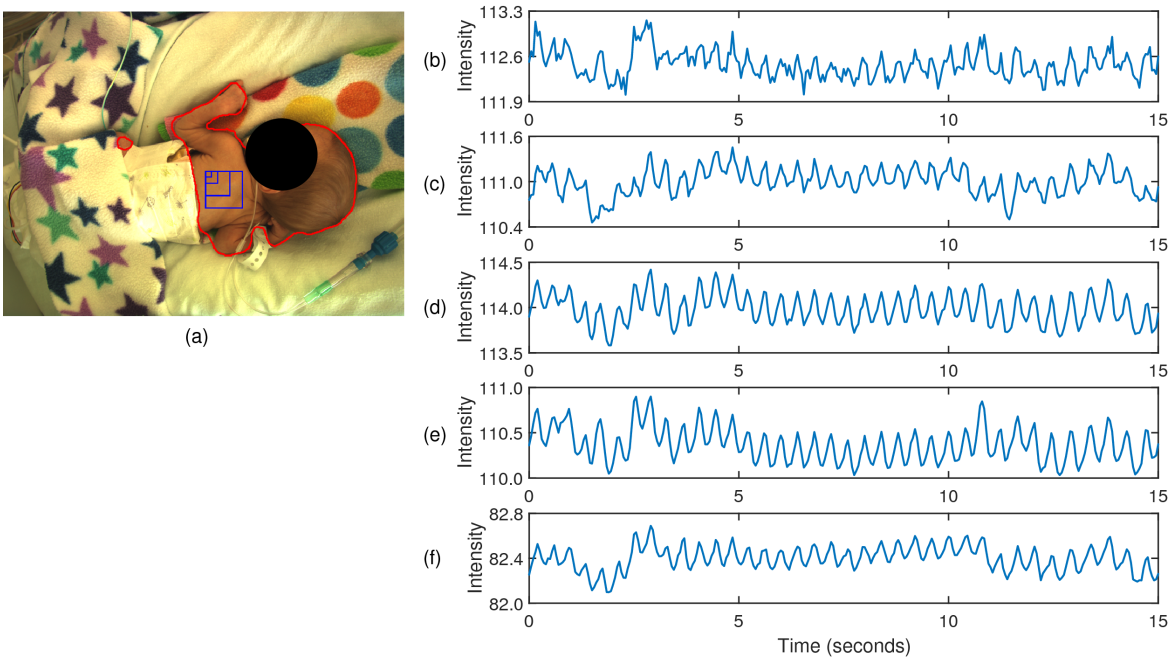


Figure 7.3: Raw PPGi signals extracted from the green channel using five different ROI sizes. (a) The first video frame of the signals where the blue squares indicate the location and size of ROIs located across the skin area and the red line indicates the segmented skin area obtained from the deep learning algorithms. (b–e) 15-second raw PPGi signals obtained from the ROIs of size:(b) 25×25 pixels, (c) 50×50 pixels, (d) 100×100 pixels (e) 150×150 pixels. (f) PPGi signal extracted from the entire skin area (red segment in (a)). Each signal contains 37 peaks, corresponding to a heart rate of 148 beats/min, equal to the reference heart rate from the Philips monitor.

by breathing. The results are consistent with those presented in [6], [14] and [54]. The signal from the whole skin area has a pulsatile nature and amplitude changes comparable to the signals from the square regions corresponding to the fixed ROIs.

7.3.2 Band-pass filtering

The PPGi signal contains pulsatile components correlated with the cardiac frequency as well as motion artefacts and often other sources of noise. Since the raw PPGi signal extracted from the skin is sensitive to disturbances such as motion artefacts and lighting changes, the signal is first detrended in order to remove any non-pulsatile DC offset, and then filtered to reduce frequency components outside the physiological range of interest.

Several studies [8, 9, 11, 14, 106, 107, 196, 197] suggest the use of a band-pass filter or a cascade of low-pass and high-pass filters to extract the frequency band of interest when attempting to estimate the heart rate. The normal heart rate of pre-term infants ranges from 90 to 180 beats/min [42], which is higher than that of healthy adults and full-term infants. Analysis of the reference heart rates in the training set showed that more than 99% of the values are concentrated in the range of 90 to 270 beats/second (1.5 to 4.5 Hz). Hence, the pre-processing filter is designed as a cascade of a 40^{th}-order low-pass FIR filter with a cut-off frequency at 4.5 Hz and a 60^{th}-order high-pass FIR filter with a cut-off frequency at 1.5 Hz. Figure 7.4 shows a 30-second PPGi signal before and after band-pass filtering. The filtering successfully attenuates the non-physiological frequency components.

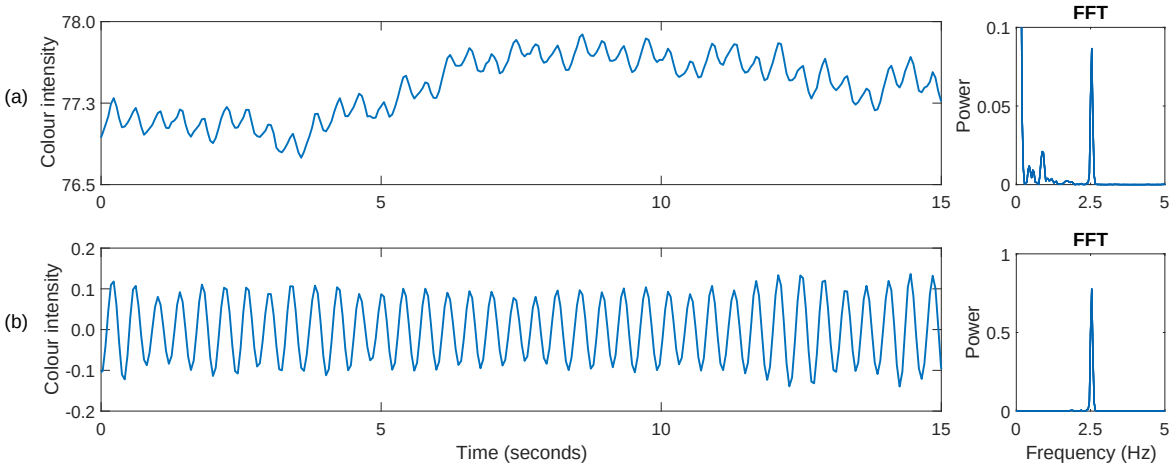


Figure 7.4: Band-pass filtering applied to the PPGi signal for a 15-second video segment. (a) The raw PPGi signal extracted from the infant's skin with the corresponding frequency components computed by the FFT. (b) Band-pass filtered signal with the non-physiological frequency components attenuated. There are 38 beats in the signal, corresponding to a heart rate of 152 beats/min. The reference heart rate was 151 beats/min.

7.3.3 Peak and onset detection

Peak and onset detection is a process whereby peaks and onsets corresponding to cardiac cycles are identified in the signal. Zong *et al.* [198] proposed an algorithm to determine the onset of

arterial blood pressure pulses based on computing the sum of the upslope of each pulse, called a slope sum function (SSF). Villarroel *et al.* [54] extended the algorithm to compute a boxed slope sum function (BSSF) and adaptively track the upslopes of the signal.

The slope sum function (SSF) aims to enhance the slope of the pulse and suppress the remainder of the waveform. Given x_k as the filtered PPGi signal at time k and u_k as a time-series signal containing positive slopes, the SSF is defined as:

$$\text{SSF}(i) = \sum_{k=i-w}^{i} u_k, \qquad u_k = \begin{cases} \Delta x_k & \text{if } \Delta x_k > 0 \\ 0 & \text{if } \Delta x_k \leq 0 \end{cases} \tag{7.2}$$

where w is the analysis window length and $\Delta x_k = x_k - x_{k-1}$. The SSF signal is the summation of the u_k over the running window w (see figure 7.5b). The window length w was chosen to be the typical duration of the upslope of the pulse; $w = 170$ ms or 3 samples for the PPGi signal with a sampling frequency of 20 Hz.

The boxed slope sum function (BSSF) is computed from the SSF signal by finding the zero-crossing points at the beginning of the upslope and at the end of the downslope for each pulse. The pulse onsets correspond to the upslope points in the BSSF signal (uBSSF). Figure 7.5a–d show the input PPGi signal, the resulting SSF, BSSF signals and the uBSSF data points respectively.

Adaptive thresholding and a decision rule are applied to the uBSSF data points to identify the peak and onset of each beat. The threshold value is initially taken to be the median uBSSF signal

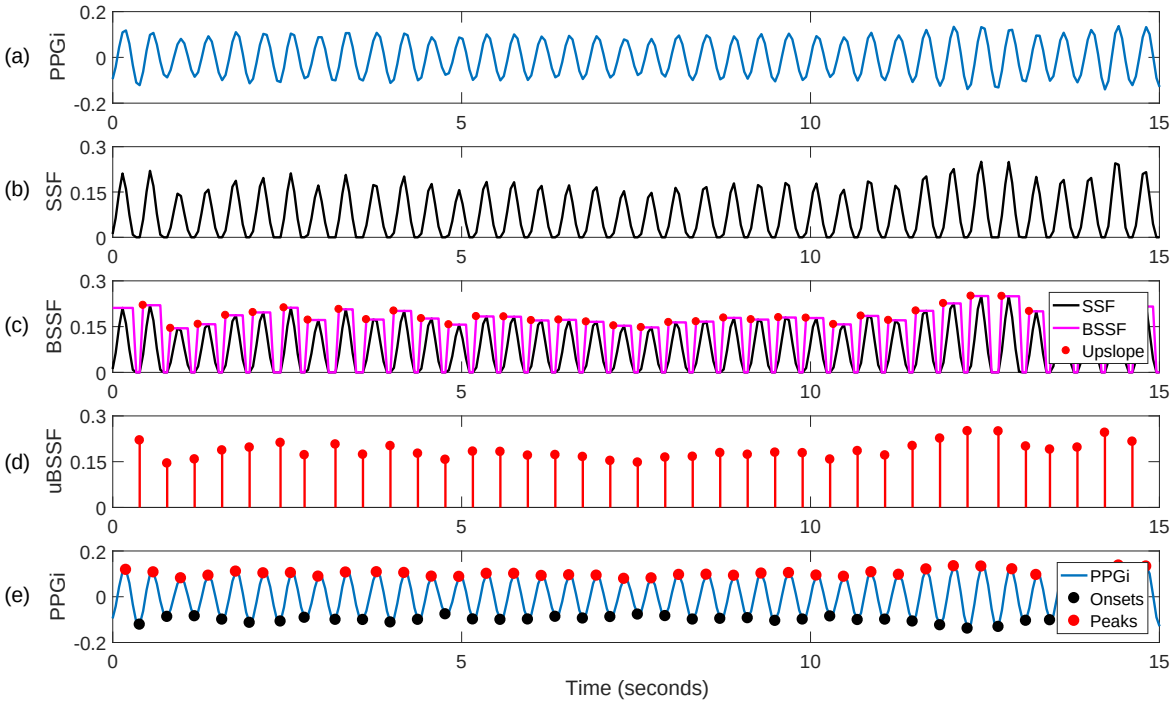


Figure 7.5: Peak and onset detection for a 15-second PPGi signal. (a) The PPGi signal. (b) SSF signal computed based on the sum of the slope of the pulses. (c) The BSSF signal is matched with each beat. (d) The uBSSF data points correspond to the upslope points in the BSSF signal. (e) The beat onsets and peaks are detected based on adaptive filtering and decision rules.

computed over the first ten seconds. For each uBSSF data point, the data point is considered a beat onset if its value crosses the threshold. The threshold is adaptively adjusted according to the uBSSF value of each detected beat onset. A 220-ms refractory period (a recovery time after the peak) is applied to prevent multiple onset detections. The peak is defined as the maximum value between a pair of successive onsets. Figure 7.5 shows an example of the peak and onset detection applied to a sample 15-second PPGi signal.

The algorithm is shown to be effective in annotating pulse peaks and onsets [54, 123, 198]. Accurate detection allows a beat-by-beat signal quality assessment of the PPGi signal and the estimation of heart rate by counting the number of beats in the time domain.

7.4 Signal quality assessment

The assessment of the quality of the PPGi signal is of high importance as data corruption by subject movements and changes in the lighting condition present considerable challenges for video analysis. Even when the infant is not active or sleeping, the video camera automatically adjusts the gain, lightness and other settings to compensate for changes in lighting conditions in the scene.

The assessment of the PPGi signal was extended from that described in Villarroel *et al.* [54, 123] by incorporating a new signal quality index measurement. Figure 7.6 outlines the signal quality assessment process. As an initial step, an activity index is computed based on changes in the segmented skin area corresponding to the movement of the subject. A Bayesian change point detection algorithm is applied to identify step changes, often caused by sudden lighting condition changes, in the raw PPGi signal. The pulses occurring during the periods of high subject motion and step changes are flagged as invalid. In order to further identify whether each detected beat is of good quality, the algorithm performs a beat-by-beat assessment by combining multiple analysis methods: frequency bounding, clipping detection, amplitude thresholding, and multi-scale dynamic time warping. Finally, the signal quality index of each detected beat (SQI_{beat}) was obtained.

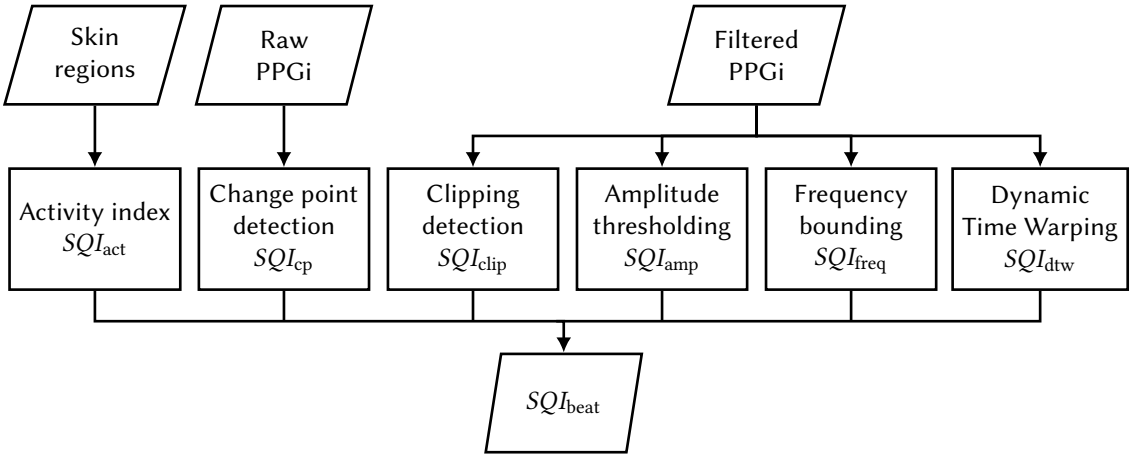


Figure 7.6: Block diagram outlines the signal quality assessment process.

7.4.1 Activity index

Suitable time periods for estimating heart rate from the video camera occur when the movement of the infant is minimal. Changes in the segmented skin area over time can be used as an indicator of the degree of subject motion. The centroid (C_x, C_y) of skin regions is defined as the average location of the predicted skin pixels in the horizontal and vertical directions. Motion $M(i)$ at frame i is defined as the Euclidean distance between centroids for two successive frames:

$$\mathrm{M}(i) = \sqrt{\big(C_x(i) - C_x(i-1)\big)^2 + \big(C_y(i) - C_y(i-1)\big)^2}. \tag{7.3}$$

The signal quality index (SQI) of the kth beat is taken to be 0 if the Euclidean distance in the kth beat and two beats both before and after (5 beats in total) is higher than a threshold of 20 pixels. Thus, the SQI_{act} is defined as

$$SQI_{\mathrm{act}}(k) = \begin{cases} 0 & \text{if } \exists\, i \in \{b_{k-2}, \ldots, b_{k+2}\}\ \mathrm{M}(i) > 20 \\ 1 & \text{otherwise} \end{cases} \tag{7.4}$$

where b_{k-2} and b_{k+2} denote the location of the entire $(k-2)$th and $(k+2)$th beat respectively. The distance threshold is set to 20 pixels, corresponding to a distance of approximately 1 cm measured by the ruler in the colour chart (see figure 7.7a–b) placed near the subject. The camera was positioned approximately 30 cm away from the subject for all recording sessions.

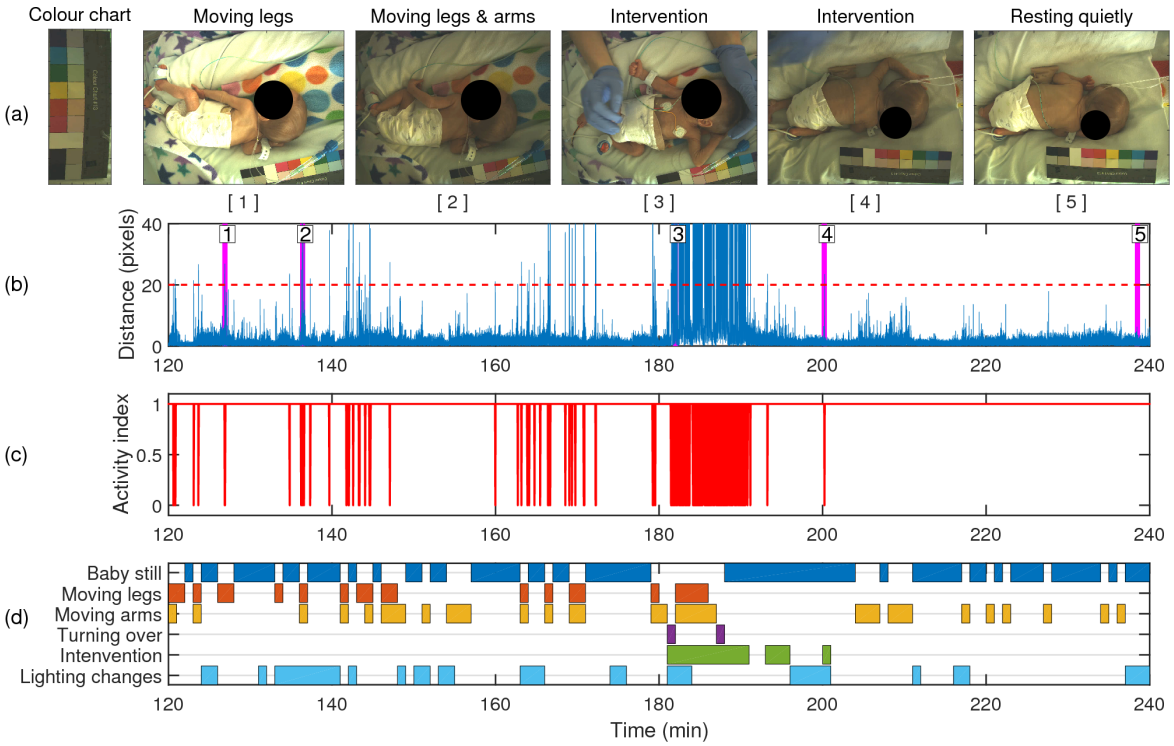


Figure 7.7: Calculation of the activity index for a typical 2-hour segment. (a) Video frames annotated with the labels corresponding to the time in the distance plot below. (b) Distance between skin centroids in successive frames. (c) Activity index computed by thresholding the distance, 0 denotes a period of high motion and 1 denotes a period during which the patient is considered to be still. (d) Manual annotation of events during the recording. Intervention indicates a period during which the infant is undergoing a procedure.

Figure 7.7 shows both the calculation of the activity index for a 2-hour video segment of a typical session and the manual annotations of activities in the video. The manual annotations were performed by the author marking the infant activities on a minute-by-minute basis. A large distance between successive skin centroids corresponds to the movement of the subject (see figure 7.7b&d).

7.4.2 Change point detection

Abrupt changes in the PPGi signal occurs due to changes in subject posture or sudden changes in the lighting conditions, for example: when the overhead light over the incubator is turned on or off; window blinds are opened or closed; or clinical staff walk pass by the incubator. Figure 7.8a shows an abrupt change as a result of a shadow cast on the infant.

In order to detect the location of these step changes, a Bayesian change point detection is applied to the raw PPGi signal [199]. Assume that data x_i of length N construct of two piece-wise constant inputs, μ_1 and μ_2, with added Gaussian noise e_i:

$$x_i = \begin{cases} \mu_1 + e_i & \text{if} \quad i < m \\ \mu_2 + e_i & \text{otherwise} \end{cases} \tag{7.5}$$

where m is the location of a step change. The likelihood of the data is defined as the joint probability of the Gaussian noise e_i:

$$P(x|\{\mu_1, \mu_2, \sigma, m\}) = \prod_{i=1}^{N} P(e_i) \tag{7.6}$$

where σ is the standard deviation of the Gaussian noise, $\{\mu_1, \mu_2, \sigma, m\}$ are the assumed values of the time series parameters, and $P(e_i)$ denotes the probability density function of the Gaussian noise samples. The posterior probability of a step change at location m is defined as [54, 199],

$$\mathrm{P}(m|x) \propto \frac{1}{\sqrt{m(N-m)}} \left[\sum_{i=1}^{N} x_i^2 - \frac{1}{m} \Big(\sum_{i=1}^{m} x_i \Big)^2 - \frac{1}{N-m} \Big(\sum_{i=m+1}^{N} x_i \Big)^2 \right]^{-\frac{N-2}{2}}. \tag{7.7}$$

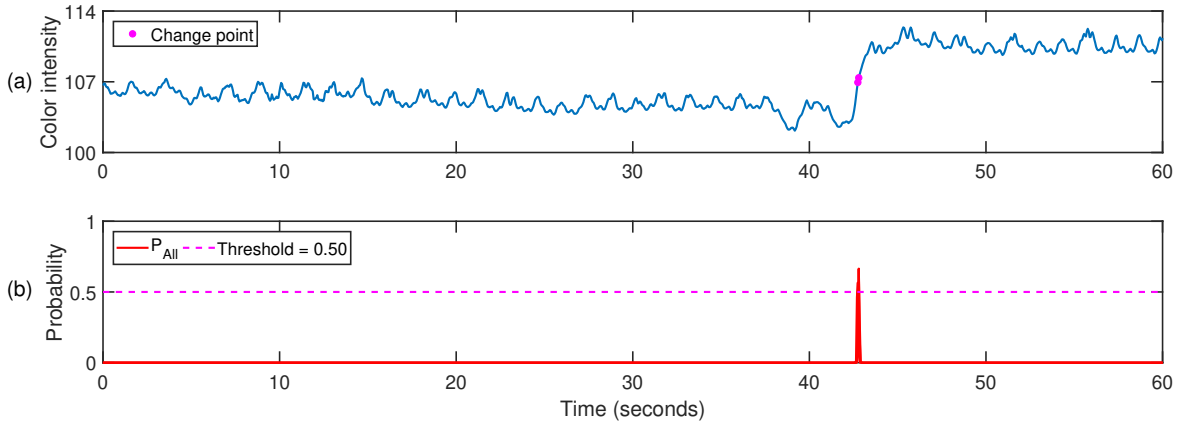


Figure 7.8: Change point detection for a 60-second raw PPGi segment. (a) Raw PPGi signal during a shadow cast on the subject causing step changes. Change points are marked as magenta dots. (b) Probability $\mathrm{P}_{\mathrm{all}}$ of the changes points detected on all window sizes. Note that the signal contains both cardiac and respiratory components (small changes correspond to heart beats and larger changes correspond to breathing).

A change point is deemed to occur if the probability $P(m|x)$ is above 0.50. Change point detection is performed on a window-by-window basis for different window sizes of 5, 10 and 15 seconds and a step size of 5 seconds, and all the change points detected are then merged together. Given $P_{\text{all}}(m)$ is the probability of a change point at m merged from the detections at multiple window sizes, the SQI_{cp} of the kth beat is defined, similarly to the SQI_{act}, as

$$SQI_{\text{cp}}(k) = \begin{cases} 0 & \text{if } \exists\, i \in \{b_{k-2}, \ldots, b_{k+2}\} \;\; P_{\text{all}}(i) > 0.50 \\ 1 & \text{otherwise} \end{cases} \tag{7.8}$$

where b_{k-2} and b_{k+2} denote the location of the entire $(k-2)$th and $(k+2)$th beats respectively. If a change point is detected at the location of the kth beat, the SQI_{cp} values of this beat and the two beats before and after are set to zero.

Figure 7.8 shows the result of the change point detection for a 60-second PPGi segment. A change point is at the time when a clinician walks past the incubator and casts a shadow.

7.4.3 Frequency bounding

Frequency bounding determines whether the instantaneous HR falls within the physiological range of typical pre-term infants, taken to be within a range of 90 and 270 beats/min. Given that HR_{inst} is the instantaneous heart rate of the kth beat, the SQI_{freq} is taken to be 0 if the HR_{inst} falls outside the valid physiological range:

$$SQI_{\text{freq}}(k) = \begin{cases} 0 & \text{if } \; HR_{\text{inst}}(k) < 90 \;\text{ and }\; HR_{\text{inst}}(k) > 270 \\ 1 & \text{otherwise} \end{cases}. \tag{7.9}$$

This heart rate range was defined according to the filter used for pre-processing the signal and the physiological data in the Neonatal HDU dataset (see section 7.3.2). Figure 7.9 shows an example of the frequency bounding detection in a 60-second PPGi segment. The beats during motion artefacts are flagged as invalid.

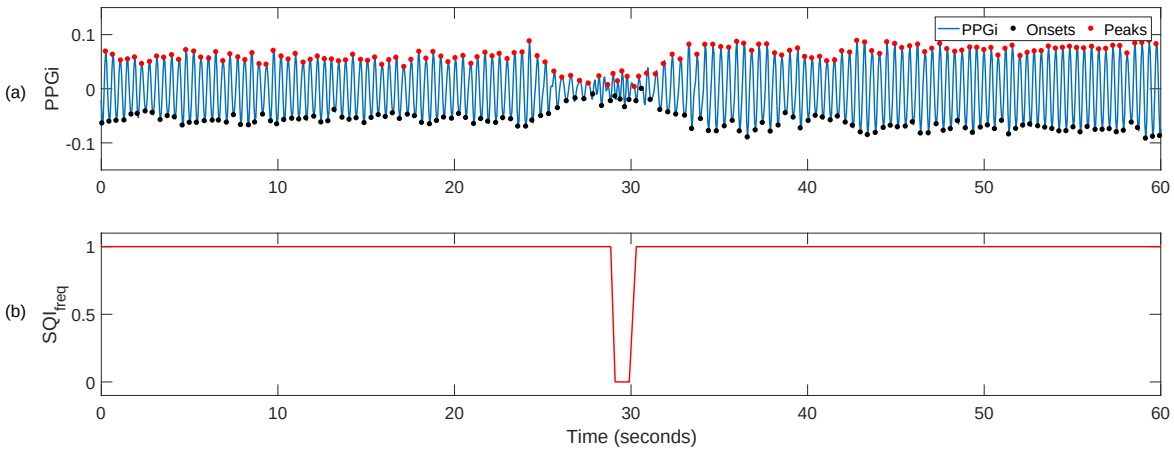


Figure 7.9: Frequency bounding detection determines whether the instantaneous HR is within the physiological bounds. (a) 60-second PPGi signal during a period of motion artefacts, (b) SQI_{freq} indicates the beats that have instantaneous HR outsided the physiological bounds, assigned a value of 0.

7.4.4 Clipping detection

Clipping generally occurrs as a result of motion artefacts. When the derivative of the signal is computed, the DC component is effectively attenuated. Signal clipping can be detected when the derivative crosses a given threshold [200]. Given that $N_{\text{length}}(k)$ is the length of the kth beat and $N_{\text{clipped}}(k)$ is the amount of the derivative of the kth beat that crosses a clipping threshold of 0.1, the SQI_{clip} of the kth beat is set to 0 when more than one-third of the derivative is clipped:

$$SQI_{\text{clip}}(k) = \begin{cases} 0 & \text{if } N_{\text{clipped}}(k)/N_{\text{length}}(k) > 1/3 \\ 1 & \text{otherwise} \end{cases}. \tag{7.10}$$

Figure 7.10 shows an example of clipping detection in a PPGi signal containing motion artefacts. The beats occurring during motion artefacts are flagged as invalid.

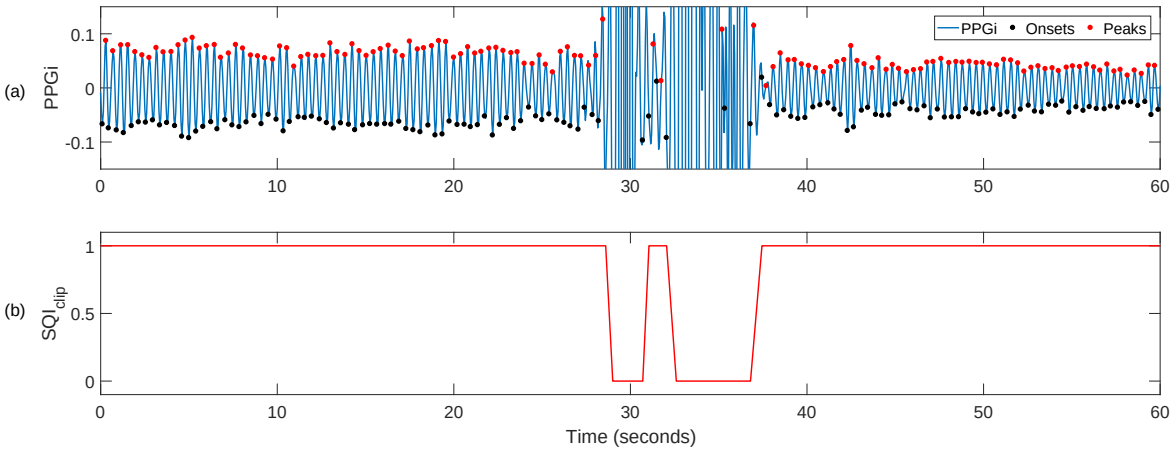


Figure 7.10: Clipping detection for a 60-second PPGi segment: (a) PPGi signal during a period of motion artefacts, (b) SQI_{clip} indicates that clipped beats are flagged as invalid (assigned a value of 0).

7.4.5 Amplitude thresholding

Amplitude thresholding is performed to determine whether the amplitude of each beat remains within three standard deviations σ_w from the mean μ_w of the window w. The statistics are calculated locally for each 15-second moving window w. The SQI_{amp} of the kth beat is set to 0 if part of the beat is outside the valid range:

$$SQI_{\text{amp}}(k) = \begin{cases} 0 & \text{if } \exists i \in \{b_k\} \quad PPG_{\text{filt}}(i) > \mu_w + 3 \cdot \sigma_w \text{ or} \\ & \qquad\qquad\quad PPG_{\text{filt}}(i) < \mu_w - 3 \cdot \sigma_w \\ 1 & \text{otherwise} \end{cases} \tag{7.11}$$

where b_k is the location of the entire kth beat. Figure 7.11 shows an example of the amplitude thresholding method on a 15-second PPGi signal obtained during a period of infant movement. The beats that are outside the valid amplitude range are flagged as invalid.

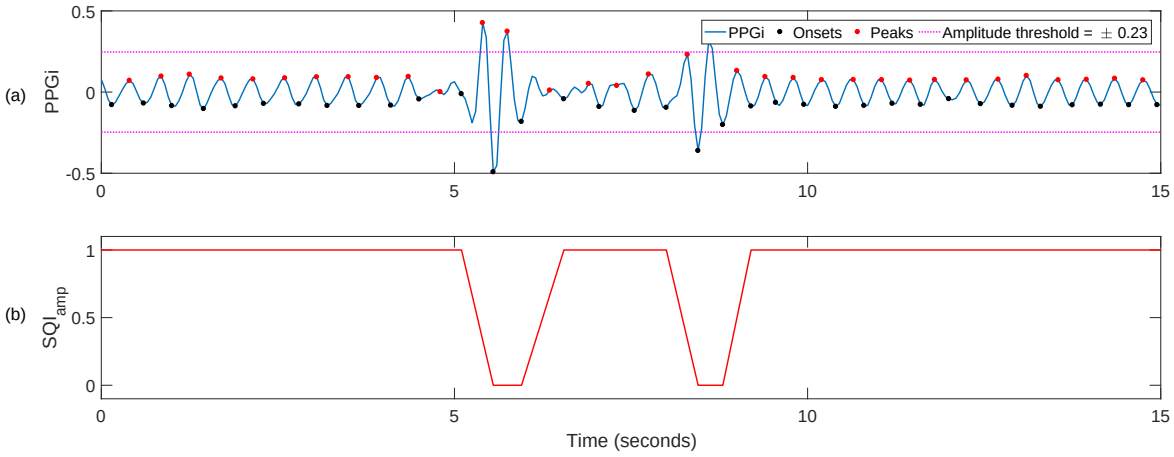


Figure 7.11: Amplitude thresholding for a 15-second PPGi segment: (a) PPGi signal, (b) SQI_{amp} indicates that the beats outside the window amplitude range are flagged as invalid by assigning a value of 0.

7.4.6 Multi-scale dynamic time warping

Dynamic time warping (DTW) is a time series technique used to determine a distance (or a degree of similarity) between two given time series based on the best possible alignment between the two [201]. The DTW technique is suitable for the time series whose characteristics may vary in time. For example, similarities in each cardiac cycle can be measured using the DTW technique, even if heart rate is increasing or decreasing during the course of an observation. The sequences are warped in the time domain to determine a degree of similarity, independent of temporal variations. The classical DTW algorithm is computationally intensive as it needs to evaluate every possible warping path in order to obtain an optimal alignment. Fitriani [202] and Salvador *et al.* [203] extended the algorithm to perform multi-scale warping by refining a search space for the optimal alignment between the two time series from a coarse to a finer resolution. Villarroel *et al.* [123][54] extended the multi-scale DTW technique for assessing the quality of pulsatile signals by determining the distance of the optimal alignment between each beat and a running beat template computed over a window.

To compute the signal quality based on the DTW method (SQI_{dtw}), the PPGi signal is first divided into 15-second moving windows with a step size of 5 seconds. Each window is assessed independently of each other. The following sections describe the process by which the SQI_{dtw} is calculated.

Template construction

For each window w, all beats in the window w are extracted and aligned, with their peaks serving as reference points. The template is computed using a three-stage technique. At the end of each stage, the template is computed by averaging all valid beats. The length of the template is set to be equal to the median length of all valid beats. For the first stage, the first template (T1) is computed from the pulses whose beat-to-beat interval corresponds to a heart rate between 90 and 270 breaths/min. In the second stage, the cross-correlation coefficient between each valid beat (after stage 1) and

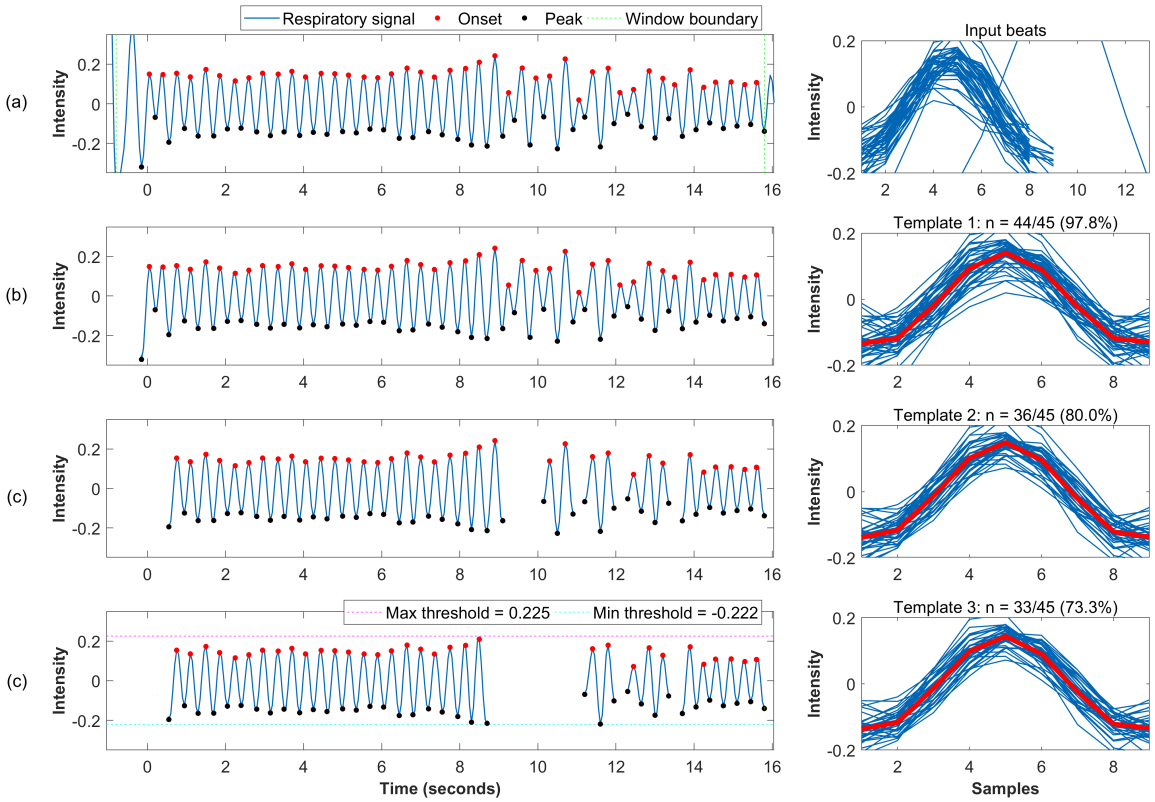


Figure 7.12: Template construction for multi-scale dynamic time warping.(a) Input time series with peak and onset markers. (b) Template 1 (T1) is constructed from the pulses whose peak-to-peak interval corresponds to a heart rate between 90 and 270 breaths/min. (c) Template 2 (T2) is constructed from the valid beats (after stage 1) whose correlation coefficient, calculated between each beat and the template T1, was larger than 0.8. (d) Template 3 (T3) was constructed from the valid beats (after stage 2) whose amplitude does not exceed the mean plus or minus three standard deviations of the time window.

the T1 template is calculated. The second template (T2) is computed from the beats that have a cross-correlation coefficient greater than 0.8. In the final stage, the third template (T3) is computed from the valid beats (after stage 2) whose amplitude is within the mean plus or minus three standard deviations for that window. The T3 template is considered valid, if there are more than 50% of the valid beats at the end of stage 3. If the template is invalid, the template from the previous window is used instead. The template is later employed to analyse the morphology of each beat in the window. Figure 7.12 shows an example of a template calculation for a sample 15-second window.

Time alignment

Once a template has been computed for a given window, time alignment is performed between each beat within the time window and the template. The DTW method compares each beat $X = \{x_1, x_2, ..., x_N\}$ of length N within the current window and the template $Y = \{y_1, y_2, ..., y_M\}$ of length M. Given an $N \times M$ cost matrix $C(X, Y)$ whose element $c(x_i, y_j)$ is the local distance measure between the current beat X at location i and the template Y at location j, the total cost $c_p(X, Y)$ of the warping path $p = \{p_1, p_2, ..., p_L\}$ of length L between X and Y by assigning the element

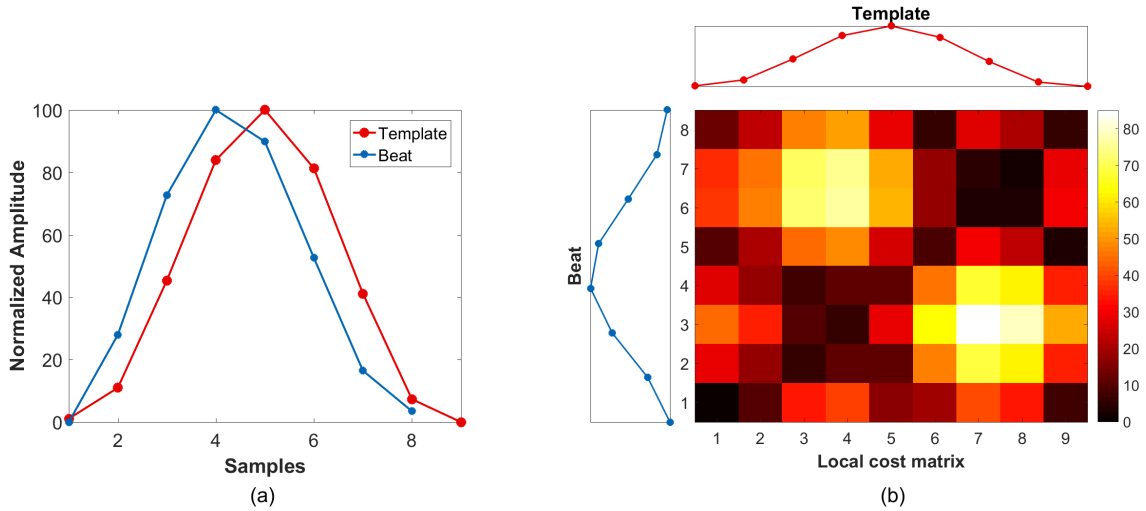


Figure 7.13: Computation of the local cost matrix. (a) Template (red) and input beat (blue). (b) Local cost matrix with lighter regions indicating locations where the template and the beat are further apart and darker regions indicating locations where they are close.

x_{n_ℓ} of X to the element y_{m_ℓ} of Y is given by

$$c_p(X, Y) = \sum_{\ell=1}^{L} c(x_{n_\ell}, y_{m_\ell}). \tag{7.12}$$

The local distance measure $c(x_{n_\ell}, y_{m_\ell})$ is calculated by comparing the beat against the template [54] (see figure 7.13). The cost has a small value if x_{n_ℓ} and y_{m_ℓ} are similar to each other, and a large value if they are different. The DTW method aims to find the optimal warping path between the beat and the template that has a minimum total cost. Thus, the DTW distance is defined as

$$DTW(X, Y) = \underset{p}{\text{argmin}}\ c_p(X, Y). \tag{7.13}$$

Dynamic programming is used to reduce the complexity of evaluating every possible warping path. A cumulative distance matrix $D(m, n)$ is computed from the cost matrix $C(X, Y)$ as

$$D(n, m) = DTW(X(1 : n), Y(1 : m)). \tag{7.14}$$

The optimal warping path passes through an area of low-cost values within the cumulative distance matrix $D(m, n)$.

In order to reduce the search space for an optimal warping path, the DTW method is performed in a multi-scale manner. Signal alignment is first performed for coarse or down-sampled versions of the time series X and Y. The optimal warping path is then iteratively projected onto the finer version and the search space is constrained within the boundary of the optimal warping path from the previous scale (see figure 7.14d).

Computing SQI_{dtw}

Multi-scale DTW is described in more detail, including its mathematical derivations, in [54]. Figure 7.14a–c shows an optimal warping path computed using the iterative multi-scale DTW method. The

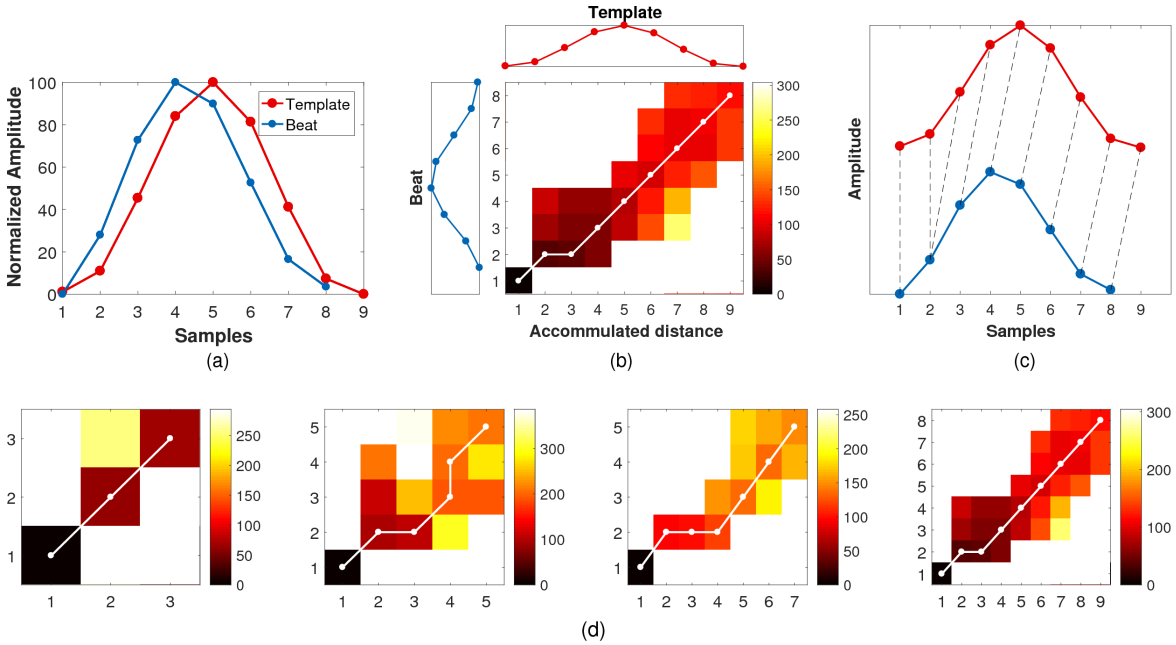


Figure 7.14: Computation of multi-scale dynamic time warping. (a) Template (red) and input beat (blue). (b) Accumulated distance matrix with darker regions indicating locations where the template and the beat are close and lighter regions indicating locations where they are apart. (c) Time alignment between the template and the beat. (d) Multi-scale dynamic time warping iteratively constrains the search space in order to find the optimal path.

DTW distance is computed for each individual beat in the window w (see figure 7.12c). It can be thought of as a measure of similarity between each beat and the average beat within the time window. The SQI_{dtw} is defined as

$$SQI_{\text{dtw}}(k) = 1 - DTW(X_k, Y_k)/100. \tag{7.15}$$

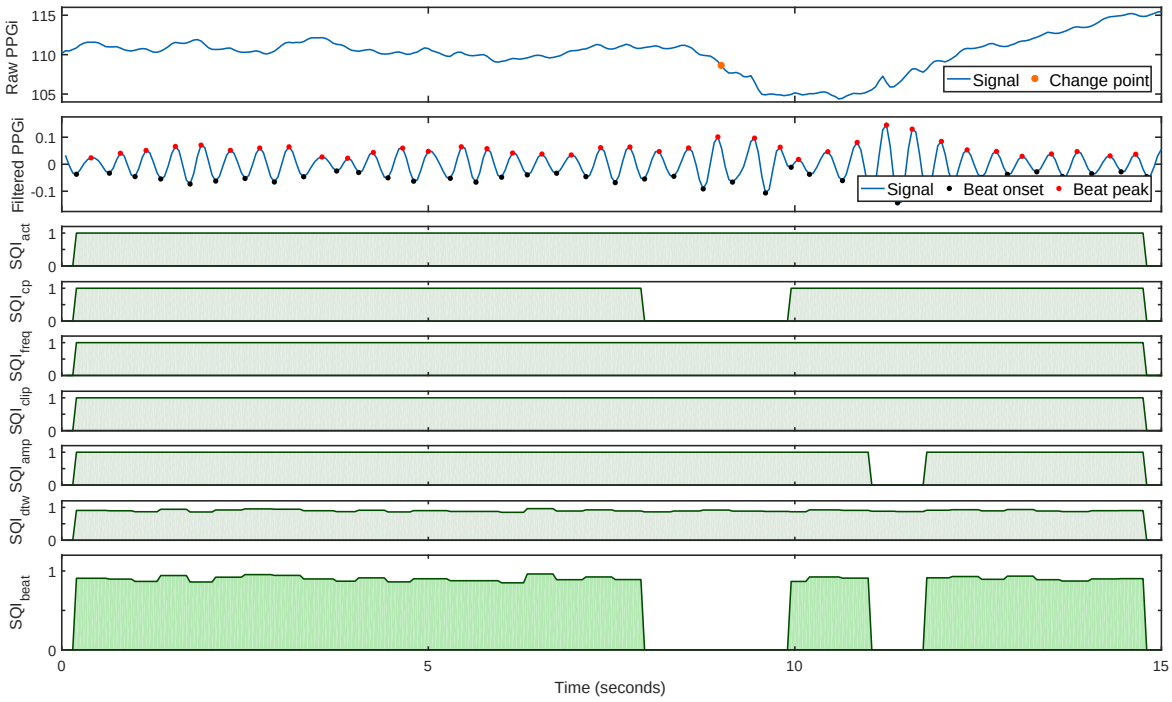
A good-quality beat will have a low value of DTW such that it has a similar shape to the average beat in the time window. On the training set, the DTW value of good-quality beats was found to be around $5 - 20$. A poor-quality beat will have a much higher value of DTW. The value of SQI_{dtw} ranges from 0 to 1 where a high value means that the beat is of good quality. Good SQI_{dtw} values usually lie between 0.80 and 0.95.

Since the calculation of SQI_{dtw} is performed using a 15-second sliding window with a step size of 5 seconds, the highest SQI_{dtw} value is taken as the quality measure for a given time when the windows overlap.

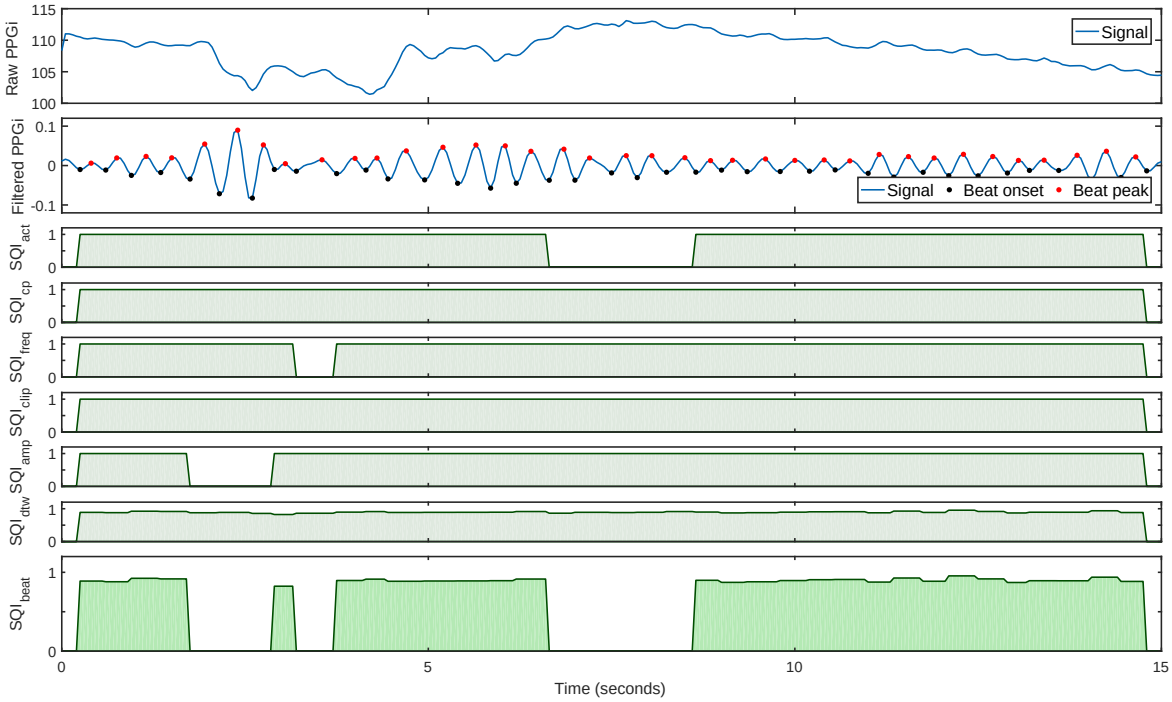
7.4.7 Beat SQI

Once all the individual SQI metrics for each beat have been calculated, the combined beat SQI (SQI_{beat}) of the kth beat is derived by simply multiplying all the SQI metrics together:

$$SQI_{\text{beat}}(k) = SQI_{\text{act}}(k) \cdot SQI_{\text{cp}}(k) \cdot SQI_{\text{freq}}(k) \cdot SQI_{\text{clip}}(k) \cdot SQI_{\text{amp}}(k) \cdot SQI_{\text{dtw}}(k). \tag{7.16}$$



(a)



(b)

Figure 7.15: Signal quality assessment for two 15-second sample windows. (a) The signal taken during a period where the lighting condition changed, causing a step change on the raw PPGi signal. SQI_{cp} is set to 0 at around second 9. (b) The signal taken from a noisy period during which the infant was awake and very active. SQI_{act} is set to 0 when the subject stretched his body (between second 7 and 9).

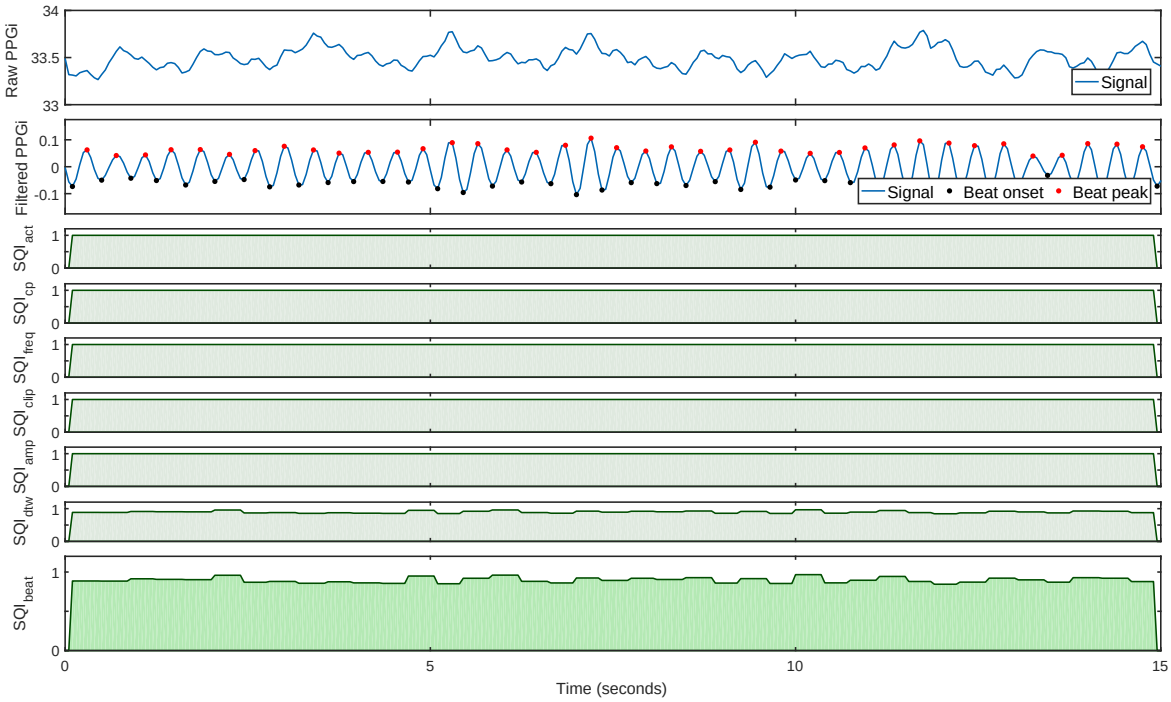


Figure 7.16: Signal quality assessment for a 15-second PPGi signal taken during a sleeping period.

SQI_{beat} is taken to be 0 if at least one of the first five SQI metrics is 0. If all of the first five SQI metrics have a value of 1, SQI_{beat} then has the value of SQI_{dtw}. Figure 7.15 shows examples of the computation of the SQIs for two periods of high activity. Figure 7.16 shows an example of the computation of the SQIs for a period during which the pre-term infant was sleeping.

7.5 Heart rate estimation methods

Once the SQI of each detected heart beat (SQI_{beat}) has been calculated, the estimation of heart rate is performed on the PPGi signal using both time and frequency-domain methods. A Kalman filter is then applied to the estimates to reduce artefacts and obtain more reliable estimates.

Heart rate is expressed as a number of beats per minute (beats/min). In this work, the heart rate is calculated using a window length of 8 seconds with a step size of 1 second, which are typical parameters in clinical practice [14, 107]. There are 160 samples or approximately 18 cardiac cycles for each time window. The SQI of the heart rate (SQI_{hr}) is defined as the mean of the SQI_{beat} values for the entire time window. Heart rate estimation is performed using one of the following algorithms: beat counting, Fast Fourier transform (FFT), dominant pole of an autoregressive model (AR dominant pole), and choosing the best model order of autoregressive models (AR best model).

7.5.1 Beat counting

In the time domain, heart rate can be estimated by counting the number of heart beats over a period of time. For this method, the time window w of 8 seconds is first expanded to include

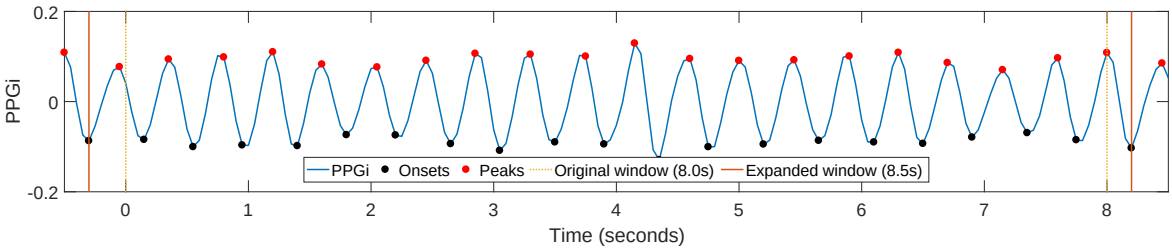


Figure 7.17: Heart rate estimation using the beat counting method. The original window w of 8 seconds is expanded to include the whole duration of the first and last beats. The expanded window w_{exp} of 8.5 seconds contains 20 beats, corresponding to a heart rate of 141.2 beats/min.

the first and last beats (see figure 7.17). The expanded window is called w_{exp}. The number of heart beats in the window w_{exp} is then counted and scaled to the unit of beats/min. Thus, the heart rate of the window w is given by

$$HR(w) = N_{\text{beats}} \cdot \frac{60}{L_{\text{exp}}} \tag{7.17}$$

where N_{beats} is the number of beats in the expanded window w_{exp} and L_{exp} is the length of the expanded window w_{exp}. Figure 7.17 shows the calculation of heart rate using the beat counting technique. This method relies solely on the accuracy of the peak and onset detector.

7.5.2 Fast Fourier Transform

In the frequency domain, heart rate can be estimated by identifying the dominant frequency with the highest power in the Fast Fourier Transform (FFT) of the signal. Heart rate estimation using the FFT can be challenging in the presence of noise or during bradycardic episodes since more than one frequency peaks may present in the time window.

The heart rate estimates obtained from the FFT method are affected by quantisation errors since the frequency resolution of the FFT depends on the sampling rate (f_s) and the number of samples used to compute the FFT (N_{FFT}). The frequency resolution is given by

$$\Delta f = \frac{f_s}{N_{\text{FFT}}}. \tag{7.18}$$

In the case of the PPGi signal with a frame rate $f_s = 20$ Hz and a window sample size $N = 160$ (8-second time window), N_{FFT} is taken to be the next power of two such that $N_{\text{FFT}} = 256$. Hence,

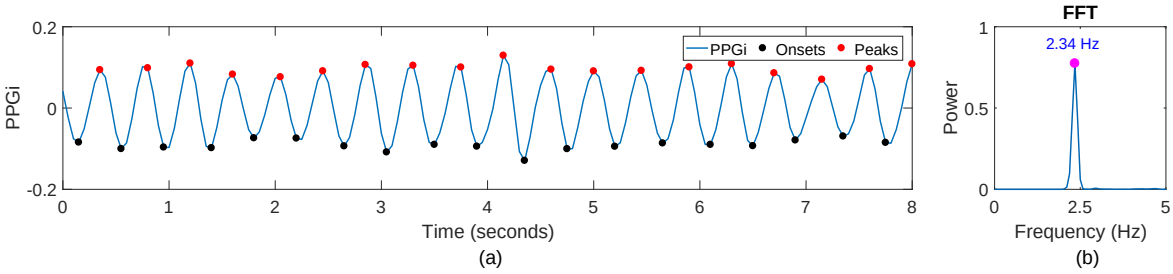


Figure 7.18: Heart rate estimation using the FFT method. (a) 8-second PPGi window. (b) The FFT of the PPGi signal shows the peak at 2.34 Hz, corresponding to a heart rate of 140.4 beats/min.

the frequency resolution is $f_s/N_{\text{FFT}} = 20/256 = 0.08$ Hz or 4.7 beats/min. Figure 7.18 shows the calculation of heart rate using the FFT method.

7.5.3 Dominant pole of an Autoregressive model

Auto-regressive (AR) modelling is a random process that linearly predicts future values based on past data. This technique is used, similarly to the FFT, to identify frequency content in a signal. Unlike the FFT technique, the AR model has no frequency resolution limitations when applied to short time series segments. The AR model is generated as a weighted sum of past values from 1 to p (the model order):

$$x_t = c + \sum_{i=1}^{p} a_i x_{t-i} + \epsilon_t \tag{7.19}$$

where c is a constant, $a_1, a_2, ..., a_p$ are the model's coefficients, and ϵ_t is Gaussian noise. The time series x_t is regressed on its previous values. The transfer function for the z-transform for the AR model is given by

$$H(Z) = \frac{1}{\sum_{k=1}^{p} a_k Z^{-k}} = \frac{Z_p}{(Z - Z_1)...(Z - Z_p)} = \frac{Z_p}{\vec{P}_1 ... \vec{P}_p} \tag{7.20}$$

where $\vec{P}_1 ... \vec{P}_p$ are vectors from any point Z to each of the poles $Z_1, ..., Z_p$ of H(Z). The AR model contains various information that are useful for time series analysis [204].

In this work, the AR model is estimated using the Burg method [205]. Heart rate is calculated by finding the dominant pole that is located inside the angle (in radians) of interest, corresponding to

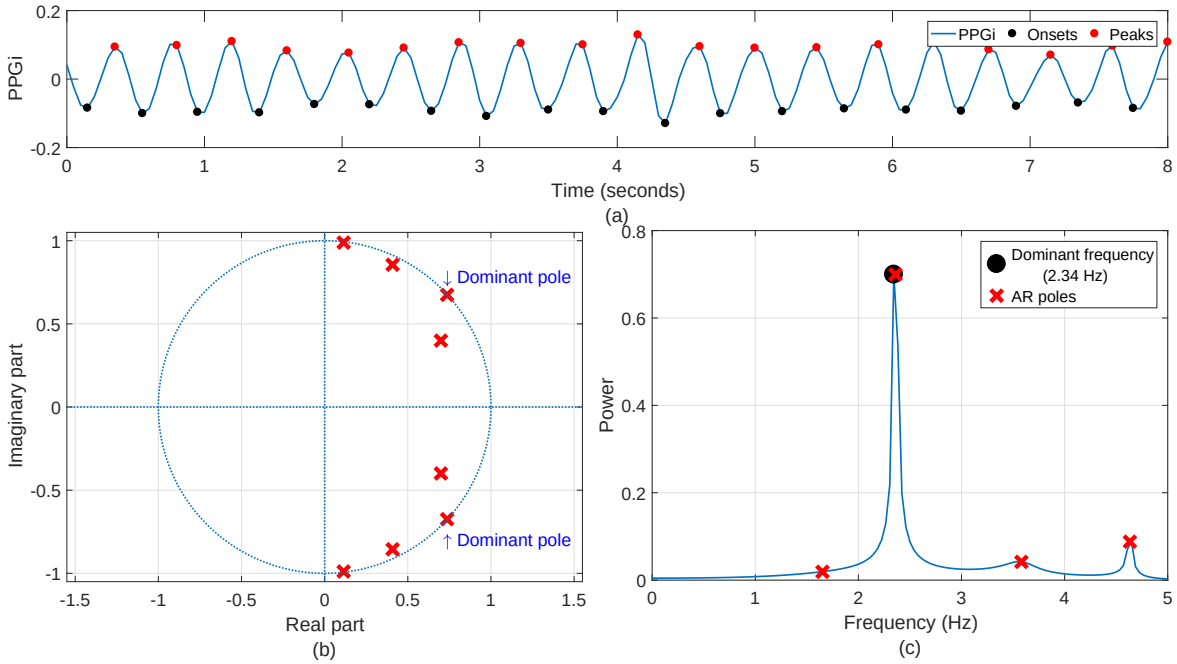


Figure 7.19: Heart rate is estimated by computing the dominant pole of an AR model. (a) 8-second PPGi window. (b) Zero-pole plot showing 8 poles inside the unit circle. The dominant pole has the highest amplitude at 0.99. (c) Frequency response of the AR model. The dominant frequency is at 2.34 Hz, corresponding to a heart rate of 140.4 beats/min.

a heart rate range between 90 and 270 beats/min, in the z-transform of the AR model (see figure 7.19). The choice of model order is a compromise between a higher model order which can provide a better approximation but can also fit the noise in the signal and a lower model order which may not be sufficient to represent the signal [206]. This work uses a fixed model order p_f of 8 which was found to achieve the lowest mean absolute error in the training set. Figure 7.19 shows the estimation of heart rate using the AR dominant pole method.

7.5.4 Autoregressive modelling with the best model

Even though autoregressive modelling is useful for describing time-varying processes, the determination of the most appropriate order for the model can be challenging. If a model order lower than the appropriate order of the process is chosen, the estimation of model parameters will not be able to describe with the process. On the other hand, if a model order higher than the appropriate order is chosen, the model will fit noise as well as signal. Although many techniques have been presented for the selection of the model order, there is no definite method that can guarantee the optimal model order [207].

With this approach, the best model is sought in a range between 6 and 12 (between $p_f - 2$ and $p_f + 4$). The size of the search space provides a good compromise between computation time and the number of model orders evaluated. The choice of the model order is made by comparing the frequency of the dominant pole and the frequency of the highest peak of the frequency response of the model.

Given an AR model with order p and noise variance σ_e^2, the frequency response of the model is given by [206]

$$S(f) = \frac{\sigma_e^2}{|\sum_{k=0}^{p} a_k e^{-i2\pi f k}|^2} \tag{7.21}$$

where a_k are the coefficients of the AR model. The best model is the model for which the difference between the frequency of the dominant pole and the highest peak of the frequency response (calculated using equation 7.21) in the frequency band between 1.5 and 4.5 Hz is less than 1 beat/min. If more than one model meets in this criterion, the model with the highest amplitude of the dominant pole was chosen. If no model meets this criterion, the heart rate for that time window is estimated by finding the highest peak of the frequency response, as described by equation 7.21.

7.5.5 Kalman filtering

Once heart rate has been computed, Kalman filtering is applied, similarly to Tarassenko *et al.* [208] and Li *et al.* [189], to adjust the heart rate estimates based on their signal quality and reduce the effect of transient changes from noise and motion artefact. The Kalman filter is an algorithm to compute an estimate of the current state of a linear dynamical system as a weighted average of the system's

previous estimate and the current measurement. The current state estimate $\hat{x}_t$ at time t is defined as

$$\hat{x}_t = F_t\hat{x}_{t-1} + B_t u_t + w_t \tag{7.22}$$

where u_t is a known control input, w_t is process noise representing uncertainty in the system, F_t is a state transition model, and B_t is a control-input model. At time t, a measurement z_t (heart rate) is made according to

$$\hat{z}_t = H_t\hat{x}_t + v_t \tag{7.23}$$

where H_t is an observation model describing the contribution of the current state to the observation and v_t is the observation noise.

The Kalman filter is a recursive estimator consisting of two stages: the prediction stage and the update stage. In the simple case of the Kalman filter for a one-dimensional signal [189], the estimate at each moment is assumed to be approximately the same as the estimate at the previous moment. The Kalman filtering algorithm is given by

Prediction stage:

$$\hat{x}_t^- = \hat{x}_{t-1} \tag{7.24}$$

$$P_t^- = P_{t-1} + Q \tag{7.25}$$

Update stage:

$$K_t = P_t^-(P_t^- + R)^{-1} \tag{7.26}$$

$$\hat{x}_t = \hat{x}_t^- + K_t(z_t - \hat{x}_t^-) \tag{7.27}$$

$$P_t = (1 - K_t)P_t^- \tag{7.28}$$

where $\hat{x}_t^-$ and $\hat{x}_k$ are the prior and posterior state estimates respectively, P_t^- and P_t are the prior and posterior error variances, K_t is the Kalman gain that minimises the posterior error variance P_t, Q is the state noise variance, and R is the measurement noise variance. Similar to Li *et al.* [189], SQI_{hr} is used to adjust the measurement noise variance R, by assigning higher weights to the measurements for clean data and lower weights for noisy data. This can reduce the influence of transient changes or artefacts. The measurement noise R is modified using the equation:

$$R = R \cdot \exp(1/SQI_{\text{hr}}^2 - 1). \tag{7.29}$$

R is initialised to 1.0 and then weighted by SQI_{hr}. Q is initialised to 1.0 which leads to the Kalman filter following the new measurements closely [54].

Figure 7.20 shows heart rate estimates from the video camera data before and after the Kalman filter is applied. The Kalman filter plays an important role in improving heart rate estimates by reducing the effect of transient changes from noise and artefacts.

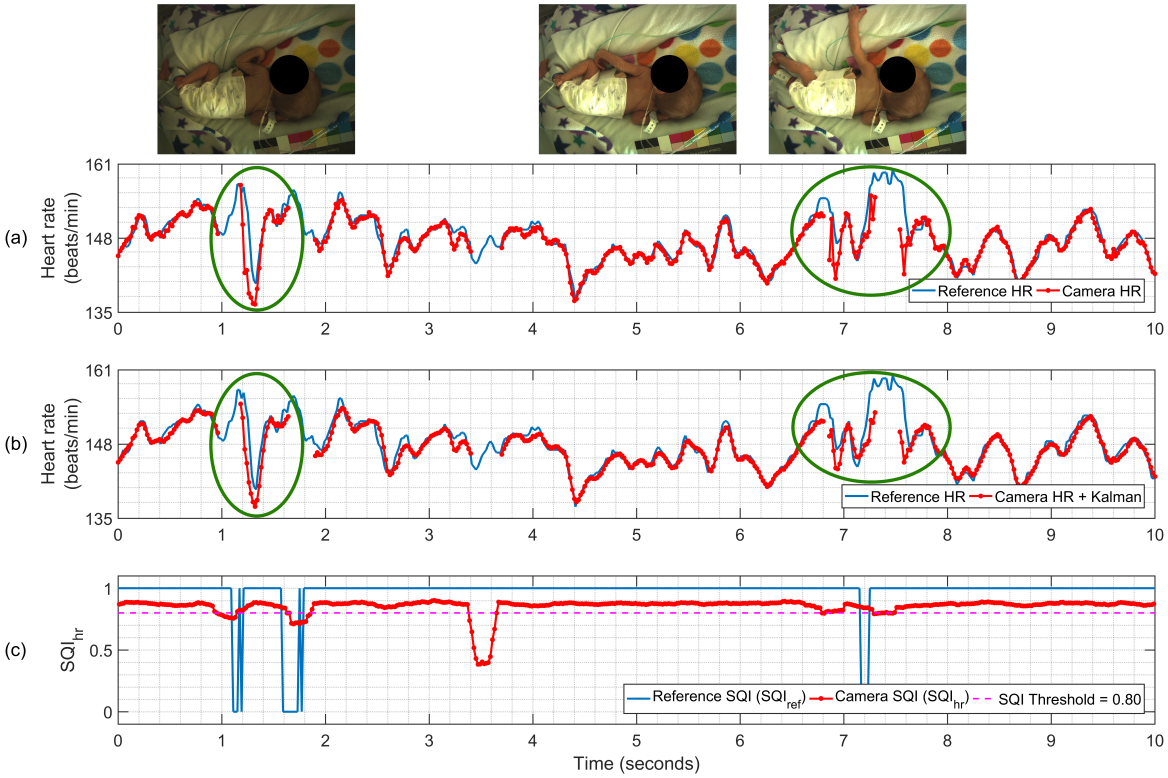


Figure 7.20: Estimation of heart rate and signal quality for a 10-minute segment (a) before applying Kalman filtering and (b) after applying Kalman filtering with (c) SQI_{hr}. The AR best model approach is used to derive heart rate estimates. The Kalman filter can reduce the effects of transient changes, noise and artefacts. The effects of the Kalman Filter can be clearly seen when there is large changes in heart rate, as highlighted with green circles.

7.6 Results

The methods presented in the previous sections are used to derive heart rate estimates for 30 pre-term infants in the Neonatal HDU study (see chapter 3). The dataset contains 426.6 hours of data recording (simultaneous recording with both reference data and camera data were available) from 90 recording sessions (one session a day). In compliance with the study protocol, the study is divided into two groups: the training set and the test set. Each group has 15 pre-term infants. The training set is used to evaluate the estimation parameters such as filter coefficients and appropriate thresholds for the assessment of signal quality, while the test set uses these parameters to assess the performance of the algorithms.

Nine recording out of ninety sessions were discarded from the heart rate estimation due to the following circumstances:

- Reference data not recorded due to equipment malfunction (1 session)
- Subject under blue-light phototherapy treatment (5 sessions)
- Video out of focus due to video camera not properly calibrated by the nurse (4 sessions)

Therefore, heart rate estimation was performed on only 80 recording sessions, corresponding to a total recoding time of 384.3 hours.

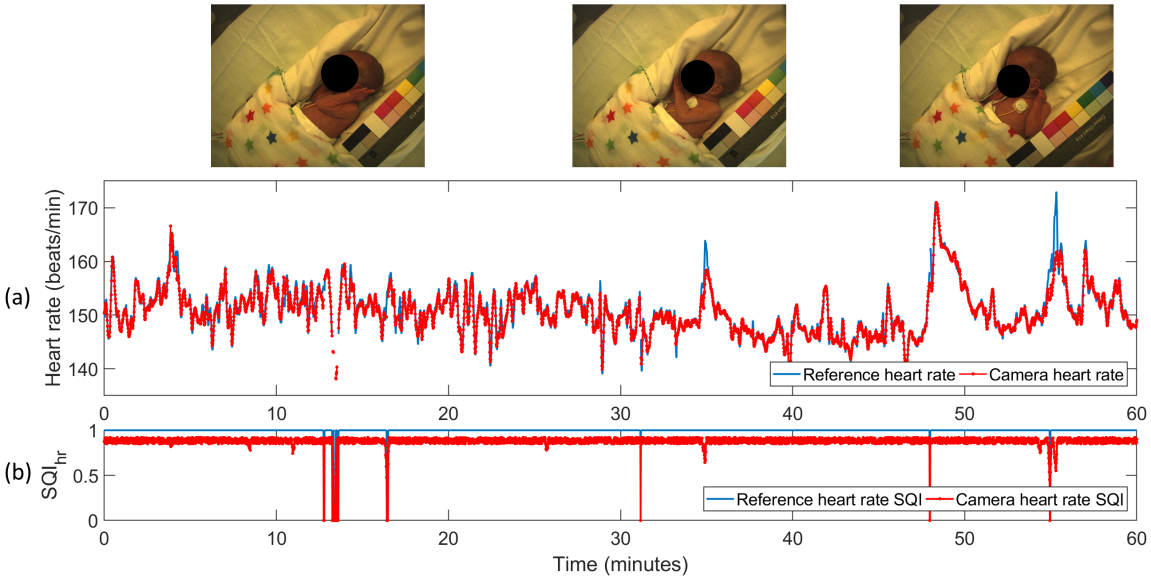


Figure 7.21: Heart rate estimation for a 60-minute video segment during which the infant was quiet. (a) Heart rate estimates from the camera data using the AR best model method with Kalman filtering applied. (b) Signal quality of the camera heart rate estimates.

The comparison of the reference heart rate and the camera-derived heart rate was performed on the valid data which was defined as the time periods during which (1) the reference heart rate was available; (2) the infant was present in front of the camera as identified by the deep learning algorithms described in chapter 4; and (3) the infant was not undergoing any clinical intervention, as identified by the intervention detection algorithm described in chapter 5. The comparison was computed from the heart rate estimates which had SQI_{hr} higher than 0.80. The threshold defines the cusp between good-quality and poor-quality heart rate estimates. It is a compromise between the accuracy of heart rate estimates and the amount of heart rate data that can be estimated.

Figure 7.21 shows the result from the automatic heart rate estimation process for a 60-minute video segment during a quiet period. Heart rate is derived using the AR best model approach with Kalman filtering. For most of the segment, a good agreement is found between the reference signal and the camera-derived heart rate. Episodes of short-term fluctuations are often associated with rapid infant movement motion artefact [42]; therefore, there is a decrease of the SQI_{hr} values during these rapid variation periods.

7.6.1 Comparison of heart rate estimation methods

Table 7.1 summarises the heart rate estimation results for the training and test sets (a total of 80 recording sessions) using the time- and frequency-domain algorithms described in section 7.5. For the training set, 72.9% (139.9 hours out of 192.0 hours) of the total recording time was considered valid. For the test set, 63.6% (122.3 hours out of 192.3 hours) of the total recording time was considered valid. The AR best model method slightly outperforms all the other methods. For the training set, heart rate could be estimated for 62.8% (87.8 hours out of 139.9 hours) of the valid data with a MAE

Table 7.1: Summary of heart rate estimation results for both training and test sets

Method	Without Kalman filtering		With Kalman filtering		Estimated time (%, hours)‡
	MAE †	MAD†	MAE†	MAD†	
Training set					
Beat counting	4.2	4.6	4.1	4.5	69.1%, 96.7 hours
FFT	3.9	4.1	3.4	3.8	69.1%, 96.7 hours
AR dominant pole	4.8	4.9	4.7	4.8	69.1%, 96.7 hours
AR best model	**3.4**	**3.4**	**2.9**	**2.9**	62.8%, 87.8 hours
Test set					
Beat counting	3.3	3.9	3.3	3.8	79.4%, 97.1 hours
FFT	2.9	3.1	2.6	2.8	79.4%, 97.1 hours
AR dominant pole	4.0	4.3	4.0	4.2	79.4%, 97.1 hours
AR best model	**2.7**	**2.7**	**2.3**	**2.3**	75.4%, 92.2 hours

† Heart rate in beats/min; ‡ Percentage of the estimated time to the valid time.

of 2.9 beats/min and a MAD of 2.9 beats/min. For the test set, heart rate could be estimated for 75.4% (92.2 hours out of 122.3 hours) of the valid data with a MAE of 2.3 beats/min and a MAD of 2.3 beats/min. Errors are reduced in all cases when Kalman filtering is applied. The estimated time of the AR best model method is slightly lower than the other methods as a result of rejecting the time periods during which an AR model order cannot be determined.

7.6.2 Overall results

Figures 7.22 and 7.23 show the comparison between the reference heart rate and camera-derived heart rate using the the AR best model method for the training set and the test set respectively. The mean difference between the two measurements is 0.2 beats/min for the training set and 0.3 beats/min for the test set. 95% of the differences between the two measurements lie within [-11.1, 10.6] beats/min for the training set and [-8.5, 9.0] beats/min for the test set. The histogram of the differences for both sets shows that the majority of errors are clustered around 0 beats/min. The mean values are distributed across the expected physiological range of the neonatal population, taken to be from 90 to 270 beats/min. The correlation plot shows that data points are located mostly around the identity line with strong correlation coefficients r of 0.86 for the training set and 0.93 for the test set. Estimation errors are mostly due to short periods of motion artefact and changes in the lighting condition during video recording. Results from the training and test sets are comparable.

7.6.3 Results by recording session

Tables 7.2–7.5 detail heart rate estimation results for each recording session in the training set (tables 7.2 and 7.3) and the test set (tables 7.4 and 7.5). Heart rate is estimated using the AR best model method with Kalman filtering. For each session, the tables also show the results of infant detection from chapter 4, intervention detection from chapter 5, and reference analysis from chapter 6. For both the training and test sets, the MAE varies from 1.3 to 7.3 beats/min and the estimated time ranges from 41.4% to 92.8%, with comparable statistics for the two datasets.

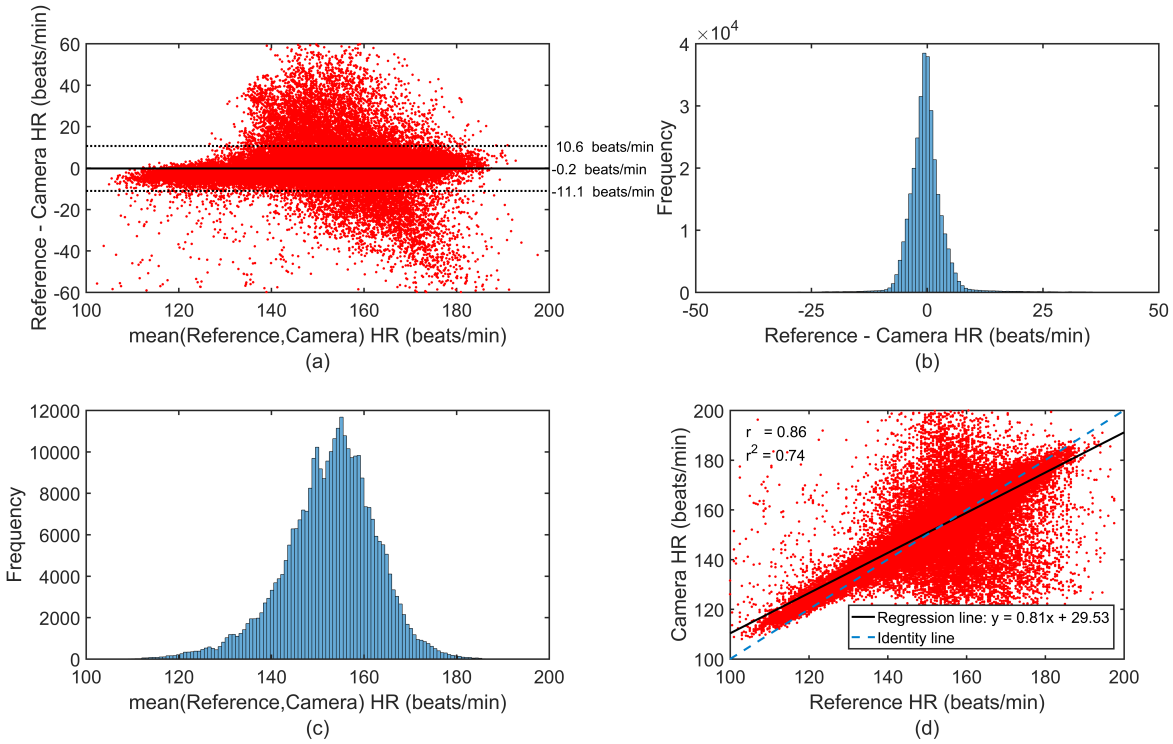


Figure 7.22: Comparison between the reference and camera-derived heart rates using the AR best model method for the training set. (a) Bland-Altman plot. (b) Histogram of the differences between the two heart rates. (c) Histogram of the mean values. (d) Correlation plot.

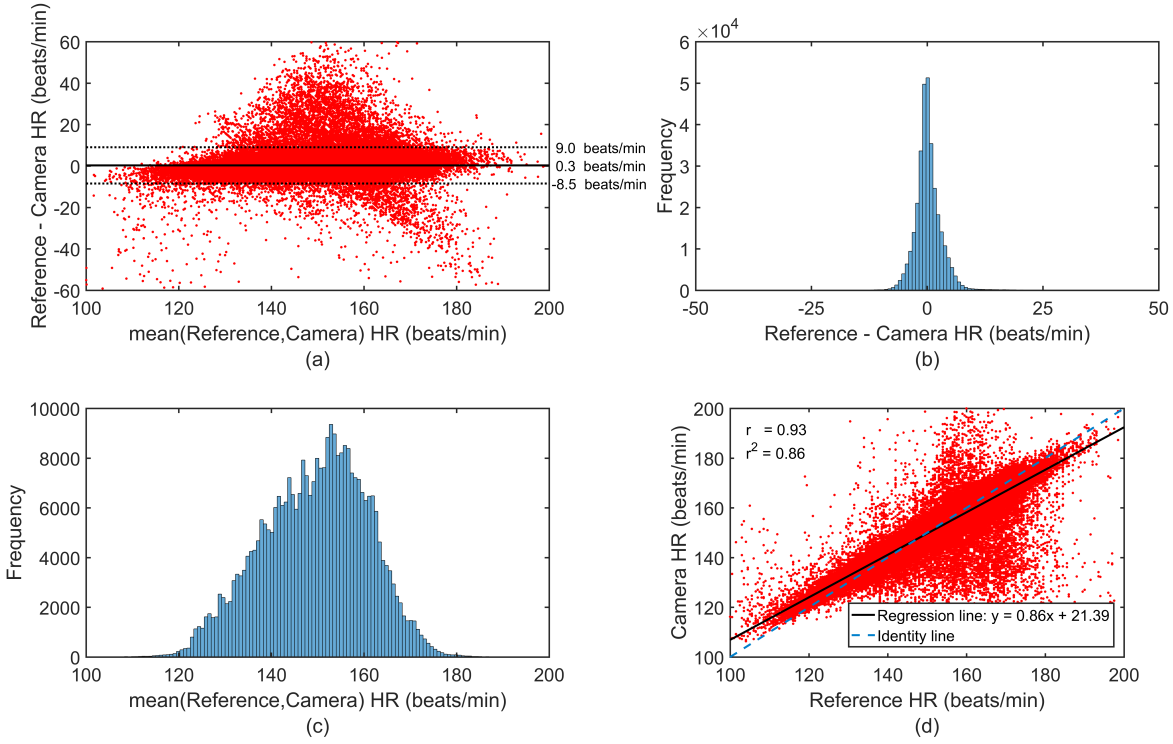


Figure 7.23: Comparison between the reference heart rate and camera-derived heart rates using the AR best model method for the test set. (a) Bland-Altman plot. (b) Histogram of the differences between the two heart rates. (c) Histogram of the mean values. (d) Correlation plot.

Table 7.2: Heart rate estimation results for each recording session in the training set — Part 1 of 2

Session	Gest. Age (weeks)	Ethnicity	Gender	Total time (hours)	Time (hours, %)					Error	
					Poor reference[1]	Subject absence[1]	Clinical Intervention[1]	Valid time[1]	Estimated time[2]	MAE (beats/min)	MAD (beats/min)
PN01 Day 1	31.3 wk	White	Male	5.4 h	0.4 h, 6.7 %	0.0 h, 0.1 %	1.0 h, 19.4 %	4.0 h, 73.9 %	2.7 h, 68.4 %	3.1	3.0
PN01 Day 2	31.4 wk	White	Male	5.3 h	0.2 h, 3.6 %	0.0 h, 0.1 %	0.4 h, 7.1 %	4.7 h, 88.8 %	3.4 h, 72.0 %	2.6	2.5
PN01 Day 3	31.6 wk	White	Male	5.0 h	0.3 h, 6.0 %	1.4 h, 27.5 %	0.8 h, 16.5 %	2.5 h, 50.0 %	2.0 h, 81.3 %	2.3	2.3
PN02 Day 1	27.6 wk	White	Female	4.0 h	0.2 h, 6.2 %	0.6 h, 14.7 %	0.3 h, 6.8 %	2.9 h, 72.2 %	2.5 h, 88.7 %	1.7	1.6
PN02 Day 2	27.7 wk	White	Female	7.1 h	0.3 h, 3.6 %	0.7 h, 10.1 %	0.8 h, 10.7 %	5.4 h, 75.6 %	4.0 h, 73.8 %	3.0	3.0
PN02 Day 3	27.9 wk	White	Female	6.8 h	1.7 h, 24.7 %	0.9 h, 13.8 %	0.2 h, 3.6 %	3.9 h, 57.9 %	3.6 h, 90.0 %	1.6	1.6
PN02 Day 4	28.0 wk	White	Female	4.7 h	0.2 h, 3.8 %	0.7 h, 14.8 %	0.5 h, 9.5 %	3.4 h, 72.0 %	2.3 h, 68.1 %	1.8	1.7
PN03 Day 1	28.7 wk	White	Female	6.7 h	0.2 h, 3.1 %	0.8 h, 11.3 %	1.2 h, 17.9 %	4.5 h, 67.7 %	2.9 h, 64.7 %	1.6	1.5
PN03 Day 2	28.9 wk	White	Female	6.1 h	0.3 h, 5.3 %	0.1 h, 1.5 %	1.4 h, 22.2 %	4.3 h, 71.1 %	1.8 h, 41.8 %	1.9	1.9
PN03 Day 3	29.0 wk	White	Female	4.5 h	0.2 h, 3.7 %	0.1 h, 1.7 %	0.4 h, 8.0 %	3.9 h, 86.5 %	0.9 h, 23.2 %	1.5	1.6
PN05 Day 1	36.4 wk	Asian	Male	6.5 h	0.6 h, 8.7 %	0.0 h, 0.1 %	0.4 h, 5.7 %	5.5 h, 85.5 %	4.1 h, 73.4 %	3.3	2.3
PN06 Day 1	33.4 wk	White	Male	*Subject under phototherapy treatment*							
PN06 Day 2	33.6 wk	White	Male	*Subject under phototherapy treatment*							
PN06 Day 3	33.7 wk	White	Male	3.0 h	0.3 h, 9.5 %	0.0 h, 0.6 %	0.7 h, 24.4 %	1.9 h, 65.5 %	1.3 h, 64.5 %	3.1	2.5
PN07 Day 1	32.0 wk	White	Female	*Subject under phototherapy treatment*							
PN07 Day 2	32.1 wk	White	Female	6.1 h	0.2 h, 3.2 %	0.0 h, 0.3 %	0.6 h, 9.0 %	5.3 h, 87.5 %	3.6 h, 67.1 %	2.7	2.5
PN07 Day 3	32.3 wk	White	Female	6.1 h	1.1 h, 17.9 %	0.0 h, 0.3 %	0.3 h, 5.5 %	4.7 h, 76.2 %	3.5 h, 75.5 %	2.7	2.2
PN09 Day 1	30.9 wk	Black	Male	6.8 h	0.7 h, 11.0 %	0.0 h, 0.1 %	0.5 h, 7.5 %	5.5 h, 81.4 %	3.0 h, 54.1 %	5.0	4.6
PN09 Day 2	31.0 wk	Black	Male	5.4 h	0.7 h, 12.5 %	0.0 h, 0.0 %	0.2 h, 3.9 %	4.5 h, 83.6 %	3.1 h, 69.5 %	2.8	2.7
PN09 Day 3	31.1 wk	Black	Male	6.0 h	0.5 h, 8.3 %	0.0 h, 0.2 %	0.3 h, 4.5 %	5.2 h, 87.0 %	2.5 h, 47.3 %	3.6	3.0
PN11 Day 1	28.7 wk	White	Female	3.9 h	0.4 h, 9.0 %	0.0 h, 0.4 %	0.4 h, 10.4 %	3.1 h, 80.2 %	1.4 h, 44.3 %	1.5	1.5
PN11 Day 2	28.9 wk	White	Female	4.6 h	0.1 h, 2.8 %	0.0 h, 0.3 %	0.5 h, 11.6 %	4.0 h, 85.3 %	1.4 h, 35.0 %	2.5	2.8
PN11 Day 3	29.0 wk	White	Female	2.9 h	0.1 h, 2.5 %	0.0 h, 0.4 %	0.3 h, 11.0 %	2.5 h, 86.1 %	1.5 h, 60.0 %	1.8	2.0

Continued on the next page (table 7.3).

[1] Percentage of the time with respect to the total duration; [2] Percentage of the estimated time with respect to the valid time; [3] Mixed White & Asian; [4] Mixed White & Black.

Table 7.3: Heart rate estimation results for each recording session in the training set – Part 2 of 2

Session	Gest. Age (weeks)	Ethnicity	Gender	Total time (hours)	Time (hours, %)					Error	
					Poor reference[1]	Subject absence[1]	Clinical Intervention[1]	Valid time[1]	Estimated time[2]	MAE (beats/min)	MAD
Continued from the previous page (table 7.2).											
PN15 Day 1	31.7 wk	Mixed W&A[3]	Male	3.0 h	0.1 h, 3.1 %	0.0 h, 0.2 %	0.3 h, 10.8 %	2.6 h, 85.9 %	1.4 h, 53.5 %	5.1	5.2
PN15 Day 2	31.9 wk	Mixed W&A[3]	Male	3.9 h	0.1 h, 2.5 %	0.0 h, 0.6 %	0.4 h, 11.6 %	3.3 h, 85.3 %	2.5 h, 75.7 %	3.0	3.0
PN15 Day 3	32.0 wk	Mixed W&A[3]	Male	3.2 h	0.0 h, 1.2 %	0.0 h, 0.4 %	0.7 h, 21.0 %	2.5 h, 77.4 %	1.3 h, 51.3 %	4.3	4.6
PN15 Day 4	32.1 wk	Mixed W&A[3]	Male	3.0 h	0.2 h, 6.2 %	0.0 h, 0.1 %	0.4 h, 11.6 %	2.5 h, 82.0 %	1.8 h, 71.3 %	3.4	3.3
PN16 Day 1	32.7 wk	White	Female		*Subject under phototherapy treatment*						
PN16 Day 2	32.9 wk	White	Female		*Video out of focus*						
PN17 Day 1	29.6 wk	White	Male	4.8 h	1.3 h, 27.6 %	0.0 h, 0.1 %	0.8 h, 15.8 %	2.7 h, 56.4 %	1.8 h, 67.5 %	2.6	2.3
PN17 Day 2	29.7 wk	White	Male	6.6 h	0.9 h, 13.2 %	1.1 h, 17.1 %	0.5 h, 7.1 %	4.1 h, 62.6 %	3.3 h, 79.1 %	1.9	1.8
PN17 Day 3	29.9 wk	White	Male	5.9 h	0.4 h, 6.1 %	2.4 h, 41.4 %	0.8 h, 12.9 %	2.3 h, 39.6 %	1.3 h, 54.3 %	2.1	2.1
PN19 Day 1	29.6 wk	Mixed Other	Female	7.0 h	0.2 h, 2.8 %	0.0 h, 0.1 %	0.8 h, 11.2 %	6.0 h, 85.9 %	3.4 h, 55.9 %	2.8	2.6
PN19 Day 2	29.7 wk	Mixed Other	Female	5.8 h	0.5 h, 7.8 %	0.7 h, 12.6 %	1.7 h, 29.3 %	2.9 h, 50.3 %	1.5 h, 52.0 %	2.5	2.5
PN19 Day 3	29.9 wk	Mixed Other	Female	5.3 h	0.2 h, 2.9 %	0.0 h, 0.4 %	0.5 h, 8.6 %	4.7 h, 88.1 %	2.6 h, 56.7 %	4.2	4.2
PN21 Day 1	33.3 wk	White	Male	2.0 h	0.1 h, 4.1 %	0.0 h, 0.8 %	0.2 h, 11.0 %	1.7 h, 84.1 %	1.4 h, 83.8 %	3.0	2.1
PN21 Day 2	33.4 wk	White	Male	6.4 h	1.0 h, 15.2 %	1.1 h, 16.5 %	0.5 h, 8.2 %	3.9 h, 60.1 %	2.3 h, 58.7 %	4.2	3.7
PN21 Day 3	33.6 wk	White	Male	3.2 h	0.2 h, 4.8 %	0.0 h, 0.3 %	0.5 h, 15.5 %	2.5 h, 79.3 %	1.7 h, 69.2 %	3.0	2.9
PN23 Day 1	30.0 wk	White	Female	5.4 h	0.2 h, 3.9 %	1.3 h, 23.1 %	0.6 h, 11.0 %	3.4 h, 61.9 %	2.6 h, 78.3 %	2.2	2.0
PN23 Day 2	30.1 wk	White	Female	4.0 h	0.1 h, 2.5 %	0.0 h, 0.2 %	0.1 h, 2.0 %	3.8 h, 95.4 %	2.6 h, 69.1 %	3.0	2.5
PN23 Day 3	30.3 wk	White	Female	6.2 h	0.2 h, 3.1 %	2.3 h, 37.2 %	0.9 h, 14.4 %	2.8 h, 45.3 %	1.8 h, 64.0 %	3.8	2.7
PN25 Day 1	32.1 wk	Mixed W&B[4]	Male	6.4 h	0.2 h, 2.6 %	1.9 h, 29.9 %	0.6 h, 10.1 %	3.7 h, 57.3 %	2.0 h, 54.1 %	4.5	4.7
PN25 Day 2	32.3 wk	Mixed W&B[4]	Male	3.0 h	0.1 h, 2.3 %	0.0 h, 0.0 %	0.2 h, 5.1 %	2.8 h, 92.6 %	1.2 h, 41.4 %	7.3	8.0
Total (38 recording sessions)				192.0 h	14.3 h, 7.5 %	16.3 h, 8.5 %	21.5 h, 11.2 %	139.9 h, 72.9 %	87.8 h, 62.8 %	2.9	2.9

[1] Percentage of the time with respect to the total duration; [2] Percentage of the estimated time with respect to the valid time; [3] Mixed White & Asian; [4] Mixed White & Black.

Table 7.4: Heart rate estimation results for each recording session in the test set — Part 1 of 2

Session	Gest. Age (weeks)	Ethnicity	Gender	Total time (hours)	Time (hours, %)					Error	
					Poor reference[1]	Subject absence[1]	Clinical Intervention[1]	Valid time[1]	Estimated time[2]	MAE (beats/min)	MAD
PN04 Day 1	33.4 wk	White	Male	2.0 h	0.0 h, 2.5 %	0.0 h, 0.6 %	0.8 h, 39.4 %	1.1 h, 57.5 %	0.6 h, 51.0 %	3.3	3.4
PN04 Day 2	33.6 wk	White	Male	4.7 h	0.1 h, 1.5 %	0.0 h, 0.4 %	1.1 h, 23.4 %	3.5 h, 74.7 %	2.3 h, 64.8 %	2.2	1.8
PN04 Day 3	33.7 wk	White	Male	5.4 h	0.3 h, 5.5 %	2.4 h, 43.9 %	0.5 h, 9.2 %	2.2 h, 41.5 %	1.8 h, 78.6 %	1.4	1.3
PN04 Day 4	33.9 wk	White	Male	8.1 h	0.4 h, 4.6 %	3.8 h, 46.9 %	0.8 h, 10.1 %	3.1 h, 38.4 %	2.3 h, 73.5 %	1.8	1.6
PN08 Day 1	30.1 wk	White	Female	5.8 h	0.4 h, 6.0 %	0.0 h, 0.3 %	0.8 h, 12.9 %	4.7 h, 80.8 %	4.0 h, 85.2 %	1.6	1.7
PN08 Day 2	30.3 wk	White	Female	5.9 h	0.2 h, 4.1 %	0.0 h, 0.2 %	0.3 h, 5.1 %	5.3 h, 90.5 %	4.7 h, 88.0 %	1.5	1.5
PN08 Day 3	30.4 wk	White	Female	3.7 h	0.2 h, 5.2 %	0.0 h, 0.3 %	1.0 h, 26.0 %	2.6 h, 68.5 %	1.7 h, 67.9 %	1.7	1.7
PN10 Day 1	30.1 wk	White	Male	*Video out of focus*							
PN10 Day 2	30.3 wk	White	Male	8.3 h	0.3 h, 3.1 %	1.6 h, 19.7 %	0.4 h, 5.0 %	6.0 h, 72.2 %	4.5 h, 75.3 %	2.2	2.1
PN10 Day 3	30.4 wk	White	Male	4.6 h	0.1 h, 2.8 %	0.0 h, 0.2 %	0.2 h, 3.9 %	4.3 h, 93.1 %	2.9 h, 66.8 %	2.7	2.7
PN12 Day 1	31.0 wk	Asian	Female	*Subject under phototherapy treatment*							
PN13 Day 1	28.0 wk	Mixed W&A[3]	Female	3.0 h	0.1 h, 4.4 %	0.0 h, 0.2 %	0.4 h, 12.7 %	2.4 h, 82.6 %	2.1 h, 86.7 %	2.0	2.1
PN13 Day 2	28.1 wk	Mixed W&A[3]	Female	9.3 h	1.1 h, 11.4 %	3.8 h, 40.6 %	1.0 h, 11.0 %	3.4 h, 36.9 %	2.8 h, 81.9 %	1.9	2.0
PN13 Day 3	28.3 wk	Mixed W&A[3]	Female	2.4 h	0.1 h, 2.6 %	0.0 h, 0.3 %	0.5 h, 20.7 %	1.8 h, 76.4 %	1.7 h, 90.4 %	1.9	1.8
PN14 Day 1	29.4 wk	White	Female	*Reference data not recorded*							
PN14 Day 2	29.6 wk	White	Female	6.3 h	0.2 h, 3.7 %	0.5 h, 7.4 %	0.8 h, 13.0 %	4.8 h, 76.0 %	4.1 h, 84.2 %	2.0	2.0
PN14 Day 3	29.7 wk	White	Female	6.0 h	0.4 h, 6.1 %	1.0 h, 17.2 %	1.7 h, 28.1 %	2.9 h, 48.6 %	2.4 h, 83.2 %	1.9	1.9
PN14 Day 4	29.9 wk	White	Female	4.7 h	0.4 h, 8.1 %	0.0 h, 0.2 %	0.3 h, 5.5 %	4.0 h, 86.2 %	3.2 h, 80.4 %	2.2	2.2
PN18 Day 1	30.4 wk	White	Male	3.1 h	0.3 h, 8.8 %	0.7 h, 21.5 %	0.3 h, 10.6 %	1.8 h, 59.1 %	1.2 h, 66.7 %	2.9	2.7
PN18 Day 2	30.6 wk	White	Male	2.2 h	0.2 h, 11.2 %	0.0 h, 0.7 %	0.3 h, 12.0 %	1.7 h, 76.1 %	1.1 h, 65.0 %	3.7	3.8
PN18 Day 3	30.7 wk	White	Male	4.2 h	1.8 h, 43.4 %	0.0 h, 0.4 %	0.4 h, 8.9 %	2.0 h, 47.3 %	1.3 h, 65.3 %	3.0	2.8
PN20 Day 1	30.4 wk	Mixed Other	Male	4.3 h	0.5 h, 11.9 %	0.4 h, 10.3 %	0.3 h, 7.9 %	3.0 h, 69.9 %	2.7 h, 90.1 %	1.4	1.4
PN20 Day 2	30.6 wk	Mixed Other	Male	4.5 h	0.2 h, 3.4 %	0.0 h, 0.2 %	0.6 h, 13.2 %	3.7 h, 83.2 %	3.0 h, 81.8 %	2.1	2.0
PN20 Day 3	30.7 wk	Mixed Other	Male	4.3 h	0.3 h, 6.5 %	0.0 h, 0.2 %	0.2 h, 4.0 %	3.8 h, 89.4 %	3.4 h, 88.8 %	1.6	1.6
PN20 Day 4	30.9 wk	Mixed Other	Male	2.5 h	0.2 h, 7.8 %	0.0 h, 0.4 %	0.1 h, 4.0 %	2.2 h, 87.8 %	2.0 h, 92.8 %	1.3	1.2

Continued on the next page (table 7.5).

[1] Percentage of the time with respect to the total duration; [2] Percentage of the estimated time with respect to the valid time; [3] Mixed White & Asian; [4] Mixed White & Black.

Table 7.5: Heart rate estimation results for each recording session in the test set — Part 2 of 2

Session	Gest. Age (weeks)	Ethnicity	Gender	Total time (hours)	Time (hours, %)					Error	
					Poor reference[1]	Subject absence[1]	Clinical Intervention[1]	Valid time[1]	Estimated time[2]	MAE (beats/min)	MAD (beats/min)
Continued from the previous page (table 7.4).											
PN22 Day 1	29.6 wk	White	Male		*Video out of focus*						
PN22 Day 2	29.7 wk	White	Male		*Video out of focus*						
PN22 Day 3	29.9 wk	White	Male	2.5 h	0.3 h, 11.5 %	0.0 h, 0.3 %	0.3 h, 10.1 %	2.0 h, 78.1 %	1.4 h, 72.4 %	2.1	2.2
PN24 Day 1	31.7 wk	Mixed W&B[4]	Male	5.0 h	0.1 h, 2.1 %	0.9 h, 18.1 %	0.2 h, 3.3 %	3.8 h, 76.6 %	2.3 h, 59.4 %	3.2	2.8
PN24 Day 2	31.9 wk	Mixed W&B[4]	Male	5.7 h	1.0 h, 17.3 %	1.1 h, 19.8 %	0.4 h, 6.5 %	3.2 h, 56.4 %	1.8 h, 57.3 %	3.8	3.7
PN24 Day 3	32.0 wk	Mixed W&B[4]	Male	3.1 h	0.4 h, 14.6 %	0.0 h, 0.1 %	0.2 h, 5.2 %	2.4 h, 80.1 %	1.6 h, 67.5 %	3.0	2.4
PN26 Day 1	34.3 wk	White	Male	1.8 h	0.0 h, 2.6 %	0.0 h, 0.2 %	0.1 h, 3.0 %	1.7 h, 94.2 %	1.2 h, 71.3 %	3.1	3.1
PN26 Day 2	34.4 wk	White	Male	3.1 h	2.2 h, 68.9 %	0.0 h, 0.5 %	0.2 h, 5.3 %	0.8 h, 25.3 %	0.5 h, 67.6 %	3.0	3.1
PN26 Day 3	34.6 wk	White	Male	3.4 h	0.4 h, 12.8 %	0.0 h, 0.2 %	0.5 h, 15.5 %	2.5 h, 71.5 %	1.7 h, 67.3 %	3.4	3.5
PN26 Day 4	34.7 wk	White	Male	3.4 h	1.1 h, 31.3 %	0.0 h, 1.3 %	1.0 h, 28.0 %	1.4 h, 39.3 %	0.8 h, 59.7 %	3.6	3.3
PN27 Day 1	31.3 wk	Mixed W&A[3]	Male	4.6 h	0.5 h, 11.2 %	1.0 h, 21.1 %	0.3 h, 7.0 %	2.8 h, 60.8 %	2.0 h, 70.4 %	2.7	2.7
PN27 Day 2	31.4 wk	Mixed W&A[3]	Male	3.9 h	0.9 h, 23.7 %	0.0 h, 0.2 %	0.4 h, 11.0 %	2.5 h, 65.1 %	1.8 h, 69.5 %	2.9	2.9
PN28 Day 1	31.0 wk	White	Female	6.4 h	0.2 h, 2.8 %	3.7 h, 58.1 %	0.7 h, 11.5 %	1.8 h, 27.6 %	1.3 h, 70.8 %	2.7	2.8
PN28 Day 2	31.1 wk	White	Female	5.5 h	0.2 h, 3.3 %	0.9 h, 16.4 %	0.9 h, 16.6 %	3.5 h, 63.7 %	2.5 h, 71.8 %	2.8	2.6
PN28 Day 3	31.3 wk	White	Female	5.8 h	0.7 h, 13.0 %	1.5 h, 26.6 %	0.4 h, 6.5 %	3.1 h, 54.0 %	2.8 h, 88.9 %	1.7	1.4
PN29 Day 1	30.6 wk	White	Male	6.1 h	0.6 h, 10.1 %	1.4 h, 22.1 %	0.5 h, 8.2 %	3.7 h, 59.7 %	2.8 h, 76.5 %	2.7	2.4
PN29 Day 2	30.7 wk	White	Male	6.2 h	1.4 h, 22.3 %	1.5 h, 24.0 %	0.9 h, 13.8 %	2.5 h, 40.0 %	1.2 h, 47.5 %	4.4	4.2
PN29 Day 3	30.9 wk	White	Male	5.8 h	0.7 h, 12.1 %	0.5 h, 8.2 %	0.6 h, 9.8 %	4.1 h, 70.0 %	2.4 h, 59.0 %	3.9	3.3
PN29 Day 4	31.0 wk	White	Male	3.1 h	0.4 h, 13.4 %	0.8 h, 27.2 %	0.4 h, 14.3 %	1.4 h, 45.1 %	0.7 h, 52.7 %	3.5	3.2
PN30 Day 1	32.1 wk	White	Male	4.0 h	0.5 h, 12.5 %	0.0 h, 0.6 %	0.2 h, 5.5 %	3.3 h, 81.4 %	3.0 h, 91.3 %	1.6	1.5
PN30 Day 2	32.3 wk	White	Male	3.6 h	0.6 h, 15.8 %	0.0 h, 0.4 %	0.4 h, 10.8 %	2.6 h, 73.0 %	2.4 h, 91.3 %	2.0	2.0
PN30 Day 3	32.4 wk	White	Male	3.9 h	0.1 h, 2.6 %	0.1 h, 1.6 %	1.0 h, 25.3 %	2.7 h, 70.6 %	2.2 h, 81.0 %	2.4	2.3
Total (42 recording sessions)				192.3 h	20.1 h, 10.2 %	27.9 h, 14.5 %	22.0 h, 11.5 %	122.3 h, 63.6 %	92.2 h, 75.4 %	2.3	2.3

[1] Percentage of the time with respect to the total duration; [2] Percentage of the estimated time with respect to the valid time; [3] Mixed White & Asian; [4] Mixed White & Black.

7.6.4 Results by ethnicity

Table 7.6 shows heart rate estimation results for each ethnic group, as recorded in the clinical data management system. Figure 7.24 shows images of each non-White subject corresponding to the data in table 7.6. One White subject in the training set (PN16) and one Asian subject in the test set (PN12) were discarded since the infant was under phototherapy treatment since they fell into the exclusion criteria (they were under phototherapy treatment). Compared to the test set, the training set had a slightly wider range of ethnic groups. Subjects in the non-White groups in the test set had lighter skin tone than those in the training set. Errors in heart rate estimation in White subjects were lower than the other ethnic groups.

Table 7.6: Heart rate estimation results according to ethnicity

Ethnicity	Number of subjects	Number of sessions	Error (beats/min)		Estimated time (%)
			MAE	MAD	
Training set[1]					
White	9	25	2.5	2.5	65.3%
Asian	1	1	3.3	2.3	73.4%
Black	1	3	3.8	3.5	56.4%
Mixed White & Asian	1	4	3.7	3.9	63.9%
Mixed White & Black	1	2	5.6	6.0	48.6%
Mixed Others	1	3	3.2	3.1	55.3%
Test set[2]					
White	10	30	2.3	2.3	74.6%
Mixed White & Asian	2	3	3.3	3.0	60.8%
Mixed White & Black	1	5	2.2	2.3	79.1%
Mixed Others	1	4	1.7	1.6	87.7%

[1] One subject (White) in the training set was discarded.
[2] One subject (Asian) in the test set was discarded.

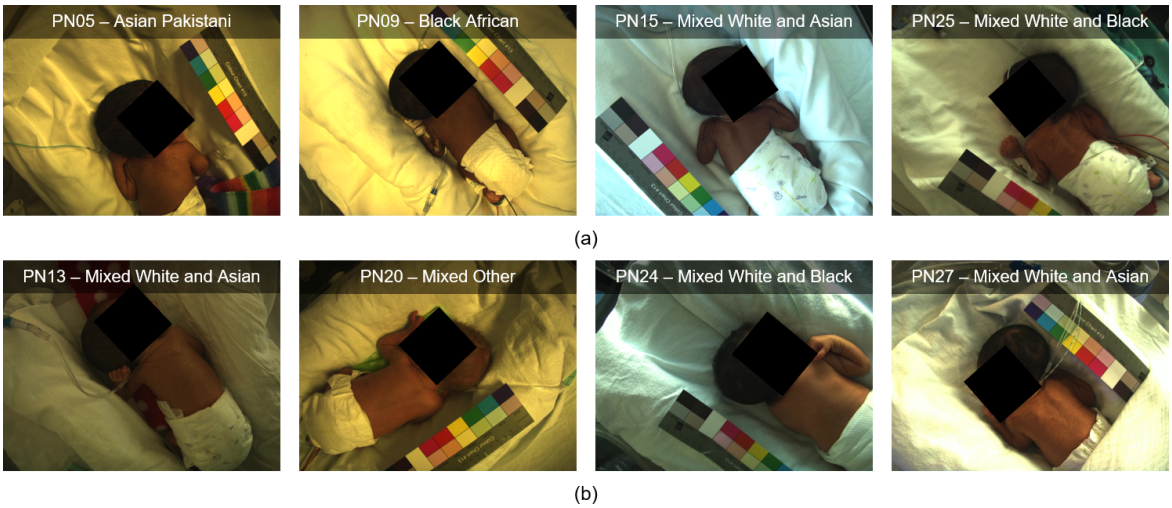


Figure 7.24: Skin colour of non-White subjects in (a) the training set and (b) the test set.

7.7 Discussion

7.7.1 Automated heart rate estimation process

To date, there is no standard procedure for selecting a measurement site when using video-based vital-sign estimation methods. Most research relies on the manual selection of an ROI containing the human skin. Several studies employ face detection algorithms to automatically select the ROI from the facial area. In this study, the whole skin area, as determined by the deep learning algorithms described in the previous chapters, was used as the ROI for heart rate estimation.

The use of the whole skin area allows the heart rate monitoring to be performed automatically in challenging clinical settings, e.g. different skin tone and posture changes. However, the estimates can be noisy or be interrupted during periods of large and rapid movement of the subject. As expected, the PPGi signal from the whole skin area is subject to interference due to artefacts caused by patient movement. As a result, we used a band-pass filter with higher cut-off frequencies of 1.5 and 4.5 Hz (90 – 270 beats/min) compared to previous published research [107]. Hence, the algorithm cannot estimate accurately heart rates below 90 beats/min, e.g. during bradycardic episodes.

7.7.2 Differences between the training set and the test set

Table 7.1 shows the comparison of heart rate estimation results for both training and test sets. The test set has better results than the training set in all cases. Using the AR best model with Kalman filtering, the MAE and MAD of the test set were both 0.6 beats/min lower than those of the training set, and the estimated time of the test set was 12.6% higher than that of the training set.

This is as a result of dividing the Neonatal HDU dataset into training and test sets by considering only patient demographics (ethnicity, gestational age and weight) and recording information (date, time and duration). It did not consider the location of an incubator, the lighting conditions of the room and the expertise of research staff which improved over time. As a result, the training set contains more patients in the first half of the dataset than the second half of the dataset, and the test set contains more patients in the second half of the dataset than the first half of the dataset. The lighting conditions of video recordings as well as the configurations of the camera (focus and aperture settings) at the beginning of the study were not controlled properly. They were improved in later recording sessions.

7.7.3 Comparison with conventional monitoring

The ANSI/AAMI EC13:2002 "Cardiac monitors, heart rate meters, and alarms" standard of the Association of the Advancement of Medical Instrumentation (AAMI) specifies the maximum heart rate measurement error for medical equipment to be 5 beats/min [176]. In chapter 6, the comparison between two heart rate measurements provided by the patient monitor was performed. The MAE

between the two measurements was 1.7 beats/min. The heart rate estimates derived from the camera have an MAE of 2.9 beats/min for the training set and an MAE of 2.3 beats/min. The camera-derived heart rate can therefore be considered accurate under the EC13:2002 standard.

From the data in table 7.1, heart rate estimates could be obtained for approximately 63% and 76% of the valid time for the training set and the test set respectively. Without the Kalman filter applied, the AR best model approach produces the lowest MAE of 2.9 beats/min for the training set and 2.3 beats/min for the test set. Nevertheless, the results from the FFT method are close to those from the AR best model approach. When the Kalman filter is applied, the FFT produces the lowest MAE of 3.4 beats/min for the training set and 2.6 beats/min for the test set. However, its frequency resolution for an 8-second estimation window and a camera sampling rate of 20 Hz is only approximately 5 beats/min, which is unsuitable for the estimation of heart rate in this clinical study. The AR best model with Kalman filtering is the best performing method with the lowest MAE and MAD.

7.7.4 Comparison to other clinical studies

The errors obtained from the estimation of the Neonatal HDU study are similar to those presented in the initial analysis of the first two pre-term infants in 2014 [107]. For an adult population, Villarroel *et al.* [123] 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. As reported in this chapter, these errors are also similar to those obtained in this study of pre-term infants. Compared to the adult population, pre-term infants generally have a higher range of heart rate, depending on age and clinical condition. Episodes of short-term fluctuations in heart rate (increasing or decreasing by 10 beats/min in a few seconds) are found even when the infant is asleep (see figure 7.21).

7.7.5 Continuous heart rate estimation

For hospitalised pre-term infants, it is important for the estimates of vital signs to be continuous. Even though the estimation of heart rate can be challenging in the clinic due to the active nature of the infants, heart rate could be estimated with good accuracy during quiet and stable periods (see figure 7.21). Figure 7.25 shows the histogram of the periods for which heart rate could not be estimated. The gap durations are mostly less than 30 seconds. If gaps of up to 30 seconds are allowed, the percentage of the estimated time with respect to the valid time increases from 62.8% to 90.0% for the training set and from 75.4% to 96.9% for the test set. The frequency of obtaining vital sign measurements in the hospital can vary between continuous second-by-second estimates to values measured by clinical staff every 15 minutes, every hour or every several hours, as shown in table 5.1. Clinical staff are interested in vital signs and patterns rather than second-by-second estimates. Hence, the interruption of heart rate estimates for a few seconds is not considered to be

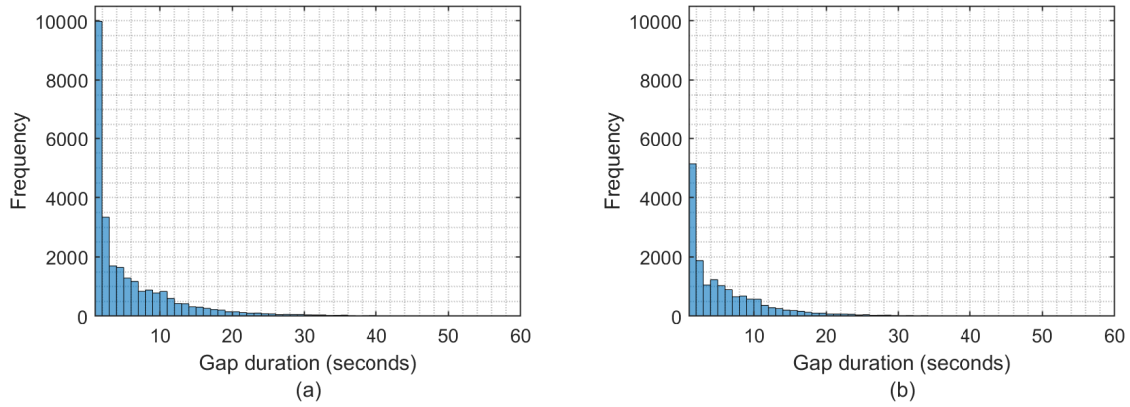


Figure 7.25: Gaps in heart rate estimation using the the AR best model with Kalman filtering on (a) the training set and (b) the test set during valid time periods. Most of the gaps are under 30 seconds. If gaps of up to 30 seconds are allowed, the percentage of the estimated time increases from 62.8% to 90.0% for the training set and from 75.4% to 96.9% for the test set. Therefore, heart rate estimation using a video camera can be considered to be continuous in a practical clinical context.

a major problem in a practical clinical environment. The results of heart rate estimation could be further processed by using Gaussian process regression [209] for modelling a time series of heart rate estimates in order to cope with noisy or incomplete periods of data.

7.7.6 Kalman filtering

Li *et al.* [189] suggested using Kalman filtering to obtain a more robust estimation of vital signs. In this work, as shown in table 7.1, the Kalman filter did improve the estimation of heart rate obtained using frequency-domain approaches, i.e. FFT and AR. The Kalman filter was able to suppress spikes occurring during noisy periods in the PPGi signal and smooth heart rate estimates (see Figure 7.20). This led to the highest improvement in the values of the MAE and MAD.

7.7.7 Effects of skin colour and body hair texture

Table 7.6 shows that the estimation errors of heart rate estimates from the White ethnic group were generally lower than the other ethnic groups with darker skin tone. The results were similar to previous studies [84, 85], which evaluated the use of pulse oximeters with dark skin tone. The studies reported poor-quality signal, and greater inaccuracies when monitoring ethnic groups with dark skin tone.

For our clinical study, a PPGi signal from a subject with darker skin tone as well as a subject filmed under low-light condition was noisier than that from a subject with lighter skin tone under normal ambient light. The amplitude of the pulsatile component (AC component) as well as the strength of cardiac component in the signal from the dark skin-tone subject were lower than those from the light skin-tone subjects. This is probably due to the lower level of colour intensity in

the dark skin-tone subjects and the quantum efficiency of the green channel of the camera which favours the subject with lighter skin tone (see chapter 2).

Subject PN25 (see table 7.5) is a dark-skinned infant with his whole body covered with hair. There were difficulties in the estimation of heart rate from this subject. All sessions from this subject yield very high errors (MAE of 4.5 beats/min for PN025 Day 1 and MAE of 7.3 beats/min for PN025 Day 2).

7.7.8 Challenges for heart rate estimation in the clinic

Motion artefacts

Figure 7.26 shows heart rate estimates for a 60-minute video segment with annotated patient and clinical activities. During periods of patient movement, heart rate estimates are unreliable with low signal quality, and the absolute error between the reference values and the camera-derived heart rate estimates increases substantially. One important disadvantage of using the whole skin area to estimate heart rate is that a small hand movement can result in a large distortion in the PPGi signal. This could be improved by using a localised ROI, e.g. face or torso, instead of the whole skin to avoid the effects of small movements from hands and legs. Computer vision techniques

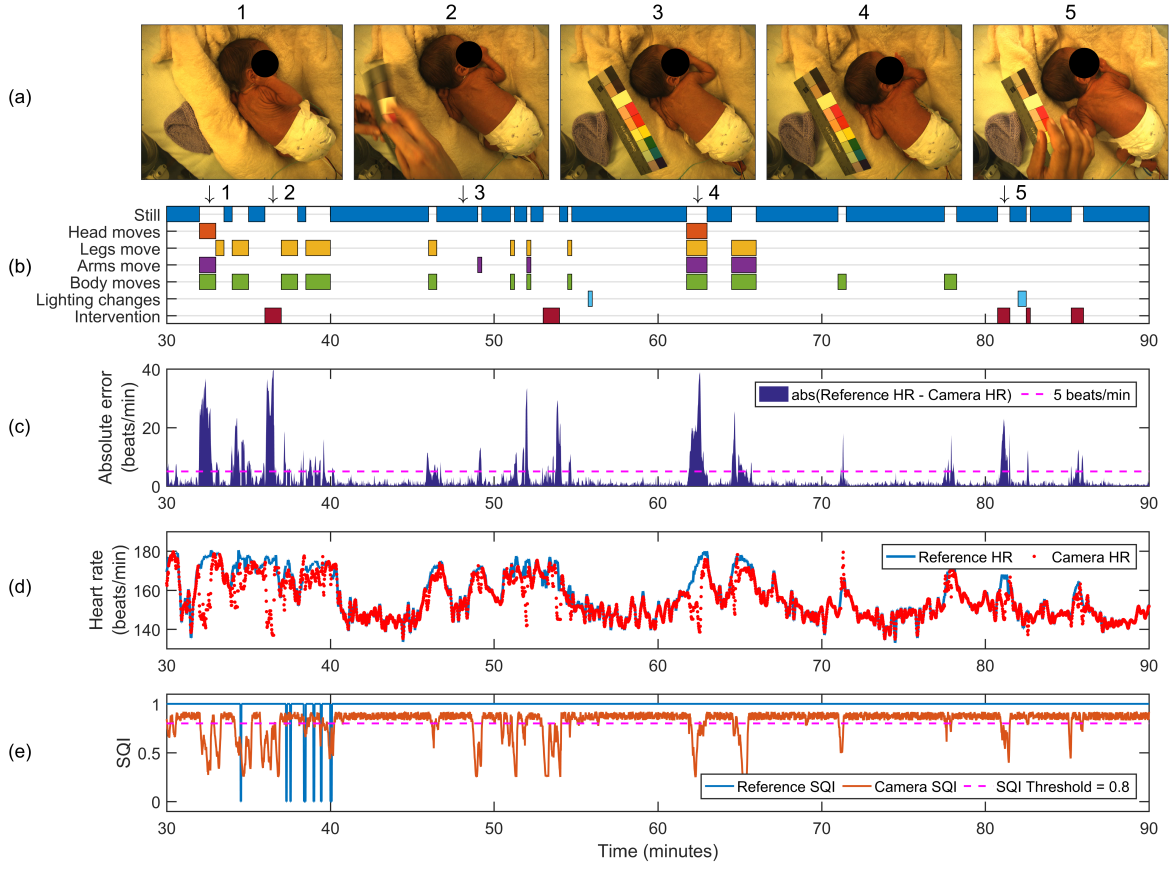


Figure 7.26: Typical activities of a patient and clinical staff can contribute to errors in estimating vital signs. (a) Video frames corresponding to the time in the plot below. (b) Timeline of patient activities over a 60-minute segment for a typical recording session, manually annotated minute-by-minute by the author. (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 estimates obtained using the AR best model with Kalman filtering. (e) Heart rate SQIs.

could be used to derive features relating to the activity and movement of the subject such that the signal quality assessment could potentially be improved.

Clinical intervention

In an intensive care environment, clinical staff perform routine clinical interventions several times a day, such as: physical examination, changing nappies, or administering medications. During these periods, the substantial motion caused by clinical staff makes it challenging to compute accurate heart rate estimates from video. However, during kangaroo care (holding the infant silently for a period of time) the intervention detection algorithm flagged these periods and discarded them from the estimation of heart rate. Since the skin detection algorithm was trained only with the pre-term infant's skin data, it could not recognise whether the skin belonged to the infant or clinical staff. An algorithm that is capable of identifying the subject of interest when he or she is being held, for example during kangaroo care, is needed.

Lighting conditions

One important factor for continuous non-contact heart rate estimation is the amount of lighting in the scene. This clinical study shows that heart rate estimation can be performed in the clinic at daytime under regular lighting of the HDU room. The algorithm can cope with small variations in lighting; however, heart rate cannot be derived when the light is dimmed or blocked, e.g. when clinical staff cover the incubator to facilitate quiet sleeping periods.

In this clinical study, the aperture of the camera was fixed at the beginning of each recording. Since the camera was equipped with manual focus lenses, the setting could not be programmatically changed according to the changes in the lighting conditions. It may be possible to estimate heart rate when the light is dimmed if the aperture of the lens is adjusted dynamically to allow more light reaching the camera's sensor. More studies are needed in order to define the interaction between the heart rate estimation algorithm and the aperture setting of a video camera.

Camera calibration

The detailed results of tables 7.2–7.5 show that some recording sessions had very low estimation error (MAE < 2.0 beats/min) and very high estimated time (> 90%). Example sessions are PN02 Day 3, PN08 Day 2, and PN20 Day 3. These were typical sessions, and the infants in these sessions were having clinical interventions, parent visits, and multiple position changes during the day. These sessions, however, have two important conditions in common: the videos are in focus and very sharp, with crisp detail and a high dynamic range. The lighting levels in the scene were fairly stable throughout the video.

Heart rate cannot be estimated if the video is out of focus. The strength of the cardiac component derived from the skin region in those cases is very low. The signal quality assessment flags most beats as invalid. Therefore, the sessions that were out of focus were excluded from the estimation. For each subject, in addition, the camera settings (focus and aperture) were normally set on the first day of the study and retained for the following days. Improper settings on the first day might affect all sessions of that particular subject. For example, all sessions from PN12 and PN16 were excluded from the estimation.

For the deployment of the system in the clinic, the heart rate estimation algorithm should be able to control the video camera and its lenses in order to ensure that suitable settings (focusing, aperture, shutter speed, ISO speed, and gain) are correctly configured. The algorithm should also be able to control the exposure dynamically to compensate for lighting changes during video recording.

7.8 Conclusion

This chapter has presented algorithms for the estimation of heart rate on hospitalised pre-term infants nursed in the Neonatal HDU, with an emphasis on the investigation of the feasibility of using the whole skin area for computing heart rate estimates. The skin regions of interest derived using the deep learning algorithm can be used for the estimation of heart rate. The mean absolute error between the reference heart rate and the camera-derived heart rate is lower than 2.3 beats/min for over 76% of the time. The accuracy of the camera-estimated heart rate is comparable to the accuracy of heart rates obtained from conventional monitoring. Estimation errors mostly occur because of artefacts caused by subject motion.

Future work may involve developing an algorithm to search for an optimal and localised skin area for estimating the vital signs. The PPGi signal acquired from optimal skin areas is likely to be less noisy, which would allow the use of lower cut-off frequencies for the detection of clinical conditions associated with a decrease in heart rate such as bradycardia. The use of localised skin areas could also allow the vital sign estimation to be performed during periods of small movements of the hands and legs. Finally, Gaussian process regression could be used to predict heart rate estimates during short-time periods where the data are incomplete.

Chapter 8
Reference respiratory rate

8.1 Introduction

Respiratory rate is an important vital sign that provides information regarding the patient's current condition and physiological stability. The standard technology commonly used for monitoring respiratory rate in pre-term infants is impedance pneumography (IP) which requires electrodes to be placed on the chest to measure changes of transthoracic impedance during breathing. The IP signal is known to be affected by noise and artefacts which could lead to errors in the estimation of respiratory rate [21, 210]. In sick newborn infants, the movement of the upper body during active awake periods often causes large motion artefacts in the IP signal which prevents a reliable estimation of respiratory rate [210, 211]. The shallow and irregular breathing patterns of pre-term infants make it difficult to measure respiratory rate using patient monitoring equipment. Even trained clinical staff found it difficult to estimate the respiratory rate of pre-term infants by counting chest wall movements [212].

In the Neonatal HDU study, the Philips IntelliVue patient monitor provides respiratory rate estimates derived from the IP signal using proprietary algorithms, without a corresponding quality measure. In the study by Jorge *et al.* [213], large discrepancies were found when comparing the respiratory rate reported by the patient monitor with that provided by manual breath counting. This finding suggests that the respiratory rates provided by the patient monitor are not suitable to be used as reference values for comparing with camera-derived estimates. Therefore, a new gold-standard reference respiratory rate is needed.

This chapter seeks to compute the gold-standard reference respiratory rate from the IP signal and a respiratory signal extracted from the ECG. It begins with an outline of methodology in section 8.2. The nature of the IP signal is described in section 8.3. The extraction of respiratory signals from the ECG is explained in section 8.4. The detection of peaks and onsets in the respiratory signals is described in section 8.6, with signal quality assessment covered in section 8.7. The methods for estimating respiratory rate and computing the new reference respiratory rate are described

in section 8.8. Subsequently, results are presented in section 8.9 and followed by a discussion in section 8.10. This chapter ends with conclusion in Section 8.11.

8.2 Overview of the estimation process

Figure 8.1 shows the overview of the process for estimating the reference respiratory rate. In the Neonatal HDU study dataset, there are two source signals which can be used to compute respiratory

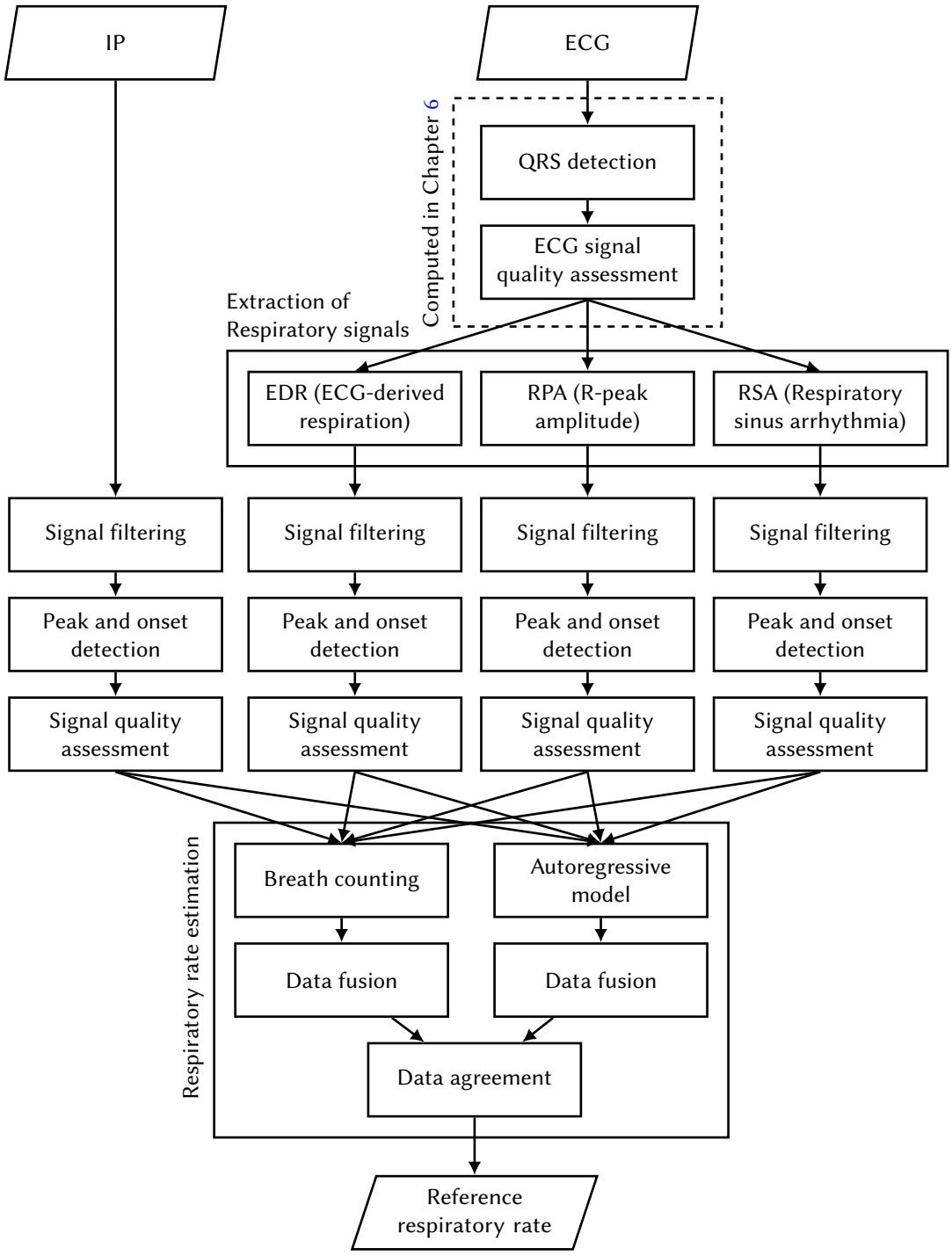


Figure 8.1: Block diagram outlining the reference respiratory rate estimation process.

rate: the IP and the ECG. The IP signal records a respiratory signal from which respiratory rate can be directly estimated straight away. Respiratory signals can also be derived from the ECG by extracting the amplitude of frequency that modulates in the ECG signal caused by breathing. The process of extracting respiratory signals continues from chapter 6 in which each QRS cycle in the ECG waveform was detected and the signal quality of each QRS cycle was assessed. Three respiratory signals are extracted from the ECG signal: ECG-derived respiration (EDR), respiratory sinus arrhythmia (RSA) and R-peak amplitude (RPA).

The four respiratory signals (IP, EDR, RSA and RPA) are fed into the same estimation pipeline. Each respiratory signal is first filtered within the frequency range of interest, which is between 18 and 120 breaths/min. Subsequently, the peak and onset of each breath are identified. The assessment of signal quality is then performed for each breath cycle. Respiratory rate is estimated using two techniques: a time-domain technique, by counting the number of breaths within a window; and a frequency-domain method; by finding the frequency of the dominant pole of an autoregressive model. For each estimation method, the respiratory rate estimates from each respiratory signal are combined together using a data fusion technique. The new reference respiratory rate is computed using the agreement of the combined respiratory rate estimates from both methods. The reference respiratory rate is computed as the mean of the respiratory rates from both methods if they are close enough. The reference respiratory rate is discarded if the two respiratory rates are far apart.

This process is performed using well-known algorithms that have been extensively validated by the research community on publicly available physiological databases.

8.3 Impedance Pneumography (IP)

The impedance pneumography (IP) signal records a respiratory signal by observing changes in the thoracic electrical impedance during each respiratory cycle [214]. In the Neonatal HDU study, the

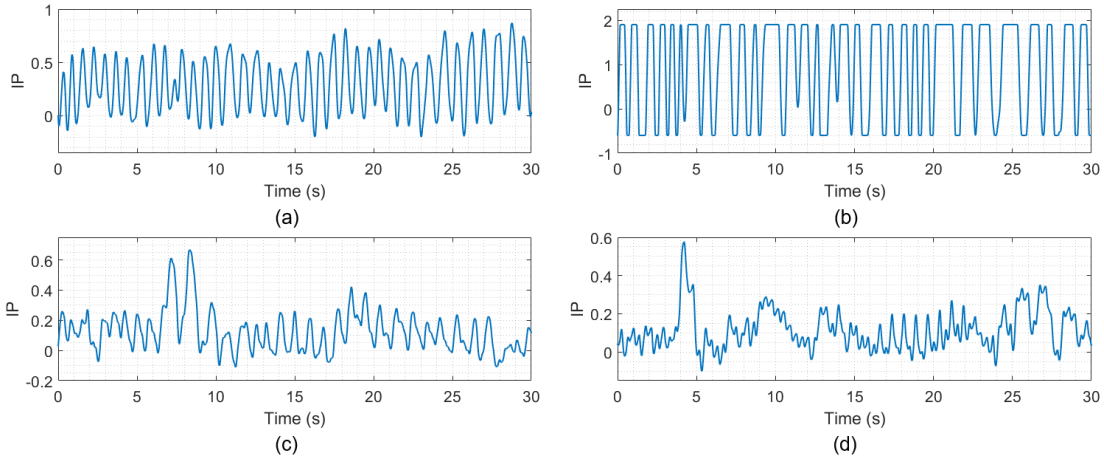


Figure 8.2: Typical IP waveforms extracted from the Neonatal HDU dataset. (a) IP signal not affected by noise. (b) IP signal clipped at extreme inspiratory and expiratory levels (deep breathing). (c–d) IP signal containing artefacts from the movement of the entire body.

IP signal was acquired at 64 Hz. Several conditions may have negative effects on the recording of the IP signal: improper contact of electrodes with the skin, electrodes peeling off the skin as the infant moves around, very dry skin causing low conductance between electrodes and the skin, shallow and irregular breathing, or infant orientation position, such as supine (face up, lying on the back) or prone (face down, lying on the stomach).

Figure 8.2 shows samples of the IP signal from typical recording sessions in the Neonatal HDU study. The IP signal does not have a distinct morphology and shape like the ECG or PPG signal. Shallow breathing may result in a low-amplitude respiration cycle. The IP signal is very sensitive to the movement of the upper body (the chest, arms and hands) [215].

8.4 Respiratory signal extraction from ECG

This section builds on the work presented in chapter 6 where QRS complexes were detected and the signal quality of the ECG waveform was assessed. Three techniques are used to extract respiratory signals from the ECG: ECG-derived respiration (EDR), respiratory sinus arrhythmia (RSA) and R-peak amplitude (RPA). Since these techniques generally provide an unevenly-sampled time series, all the extracted respiratory signals are resampled at 64 Hz, the same sampling rate as the IP signal, using spline interpolation. Figure 8.3 shows the respiratory signals extracted from a 60-

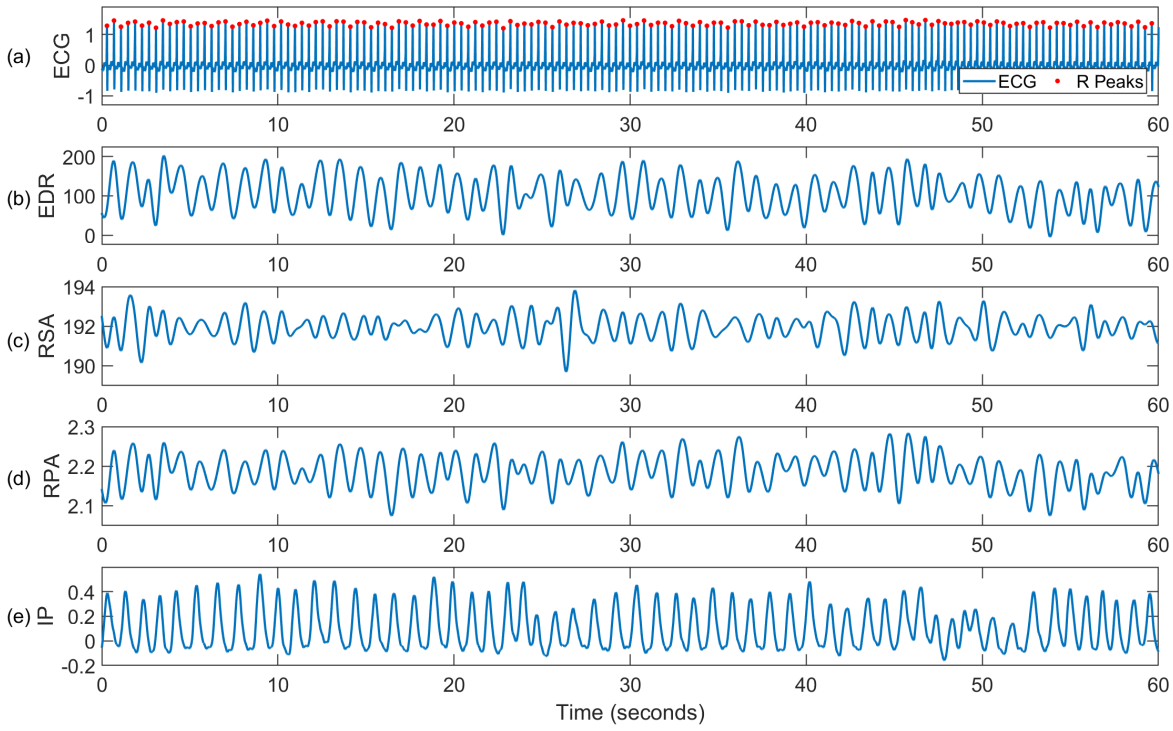


Figure 8.3: Respiratory signals extracted from a 60-second ECG segment. (a) ECG waveform with detected R peaks. (b) EDR signal extracted from the amplitudes of successive R peaks. (c) RSA signal extracted from R-R intervals. (d) RPA signal extracted from the distance from the R peak to the S peak. (e) IP signal during the same period as the ECG-derived respiratory signals. There were 55 breath pulses in the respiratory signal, corresponding to a respiratory rate of 55 breaths/min.

second ECG segment. The respiratory signals are only extracted from the periods during which the ECG is of good quality (see Figure 8.4).

8.4.1 ECG-derived respiration (EDR)

The motion of the heart during the cardiac cycle and the lungs during the respiratory cycle has an influence on the ECG signal measured from the chest surface. The electrical impedance of the thoracic cavity varies due to the filling and emptying of the lungs producing fluctuations accompanying respiration in the mean cardiac electrical axis [216]. EDR is a technique used to derive a respiratory signal from these fluctuations. An EDR signal is extracted by interpolating the amplitudes of successive R peaks. Figure 8.3b shows the EDR signal derived from a 60-second ECG segment.

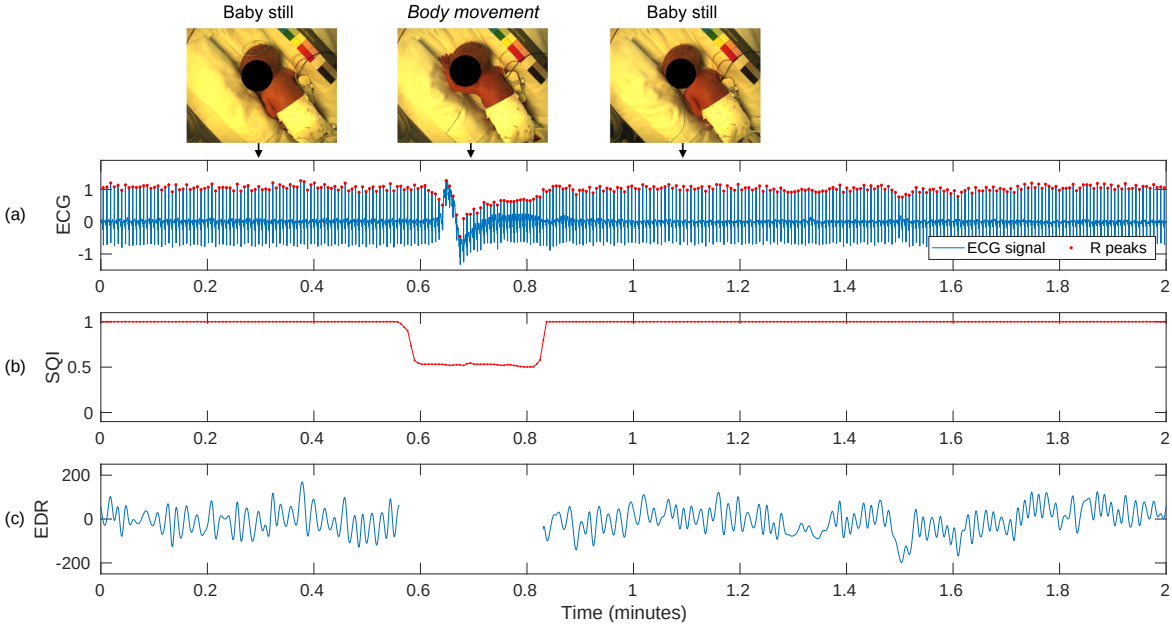


Figure 8.4: The extraction of a EDR signal on a 2-minute ECG segment during the presence of body movements. The signal quality of the ECG was low during the period of body movement such that the EDR signal was not extracted.

8.4.2 Respiratory sinus arrhythmia (RSA)

In the ECG signal, the beat-to-beat variations in R-R intervals occurr primarily due to respiratory sinus arrhythmia (RSA) [217]. The R-R interval is shortened during inspiration and prolonged during expiration. RSA has been shown to be present in newborns and found to be more prominent during sleeping [218]. An RSA signal is extracted by computing the R-R intervals from the ECG signal. Figure 8.3c shows the RSA signal extracted from a 60-second ECG segment.

8.4.3 R-peak amplitude (RPA)

During the respiration cycle, the movement of the chest causes a rotation of the electrical axis of the heart which, in turn, modulates the QRS complex in the ECG [219]. An RPA signal can be obtained by tracking the modulation in the amplitude of the R peak. This was done by measuring the absolute

distance from the R peak to the S peak for each cardiac cycle. RPA has been regarded as reliable in young populations due to the prominent movement of the chest during the respiratory cycle [220]. Figure 8.3d shows the RPA signal extracted from a 60-second ECG segment.

8.5 Signal filtering

The respiratory signal is usually contaminated by noise over the course of the recording. The sources of the noise in the respiratory signals are similar to those that affect the ECG signal since they share the same set of electrodes. In the Neonatal HDU dataset, the respiratory rate in pre-term infants varies from 20 to 120 breaths/min, depending on the gestational age and health condition of the infant. Pre-term infants tend to have higher respiratory rates [47]. Hence, frequencies outside the range of 0.3 and 2.0 Hz are attenuated using an 8th-order band-pass infinite impulse response (IIR) filter with cut-off frequencies of 0.3 Hz and 2.0 Hz. Figure 8.5 shows a 30-second IP waveform before and after filtering.

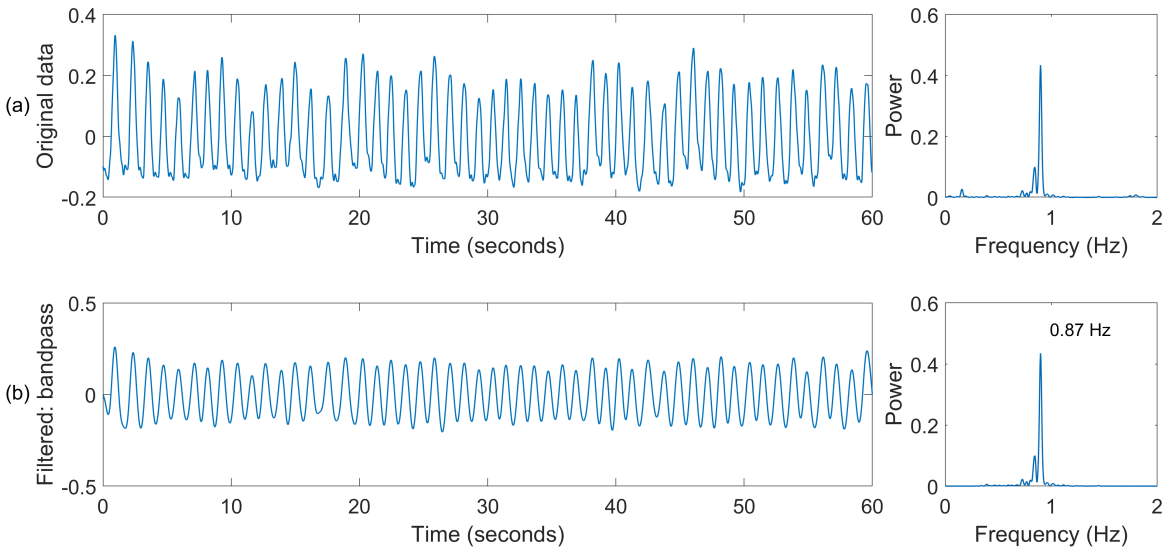


Figure 8.5: Signal filtering of a 60-second IP waveform. (a) Original waveform. (b) Waveform after the band-pass filter with cut-off frequencies of 0.3 Hz and 2.0 Hz. This waveform has 52 breaths in 60 seconds, corresponding to a frequency of 0.87 Hz.

8.6 Peak and onset detection

To identify the breathing cycle, the peak and onset corresponding to each cycle in the filtered respiratory signal are first detected using a method based on calculation of the mean average curve (MAC) proposed by Ruangsuwana *et al.* [221] and Lu *et al.* [222]. The MAC is computed by sliding a mean filter over the respiratory signal. The length of the mean filter is computed by first dividing the respiratory signal into ten small learning periods of 5 minutes distributed sparsely over the signal for each recording session. The respiratory frequency is computed as the peak of the FFT power spectrum over each of the learning period. Finally, the length of the mean filter is set to be the median of respiratory durations (calculated from the respiratory frequency) from all the learning periods.

The MAC signal is used as a dynamic threshold to find peaks and onsets in the respiratory signal. The peak is defined as the local maximum of values above the threshold, and the onset is defined as the local minimum between two successive peaks that have already been detected. A peak is rejected if its amplitude is below 10% of the median amplitude calculated from the learning periods. Figure 8.6 shows the results of the peak and onset detection for a 30-second respiration signal.

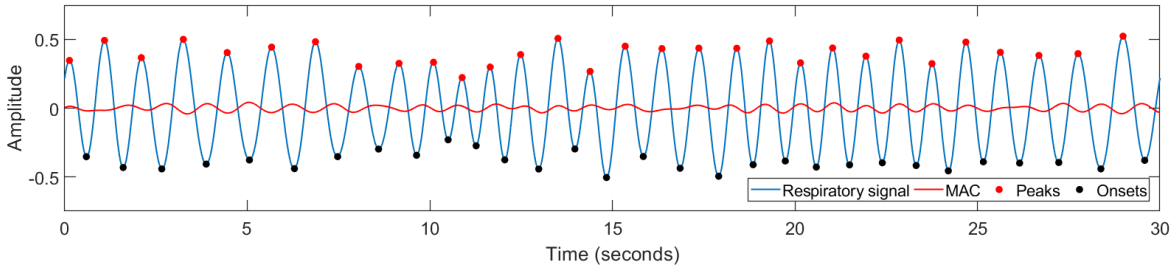


Figure 8.6: Peak and onset detection on a 30-second respiration signal (IP) extracted from a typical recording session in the Neonatal HDU study. The MAC signal, shown in red, is computed as a dynamic threshold. Peaks correspond to the local maxima above the threshold and onsets correspond to the local minima found between two successive peaks.

8.7 Signal quality assessment

The assessment of signal quality is performed based on the work of Li *et al.* [189, 223] whose signal quality was derived from multiple signal quality measures, namely: physiological bounding (SQI_{phys}), spectral density distribution (SQI_{bin}) and signal purity index (SPI). Each signal quality measure is computed for each each respiratory cycle, based on the peak and onset detected in the previous section. These four individual SQIs are later combined to compute the signal quality index of the respiratory signal, SQI_{breath}. This process assigns a score from 0 (poor quality) to 1 (good quality) to each respiratory cycle.

8.7.1 Physiological bounding

Physiological bounding measures whether the instantaneous respiratory rate calculated from the duration of each respiratory cycle is within the expected range for pre-term infants, taken to be from 18 to 120 breaths/min. Given $RR_{\text{i}}(k)$ is the instantaneous respiratory rate of the kth peak, the measure of physiological bounding, SQI_{phys}, is defined as

$$SQI_{\text{phys}}(k) = \begin{cases} 1 & \text{if} \quad 18 \leq RR_{\text{i}}(k) \leq 120 \text{ breaths/min} \\ 0 & \text{otherwise} \end{cases} . \tag{8.1}$$

8.7.2 Spectral density distribution

Spectral density distribution measures how unimodal the distribution of the power spectrum is. Since the energy of the respiratory signal is often concentrated around the peak of the power spectrum, the ratio of frequencies around the power spectrum peak to the overall power spectrum can be used

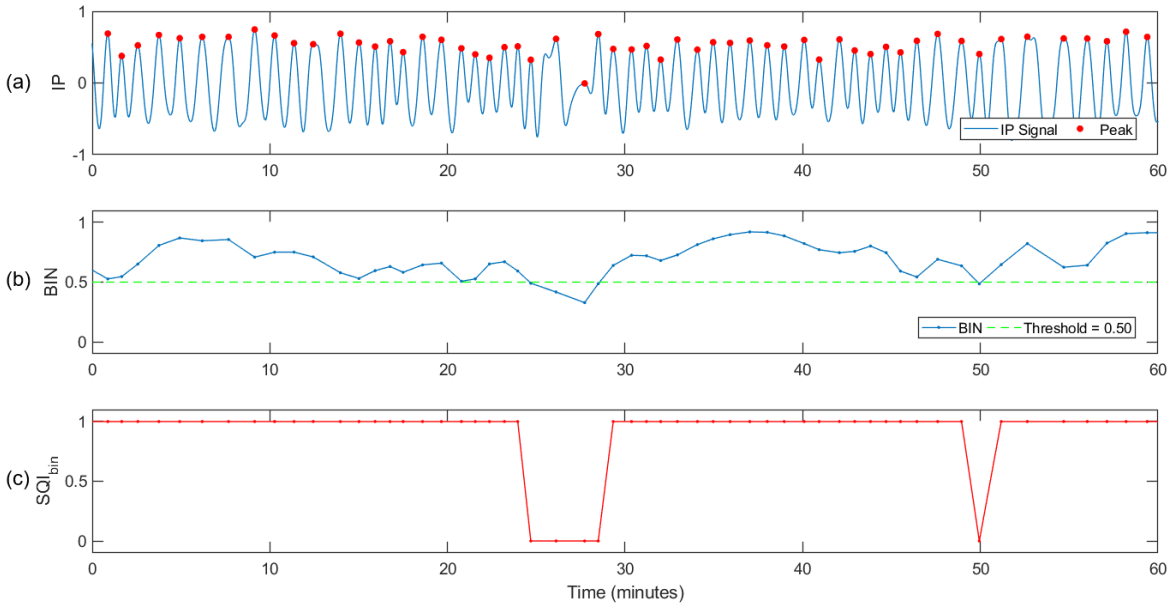


Figure 8.7: Computation of spectral density distribution for a 60-second IP segment: (a) IP signal, (b) *BIN* values, (c) SQI_{bin} indicates the breaths that have BIN values less than 0.50.

as a measure of signal quality. Given that $P(w_k)$ is the power spectrum of the window w_k of 15 seconds centred around the kth peak and f_d is the dominant peak frequency in $P(w_k)$, the ratio of the power spectrum $BIN(w_k)$ of f_d to the overall power spectrum is defined as

$$BIN(w_k) = \frac{\int_{f_k - f_{bw}/2}^{f_k + f_{bw}/2} P(w_k) df}{\int_0^{f_s/2} P(w_k) df} \tag{8.2}$$

where f_{bw} is the bandwidth around the dominant peak frequency f_d defined as 0.3 Hz, and f_s is the sampling frequency. The measure of spectral density distribution, SQI_{bin}, is then defined as

$$SQI_{\text{bin}}(k) = \begin{cases} 1 & \text{if} \quad BIN(w_k) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}. \tag{8.3}$$

The SQI_{bin} is set to 1 (good quality) if more than half of the power spectrum is concentrated within a bandwidth of 0.3 Hz around the dominant peak (see figure 8.7).

8.7.3 Spectral Purity Index (SPI)

Spectral Purity Index (*SPI*) is a measure describing the degree of that signal can be approximately a single sinusoid. It reflects the stability of the signal. The *SPI* is computed based on Hjorth descriptors [224]. Given that $S_x(e^{j\omega})$ is the signal power spectrum as a function of angular frequency ω, the nth-order spectral moment $\bar{\omega}_n$ is defined as:

$$\bar{\omega}_n = \int_{-\pi}^{\pi} \omega^n S_x(e^{j\omega}) d\omega. \tag{8.4}$$

The zero-order, second-order, and fourth-order spectral moments are equivalent to the Hjorth's parameters of Activity, Mobility, and Complexity respectively [224]. Given a discrete time series

$x(n)$ of length N, the spectral moments can be estimated as:

$$\bar{\omega}_i \approx \frac{2\pi}{N} \sum_{n=0}^{N-1} \left(x^{\frac{i}{2}}(n)\right)^2, \qquad i = 0,\ 2,\ 4. \tag{8.5}$$

The *SPI* of a signal in a window w_k of 15 seconds centred around the kth peak is defined as:

$$SPI(k) = \frac{\bar{\omega}_2^2(w_k)}{\bar{\omega}_0(w_k)\bar{\omega}_4(w_k)}. \tag{8.6}$$

The *SPI* ranges from 0 for non-sinusoidal noisy signals to 1 for pure sine wave signals.

8.7.4 Combining signal quality measures

Once all the signal quality measures have been computed, the $SQI_{\text{breath}}(k)$ describing the overall signal quality of the kth breath in the respiratory signal is estimated by combining all the signal quality measures:

$$SQI_{\text{breath}}(k) = SQI_{\text{phys}}(k) \cdot SQI_{\text{bin}}(k) \cdot SPI(k). \tag{8.7}$$

SQI_{phys} sets the SQI_{breath} to be zero if an instantaneous respiratory rate is not in the expected range for pre-term infants. In the same way, if there are more than one dominant frequencies in the power spectrum, SQI_{bin} penalises the SQI_{breath} to be zero. Therefore, the *SPI* is trusted only when the spectrum is likely to be unimodal.

Figures 8.8 and 8.9 show the assessment of the signal quality for two respiratory waveforms extracted from a typical recording session. The SQI_{breath} is high when the waveform is sinusoidal. In Figure 8.9, the SQI_{bin} is 0 and the *SPI* decreases during the presence of body movement, therefore the SQI_{breath} during this period is set to 0, corresponding to a period of poor quality for which respiratory rate cannot be reliably computed.

8.8 Estimation of reference respiratory rate

Once the peaks and onsets of all respiratory signals (IP, EDR, RSA and RPA) have been detected and the signal quality has been assessed for each breath, respiratory rate is first estimated for each respiratory signal using the breath counting and autoregressive model methods over a 10-second sliding window with a step size of 1 second. Subsequently, for each method, the respiratory rate estimates (from IP, EDR, RSA and RPA) are combined together using the Kalman filter-based data fusion technique. Finally, the new reference respiratory rate is computed as the mean of the combined respiratory rate estimates from both methods during which the data are of good quality and their difference is less than 5 breaths/min. The maximum error of 5 breaths/min has been used as a primary outcome measure in many clinical trials involving the measurement of respiratory rate that have been approved by the U.S. National Institutes of Health (NIH).

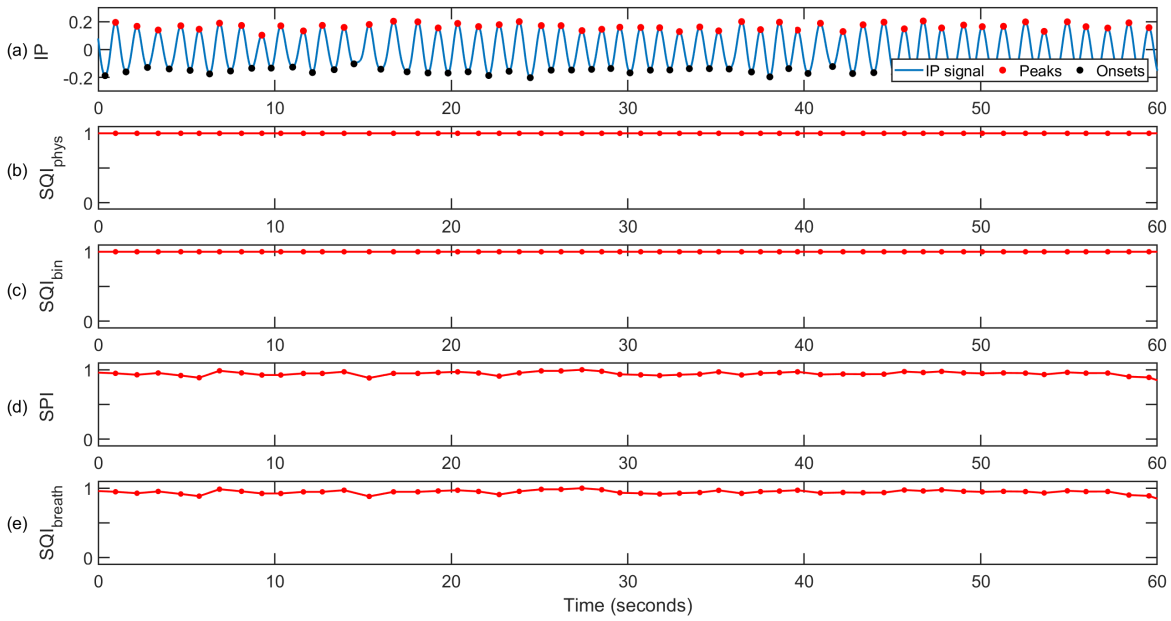


Figure 8.8: Signal quality assessment of a 60-second IP waveform from a quiet period. (a) IP signal with the peaks and onsets detected. (b) SQI_{phys} checks whether instantaneous respiratory rate is valid. (c) SQI_{bin} measures the waveform being unimodal. (d) SPI assesses the purity of the waveform. (d) SQI_{breath} combines all the previous measures.

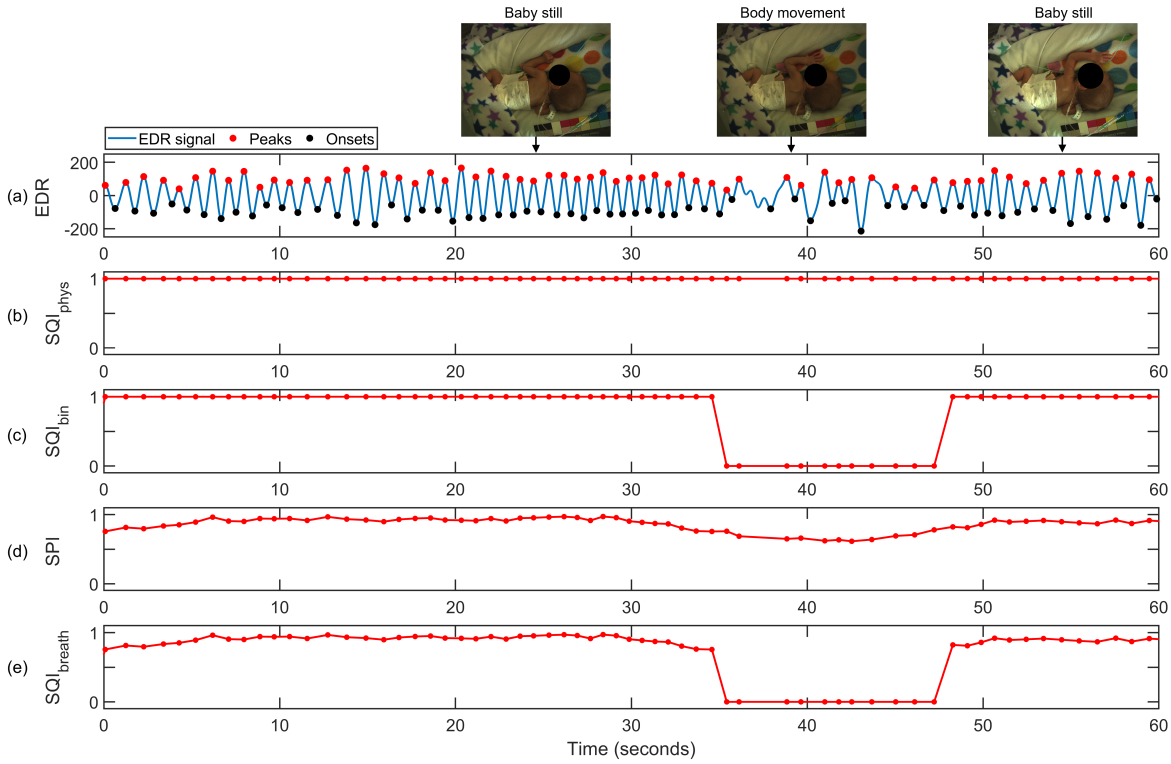


Figure 8.9: Signal quality assessment of a 60-second EDR waveform from a noisy period. (a) EDR signal with the peaks and onsets detected. Noise is present at around second 40. (b) SQI_{phys}. (c) SQI_{bin} (d) SPI (d) SQI_{breath} combines all the previous measures. The values of the SQI_{bin} and SPI decrease during the presence of motion artefacts.

8.8.1 Breath counting

The breath counting method is used to estimate respiratory rate by counting the number of breaths in a 10-second sliding window. During the estimation, the window size is slightly expanded to include the complete first and last breaths. The number of breaths in the window is counted (see section 7.5.1). The signal quality of the respiratory rate (SQI_{RR}) is defined as the mean of the SQI_{breath} values in the window. Figure 8.10 shows the estimation of respiratory rate for a 5-minute respiratory signal comparing with values from the Philips patient monitor. The SQI_{breath} values decreased during noisy periods.

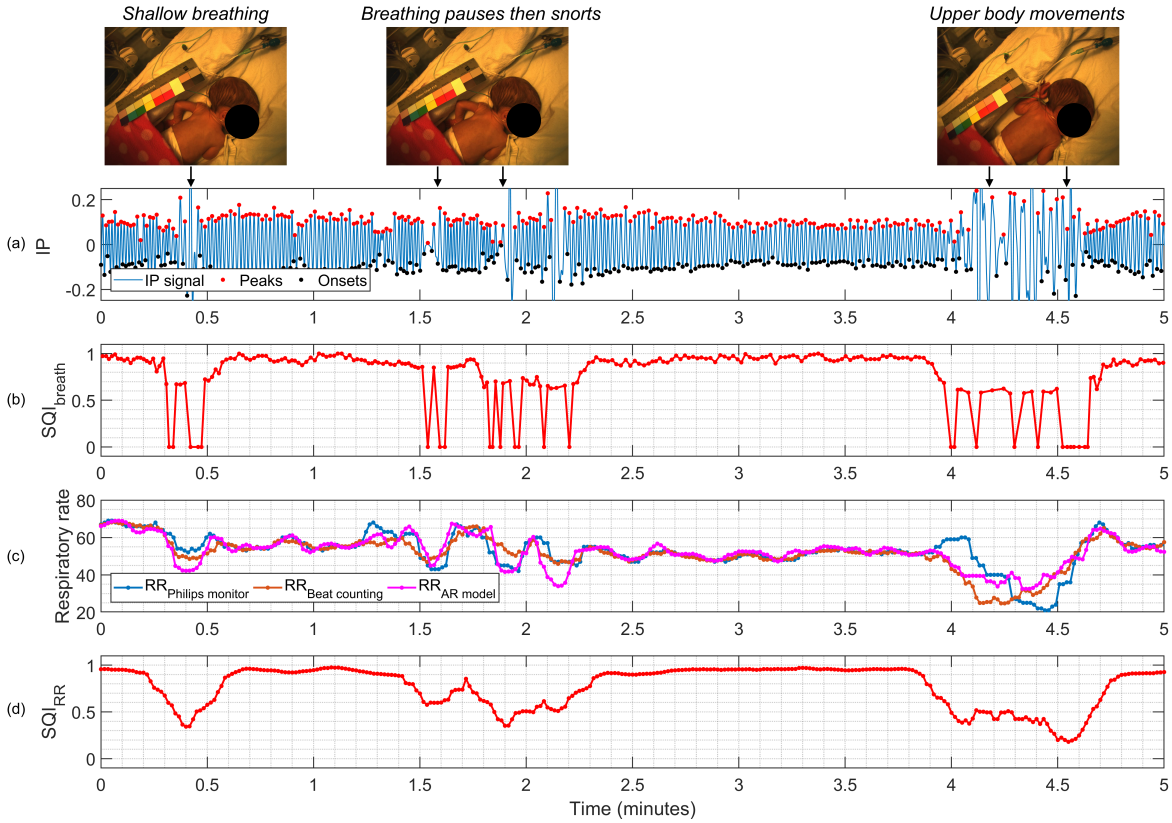


Figure 8.10: Respiratory rate estimation using the breath counting and autoregressive model methods on a 5-minute IP signal: (a) IP signal, (b) SQI_{breath}, (c) Respiratory rate estimates, and (d) SQI_{RR}. The SQI_{RR} decreases during the presence of irregular breathing patterns and body movements.

8.8.2 Autoregressive model

The autoregressive model is a frequency-domain technique that can provide an estimate of respiratory rate without the limit of frequency resolution that traditional methods, such as the Fast Fourier Transform (FFT), have (see section 7.5.2). The estimation is performed by fitting a 9th-order autoregressive model to 10-second windows using Burg's algorithm [205] (see section 7.5.3 for more detail). The respiratory frequency is defined as the frequency of the highest-magnitude pole of the autoregressive model that corresponded to the physiological range of 18 to 120 breaths/min.

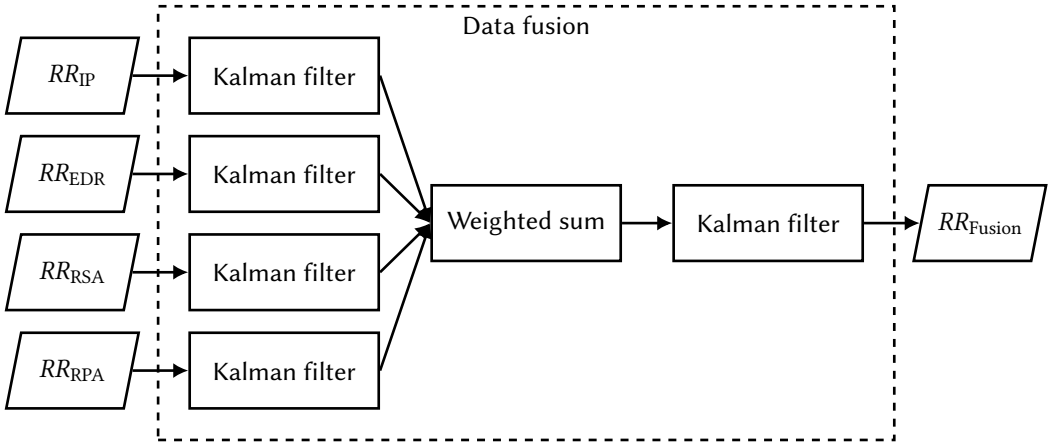


Figure 8.11: Multiple Kalman filters were used to combine multiple respiratory rate estimates through a weighted-sum data fusion technique. Data fusion is applied for respiratory rate estimates computed using the breath counting and autoregressive model techniques separately.

Figure 8.10 shows the estimation of respiratory rate for a 5-minute respiratory signal using the autoregressive model method, compared to that using the breath counting method. The respiratory rate estimates from both methods during the periods of high signal quality are consistent with the respiratory rate values reported by the Philips monitor during the same periods.

8.8.3 Data fusion

Once respiratory rate has been computed for every respiratory signal, a data fusion technique based on multiple Kalman filters [189] is applied to produce a combined respiratory rate from each of the two estimation methods (breath counting and autoregressive model). In contrast to the technique based on a single Kalman filter used in section 7.5.5 for improving a single physiological estimate, the multiple Kalman filtering technique combines multiple respiratory rate estimates by weighting the physiological values based on their signal quality indices.

Figure 8.11 illustrates the data fusion process using multiple Kalman filters for any of the two estimation methods. Each respiratory rate first is passed through an individual Kalman filter to adjust its values based on its signal quality in order to reduce the effects of transient changes, artefacts and noise (see section 7.5.5). Subsequently, the output from each Kalman filter is combined based on the methods of Tarassenko *et al.* [208] and Li *et al.* [189]. The combined respiratory rate estimate RR_{fusion} is computed for each time window by weighting each Kalman-filtered respiratory rate estimate RR_{k} as [123, 189]:

$$RR_{\text{Fusion}} = \sum_{k=1}^{N} \left(\frac{\prod_{i=1, i\neq k}^{N} \sigma_i^2}{\sum_{i=1}^{N} (\prod_{j=1, j\neq i}^{N} \sigma_i^2)} \cdot RR_{\text{k}} \right) \tag{8.8}$$

where $k = [1, 2, 3, 4]$ corresponds to the Kalman-filtered respiratory rate estimates from the IP, EDR, RSA and RPA signals respectively. Given the SQI of respiratory rate estimate SQI_{k} and Kalman

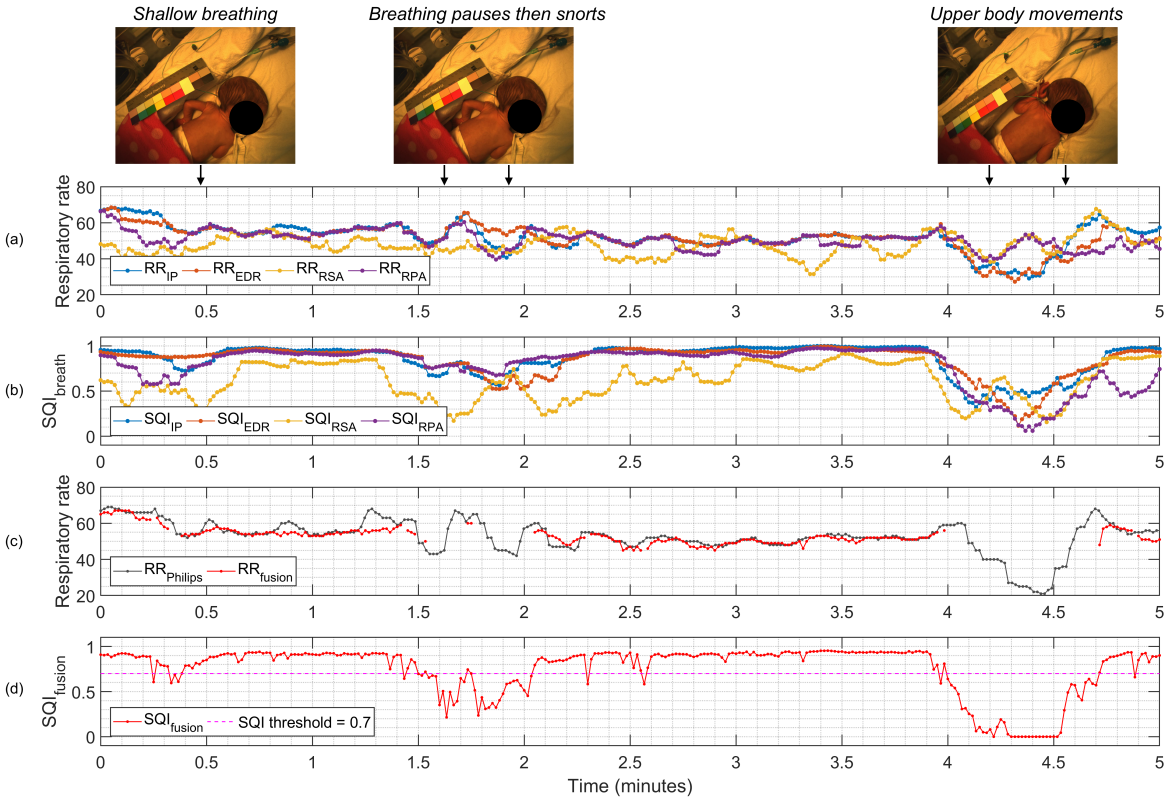


Figure 8.12: Data fusion of respiratory rate estimates: (a) Respiratory rate estimates (RR) from IP, EDR, RSA and RPA using the breath counting method, (b) SQI_{breath} associated with the respiratory rate estimates above, (c) Combined respiratory rate estimate (RR_{Fusion}), and (d) Signal quality of RR_{Fusion} (SQI_{Fusion}). The disagreement among respiratory rate estimates computed from different respiratory signals often occurs during the periods of low SQI_{breath}, e.g. the presence of irregular breathing patterns or artefacts.

residual error r, the weighted innovation σ^2 is defined as:

$$\sigma^2 = \left(\frac{r}{SQI_{\text{k}}}\right)^2. \tag{8.9}$$

This procedure gives the respiratory rate estimates from cleaner data (high SQI_{k} and low r) higher weights than noisy data (low SQI_{k} and high r) [189]. The combined signal quality SQI_{Fusion} is calculated in the same way, using the same weighting factors, as the combined respiratory rate. Finally, the RR_{Fusion} values are processed through another Kalman filter in order to improve the overall respiratory rate estimates. Figure 8.12 shows the data fusion of the respiratory rate estimates for a 5-minute segment. The respiratory rate estimates from the IP, EDR, RSA and RPA signals generally agree with each other when their signal quality is high (see figure 8.12a–b). The resulting combined respiratory rate estimates during the periods of high signal quality were consistent with the values from the patient monitor (see figure 8.12c–d).

8.8.4 Data agreement

Table 8.1 summarises the comparison of the respiratory rate from the manufacturer (RR_{Philips}), the breath counting method ($RR_{\text{Fusion BC}}$) and the autoregressive model method ($RR_{\text{Fusion AR}}$) for the

Table 8.1: Comparison of the respiratory rate from the patient monitor (RR_{Philips}), the breath counting method ($RR_{\text{Fusion BC}}$) and the autoregressive model method ($RR_{\text{Fusion AR}}$) for periods during which the SQI_{Fusion} was higher than 0.7 for the entire Neonatal HDU dataset, comprising 426.6 hours.

Sources	MAE (breaths/min)	MAD (breaths/min)	Estimated time	
			%	hours
RR_{Philips} vs $RR_{\text{Fusion BC}}$	5.8	6.5	74.5%	318.0
RR_{Philips} vs $RR_{\text{Fusion AR}}$	7.3	7.1	74.4%	317.2
$RR_{\text{Fusion BC}}$ vs $RR_{\text{Fusion AR}}$	**5.5**	**5.2**	**75.0%**	**320.0**

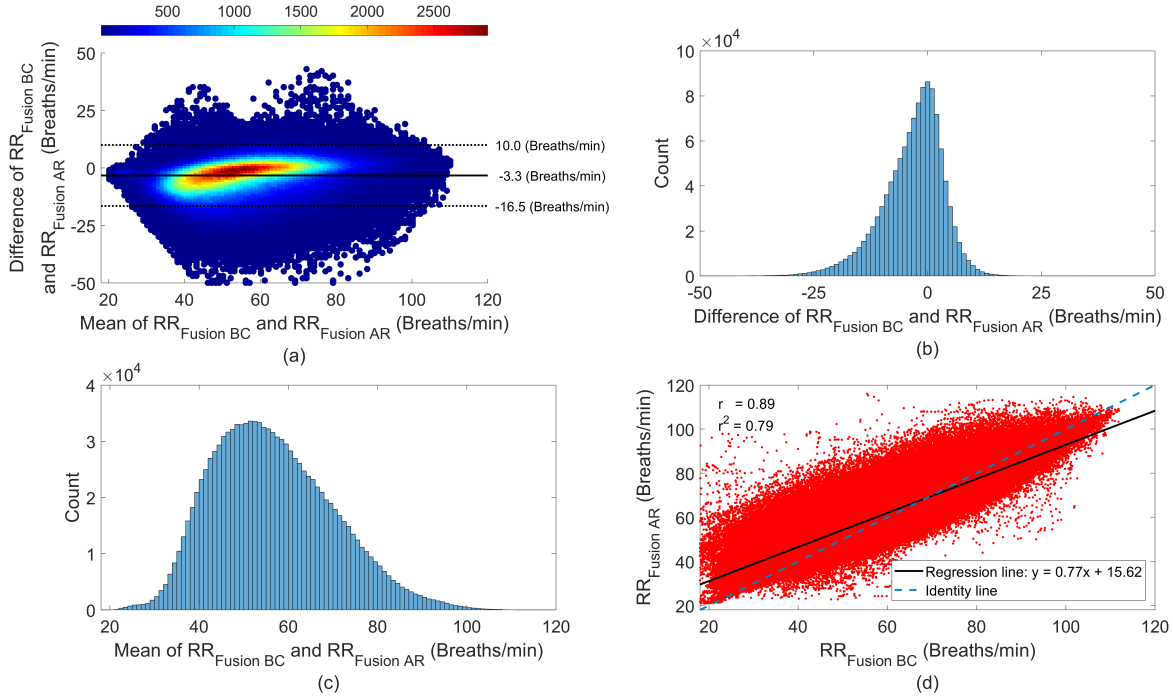


Figure 8.13: Comparison of the respiratory rate estimates between the breath counting ($RR_{\text{Fusion BC}}$) method and the autoregressive model method ($RR_{\text{Fusion AR}}$) for the entire Neonatal HDU dataset. (a) The Bland-Altman plot shows a bias of -3.3 breaths/min. (b) The distribution of the difference between the two time series. Most errors are concentrated around zero. (c) The mean of the two time series. (d) The correlation plot shows a positive correlation between the two time series with a correlation coefficient r of 0.89.

entire Neonatal HDU dataset of 426.6 hours. Only the time windows during which the signal quality ($SQI_{\text{Fusion BC}}$ or $SQI_{\text{Fusion AR}}$) is higher than 0.7 are retained. The lowest errors are observed when comparing between the $RR_{\text{Fusion BC}}$ and $RR_{\text{Fusion AR}}$ estimates, with a mean absolute error (MAE) of 5.5 breaths/min and a mean absolute deviation (MAD) of 5.2 breaths/min (see figure 8.13 for a graphical comparison). Higher errors are observed when comparing the RR_{Philips} value with the $RR_{\text{Fusion BC}}$ or $RR_{\text{Fusion AR}}$ values. The proportion of the estimated time to the total recording time for the three comparisons are similar, at around 74%. Figure 8.13 illustrates the comparison between the $RR_{\text{Fusion BC}}$ and $RR_{\text{Fusion AR}}$. The Bland-Altman plot shows a bias of −3.3 breaths/min. The estimation errors are mostly concentrated around zero. The correlation coefficient r was 0.89.

Given the comparison above and the noisy nature of the respiratory signals, the use of a reference respiratory rate from only one source (RR_{Philips}, $RR_{\text{Fusion BC}}$ or $RR_{\text{Fusion AR}}$) would not be

sufficient and may lead to the inaccurate results when comparing with camera-derived respiratory estimates. Given that the comparison between the $RR_{\text{Fusion BC}}$ and $RR_{\text{Fusion AR}}$ yielded the lowest estimation errors for the highest estimated time, the new reference respiratory rate was, therefore, computed by averaging the values from both sources during the periods that both values are in agreement, within 5 breaths/min.

8.8.5 Computing the reference respiratory rate

The new reference respiratory rate ($RR_{\text{Reference}}$) is computed from the agreement of the $RR_{\text{Fusion BC}}$ and $RR_{\text{Fusion AR}}$. Firstly, the two time series are aligned in time using the cross-correlation method as described in section 6.4.5. The $RR_{\text{Reference}}$ is later computed as the mean of the two time series for which the two respiratory values do not differ by more than 5 breaths/min ($|RR_{\text{Fusion BC}} - RR_{\text{Fusion AR}}| \leq 5$ breaths/min) and their signal quality indices ($SQI_{\text{Fusion BC}}$ and $SQI_{\text{Fusion AR}}$) are higher than 0.7, which implies a good signal quality for both respiratory rate estimates.

8.9 Results

Table 8.2: Comparison of the respiratory rate estimates between the breath counting method ($RR_{\text{Fusion BC}}$) and the autoregressive model method ($RR_{\text{Fusion AR}}$) for the time during which their difference was less than 5 breaths/min and their signal quality indices are higher than 0.7, for the entire Neonatal HDU dataset

MAE (breaths/min)	MAD (breaths/min)	Total reference recording time	Valid time	
			%	hours
2.2	1.4	426.6	44.3	189.0

Table 8.2 shows the comparison between the $RR_{\text{Fusion BC}}$ and the $RR_{\text{Fusion AR}}$ for the entire Neonatal HDU dataset. The total recording time of the reference clinical data was over 426.6 hours. The valid time corresponds to the duration for which the two respiratory rate values do not differ by more than 5 breaths/min and their signal quality indices ($SQI_{\text{Fusion BC}}$ and $SQI_{\text{Fusion AR}}$) are higher than 0.7. The valid time accounted for over 189.0 hours or 44.3% of the total recording time, with a mean absolute error of 2.2 breaths/min and a mean absolute deviation of 1.4 breaths/min.

Figure 8.14 shows the graphical comparison between the $RR_{\text{Fusion BC}}$ and the $RR_{\text{Fusion AR}}$ for the entire Neonatal HDU dataset. The Bland-Altman plot shows a small bias of -0.3 breaths/min. The distribution of the differences between the two time series shows that the values are concentrated around zero. The mean values of the two respiratory rate values are distributed across the expected range for pre-term infants. The correlation plot shows a very high correlation between the two values with a positive correlation coefficient r of 0.98.

Table 8.3 details the amount of the valid reference data for each recording session, listed separately for the training set and the test set. In the training set, over 95.3 hours out of the total reference

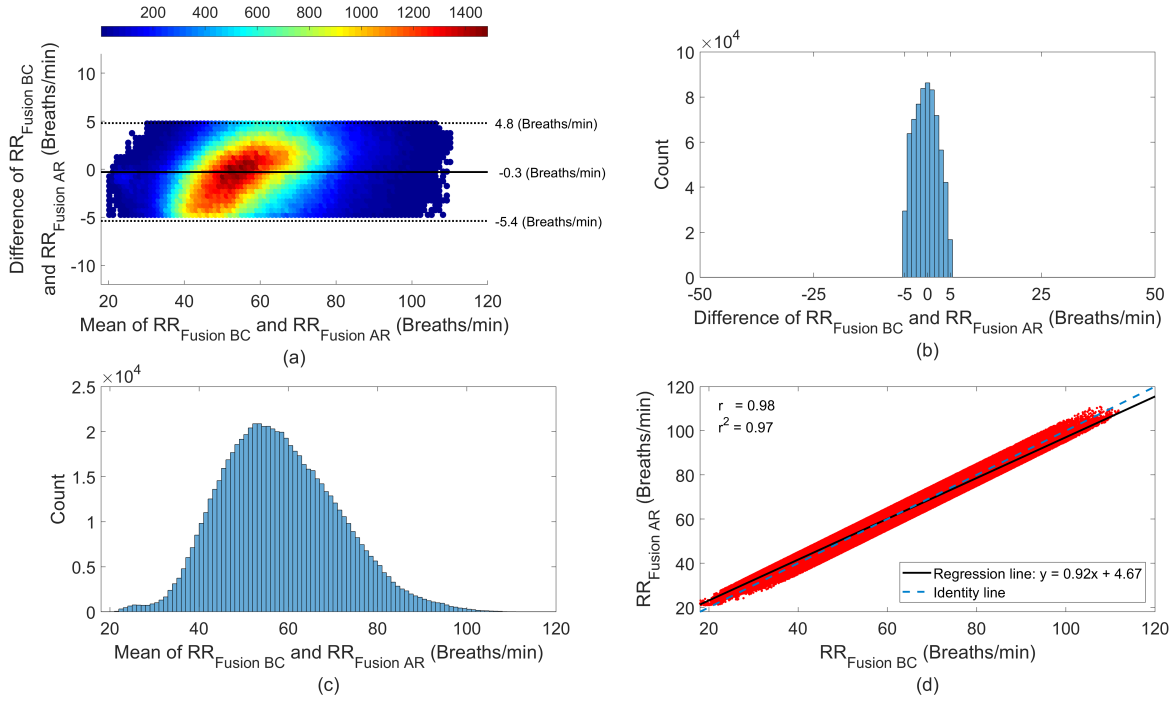


Figure 8.14: Comparison of the $RR_{\text{Fusion BC}}$ and $RR_{\text{Fusion AR}}$) for the entire Neonatal HDU dataset when their difference was less than 5 breaths/min and their signal quality indices are higher than 0.7. (a) The Bland-Altman plot shows a bias of -0.3 breaths/min. (b) The distribution of the differences between the two time series are concentrated around zero. (c) The mean of the two values. (d) The correlation plot shows a positive correlation coefficient r of 0.98.

recording time of 216.5 hours (44.0%) are valid; similarly, in the test set, over 93.7 hours out of the total recording time of 209.7 hours (44.7%) are valid. The valid time for each session varied from 22% to 68%. Note that there was one session for which the reference data were not recorded.

8.10 Discussion

8.10.1 Noise in respiratory signals

Respiratory information can be obtained through direct and indirect measurements. The direct measures include a respiratory airflow sensor and a chest band worn around the upper torso, while the indirect measures include IP and the respiratory signals derived from the ECG and PPG. The patient monitoring systems in a neonatal ward generally report the respiratory rate based on the analysis of the IP signal. The measurement of the IP signal requires the injection of a low-amplitude, high-frequency electrical current which results in respiration-modulated electrical impedance measurements. The changes in electrical impedance are mainly due to the high-resistivity air entering and leaving the lungs and partly due to the movement of the chest wall. Often, the IP signal is considerably noisy and is disturbed by patient motion [208]. Hence, respiratory monitoring using the IP signal is more reliable when the patient is quiet or sleeping rather than when the patient is active and moving [21].

Table 8.3: The amount of valid reference data for respiratory rate estimation by recording session

(a) Training set

Session	Total reference recording time	Valid time	
		hh:mm	%
PN01 Day 1	05:23	03:07	58.0
PN01 Day 2	05:15	03:03	58.3
PN01 Day 3	04:59	02:12	44.3
PN02 Day 1	03:59	02:24	60.3
PN02 Day 2	07:07	03:44	52.3
PN02 Day 3	06:49	03:59	58.5
PN02 Day 4	04:45	02:22	49.9
PN03 Day 1	06:40	03:17	49.2
PN03 Day 2	06:06	02:41	43.9
PN03 Day 3	04:32	02:07	46.9
PN05 Day 1	06:28	02:07	32.8
PN06 Day 1	04:13	01:16	30.1
PN06 Day 2	04:33	02:34	56.4
PN06 Day 3	02:58	00:52	29.3
PN07 Day 1	06:05	02:60	49.3
PN07 Day 2	06:07	02:10	35.5
PN07 Day 3	06:07	03:44	61.2
PN09 Day 1	06:47	03:31	51.9
PN09 Day 2	05:25	02:25	44.5
PN09 Day 3	05:58	02:43	45.6
PN11 Day 1	03:54	01:26	36.7
PN11 Day 2	04:39	01:49	38.9
PN11 Day 3	02:54	01:15	43.0
PN15 Day 1	03:00	00:55	30.5
PN15 Day 2	03:53	02:27	63.1
PN15 Day 3	03:10	01:37	51.2
PN15 Day 4	03:03	01:38	53.4
PN16 Day 1	03:57	01:13	30.8
PN16 Day 2	05:49	01:59	34.0
PN17 Day 1	04:47	02:50	59.3
PN17 Day 2	06:34	03:57	60.1
PN17 Day 3	05:55	02:18	39.0
PN19 Day 1	07:00	02:53	41.3
PN19 Day 2	05:50	02:26	41.9
PN19 Day 3	05:17	02:19	43.8
PN21 Day 1	02:01	00:27	22.2
PN21 Day 2	06:27	01:45	27.2
PN21 Day 3	03:11	01:04	33.5
PN23 Day 1	05:27	01:39	30.4
PN23 Day 2	03:58	01:37	40.6
PN23 Day 3	06:11	01:36	25.9
PN25 Day 1	06:23	01:39	25.8
PN25 Day 2	03:02	01:11	39.0
Total	216:36	95:17	44.0

(b) Test set

Session	Total reference recording time	Valid time	
		hh:mm	%
PN04 Day 1	01:58	00:56	47.8
PN04 Day 2	04:43	02:50	60.2
PN04 Day 3	05:22	02:41	50.1
PN04 Day 4	08:05	03:45	46.4
PN08 Day 1	05:51	02:06	35.9
PN08 Day 2	05:52	02:49	48.0
PN08 Day 3	03:44	01:18	34.6
PN10 Day 1	03:37	01:08	31.4
PN10 Day 2	08:17	03:11	38.4
PN10 Day 3	04:37	02:05	45.2
PN12 Day 1	06:16	01:55	30.7
PN13 Day 1	02:58	01:20	44.8
PN13 Day 2	09:17	03:17	35.4
PN13 Day 3	02:24	00:34	23.3
PN14 Day 1	*Reference data not recorded*		
PN14 Day 2	06:20	02:48	44.1
PN14 Day 3	06:00	03:11	53.0
PN14 Day 4	04:41	02:16	48.4
PN18 Day 1	03:04	01:31	49.6
PN18 Day 2	02:13	01:10	52.2
PN18 Day 3	04:14	01:53	44.6
PN20 Day 1	04:16	02:01	47.3
PN20 Day 2	04:29	01:48	40.3
PN20 Day 3	04:18	01:39	38.3
PN20 Day 4	02:30	01:20	53.6
PN22 Day 1	03:49	01:34	41.0
PN22 Day 2	04:01	01:26	35.5
PN22 Day 3	02:31	00:42	28.0
PN24 Day 1	05:01	02:08	42.4
PN24 Day 2	05:42	01:29	26.0
PN24 Day 3	03:03	01:53	61.5
PN26 Day 1	01:45	01:19	74.9
PN26 Day 2	03:09	01:26	45.4
PN26 Day 3	03:26	01:44	50.3
PN26 Day 4	03:27	01:26	41.8
PN27 Day 1	04:34	02:35	56.6
PN27 Day 2	03:55	01:43	44.1
PN28 Day 1	06:27	02:34	39.9
PN28 Day 2	05:31	03:22	61.2
PN28 Day 3	05:47	02:26	42.0
PN29 Day 1	06:08	03:01	49.1
PN29 Day 2	06:12	02:13	35.7
PN29 Day 3	05:51	03:00	51.4
PN29 Day 4	03:07	01:48	62.0
PN30 Day 1	04:01	02:22	59.1
PN30 Day 2	03:38	01:57	53.6
PN30 Day 3	03:51	02:06	54.6
Total	210:02	93:44	44.7

This study found that a small movement of the infant's head or hands, often in awake infants, caused a serious disruption in the respiratory signal leading to poor signal quality and unreliable respiratory rate estimates. Similar observations were described by Morley *et al.* [211], who reported that the body movement in awake infants often caused too many artefacts masking breathing patterns in the respiratory signal and making the interpretation of respiratory rate difficult. Even during sleeping, the irregular breathing and periodic jerking [225] of pre-term infants often contaminates the respiratory signal and prevented the estimation of a reliable respiratory rate.

8.10.2 Comparison to other relevant studies

In this study, the new computed reference respiratory rate is found to be valid for 44.0% of which is consistent with the values published in the literature. Villarroel *et al.* [54] found that the respiratory rate was of high quality for 32.3% of the time in adult patients monitored during dialysis. Larsen *et al.* Johansson *et al.* [77] found that the recording of the IP signal from newborn infants in an intensive care unit was of high quality for 29% of the total recording time.

Jorge *et al.* [213] reported the comparison between direct and indirect methods of measuring respiratory rate on 165 manual breath counts obtained from the Neonatal HDU dataset. The mean absolute error between the values, as reported by the manufacturer, and the manual breath counts was 8.7 breaths/min. Respiratory rate was reported to be accurate and reliable in the range from 40 to 80 breaths/min. Larger errors were found for respiratory rates higher than 80 breaths/min because of the immaturity of pre-term infants which resulted in recurrent periods of fast and shallow breathing. Their results are consistent with the findings of this thesis which reveal large errors when comparing the re-estimated respiratory rate values against the values from the patient monitor. Their results suggest that the respiratory rate values, as provided by the patient monitor, are not the best source of data for a reference value when comparing with the camera-derived respiratory rate estimates. Figure 8.15 shows the distribution of the reference respiratory rate values ($RR_{\text{Reference}}$) compared to the values reported by the Philips monitor (RR_{Philips}).

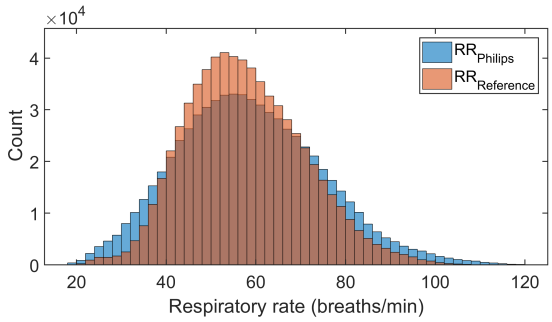


Figure 8.15: Comparison between the respiratory rate reported by the patient monitor (RR_{Philips}) and the reference respiratory rate ($RR_{\text{Reference}}$) for the periods during which the $RR_{\text{Reference}}$ value is valid.

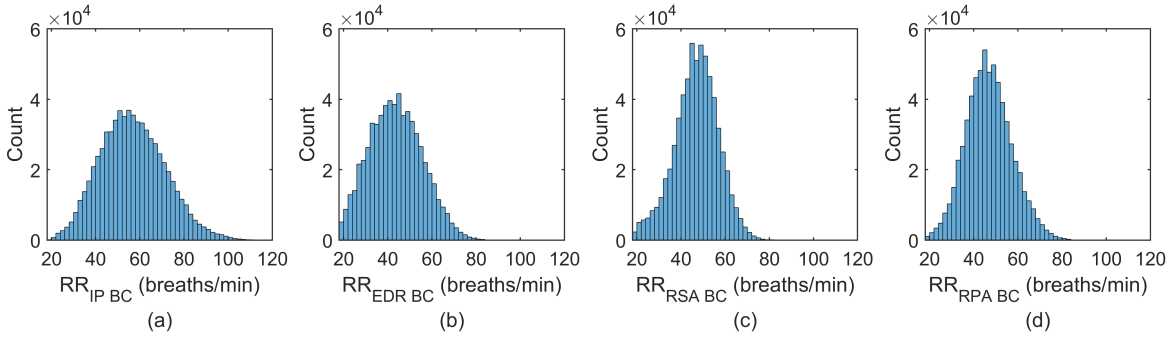


Figure 8.16: Distribution of respiratory rate estimates from the periods during which the reference respiratory rate ($RR_{\text{Reference}}$ estimate) is valid: (a) IP-derived respiratory rate (RR_{IP}), (b) EDR-derived respiratory rate (RR_{EDR}), (c) RSA-derived respiratory rate (RR_{RSA}) and (d) RPA-derived respiratory rate (RR_{RPA}).

Figure 8.16 shows the distribution of respiratory rate estimates from the periods during which $RR_{\text{Reference}}$ values were valid. The distribution of the RR_{IP} is broader on the x-axis than the distribution of other respiratory rate estimates derived from the indirect respiratory signals (RR_{EDR}, RR_{RSA} and RR_{RPA}), similar results were found by Jorge *et al.* [213]. As the respiratory signal extracted from the ECG signal is unevenly sampled once per heart beat, the fundamental frequency of the respiratory rate cannot be higher than half the frequency of the heart rate in the same time window (the Nyquist limit); as a result, aliasing occurs in the respiratory signal when the respiratory rate is higher than half the heart rate. The aliasing effect prevented the estimation of accurate respiratory rate estimates values above 80 breaths/min (see figure 8.16b–d). However, this problem only affected a small portion of the data. The periods during which aliasing is present generally has very low signal quality, therefore these periods are generally ignored in the data fusion process; during those times, the data fusion process is weighted towards the estimates from the IP rather than from the ECG-derived respiratory signals. The data agreement process then ensures that only the values for which the two estimation methods agree are used to compute the gold-standard reference respiratory rate. Figure 8.15 values shows the distribution of the $RR_{\text{Reference}}$ values which span over the entire range of the pre-term infant's respiratory rate values.

The measurement of respiratory rate in pre-term infants using contact sensors presents considerable challenges, mainly due to pathological instabilities which resulted in an inherently noisy respiratory signal. Hence, the true values of respiratory rate were not known. The pipeline presented in this chapter provided a practical and logical approach for computing the reference respiratory rate for comparison with the camera-derived respiratory rate estimates in the next chapter.

8.11 Conclusion

Respiratory signals are typically noisy and non-stationary, with interference from both biological and environmental factors. The population of the clinical study ranged from early to late pre-term

infants whose respiratory system was at different developmental stages; hence, they may have different breathing patterns. Some infants had shallow breathing interspersed with periods of high-amplitude breathing, and some infants had regular apnoeas. As a result of the active nature of pre-term infants, the respiratory signals recorded during wakefulness were often inherently noisy because of the movement of the upper body. The movement sometimes caused by the electrodes on the chest to become detached leading to disruption during recording. Sudden movements often occurred due to the infant's discomfort or pain. Even during sleep, the infant's small movements, e.g. moving the head or hands, also caused artefacts in the respiratory signal.

This chapter computed the reference respiratory rate by taking into account multiple respiratory signals and different estimation methods. The estimation errors and the proportion of time for which there is good-quality data for the reference respiratory rate are consistent with the values reported in the literature [21, 54, 77]. The reference respiratory rate will now be used for comparison with the camera-derived estimates of respiratory rate in the next chapter.

Chapter 9
Respiratory rate estimation from the video camera

9.1 Introduction

Respiratory rate is an important indicator for the health status of pre-term infants. Abnormal respiratory rate is often associated with potentially serious clinical events such as respiratory tract infections and pneumonia [211]. The monitoring of respiratory rate by conventional patient monitors is difficult when the infant is awake and active because the infant's movement often causes gross motion artefacts which mask breathing patterns [210]. This chapter explores the feasibility of using a video camera for the monitoring of respiratory rate in pre-term infants. The respiratory rate estimates are compared with reference respiratory rate estimates, as derived with methods as described in the previous chapter.

In an intensive care unit (NICU) environment, the monitoring of respiratory rate is generally performed by impedance pneumography (IP). Respiratory signals can also be derived from electrocardiography (ECG) or photoplethysmography (PPG) whose amplitude and frequency oscillations are synchronised with the respiration cycle [213, 226–228].

In the video camera monitoring domain, a photoplethysmography imaging (PPGi) signal is generally modulated by both cardiac-synchronous and respiratory-synchronous motion [107]. The signal extracted from small regions of interest in the skin or chest regions has been shown to contain strong respiratory information [14]. Recently, Jorge *et al.* [213] have shown that respiratory rate in pre-term infants can be obtained from the analysis of skin regions provided by a convolutional neural network (CNN) model. The resulting respiratory rate estimates were reported to have the same level of accuracy as the respiratory rate provided by the patient monitor [213].

This chapter presents a robust automated framework for estimating respiratory rate from multiple respiratory signals derived from the analysis of the video data. During each respiratory cycle, the infant's chest and abdomen expands and contracts with the breathing. This phenomenon

causes movement of the body and changes the reflection of the light recorded by a video camera from areas containing the skin. The skin segmentation algorithm presented in chapter 4 can be used as the first step in the extraction of various respiratory signals. In this work, the respiratory signals are extracted from the colour changes of the skin, the shape properties of the skin area and the morphology analysis of the skin area. This framework can help compensate for the problems of motion artefacts and missing data, which deviate when the respiratory rate is estimated from just one source, and identify suitable time periods for which respiratory rate can be reliably estimated.

This chapter begins with an overview of the respiratory rate estimation process in section 9.2. Section 9.3 presents the extraction of multiple respiratory signals from the video recordings. Section 9.5 details the process of peak and onset detection in the respiratory signals. Section 9.6 describes the assessment of signal quality for each respiratory signal. Section 9.7 explains the process by which respiratory rate estimates are computed from the analysis of multiple respiratory signals. Section 9.8 presents the comparison between the respiratory rate estimates and the reference respiratory rate for the entire Neonatal HDU dataset. Section 9.9 analyses the results in depth and discusses the advantages and limitations of continuous camera-based respiratory rate monitoring in the clinic. Section 9.10 concludes the chapter and presents possible directions for future research.

9.2 Overview of the estimation process

This section outlines the process for estimating respiratory rate from the video camera data (see figure 9.1). From the video camera data, the convolutional neural network (CNN) algorithm presented in chapter 4 is first applied to detect the patient's presence and segment the patient's skin regions. The intervention detection algorithm described in chapter 5 is then applied to detect the periods during which the infant is not undergoing a clinical intervention. Next, twelve respiratory signals are extracted from the analysis of the video camera data and the segmented skin regions. The respiratory signals are divided into three groups. The first group contains the PPGi signals for each colour channel extracted from the entire skin regions. The second group comprises the respiratory signals extracted from the properties of the entire skin regions. The last group consists of the respiratory signals extracted from the geometrical properties of the ellipse fitted to the skin regions.

After respiratory signal extraction, all the respiratory signals are processed through the same respiratory rate estimation pipeline: signal filtering, peak and onset detection, signal quality assessment and respiratory rate estimation. Unlike the estimation pipeline used in other chapters, the respiratory rate estimation pipeline utilises two peak and onset detection algorithms, each of which has a different sensitivity to different types of noise [189]. Both the signal quality assessment and the respiratory rate estimation are based on the agreement between the outputs of these two

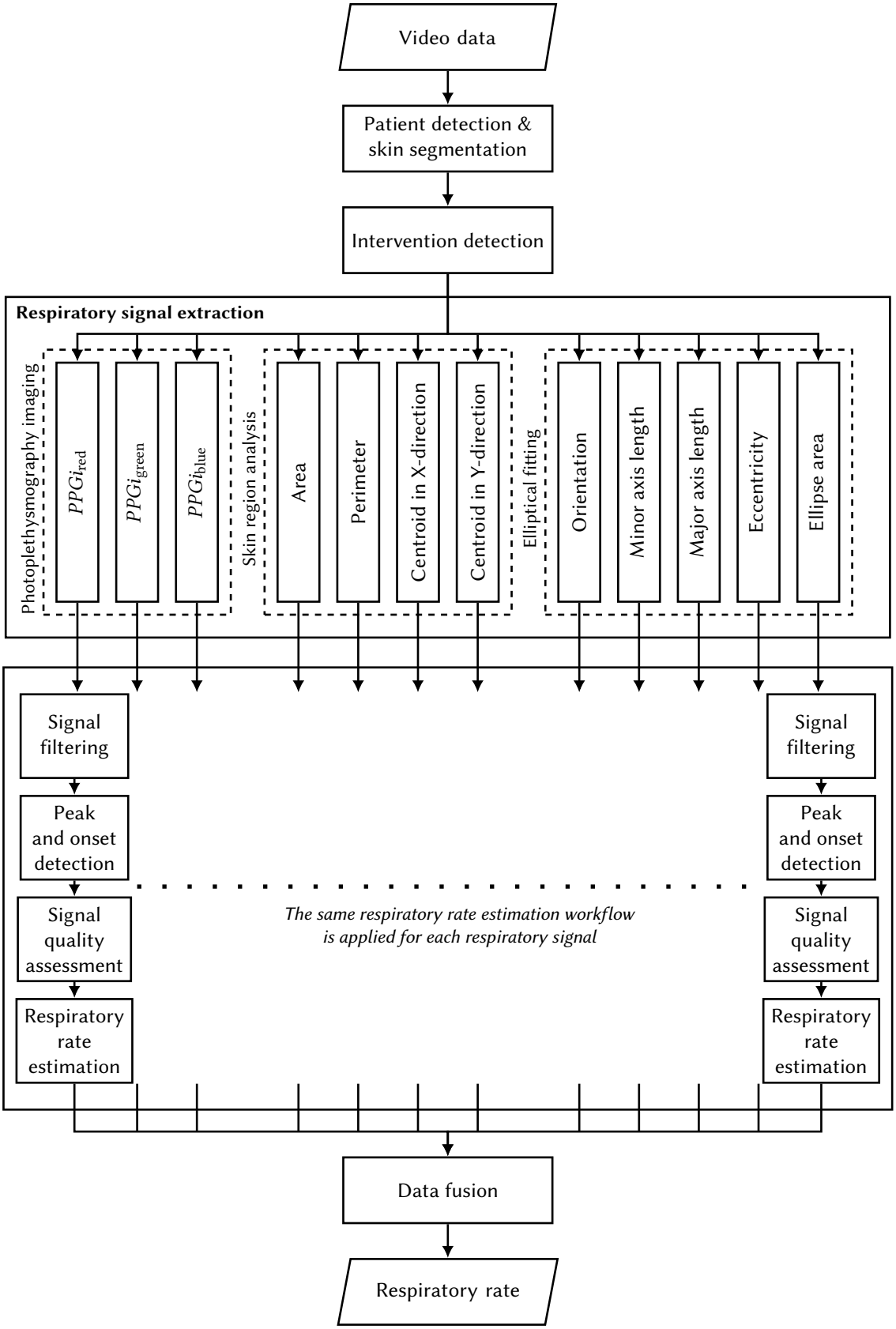


Figure 9.1: Overview of the respiratory rate estimation framework. Multiple respiratory signals are extracted from the analysis of video camera data. Respiratory rate is obtained by combining respiratory rate estimates from all the respiratory signals.

algorithms. This technique can effectively prevent respiratory rate estimates from being computed during periods when the respiratory signal is noisy.

Once the respiratory rate of each signal has been estimated, the data fusion technique based on multiple Kalman filters is applied to combine all respiratory rate estimates and produce a consensus estimation of respiratory rate. The data fusion process can compensate for the problems of noise and missing data, which occur when respiratory rate is estimated from a single source. Finally, the resulting respiratory rate estimates are compared with the reference respiratory rate estimates computed as described in chapter 8.

9.3 Respiratory signal extraction

Respiratory rate can be manually measured by observing the rise and fall of the chest which indicate inhalation and exhalation [229]. Apart from the chest, the movement due to respiration in infants and young children is also present in the abdomen [230]. From observations of the Neonatal HDU dataset, the movement caused by respiration is likely to affect the entire upper body and sometimes, for early pre-term infants, the entire body.

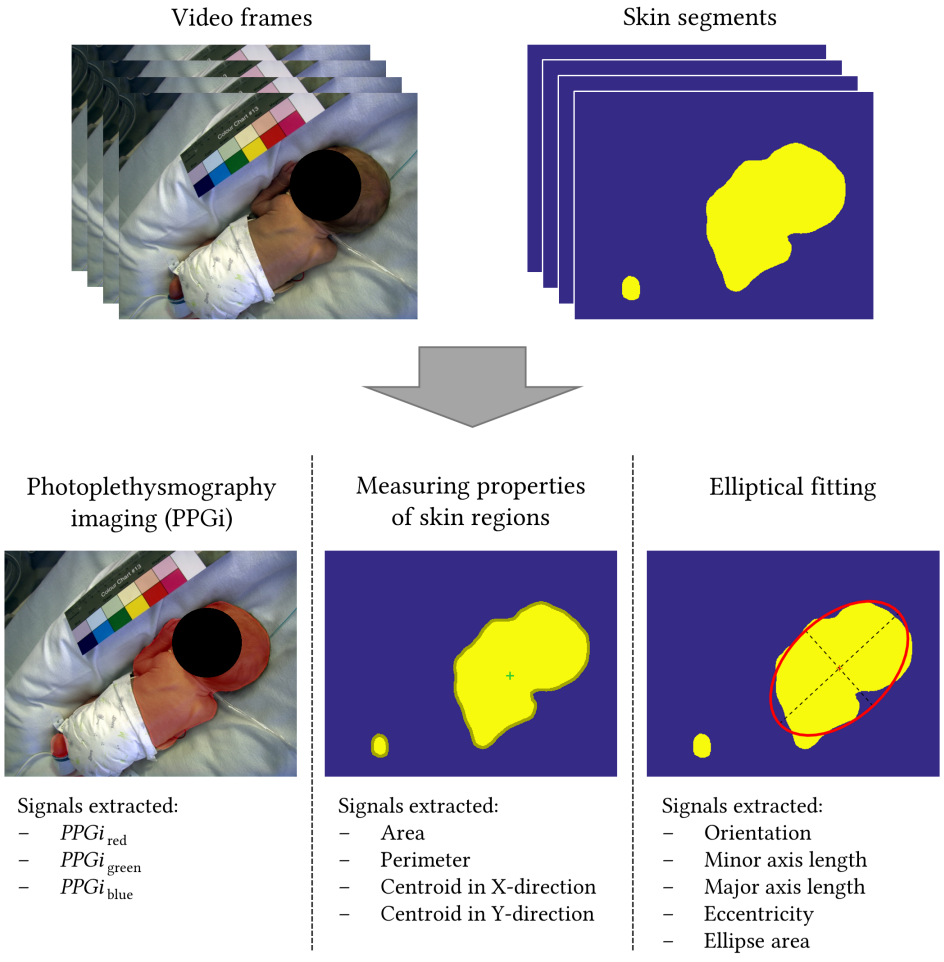


Figure 9.2: Outline of the respiratory signal extraction process.

The algorithms presented in chapter 4 provides accurate skin segmentation which, over time, captures the periodic movement during respiration. The size and morphology of the segmented skin regions oscillate as a result of the expansion and contraction of the upper body. Numerous studies have attempted to use object detection and feature tracking algorithms to estimate respiratory rate by tracking the movement of the chest [231–233]. Several studies have used a stereo video camera or depth sensor to obtain the movement of the chest and compute respiratory rate [233, 234].

In this work, the segmented skin regions, which can be seen as high-level feature descriptors provided by the CNN, are used to derive various respiratory signals. Since these respiratory signals are often corrupted by noise and artefacts, respiratory rate derived from just one respiratory signal would contain numerous periods of missing data. The combination of respiratory rate estimates from different respiratory signals will compensate for this and provide more reliable respiratory rate estimates.

The respiratory signals derived in this chapter can be grouped into three categories, according to the techniques used to extract the signals: PPGi, skin region properties and elliptical fitting. Figure 9.2 shows the outline of the respiratory rate estraction process.

9.3.1 Photoplethysmography imaging (PPGi)

The movement of the upper body during respiration causes periodic changes in the angle of reflection of the skin surface, thus modify the reflection of the incident light off the skin. This phenomenon causes changes in the skin colour during the expansion and contraction of the upper body. The PPGi signal, computed by averaging the colour of the skin regions over time (see section 7.3.1), contains both cardiac and respiratory pulsatile information. The relative size of the respiration-correlated pulsatile component in the PPGi signal depends on the the gestational age (reflecting the size and developmental stage of the infant) and the breathing pattern (shallow or deep breathing).

In this work, three PPGi signals are extracted from the skin regions: $PPGi_{\text{red}}$ derived from the red channel, $PPGi_{\text{green}}$ derived from the green channel and $PPGi_{\text{blue}}$ derived from the blue channel. Figure 9.3 shows the PPGi signals extracted from each colour channel in a typical video recording. Respiratory information is present in all three colour channels. Figure 9.4 shows PPGi signals extracted from three colour channels during different body postures. Respiratory pulsation is usually prominent when the chest and abdomen is visible to the camera. Respiratory pulsation may not be prominent when part of the chest and abdomen are covered by a blanket.

9.3.2 Skin region properties

The contraction and relaxation of the muscles during respiration causes the volume of the chest cavity to increase or decrease, thus resulting in motion of the chest and abdomen [235]. Respiratory signals

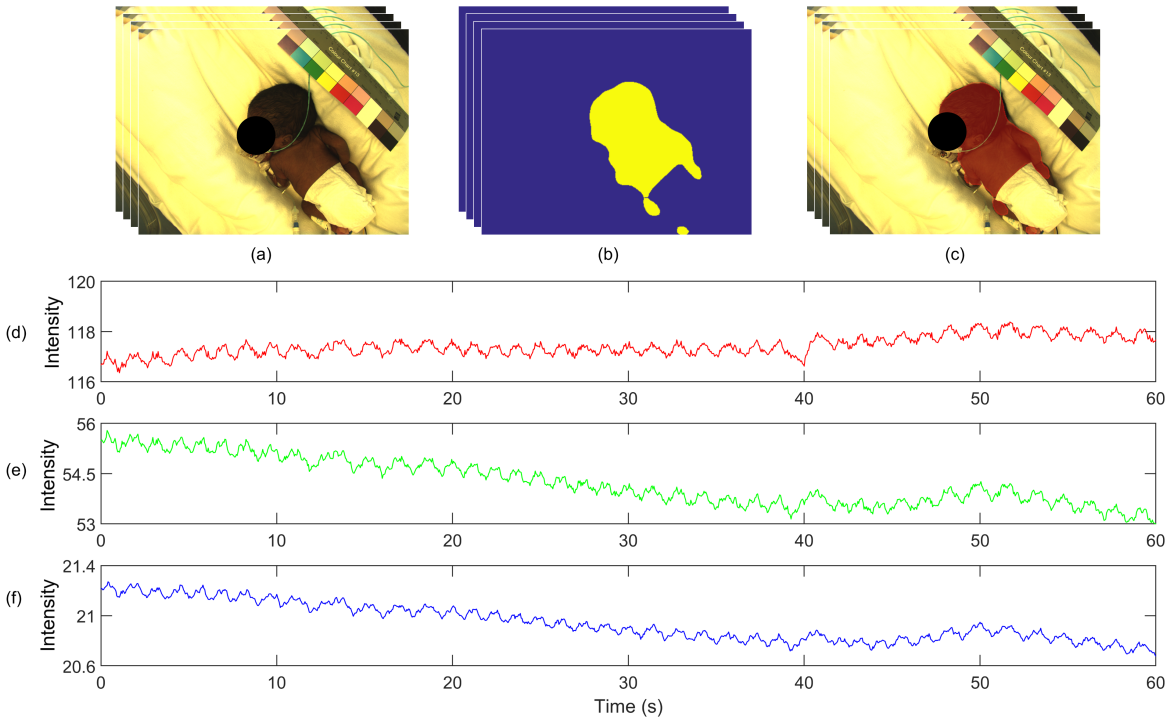


Figure 9.3: PPGi signals extracted from three colour channels contain strong respiratory oscillation. With an appropriate set of filters, a respiratory signal can be extracted from the PPGi signals. (a) Video frames. (b) Skin segmentation results. (c) Video frames with the skin regions segmented in red. (c-f) PPGi signals extracted from the red, green and blue channels respectively. 42 breath pulses can be counted from the green channel, corresponding to a respiratory rate of 42 breaths/min.

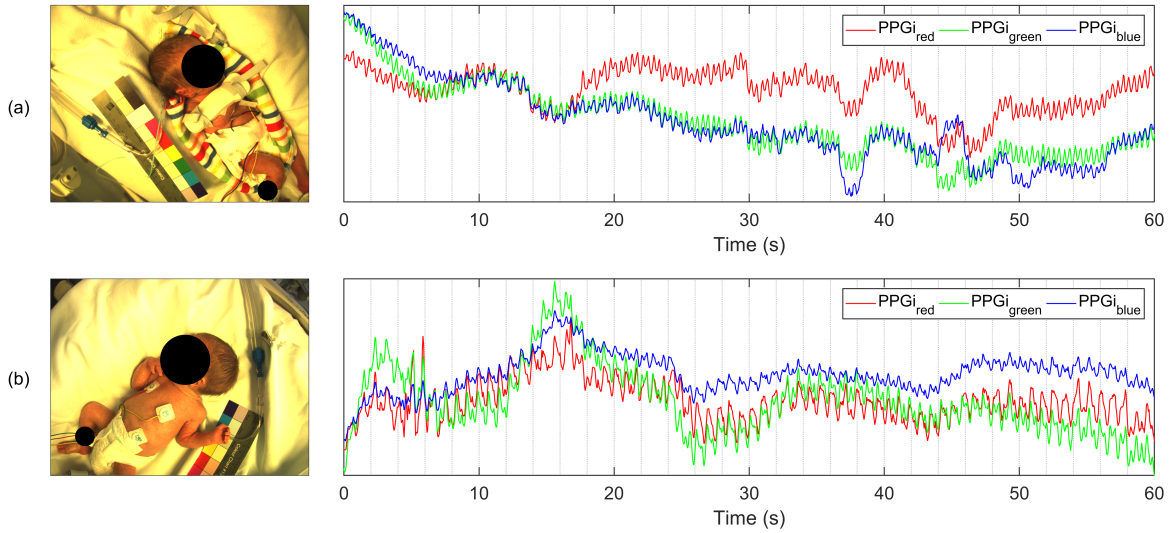
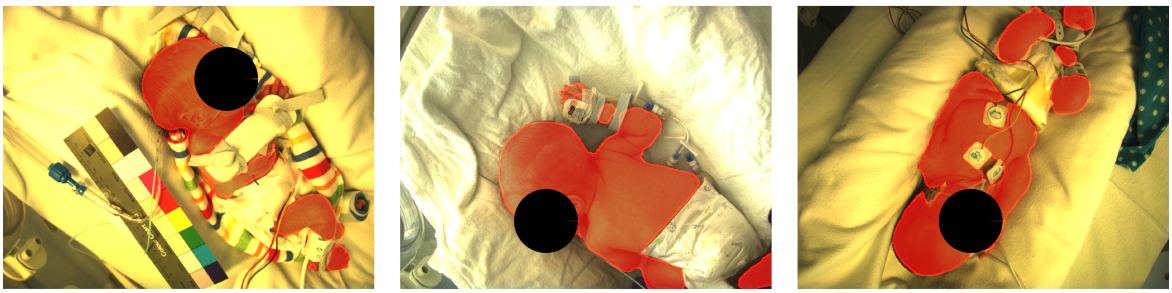


Figure 9.4: PPGi signals extracted from three colour channels during different body postures. (a) PPGi signal extracted from a period during which the infant is facing the video camera and part of the chest and abdomen are covered by the infant's hand splint. Respiratory oscillations are not found to be prominent during this period. (b) PPGi signal extracted from the period during which the infant is facing the video camera and the face and abdomen are fully visible. Respiratory pulsation is found to be prominent during this period.



(a) (b) (c)

Figure 9.5: Examples of segmented skin labels as provided by the skin segmentation algorithm described in chapter 4. (a) The skin label may include one large skin region and (b–c) several smaller skin regions, depending on the subject's body posture and clothing.

can be acquired by tracking the shape properties of the subject's skin regions, which periodically changes according to the breathing rhythm.

Computing the skin properties

Since the infant is nursed naked in the incubator while recording, changes in the shape properties of the skin are synchronous with respiration. The skin label provided by the patient detection and skin segmentation algorithm may contain one large continuous skin region or make up one large region with a series of smaller skin regions (see figure 9.5). The following shape properties are computed from the entire skin label: area, perimeter and the x and y coordinates of the centroid.

Let $I(x, y)$ be a binary image containing segmented skin regions where

$$I(x,y) = \begin{cases} 1 & \text{if the pixel at } (x,y) \text{ is skin} \\ 0 & \text{if the pixel at } (x,y) \text{ is background} \end{cases} \tag{9.1}$$

The area of the skin regions is the number of skin pixels that can be calculated by the summation of all elements in the binary skin image $I(x, y)$:

$$\text{Area} = \sum_X \sum_Y I(x,y). \tag{9.2}$$

The centroid of the skin regions with respect to the x and y axes can be then calculated by finding the average of all points within the skin regions by as follows:

$$\text{Centroid}_X = \frac{\sum_X \sum_Y xI(x,y)}{\text{Area}} \tag{9.3}$$

$$\text{Centroid}_Y = \frac{\sum_X \sum_Y yI(x,y)}{\text{Area}}. \tag{9.4}$$

Finally, the perimeter of the skin regions is calculated as follows. The boundaries of each skin region in $I(x, y)$ are first extracted using the Moore-Neighbour tracing algorithm [236] to trace the contour between the skin and background. Figure 9.6 shows the skin boundaries extracted



(a) (b) (c)

Figure 9.6: Region boundary extraction. (a) Video frame. (b) Segmented skin label. (c) Boundary of the segmented skin label calculated using the Moore-Neighbour tracing algorithm [236]. The perimeter of the binary skin label is calculated by simply counting the number of boundary pixels.

from a binary skin label. Let $B(x, y)$ be the binary image containing the boundaries of the skin regions computed using the Moore-Neighbour approach. The perimeter of the skin regions is then the number of all boundary pixels:

$$\text{Perimeter} = \sum_{X} \sum_{Y} B(x, y). \tag{9.5}$$

Extraction of the respiratory signals

The skin region properties are computed for all video frames during the periods when the subject is present and clinical intervention does not take place (as determined by the algorithms presented in chapter 4-5). Thus, four respiratory signals are extracted: area, perimeter and centroids in x and y axes. Figure 9.7 shows sample respiratory signals for a 60-second period.

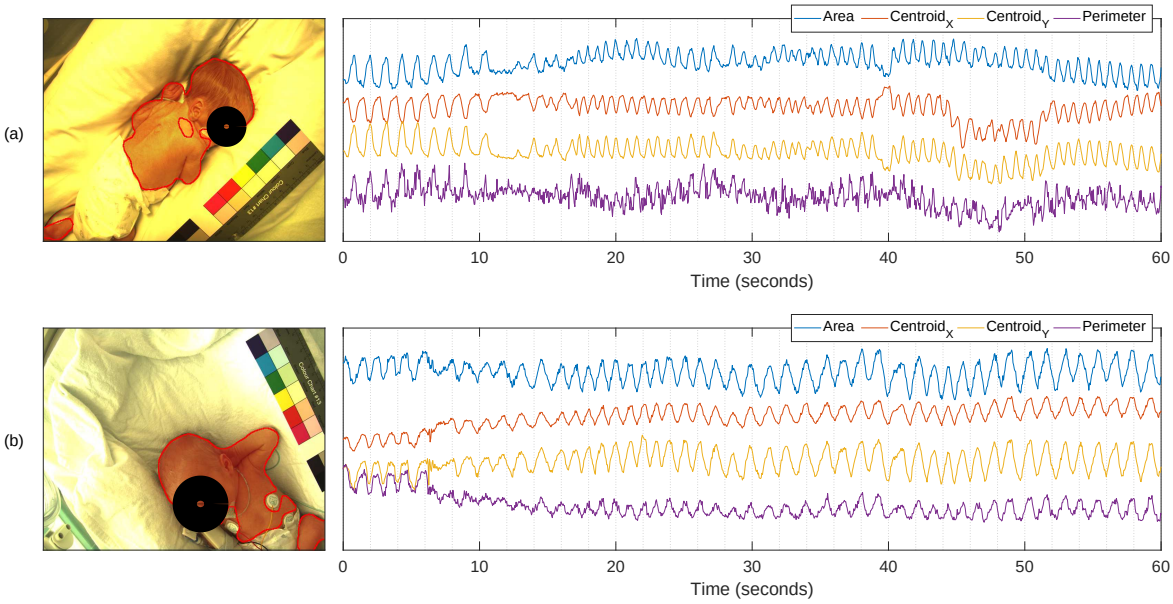


Figure 9.7: Respiratory signals extracted from the shape properties of skin regions. The respiratory signals are extracted by computing the area, centroid (x and y directions) and perimeter over the entire skin. (a) Respiratory signals derived while the infant is in a supine position. (a) Respiratory signals derived while the infant is in a prone position.

9.3.3 Elliptical fitting

Although changes in the skin region properties over time can reflect the motion of the subject which in turn can be used to estimate respiratory rate, a small degree of non-respiratory motion can easily introduce motion artefacts and corrupt the respiratory signal. Similarly, the subject's posture and clothing sometimes split the skin into smaller regions (see figure 9.5), so the respiratory signal extracted from the properties of multiple skin regions often contain abrupt changes from motion artefacts as well.

The result of skin segmentation often has a elliptical shape. Changes in the shape parameters of the ellipse (e.g. orientation and axis lengths) over time as a result of breathing can be used as a respiratory signal. In the case that the skin segmentation result consists of several separate regions, the upper body region which has a prominent respiratory motion corresponds to the largest continuous region (see figure 9.5); therefore, the respiratory signal can be extracted from the ellipse fitted to the largest continuous region. One major advantage of using the elliptical parameters is that the effect of small non-respiratory movements can be potentially ignored.

In this section, five respiratory signals are extracted from the shape parameters of the ellipse fitted to the whole skin region (if the segmentation result contains only one skin region) or to the largest continuous skin region (if the segmentation result is composed of several separate skin regions). The elliptical parameters of interest include orientation, minor axis length, major axis length, eccentricity and ellipse area. In some cases, one respiratory signal may be better than the others depending on the infant's orientation and the maturity of the infant's breathing mechanisms.

Representation of an ellipse

An ellipse is a closed shape curve (see figure 9.8). The sum of the distances from all points on the curve to the two focal points (F_1 and F_2) is a constant [237]. The major axis is the straight

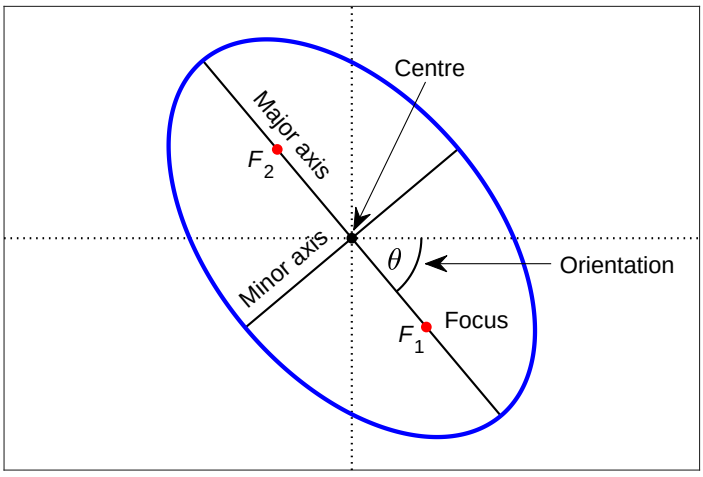


Figure 9.8: Representation of an ellipse with shape parameters: major axis, minor axis, centre point, focal points (F_1 and F_2) and orientation (θ).

line that passes through the two focal points to the endpoints on the edge of the ellipse. The midpoint of the major axis is the centre of the ellipse. The minor axis is the straight line that is perpendicular to the major axis and passes through the centre to the endpoints on the ellipse. The orientation of the ellipse is the angle between the major axis and the x-axis. The shape of an ellipse that is elongated can be expressed by its eccentricity.

Let us suppose that the semi-major axis (one half of the major axis) is a and the semi-minor axis is b. The standard equation of the ellipse in Cartesian coordinates is:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \tag{9.6}$$

This represents an ellipse centred on the origin and aligned with the x and y axes. Suppose that the ellipse is now rotated by an angle θ. The relationship between the coordinates (x, y) and the coordinates (x', y') after the rotation is given by:

$$x' = x\cos(\theta) - y\sin(\theta) \qquad \text{and} \qquad y' = x\sin(\theta) + y\cos(\theta)\,. \tag{9.7}$$

Then, the equation of an ellipse which has been rotated by an angle θ is:

$$\frac{\big(x\cos(\theta) - y\sin(\theta)\big)^2}{a^2} + \frac{\big(x\sin(\theta) + y\cos(\theta)\big)^2}{b^2} = 1. \tag{9.8}$$

In general, an ellipse may not be orientated in parallel with the coordinate axes or centred at the origin. The equation of an ellipse with a rotation angle θ and its centre at $(\bar{x}, \bar{y})$ is:

$$\frac{\big((x-\bar{x})\cos(\theta) - (y-\bar{y})\sin(\theta)\big)^2}{a^2} + \frac{\big((x-\bar{x})\sin(\theta) + (y-\bar{y})\cos(\theta)\big)^2}{b^2} = 1. \tag{9.9}$$

Moment-based fitting approach

Ellipse fitting is a fundamental task in computer vision. Several methods have been proposed for fitting an ellipse into an arbitrary region of interest such as the moment-based method, the direct-fitting method and the least-squares method [238, 239]. The ideal fitting method should be fast, efficient and invariant to rotation and translation. Hence, the moment-based method is chosen.

This work begins by finding the centre of mass and the second central moments of the skin region. The orientation, axes lengths, eccentricity and area of the ellipse are then computed from the second central moments [240].

Let R be the largest continuous skin region with uniform density in a binary skin label I. The centre of mass or centroid $(\bar{x}, \bar{y})$ of the region R over the area A with respect to the x or y axes is defined, similarly to equation 9.2, as:

$$\bar{x} = \frac{1}{A}\iint_R x\,\mathrm{d}A \qquad \text{and} \qquad \bar{y} = \frac{1}{A}\iint_R y\,\mathrm{d}A. \tag{9.10}$$

The second central moment of area provides an insight in how the points are distributed with respect to a particular axis. The second central moments of the region R with regard to only the x or y axes can be computed respectively as:

$$\mu_{xx} = \iint_R (x - \bar{x})^2 \, dA \qquad \text{and} \qquad \mu_{yy} = \iint_R (y - \bar{y})^2 \, dA. \tag{9.11}$$

The second central moments of the region R with respect to both the x and y axes can be computed as:

$$\mu_{xy} = \iint_R (x - \bar{x})(y - \bar{y}) \, dA. \tag{9.12}$$

Assuming that an ellipse has the same second central moments as the skin region R, the shape parameters of the ellipse can be computed via the second central moments as follows [240]:

$$\text{Major axis length} = 2\sqrt{2} \cdot \sqrt{\mu_{xx} + \mu_{yy} + \sqrt{(\mu_{xx} - \mu_{yy})^2 + 4\mu_{xy}^2}} \tag{9.13}$$

$$\text{Minor axis length} = 2\sqrt{2} \cdot \sqrt{\mu_{xx} + \mu_{yy} - \sqrt{(\mu_{xx} - \mu_{yy})^2 + 4\mu_{xy}^2}} \tag{9.14}$$

$$\text{Orientation, } \theta = \begin{cases} \tan^{-1}\left(\dfrac{\mu_{xx} - \mu_{yy} + \sqrt{(\mu_{xx} - \mu_{yy})^2 + 4\mu_{xy}^2}}{-2\mu_{xy}}\right) & \text{if } \mu_{xx} > \mu_{yy} \\ \tan^{-1}\left(\dfrac{-2\mu_{xy}}{\mu_{yy} - \mu_{xx} + \sqrt{(\mu_{yy} - \mu_{xx})^2 + 4\mu_{xy}^2}}\right) & \text{if } \mu_{xx} \leq \mu_{yy} \end{cases} \tag{9.15}$$

$$\text{Eccentricity} = \sqrt{1 - \left(\frac{\text{Minor axis length}}{\text{Major axis length}}\right)^2} \tag{9.16}$$

$$\text{Elliptical area} = \pi \cdot \left(\frac{\text{Major axis length}}{2}\right) \cdot \left(\frac{\text{Minor axis length}}{2}\right) \tag{9.17}$$

Figure 9.9 shows examples of an ellipse fitted to the largest continuous skin regions for multiple sample images. Respiratory signals can be extracted by observing the changes of the ellipse properties over time.



Figure 9.9: Examples of an ellipse fitted to the largest continuous skin region. The elliptical fitting approach ignores smaller skin regions and thus can reduce noise occurring due to the non-respiratory motion of the lower body.

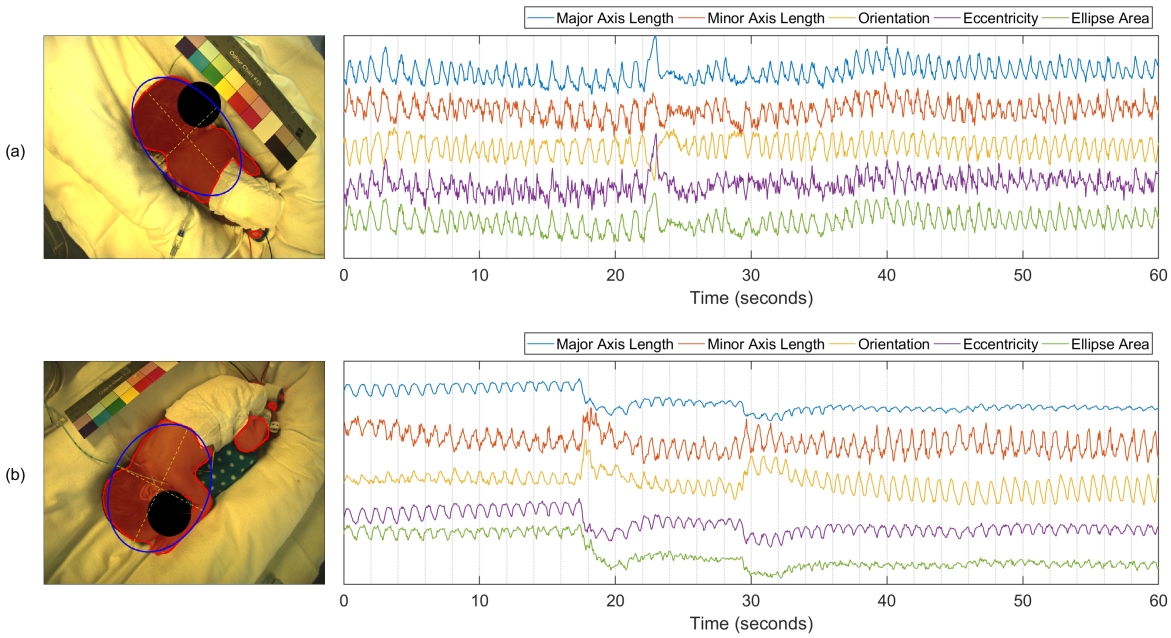


Figure 9.10: Respiratory signals extracted from the shape parameters of an ellipse fitted to the largest skin region during (a) quiet and (b) stable periods. The respiratory signals are derived from the following parameters: major axis length, minor axis length, orientation, eccentricity and ellipse area.

Extraction of the respiratory signals

Now that we have defined how an ellipse is fitted to the skin segment and how the properties of the ellipse are computed, it is necessary to explain how respiratory signals are extracted. For each video frame, an ellipse is fitted to the largest continuous skin region by matching the second central moments of the skin region to the ellipse. With the resulting shape parameters, five respiratory signals are then extracted: major axis length, minor axis length, orientation, eccentricity and elliptical area. Figure 9.10 shows examples of respiratory signals derived using the ellipse-fitting approach.

9.3.4 Summary

Figure 9.11 shows all twelve respiratory signals extracted using all the three techniques presented in this chapter. All respiratory signals include respiratory-synchronous motion information, although respiration in one particular signal may be prominent than in the others. As expected, the respiratory signals extracted from the properties of skin regions and the ellipse-fitting approach do not experience a step change due to the shadow cast on the skin, as happened with the PPG signals (see figure 9.11a). Motion artefacts are present in all respiratory signals during the periods of head and body movements (see figure 9.11b–c).

9.4 Signal filtering

The respiratory signals extracted from the video data are inherently noisy and are often contaminated by baseline drifts and high-frequency noise. To remove these artefacts, each respiratory signal

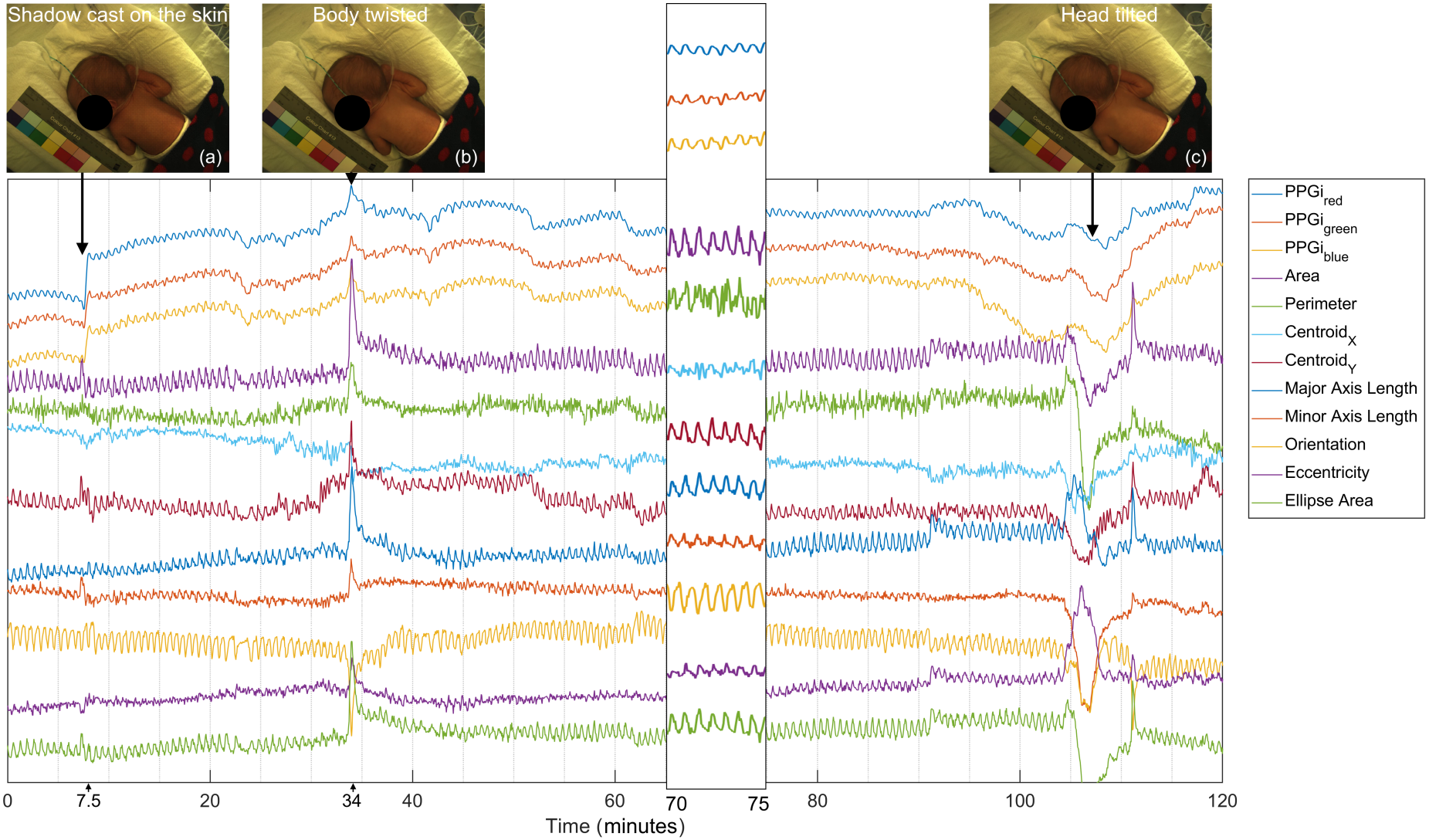


Figure 9.11: Twelve respiratory signals extracted from the analysis of video data and segmented skin regions in a typical recording session. (a) A step change is found in the PPGi signals as clinical staff walk past the infant during minute 7. (b) The infant's body twists, during minute 34, causing an abrupt change in the respiratory signals. (c) The movement of the head and shoulder causes a large motion artefact lasting for several seconds during minute 107.

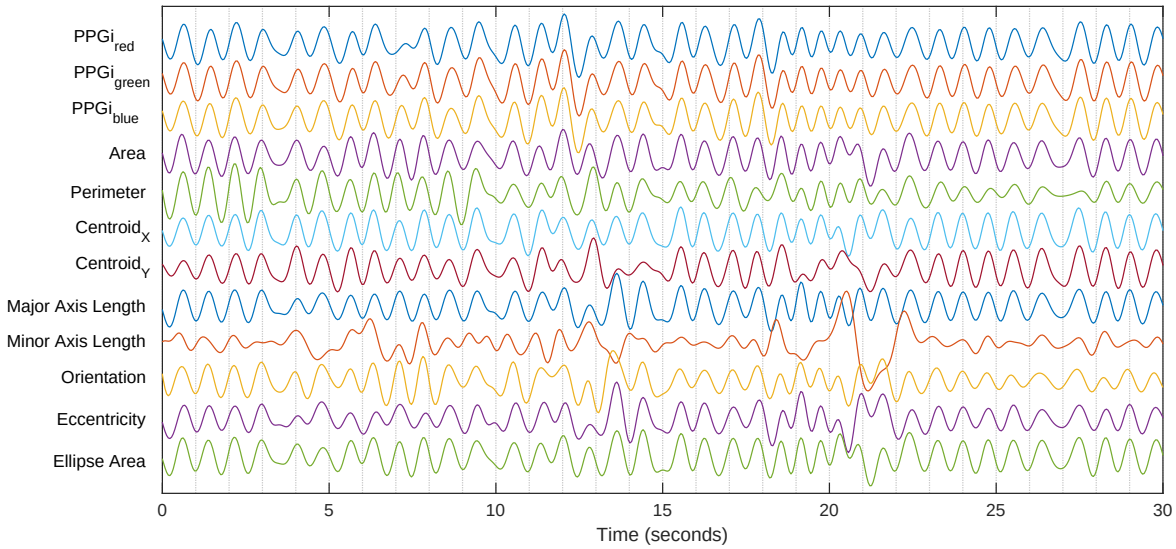


Figure 9.12: Signal filtering of all twelve respiratory waveforms extracted from video camera data

is detrended and filtered using a series of a high-pass finite impulse response (FIR) filter with a cut-off frequency of 0.3 Hz and an order of 100 and a low-pass FIR filter with a cut-off frequency of 2.0 Hz and an order of 80. These filters encompass respiratory rates in the range of 18 to 120 breaths/min, the same range as in chapter 8. Figure 9.12 shows the 30-second waveforms of all respiratory signals after signal filtering. The strength of respiratory modulation in the respiratory signals extracted from centroid positions (Centroid_X and Centroid_Y) and axis lengths (Major axis length and Minor axis length) depend on the location and posture of the subject. For example, when the subject's body posture is aligned with the x-axis, the infant's upper body tends to move more in the direction of the x-axis than that of the y-axis; hence the strength of respiratory modulation in the Centroid_X signal is higher than that in the Centroid_Y signal.

9.5 Peak and onset detection

Once the respiratory signals have been extracted, the process of peak and onset detection is applied to all respiratory signals. Inspired by Li *et al.* [189], two algorithms for peak and onset detection are applied to each respiratory signal: the first one is based on the mean average curve (MAC) (described in section 8.6) [221, 222] and the other one is based on the boxed slope sum function (BSSF) [198] (described in section 7.3.3). Both detection algorithms have high accuracy but different sensitivities to different types of noise [189]. The subsequent assessment of signal quality and estimation of respiratory late are based on the consensus of the results from both algorithms.

The process of peak and onset detection works as follows. Each respiratory signal is processed with both MAC and BSSF algorithms. Since the MAC-based algorithm (see section 8.6) is designed for respiratory data, it is applied to the respiratory signal without any modifications. Originally, the BSSF algorithm (see section 7.3.3) is designed for detecting peaks and troughs in a cardiac

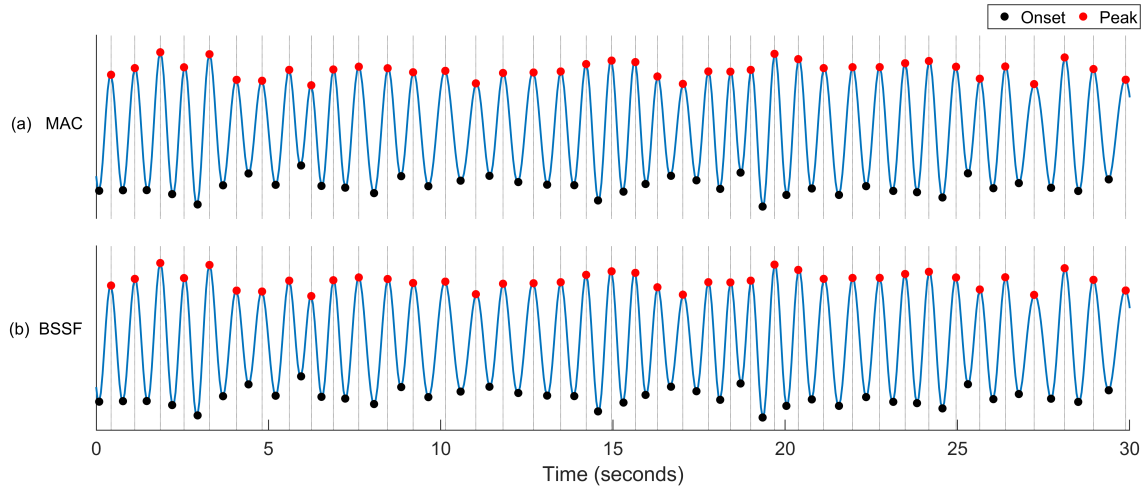


Figure 9.13: The detection of peaks and onsets for a 30-second clean respiratory waveform by (a) the MAC algorithm and (b) the BSSF algorithm. The same waveform is shown twice with markers indicating each detected peak and trough.

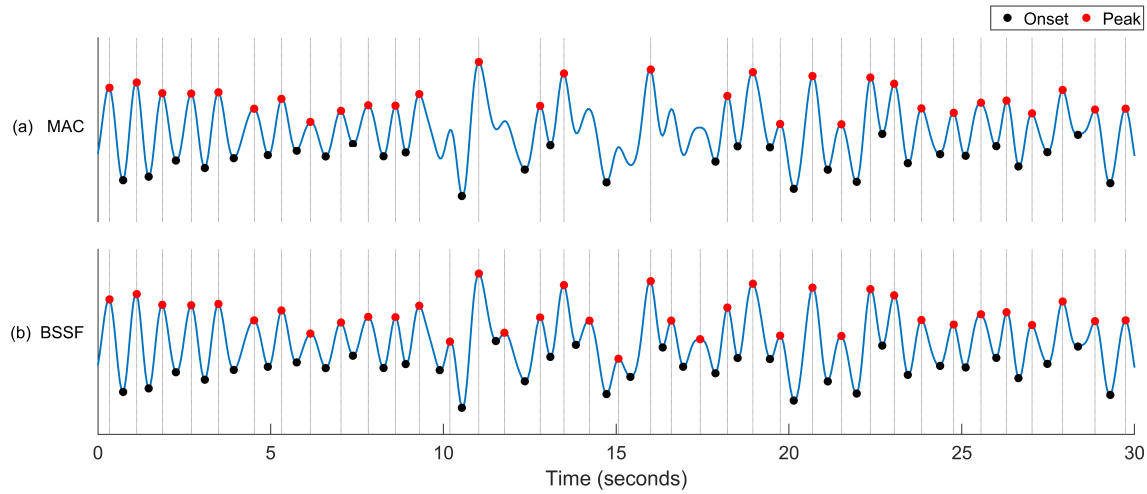


Figure 9.14: The detection of peaks and onsets for a 30-second noisy respiratory waveform by (a) the MAC algorithm and (b) the BSSF algorithm.

signal. Several parameters have been changed to make the algorithm work with a respiratory signal. Instead of using the typical duration of the upslope of the cardiac pulse for computing the BSSF signal, the upslope duration of the respiratory pulse is changed to 300 ms. To prevent mulitple detections, a 500-ms refractory period is applied such that any successive peaks with the interval from the previous one less than 500 ms are ignored.

Figure 9.13 shows the comparison of peaks and onsets detected by the two different algorithms on a clean respiratory waveform. Figure 9.14 shows the resulting peaks and onsets detected during a noisy period. Erroneous detections are made in the middle section of the waveform (between 10 and 20 seconds). Despite the fact that both algorithms have been reported to have high accuracy [198, 222], only the MAC algorithm is specifically designed and validated for respiratory waveforms. Therefore, the MAC algorithm is chosen to segment each breath in the respiratory signal. The BSSF algorithm is subsequently used later as part of the signal quality assessment and the respiratory rate estimation.

9.6 Signal quality assessment

The signal quality assessment employs a series of different measures ranging from subject motion to signal morphology analysis to assign a signal quality index to each breath in the respiratory signal.

The shape and morphology of the non-contact respiratory signals are generally different from that of PPGi signals. The amplitude of each breath generally depends on how much the skin colour changes or how much the subject moves during breathing. During a quiet and stable period, the amplitude of each peak generally reflects the depth of each breath: shallow breathing results in a low-amplitude waveform while deep breathing results in a high-amplitude waveform. Pre-term infants generally have different breathing patterns than those from term infants and adults, as a result of weak ribs, weak muscles, lack of surfactant in the lungs and low respiratory effort [241]. Figure 9.15 shows different respiratory signals extracted from typical recordings in the Neonatal HDU dataset.

The movement of the upper body often leads to drastic changes masking respiratory cycles in the signal. Respiration is more prominent and interpretable when the subject is quiet with minimal body motion. The primary aim of the signal quality assessment is to provide a quality measure from 0 (poor quality) to 1 (good quality) for each breath. The breath's signal quality is expected be zero during noisy periods. Good-quality breaths should be found in periods with minimal motion, have physiologically relevant respiratory rates and have similar morphology to the surrounding breaths.

For each respiratory signal, the following signal quality measures are estimated. All of them have already been introduced in previous chapters:

1. Activity Index (SQI_{act}) — described in section 7.4.1
2. Physiological bounding (SQI_{phys}) — described in section 8.7.1

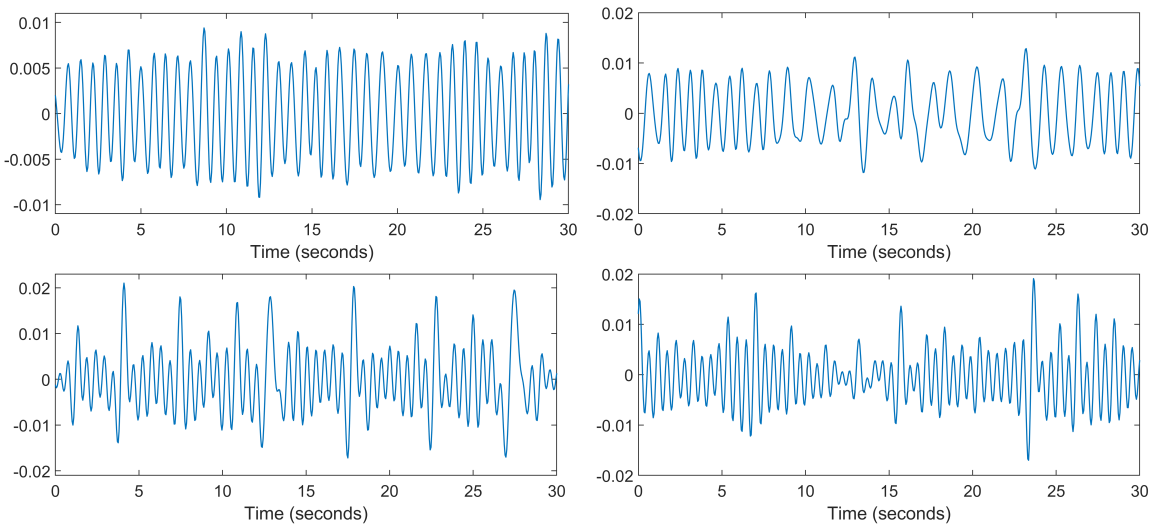


Figure 9.15: Sample respiratory signals extracted from typical recording sessions. (a) Normal and clean respiratory pattern. (b) Normal respiratory pattern followed by a slow respiratory pattern for 15 seconds. (c) Normal respiratory pattern periodically followed by a single respiratory peak. (d) Normal respiratory pattern followed by a low-amplitude segment in the middle.

3. Peak agreement (SQI_{peak}) — described in section 6.4.4 for ECG signals
4. Multi-scale dynamic time warping (SQI_{dtw}) — described in section 7.4.6 for PPGi signals

The calculation of the first two quality measures (SQI_{act} and SQI_{phys}) follows what was described in the previous chapters. This section describes the adaptation of the last two measures (SQI_{peak} and SQI_{dtw}) for non-contact respiratory signals. Each signal quality measure is computed for each respiratory cycle. Once all the signal quality measures have been computed, a combined signal quality index (SQI_{breath}) is calculated for each breath, based on all of these measures.

9.6.1 Peak agreement

Peak agreement is a measure of how much the peaks identified by the two peak-and-onset detectors agree with each other over a given time window. Both detectors usually agree with each other when the respiratory signal is clean and disagree in the presence of noise and artefacts. Here, agreement is considered to occur when the peaks identified by the two detectors are not located away from each other by more than 5 samples (or 0.25 second – half of the duration of the highest frequency rate that can be estimated). The measure of peak agreement or SQI_{peak} for the kth breath is calculated as the proportion of the number of peaks in agreement with the number of all peaks detected in the window of 10 seconds centred around the kth breath:

$$SQI_{\text{peak}}(k) = \frac{N_{\text{Agreed peaks}}}{N_{\text{All peaks}}}. \tag{9.18}$$

The SQI_{peak} ranges from 0 (poor quality) to 1 (good quality). Figure 9.16 shows the peaks and onsets detected by both techniques and the SQI_{peak} computed from the peak detectors. As expected, the SQI_{peak} decreases in the presence of noise.

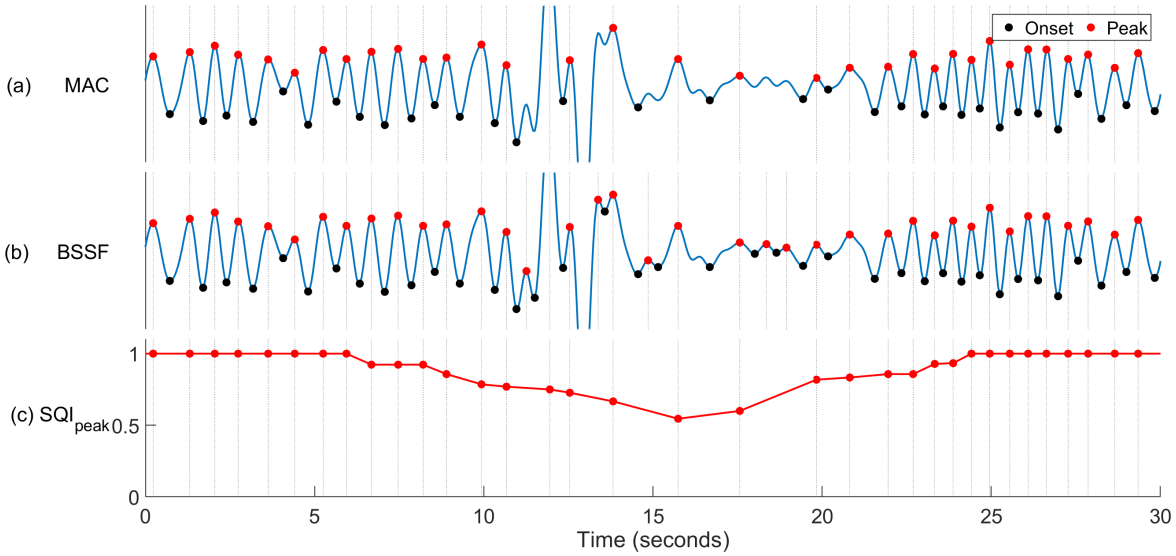


Figure 9.16: The calculation of SQI_{peak}. (a) Peaks and onsets detected by the MAC technique. (b) Peaks and onsets detected by the BSSF technique. (c) SQI_{peak} values. Each breath is segmented using peaks and onsets from the MAC technique. The SQI_{peak} value of each breath is computed from a 15-second running window, centred on each breath.

9.6.2 Multi-scale dynamic time warping

The multi-scale dynamic time warping technique has been introduced in Section 7.4.6 for the assessment of PPGi signals. This technique is used to determine the optimal alignment between each peak in the signal and the template calculated by averaging the nearby peaks over a window. Since the technique is designed for PPGi signals, several modifications are needed to make it suitable for respiratory signals.

Unlike PPGi signals, the morphology of the respiratory signal varies greatly according to the subject's breathing patterns. Pre-term infants are known to have spontaneous breathing patterns [241]. Some infants, for example, may have a short inspiration phase followed by a prolonged expiratory phase; others may have a period of hold expiration followed by multiple expiratory flow peaks. The amplitude of each breath in the respiratory signal mainly depends on the depth of breathing or the volume of air inspired into the lungs. Hence, the criteria that are originally used for constructing a peak template from PPGi signals (see section 7.4.6 for more details) are too strict and not appropriate for respiratory signals.

Multi-scale dynamic time warping is carried out using a 15-second moving window w with a step size of 5 seconds. The assessment is performed independently for each window w. In order to

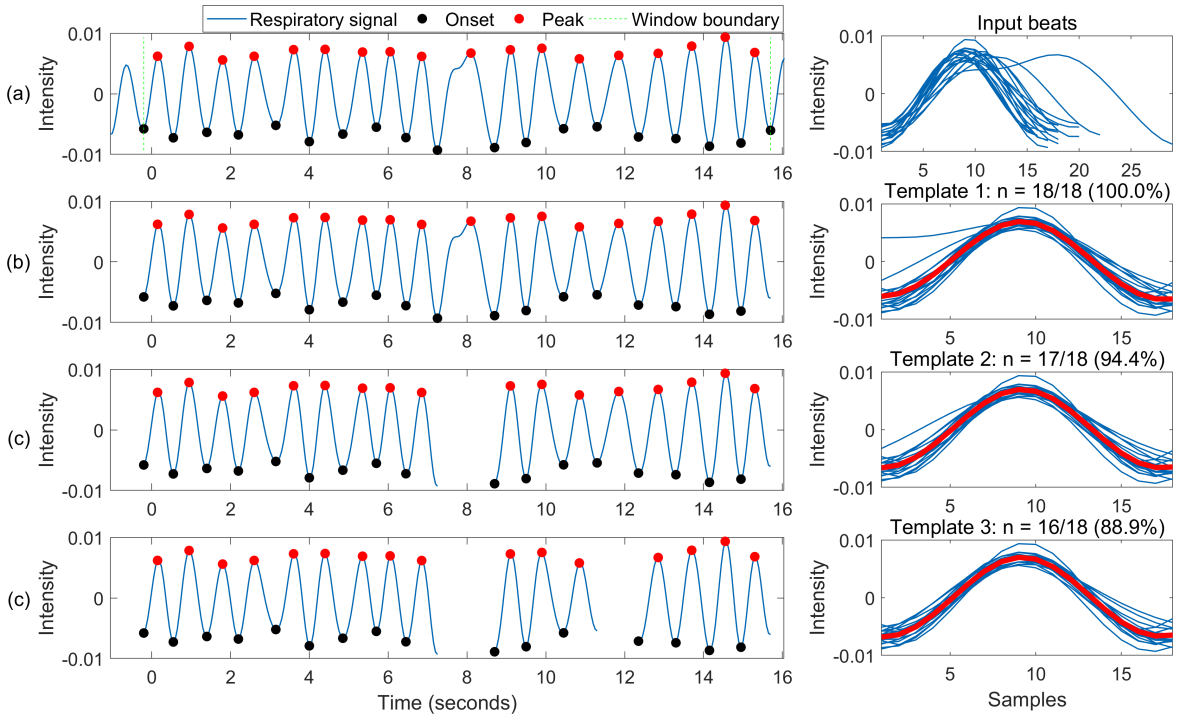


Figure 9.17: Computation of the average breath template for a 15-second respiratory waveform. (a) Input respiratory timeseries with peak and onset markers. Each breath is stacked on top of each other. (b) Template 1 (T1) is constructed from the breaths whose peak-to-peak interval corresponded to the respiratory rate range of 18 to 120 breaths/min. (c) Template 2 (T2) is computed from the remaining breaths whose correlation coefficient is larger than 0.8. (d) Template 3 (T3) is computed from the remaining breaths whose peak-to-peak intervals are within median and one-half interquartile range. Since more than half of the breaths are retained, the T3 template is considered to be a valid template.

measure the signal quality of each breath in the window, an average breath template is constructed as follows (see figure 9.17). Initially, all breaths in the window w are extracted and aligned using their peaks. The template is computed in three stages. For each stage, the template is computed by averaging all valid beats. In the first stage, the peak-to-peak interval of each breath is calculated as the time between two successive peaks. The first template (T1) is computed from the valid breaths whose beat-to-beat interval is within the range from 0.5 to 3.3 seconds (18 - 120 breaths/min). In the second stage, a cross-correlation coefficient between each valid breath in the first stage and the T1 template is calculated. The second template (T2) is computed from the valid breaths with a cross-correlation coefficient higher than 0.8. In the third stage, the breath-to-breath interval of each valid breath in the previous stage is calculated. The median plus or minus interquartile range of the breath-to-breath intervals are computed. The third template (T3) is constructed from the breaths whose beat-to-beat interval is within the median and one-half interquartile range. If more than 50% of the breaths pass all the stages, the T3 template is considered as the final window template; otherwise, the window template is not calculated.

Once the template has been calculated, the DTW distance is computed for each breath in the time window w using the same technique as described in section 7.4.6. Let DTW be the distance between each breath and the window template. The SQI_{dtw} is defined as:

$$SQI_{\text{dtw}}(k) = \begin{cases} 1 - DTW(k)/10 & \text{if } DTW(k) \leq 10 \\ 0 & \text{otherwise} \end{cases}. \tag{9.19}$$

Wherever the window template is not calculated, the SQI_{dtw} of all breaths in the window w is set to 0 (poor quality). This is different to what was described in section 7.4.6, where the

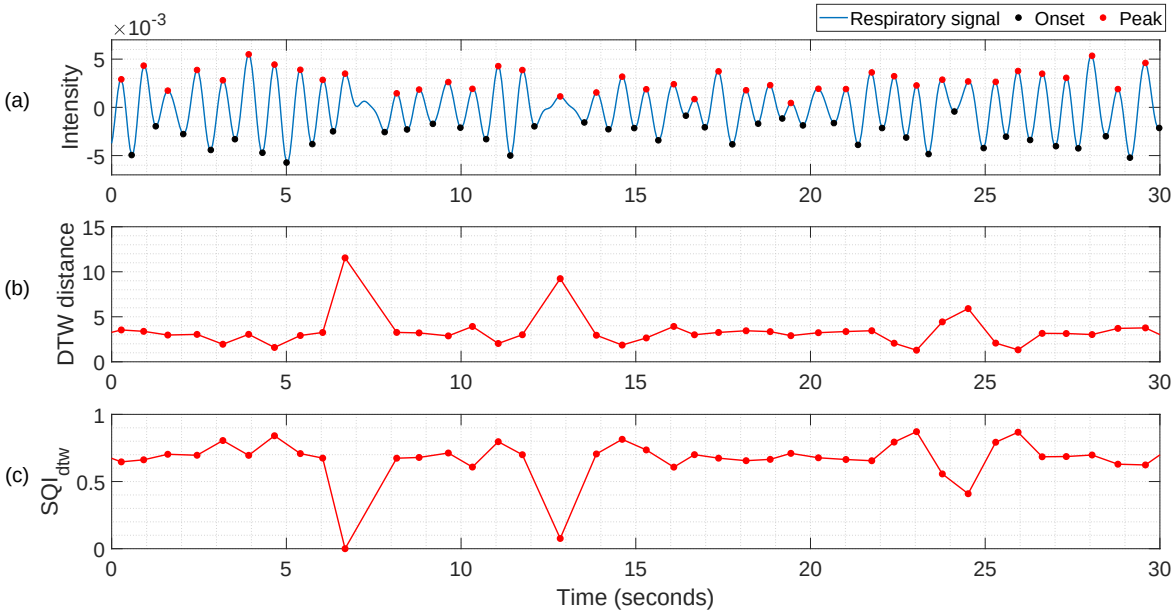


Figure 9.18: Computation of SQI_{dtw} for a 30-second respiratory signal. (a) Respiratory signal. (b) DTW distance between each beat and its running template. (c) SQI_{dtw} calculated by scaling the DTW distance. Breaths with irregular morphology have low SQI_{dtw}.

template of the previous window was used as the current window's template if the current template could not be calculated. This modification is needed as variations in respiratory rate are much greater than variations in heart rate.

Since SQI_{dtw} is calculated using a 15-second moving window with a step size of 5 seconds, the highest SQI_{dtw} value is taken where two windows overlap. Figure 9.18 shows an example of SQI_{dtw} computation for a 30-second respiratory signal.

9.6.3 Combining signal quality measures

Once all the signal quality measures have been calculated, a combined signal quality index (SQI_{breath}) is computed for the kth breath, in the same way as in previous chapters:

$$SQI_{\text{breath}}(k) = SQI_{\text{act}}(k) \cdot SQI_{\text{phys}}(k) \cdot SQI_{\text{peak}}(k) \cdot SQI_{\text{dtw}}(k). \tag{9.20}$$

SQI_{breath} is taken to be 0 (poor quality) during the periods of high subject motion ($SQI_{\text{act}} = 0$) and abnormal instantaneous respiratory rate ($SQI_{\text{phys}} = 0$). SQI_{peak} modifies SQI_{breath} so that SQI_{dtw} is trusted less when the two peak detectors disagree. During a quiet and stable period, SQI_{breath} relies mainly on SQI_{dtw}. Figure 9.19 shows the process of deriving an SQI_{breath} value from all signal quality indices for a 30-second respiratory waveform, taken from a quiet and stable period.

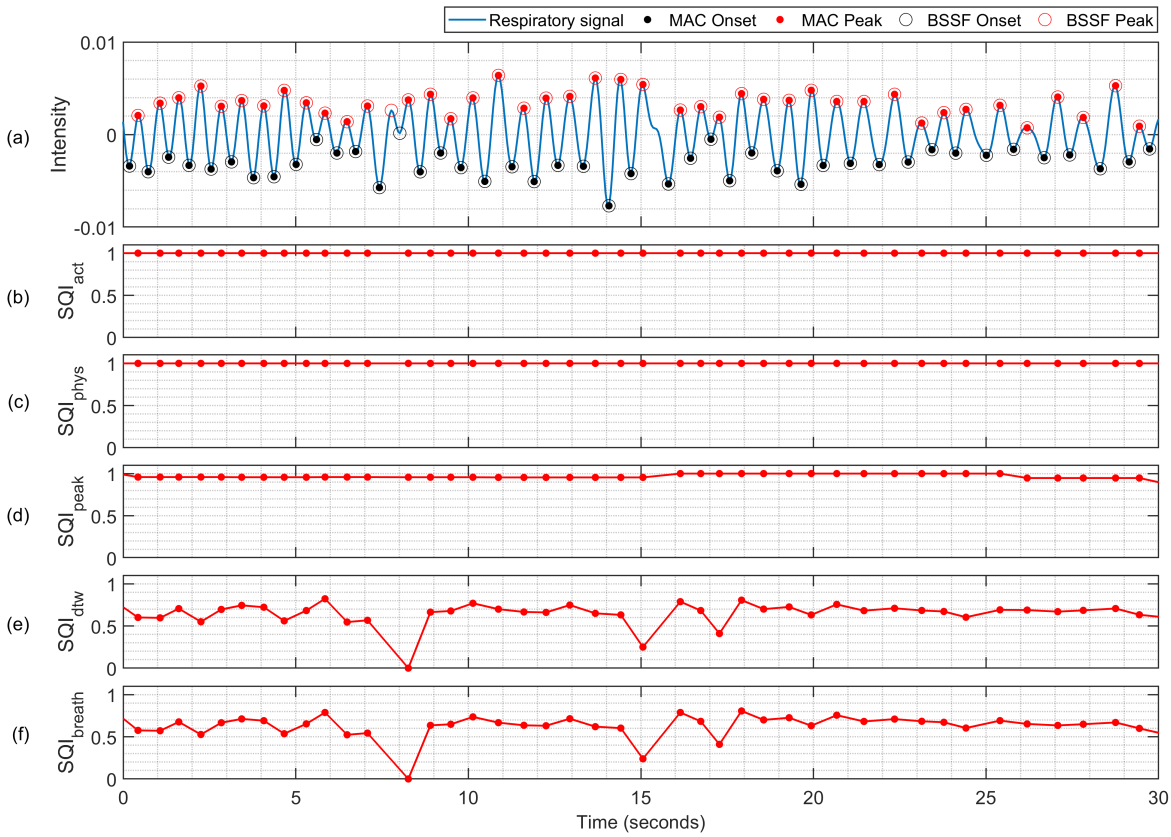


Figure 9.19: Computation of SQI_{breath} for a 30-second respiratory signal: (a) Respiratory signal extracted from a quiet and stable period, (b) SQI_{act}, (c) SQI_{phys}, (d) SQI_{peak}, (e) SQI_{dtw} and (f) SQI_{breath}. The SQI_{breath} is calculated by multiplying all the previous metrics together, as in equation 9.20.

9.7 Respiratory rate estimation methods

Following the signal quality assessment, the respiratory rate for each individual respiratory signal is estimated. These estimates are then combined using the data fusion technique based on Kalman filters discussed in section 9.7.2.

9.7.1 Estimating respiratory rate

For each respiratory signal, respiratory rate is estimated using a 10-second sliding window with a step size of 1 second. Prior to the estimation, the time window is expanded to include the complete first and last breath pulse. Estimation is performed by counting the number of breaths over the time window and scaling it to the unit of breaths/min (see Section 7.5.1). Respiratory rate is reported every second.

The signal quality index of the respiratory rate (SQI_{RR}) is calculated as the mean of the SQI_{breath} values over a given window when the numbers of breaths detected by both detectors are equal. When the numbers of breaths detected by both detectors are not equal, the SQI_{RR} is taken to be 0. Figure 9.20 shows the estimation of respiratory rate for a 2-minute respiratory signal, taken from a typical recording session.

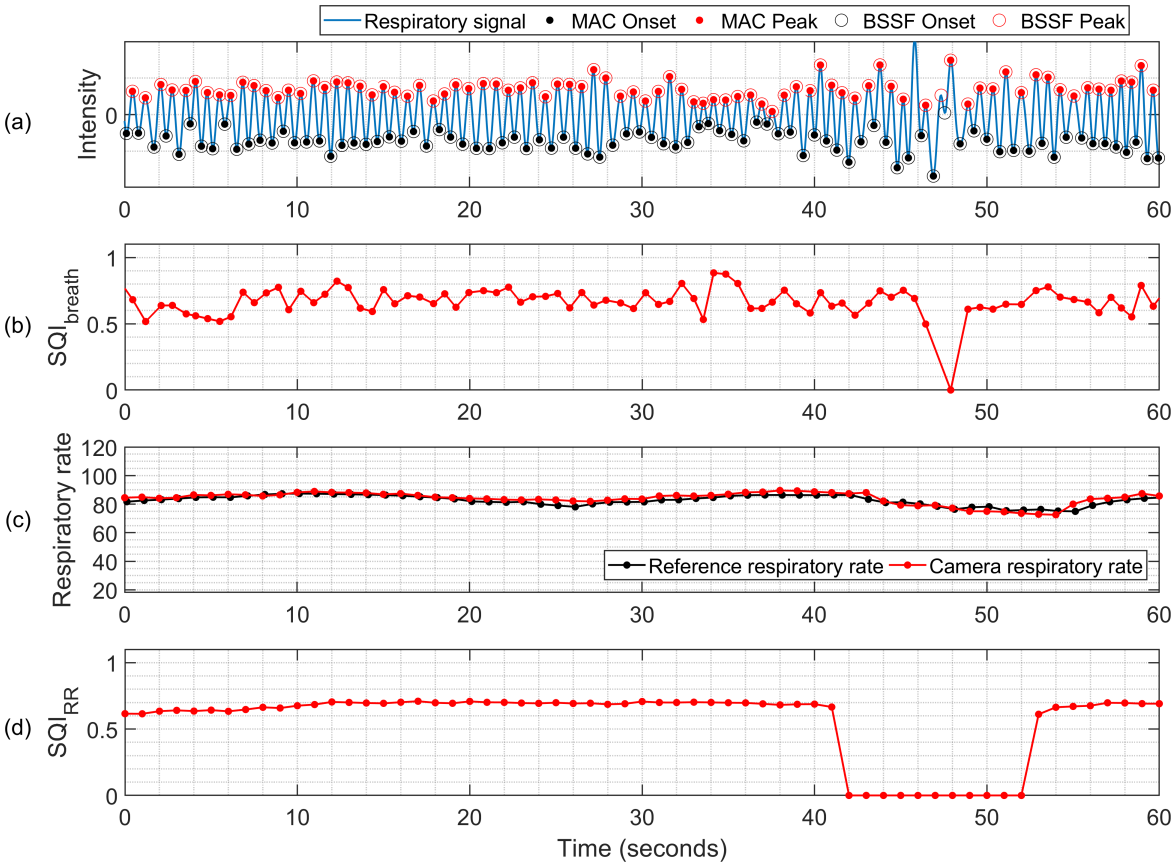


Figure 9.20: Respiratory rate estimation for a 60-second respiratory signal ($PPGi_{\text{green}}$). (a) Respiratory signal, (b) SQI_{breath}, (c) Respiratory rate estimates and (d) SQI_{RR}.

9.7.2 Data fusion

Once respiratory rate has been estimated for all respiratory signals, the data fusion technique based on multiple Kalman filters described in section 8.8.3 is applied to combine the multiple respiratory estimates from the same time window and produce a final respiratory rate RR_{fusion} and a final signal quality index SQI_{fusion}. Since the respiratory rate estimates from multiple signals are combined together, data fusion plays an important role in improving the value of respiratory rate estimates and the percentage of time that respiratory rate is estimated.

Figure 9.21 shows the result of fusing respiratory rate estimates for a 5-minute segment. The fused respiratory rate estimates are consistent with the values of the reference respiratory rate.

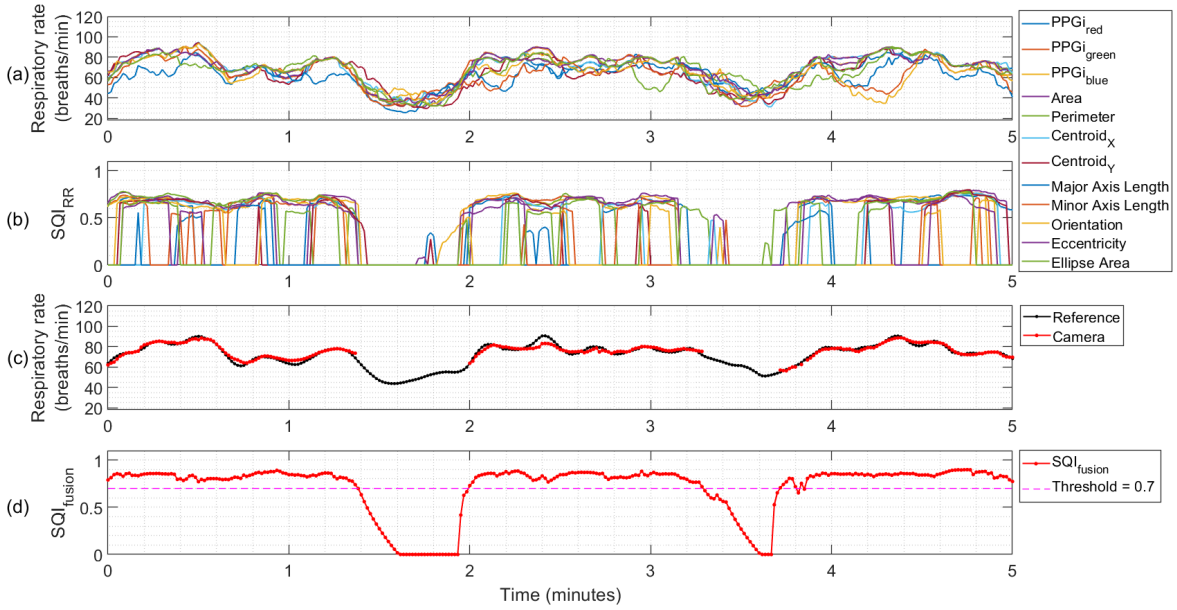


Figure 9.21: Data fusion of respiratory rate estimates for a 5-minute segment: (a) Respiratory rate estimates from different respiratory signals, (b) SQI_{RR} of respiratory rate estimates from each signal, (c) Fusion of respiratory rate estimates (RR_{fusion}) and (d) Signal quality index of RR_{fusion} (SQI_{fusion}). The fusion respiratory rate estimates agree with the reference respiratory rate during good-quality periods.

9.8 Results

This section reports respiratory rate estimation results on the Neonatal HDU dataset. According to the analysis of reference respiratory rate in chapter 8, the reference respiratory rate is available for 189.0 of 426.6 hours total recording time (both video and reference data were simultaneously recorded) or 44.3% of the time. The Neonatal HDU dataset is divided into two sets of data, each from 15 infants. The training set is used to find appropriate parameters and thresholds, whereas the test set is used to validate the method. Respiratory rate estimation results are reported for both sets.

Of 90 recording sessions, 6 recording sessions are discarded based on the following criteria:

- Reference data not recorded due to equipment malfunction (1 session)
- Infant under blue-light phototherapy treatment (5 sessions)

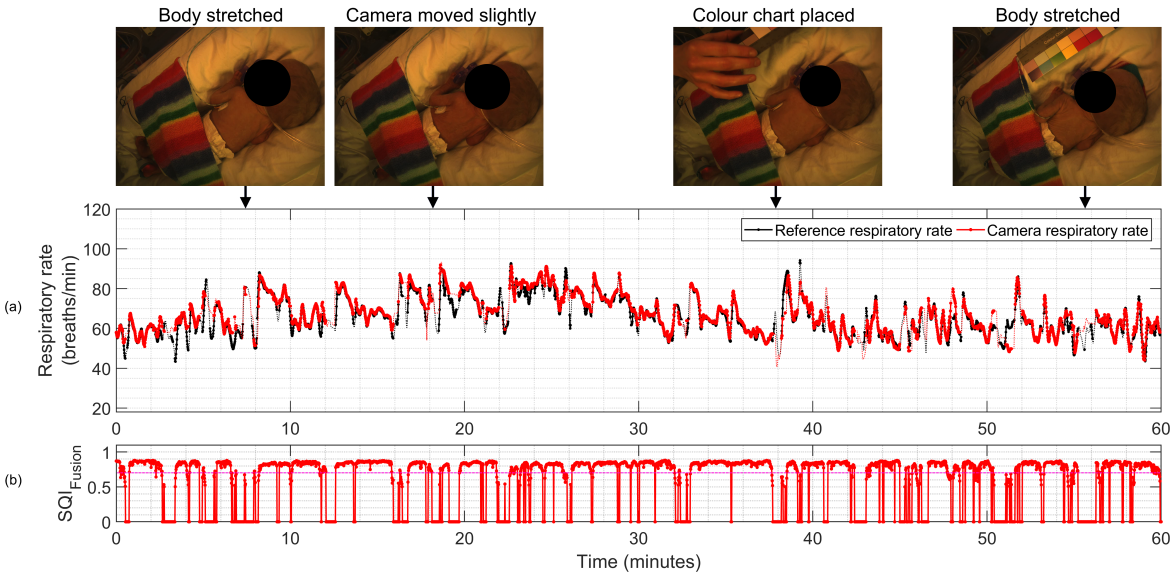


Figure 9.22: Respiratory rate estimation results for a one-hour video segment: (a) Respiratory rate estimates (RR_{fusion}) and (b) SQI_{fusion}. The reference heart rate is shown in black. Good agreement is found between the reference and the camera-derived respiratory rates.

Respiratory rate estimation is performed for the recording sessions that were out of focus. Thus estimation is performed on 84 sessions, comprising a total of 401.2 hours of recording time.

Similarly to Chapter 7, the comparison between the reference respiratory rate and the camera-derived respiratory rate is performed on periods of valid data during which: (1) the reference respiratory rate is available; (2) the infant is present; and (3) the infant is not undergoing any clinical intervention. The comparison is performed on respiratory rate estimates RR_{fusion} whose signal quality index SQI_{fusion} is higher than 0.70. This threshold is carefully selected to achieve low error rates on the training set while retaining a meaningful proportion of respiratory rate estimates.

Figure 9.22 shows the respiratory rate estimation results for a one-hour video segment. The respiratory rate estimates vary between 40 and 100 breaths/min and generally agree with the reference respiratory rate. Pre-term infants occasionally show spontaneous movement patterns such as body stretching or twitching during sleeping [242]. These movements cause motion artefacts which prevent the estimation of respiratory rates from both the reference and the camera.

9.8.1 Overall results

Table 9.1 summarises the results of respiratory rate estimation for all recording sessions in the training set and the test set of the Neonatal HDU study. The estimation errors are computed with respect to the reference respiratory rates of chapter 8. Of the six recording sessions that are excluded from the respiratory rate estimation (see the previous section), four sessions are in the training set and the other two in the test set. For the training set, 37.0% (73.2 hours out of 197.7 hours) of the total recording time is considered valid (see the validity criteria in the previous

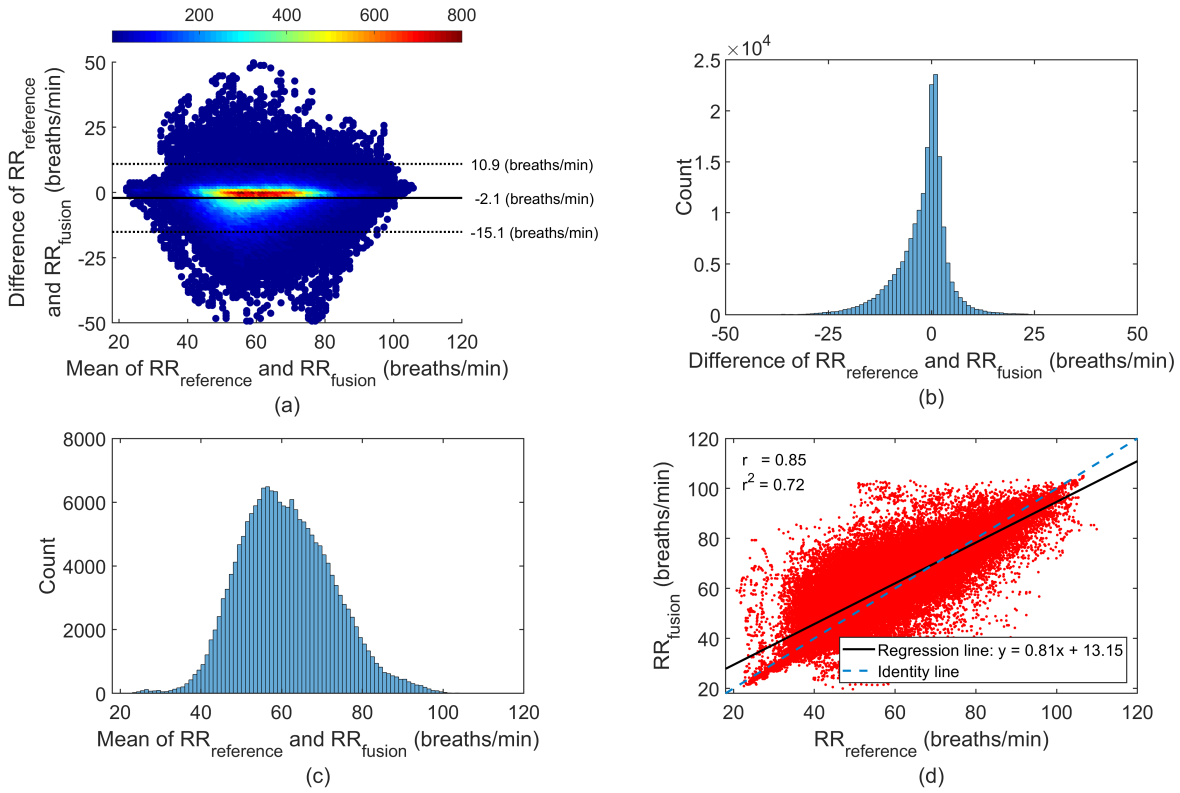


Figure 9.23: Comparison between the reference and camera-derived respiratory rates for the entire training set. (a) Bland-Altman plot. (b) Histogram of the differences between the two values. (c) Histogram of the mean values. (d) Correlation plot showing a correlation coefficient r of 0.85.

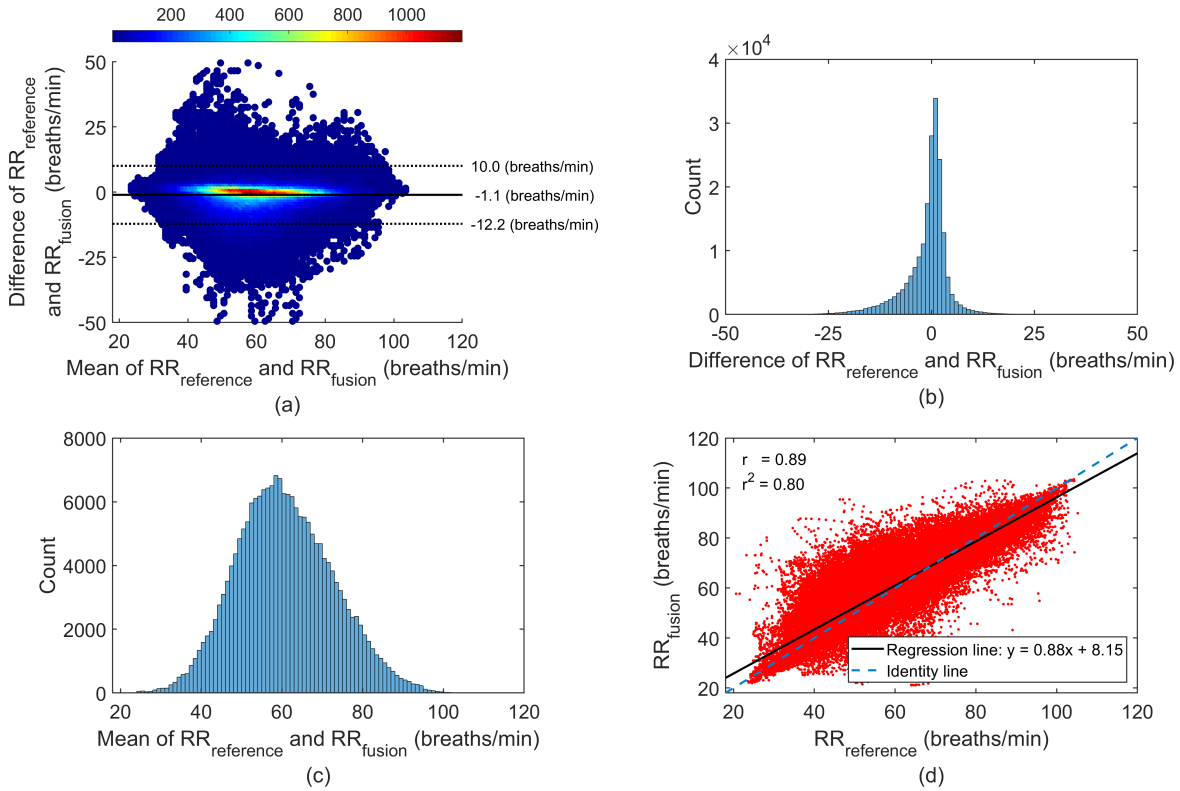


Figure 9.24: Comparison between the reference and camera-derived respiratory rates for the entire test set. (a) Bland-Altman plot. (b) Histogram of the differences between the two values. (c) Histogram of the mean values. (d) Correlation plot showing a correlation coefficient r of 0.89.

Table 9.1: Summary of respiratory rate estimation results for all sessions in the training set and the test set of the Neonatal HDU study. Four sessions are excluded from the training set, and two from the test set.

	Total time (hours)[1]	Valid time		Estimated time		Error (breaths/min)		
		hours	% total[2]	hours	% valid[3]	MAE	MAD	r
Training set	197.7	73.2	37.0%	51.4	70.2%	4.5	4.7	0.85
Test set	203.5	70.1	34.5%	54.8	78.1%	3.6	3.8	0.89

[1] Total duration when the camera and reference data are simultaneously recorded.
[2] Percentage of the estimated time with respect to the total duration
[3] Percentage of the estimated time with respect to the valid time

section). Respiratory rate can be estimated for 70.2% (51.4 hours out of 73.2 hours) with a mean absolute error (MAE) of 4.5 breaths/min and a mean absolute deviation (MAD) of 4.7 breaths/min. For the test set, 34.5% (70.1 hours out of 203.5 hours) of the total recording time is considered valid. Respiratory rate can be estimated for 78.1% (54.8 hours out of 70.1 hours) with a MAE of 3.6 breaths/min and a MAD of 3.8 breaths/min. Similarly to chapter 7, the test set has lower estimation errors than the training set. This might be because the training set has more variable lighting conditions than the test set (see section 7.7.2).

Figures 9.23 and 9.24 show graphical comparisons between the reference respiratory rate and the camera-derived respiratory rate for the training set and the test set respectively. The mean difference between the two values is -2.1 breaths/min for the training set and -1.1 breaths/min for the test set. 95% of the differences between the two values lie within [-15.1, 10.9] breaths/min for the training set and [-12.2, 10.0] breaths/min for the test set. The histogram of the differences for both sets shows that the majority of errors are concentrated around 0 breaths/min. The mean values are distributed across the expected physiological range for the neonatal population, taken to be from 18 to 120 beats/min. Estimation errors are mostly due to short periods of motion artefacts.

Figure 9.25 shows the relationship between the MAE and the percentage of the estimated time to the valid time for the training set and the test set. Lower error rates can be achieved

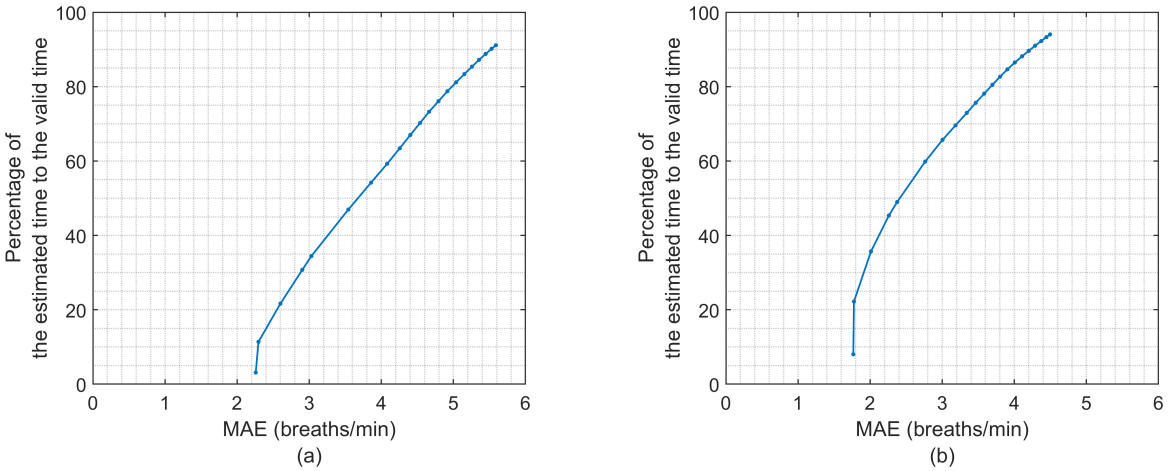


Figure 9.25: Relationship between the mean absolute error (MAE) and the proportion of the estimated time to the valid time for (a) the training set and (b) the test set.

Table 9.2: Respiratory rate estimation results for each recording session in the training set – Part 1 of 2

Session	Gest. Age (weeks)	Ethnicity	Gender	Total time (hours)	Time (hours, %)					Error	
					Poor reference[1]	Subject absence[1]	Clinical Intervention[1]	Valid time[1]	Estimated time[2]	MAE (beats/min)	MAD
PN01 Day 1	31.3 wk	White	Male	5.4 h	2.3 h, 42.0 %	0.0 h, 0.0 %	0.4 h, 7.9 %	2.7 h, 50.1 %	1.7 h, 62.9 %	4.5	4.5
PN01 Day 2	31.4 wk	White	Male	5.3 h	2.3 h, 41.7 %	0.0 h, 0.6 %	0.1 h, 3.0 %	2.9 h, 54.7 %	2.1 h, 72.4 %	5.2	4.9
PN01 Day 3	31.6 wk	White	Male	5.0 h	2.8 h, 55.7 %	0.6 h, 11.7 %	0.2 h, 4.8 %	1.4 h, 27.7 %	1.1 h, 78.7 %	6.7	6.0
PN02 Day 1	27.6 wk	White	Female	4.0 h	1.6 h, 39.7 %	0.3 h, 8.2 %	0.2 h, 4.3 %	1.9 h, 47.7 %	1.4 h, 76.2 %	4.4	4.3
PN02 Day 2	27.7 wk	White	Female	7.1 h	3.4 h, 47.7 %	0.3 h, 4.9 %	0.3 h, 3.9 %	3.1 h, 43.5 %	2.0 h, 65.3 %	4.1	4.0
PN02 Day 3	27.9 wk	White	Female	6.8 h	2.8 h, 41.5 %	0.7 h, 9.6 %	0.1 h, 1.4 %	3.2 h, 47.5 %	2.5 h, 78.3 %	4.1	3.9
PN02 Day 4	28.0 wk	White	Female	4.7 h	2.4 h, 50.1 %	0.3 h, 5.5 %	0.2 h, 3.3 %	1.9 h, 41.1 %	1.5 h, 76.5 %	4.0	4.0
PN03 Day 1	28.7 wk	White	Female	6.7 h	3.4 h, 50.8 %	0.4 h, 6.2 %	0.4 h, 5.9 %	2.5 h, 37.1 %	1.7 h, 69.0 %	5.0	4.6
PN03 Day 2	28.9 wk	White	Female	6.1 h	3.4 h, 56.1 %	0.0 h, 0.1 %	0.4 h, 6.2 %	2.3 h, 37.6 %	1.7 h, 75.1 %	5.9	5.2
PN03 Day 3	29.0 wk	White	Female	4.5 h	2.4 h, 53.1 %	0.0 h, 0.2 %	0.1 h, 2.0 %	2.0 h, 44.6 %	1.4 h, 70.5 %	4.2	4.2
PN05 Day 1	36.4 wk	Asian	Male	6.5 h	4.3 h, 67.2 %	0.0 h, 0.0 %	0.1 h, 1.8 %	2.0 h, 30.9 %	0.8 h, 40.8 %	5.7	6.4
PN06 Day 1	33.4 wk	White	Male	*Subject under phototherapy treatment*							
PN06 Day 2	33.6 wk	White	Male	*Subject under phototherapy treatment*							
PN06 Day 3	33.7 wk	White	Male	3.0 h	2.1 h, 70.7 %	0.0 h, 0.1 %	0.2 h, 6.8 %	0.7 h, 22.5 %	0.4 h, 60.4 %	4.6	4.2
PN07 Day 1	32.0 wk	White	Female	*Subject under phototherapy treatment*							
PN07 Day 2	32.1 wk	White	Female	6.1 h	3.9 h, 64.5 %	0.0 h, 0.0 %	0.1 h, 2.1 %	2.0 h, 33.4 %	1.6 h, 77.1 %	5.7	5.4
PN07 Day 3	32.3 wk	White	Female	6.1 h	2.4 h, 38.8 %	0.0 h, 0.2 %	0.2 h, 2.9 %	3.5 h, 58.1 %	2.6 h, 72.0 %	2.9	3.0
PN09 Day 1	30.9 wk	Black	Male	6.8 h	3.3 h, 48.1 %	0.0 h, 0.0 %	0.1 h, 1.7 %	3.4 h, 50.1 %	2.8 h, 82.6 %	3.7	3.7
PN09 Day 2	31.0 wk	Black	Male	5.4 h	3.0 h, 55.5 %	0.0 h, 0.0 %	0.1 h, 1.2 %	2.3 h, 43.2 %	2.0 h, 84.1 %	4.2	4.4
PN09 Day 3	31.1 wk	Black	Male	6.0 h	3.2 h, 54.4 %	0.0 h, 0.0 %	0.1 h, 1.7 %	2.6 h, 43.8 %	1.5 h, 59.0 %	4.5	4.5
PN11 Day 1	28.7 wk	White	Female	3.9 h	2.5 h, 63.3 %	0.0 h, 0.3 %	0.1 h, 3.8 %	1.3 h, 32.7 %	0.9 h, 72.1 %	4.9	4.6
PN11 Day 2	28.9 wk	White	Female	4.6 h	2.8 h, 61.1 %	0.0 h, 0.0 %	0.2 h, 3.6 %	1.6 h, 35.4 %	1.0 h, 60.0 %	4.9	4.7
PN11 Day 3	29.0 wk	White	Female	2.9 h	1.7 h, 57.0 %	0.0 h, 0.4 %	0.1 h, 4.4 %	1.1 h, 38.2 %	0.5 h, 43.3 %	5.2	5.1

Continued on the next page (table 9.3).

[1] Percentage of the time with respect to the total duration; [2] Percentage of the estimated time with respect to the valid time; [3] Mixed White & Asian; [4] Mixed White & Black.

Table 9.3: Respiratory rate estimation results for each recording session in the training set — Part 2 of 2

Session	Gest. Age (weeks)	Ethnicity	Gender	Total time (hours)	Time (hours, %)					Error	
					Poor reference[1]	Subject absence[1]	Clinical Intervention[1]	Valid time[1]	Estimated time[2]	MAE (beats/min)	MAD
Continued from the previous page (table 9.2).											
PPN15 Day 1	31.7 wk	Mixed W&A[3]	Male	3.0 h	2.1 h, 69.5 %	0.0 h, 0.0 %	0.1 h, 3.2 %	0.8 h, 27.3 %	0.6 h, 79.2 %	9.9	6.9
PN15 Day 2	31.9 wk	Mixed W&A[3]	Male	3.9 h	1.4 h, 36.9 %	0.0 h, 0.2 %	0.2 h, 5.1 %	2.2 h, 57.8 %	1.9 h, 86.1 %	2.8	2.9
PN15 Day 3	32.0 wk	Mixed W&A[3]	Male	3.2 h	1.5 h, 48.8 %	0.0 h, 0.4 %	0.3 h, 7.9 %	1.4 h, 43.0 %	1.0 h, 74.6 %	3.6	3.6
PN15 Day 4	32.1 wk	Mixed W&A[3]	Male	3.0 h	1.4 h, 46.6 %	0.0 h, 0.2 %	0.2 h, 6.7 %	1.4 h, 46.6 %	1.2 h, 86.6 %	3.7	3.8
PN16 Day 1	32.7 wk	White	Female	*Subject under phototherapy treatment*							
PN16 Day 2	32.9 wk	White	Female	5.8 h	3.8 h, 66.0 %	0.0 h, 0.1 %	0.0 h, 0.8 %	1.9 h, 33.1 %	1.1 h, 57.9 %	9.6	9.5
PN17 Day 1	29.6 wk	White	Male	4.8 h	1.9 h, 40.7 %	0.0 h, 0.1 %	0.6 h, 13.0 %	2.2 h, 46.2 %	1.6 h, 71.1 %	2.9	3.0
PN17 Day 2	29.7 wk	White	Male	6.6 h	2.6 h, 39.9 %	0.5 h, 8.1 %	0.3 h, 4.0 %	3.2 h, 48.0 %	2.4 h, 76.4 %	2.8	2.9
PN17 Day 3	29.9 wk	White	Male	5.9 h	3.6 h, 61.0 %	1.0 h, 17.4 %	0.2 h, 4.2 %	1.0 h, 17.4 %	0.5 h, 47.1 %	5.3	5.3
PN19 Day 1	29.6 wk	Mixed Other	Female	7.0 h	4.1 h, 58.7 %	0.0 h, 0.0 %	0.2 h, 2.7 %	2.7 h, 38.6 %	1.9 h, 70.8 %	4.0	4.1
PN19 Day 2	29.7 wk	Mixed Other	Female	5.8 h	3.4 h, 58.1 %	0.4 h, 6.0 %	0.7 h, 11.7 %	1.4 h, 24.1 %	0.9 h, 65.0 %	4.4	4.3
PN19 Day 3	29.9 wk	Mixed Other	Female	5.3 h	3.0 h, 56.2 %	0.0 h, 0.1 %	0.1 h, 2.6 %	2.2 h, 41.1 %	1.7 h, 77.4 %	3.4	3.5
PN21 Day 1	33.3 wk	White	Male	2.0 h	1.6 h, 77.8 %	0.0 h, 0.1 %	0.0 h, 1.3 %	0.4 h, 20.8 %	0.2 h, 51.4 %	8.4	7.0
PN21 Day 2	33.4 wk	White	Male	6.4 h	4.7 h, 72.8 %	0.5 h, 7.2 %	0.1 h, 2.1 %	1.2 h, 17.9 %	0.7 h, 61.6 %	10.3	7.7
PN21 Day 3	33.6 wk	White	Male	3.2 h	2.1 h, 66.5 %	0.0 h, 0.2 %	0.1 h, 3.0 %	1.0 h, 30.3 %	0.7 h, 69.9 %	6.7	5.5
PN23 Day 1	30.0 wk	White	Female	5.4 h	3.8 h, 69.6 %	0.4 h, 7.3 %	0.1 h, 1.7 %	1.2 h, 21.4 %	0.8 h, 71.8 %	5.6	5.1
PN23 Day 2	30.1 wk	White	Female	4.0 h	2.3 h, 59.4 %	0.0 h, 0.2 %	0.0 h, 0.2 %	1.6 h, 40.2 %	1.2 h, 73.8 %	3.7	3.7
PN23 Day 3	30.3 wk	White	Female	6.2 h	4.6 h, 74.1 %	0.9 h, 13.8 %	0.2 h, 3.0 %	0.6 h, 9.1 %	0.4 h, 72.5 %	11.8	8.0
PN25 Day 1	32.1 wk	Mixed W&B[4]	Male	6.4 h	4.7 h, 74.2 %	0.4 h, 6.7 %	0.1 h, 1.7 %	1.1 h, 17.4 %	0.7 h, 62.0 %	3.9	3.8
PN25 Day 2	32.3 wk	Mixed W&B[4]	Male	3.0 h	1.9 h, 61.0 %	0.0 h, 0.0 %	0.1 h, 1.7 %	1.1 h, 37.3 %	0.4 h, 39.2 %	6.5	6.6
Total (39 recording sessions)				197.7 h	110.5, 55.9 %	6.8 h, 3.4 %	7.3 h, 3.7 %	73.2 h, 37.0%	51.4 h, 70.2 %	4.7	4.5

[1] Percentage of the time with respect to the total duration; [2] Percentage of the estimated time with respect to the valid time; [3] Mixed White & Asian; [4] Mixed White & Black.

Table 9.4: Respiratory rate estimation results for each recording session in the test set — Part 1 of 2

Session	Gest. Age (weeks)	Ethnicity	Gender	Total time (hours)	Time (hours, %)					Error	
					Poor reference[1]	Subject absence[1]	Clinical Intervention[1]	Valid time[1]	Estimated time[2]	MAE (beats/min)	MAD
PN04 Day 1	33.4 wk	White	Male	2.0 h	1.0 h, 52.2 %	0.0 h, 0.1 %	0.3 h, 15.6 %	0.6 h, 32.1 %	0.3 h, 48.1 %	4.2	4.3
PN04 Day 2	33.6 wk	White	Male	4.7 h	1.9 h, 39.8 %	0.0 h, 0.3 %	0.6 h, 12.3 %	2.2 h, 47.6 %	1.6 h, 69.1 %	2.9	3.0
PN04 Day 3	33.7 wk	White	Male	5.4 h	2.7 h, 49.9 %	1.3 h, 24.0 %	0.2 h, 3.4 %	1.2 h, 22.6 %	0.9 h, 71.7 %	3.2	3.2
PN04 Day 4	33.9 wk	White	Male	8.1 h	4.3 h, 53.6 %	2.2 h, 27.1 %	0.4 h, 4.5 %	1.2 h, 14.8 %	0.6 h, 53.1 %	3.3	3.2
PN08 Day 1	30.1 wk	White	Female	5.8 h	3.7 h, 64.1 %	0.0 h, 0.1 %	0.1 h, 2.5 %	1.9 h, 33.2 %	1.7 h, 88.0 %	3.1	3.3
PN08 Day 2	30.3 wk	White	Female	5.9 h	3.0 h, 52.0 %	0.0 h, 0.1 %	0.0 h, 0.8 %	2.8 h, 47.1 %	2.5 h, 88.9 %	2.9	3.1
PN08 Day 3	30.4 wk	White	Female	3.7 h	2.4 h, 65.4 %	0.0 h, 0.1 %	0.3 h, 7.0 %	1.0 h, 27.5 %	0.8 h, 80.1 %	3.8	4.1
PN10 Day 1	30.1 wk	White	Male	3.6 h	2.5 h, 68.6 %	0.0 h, 0.1 %	0.0 h, 0.8 %	1.1 h, 30.5 %	0.7 h, 61.5 %	6.9	6.6
PN10 Day 2	30.3 wk	White	Male	8.3 h	5.1 h, 61.6 %	0.6 h, 7.7 %	0.1 h, 1.3 %	2.4 h, 29.4 %	1.9 h, 76.2 %	4.3	4.3
PN10 Day 3	30.4 wk	White	Male	4.6 h	2.5 h, 54.8 %	0.0 h, 0.2 %	0.0 h, 0.7 %	2.0 h, 44.4 %	1.6 h, 80.3 %	4.0	3.9
PN12 Day 1	31.0 wk	Asian	Female	*Subject under phototherapy treatment*							
PN13 Day 1	28.0 wk	Mixed W&A[3]	Female	3.0 h	1.6 h, 55.2 %	0.0 h, 0.2 %	0.1 h, 3.7 %	1.2 h, 40.9 %	1.0 h, 86.2 %	3.2	3.5
PN13 Day 2	28.1 wk	Mixed W&A[3]	Female	9.3 h	6.0 h, 64.6 %	1.4 h, 15.0 %	0.4 h, 4.2 %	1.5 h, 16.2 %	1.2 h, 80.2 %	3.9	3.9
PN13 Day 3	28.3 wk	Mixed W&A[3]	Female	2.4 h	1.8 h, 76.7 %	0.0 h, 0.2 %	0.1 h, 3.2 %	0.5 h, 20.0 %	0.3 h, 60.9 %	7.5	6.5
PN14 Day 1	29.4 wk	White	Female	*Reference data not recorded*							
PN14 Day 2	29.6 wk	White	Female	6.3 h	3.5 h, 55.9 %	0.2 h, 2.5 %	0.1 h, 2.0 %	2.5 h, 39.6 %	2.2 h, 88.6 %	3.3	3.5
PN14 Day 3	29.7 wk	White	Female	6.0 h	2.8 h, 47.0 %	0.6 h, 10.6 %	0.7 h, 11.9 %	1.8 h, 30.6 %	1.5 h, 79.4 %	3.3	3.5
PN14 Day 4	29.9 wk	White	Female	4.7 h	2.4 h, 51.6 %	0.0 h, 0.1 %	0.0 h, 0.5 %	2.2 h, 47.8 %	1.9 h, 84.2 %	3.4	3.7
PN18 Day 1	30.4 wk	White	Male	3.1 h	1.5 h, 50.4 %	0.3 h, 11.1 %	0.1 h, 4.4 %	1.0 h, 34.1 %	0.8 h, 76.4 %	3.5	3.6
PN18 Day 2	30.6 wk	White	Male	2.2 h	1.1 h, 47.8 %	0.0 h, 0.6 %	0.1 h, 4.6 %	1.0 h, 47.0 %	0.8 h, 76.5 %	3.6	3.7
PN18 Day 3	30.7 wk	White	Male	4.2 h	2.4 h, 55.4 %	0.0 h, 0.4 %	0.2 h, 4.5 %	1.7 h, 39.7 %	1.3 h, 78.8 %	5.0	5.0
PN20 Day 1	30.4 wk	Mixed Other	Male	4.3 h	2.3 h, 52.7 %	0.2 h, 4.2 %	0.2 h, 3.6 %	1.7 h, 39.5 %	1.4 h, 85.9 %	2.8	3.0
PN20 Day 2	30.6 wk	Mixed Other	Male	4.5 h	2.7 h, 59.7 %	0.0 h, 0.0 %	0.1 h, 3.1 %	1.7 h, 37.1 %	1.4 h, 86.6 %	4.8	5.0
PN20 Day 3	30.7 wk	Mixed Other	Male	4.3 h	2.6 h, 61.7 %	0.0 h, 0.0 %	0.1 h, 1.4 %	1.6 h, 36.9 %	1.4 h, 86.4 %	2.9	3.1
PN20 Day 4	30.9 wk	Mixed Other	Male	2.5 h	1.2 h, 46.4 %	0.0 h, 0.4 %	0.1 h, 2.8 %	1.3 h, 50.4 %	1.1 h, 84.5 %	2.4	2.5

Continued on the next page (table 9.5).

[1] Percentage of the time with respect to the total duration; [2] Percentage of the estimated time with respect to the valid time; [3] Mixed White & Asian; [4] Mixed White & Black.

Table 9.5: Respiratory rate estimation results for each recording session in the test set — Part 2 of 2

Session	Gest. Age (weeks)	Ethnicity	Gender	Total time (hours)	Time (hours, %)					Error	
					Poor reference[1]	Subject absence[1]	Clinical Intervention[1]	Valid time[1]	Estimated time[2]	MAE (beats/min)	MAD (beats/min)
Continued from the previous page (table 9.4).											
PN22 Day 1	29.6 wk	White	Male	3.8 h	2.3 h, 59.0 %	0.0 h, 0.2 %	0.1 h, 1.7 %	1.5 h, 39.1 %	0.8 h, 55.7 %	5.8	5.8
PN22 Day 2	29.7 wk	White	Male	4.0 h	2.6 h, 64.5 %	0.0 h, 0.1 %	0.3 h, 7.0 %	1.1 h, 28.4 %	0.6 h, 51.5 %	5.8	5.5
PN22 Day 3	29.9 wk	White	Male	2.5 h	1.8 h, 72.0 %	0.0 h, 0.0 %	0.1 h, 2.5 %	0.6 h, 25.5 %	0.4 h, 64.9 %	4.5	4.7
PN24 Day 1	31.7 wk	Mixed W&B[4]	Male	5.0 h	2.9 h, 57.6 %	0.5 h, 9.3 %	0.0 h, 0.8 %	1.6 h, 32.2 %	1.2 h, 71.6 %	3.3	3.4
PN24 Day 2	31.9 wk	Mixed W&B[4]	Male	5.7 h	4.2 h, 74.0 %	0.3 h, 5.6 %	0.1 h, 1.1 %	1.1 h, 19.3 %	0.7 h, 62.2 %	3.7	3.8
PN24 Day 3	32.0 wk	Mixed W&B[4]	Male	3.1 h	1.2 h, 38.5 %	0.0 h, 0.0 %	0.1 h, 2.4 %	1.8 h, 59.1 %	1.5 h, 82.6 %	3.0	3.1
PN26 Day 1	34.3 wk	White	Male	1.8 h	0.4 h, 25.1 %	0.0 h, 0.0 %	0.0 h, 1.3 %	1.3 h, 73.6 %	1.2 h, 91.8 %	1.9	1.9
PN26 Day 2	34.4 wk	White	Male	3.1 h	1.7 h, 54.6 %	0.0 h, 0.0 %	0.1 h, 2.2 %	1.4 h, 43.2 %	1.2 h, 91.4 %	4.1	4.4
PN26 Day 3	34.6 wk	White	Male	3.4 h	1.7 h, 49.7 %	0.0 h, 0.2 %	0.2 h, 4.7 %	1.6 h, 45.4 %	1.3 h, 84.1 %	3.2	3.3
PN26 Day 4	34.7 wk	White	Male	3.4 h	2.0 h, 58.2 %	0.0 h, 0.2 %	0.5 h, 13.6 %	1.0 h, 28.0 %	0.6 h, 63.3 %	2.9	3.0
PN27 Day 1	31.3 wk	Mixed W&A[3]	Male	4.6 h	2.0 h, 43.4 %	0.5 h, 11.6 %	0.2 h, 4.0 %	1.9 h, 41.1 %	1.6 h, 84.8 %	2.8	2.9
PN27 Day 2	31.4 wk	Mixed W&A[3]	Male	3.9 h	2.2 h, 55.9 %	0.0 h, 0.2 %	0.2 h, 4.5 %	1.5 h, 39.3 %	1.3 h, 82.7 %	5.1	5.2
PN28 Day 1	31.0 wk	White	Female	6.4 h	3.9 h, 60.1 %	1.6 h, 25.3 %	0.3 h, 4.7 %	0.6 h, 9.9 %	0.4 h, 67.0 %	5.6	5.6
PN28 Day 2	31.1 wk	White	Female	5.5 h	2.1 h, 38.8 %	0.6 h, 11.6 %	0.5 h, 8.4 %	2.3 h, 41.2 %	1.8 h, 77.3 %	2.5	2.5
PN28 Day 3	31.3 wk	White	Female	5.8 h	3.3 h, 58.0 %	0.6 h, 10.5 %	0.1 h, 2.4 %	1.7 h, 29.2 %	1.6 h, 95.4 %	2.8	2.8
PN29 Day 1	30.6 wk	White	Male	6.1 h	3.1 h, 50.9 %	0.6 h, 10.3 %	0.2 h, 3.7 %	2.2 h, 35.1 %	1.5 h, 68.8 %	4.5	4.4
PN29 Day 2	30.7 wk	White	Male	6.2 h	4.0 h, 64.3 %	0.5 h, 8.4 %	0.3 h, 4.8 %	1.4 h, 22.4 %	0.9 h, 65.1 %	4.1	4.1
PN29 Day 3	30.9 wk	White	Male	5.8 h	2.8 h, 48.6 %	0.2 h, 4.1 %	0.3 h, 4.6 %	2.5 h, 42.6 %	1.5 h, 61.7 %	5.4	5.4
PN29 Day 4	31.0 wk	White	Male	3.1 h	1.1 h, 38.0 %	0.6 h, 20.2 %	0.3 h, 9.9 %	0.9 h, 31.8 %	0.6 h, 65.9 %	4.6	4.4
PN30 Day 1	32.1 wk	White	Male	4.0 h	1.6 h, 40.9 %	0.0 h, 0.2 %	0.1 h, 1.9 %	2.3 h, 57.0 %	2.1 h, 90.0 %	2.6	2.6
PN30 Day 2	32.3 wk	White	Male	3.6 h	1.7 h, 46.4 %	0.0 h, 0.1 %	0.2 h, 4.7 %	1.8 h, 48.7 %	1.6 h, 90.9 %	3.0	3.2
PN30 Day 3	32.4 wk	White	Male	3.9 h	1.7 h, 45.4 %	0.0 h, 0.1 %	0.3 h, 7.7 %	1.8 h, 46.8 %	1.5 h, 85.9 %	3.1	3.4
Total (45 recording sessions)				203.5 h	111.6, 54.9 %	13.1 h, 6.4 %	8.6 h, 4.2 %	70.1 h, 34.5%	54.8 h, 78.1 %	3.6	3.8

[1] Percentage of the time with respect to the total duration; [2] Percentage of the estimated time with respect to the valid time; [3] Mixed White & Asian; [4] Mixed White & Black.

by tuning the SQI threshold (currently set at 0.7 for respiratory rate estimation), but this would reduce the proportion of valid estimates.

9.8.2 Results by recording sessions

Tables 9.2–9.5 detail respiratory rate estimation results for each recording session in the training set and the test set respectively. For each session, MAE varies from 1.9 to 11.8 breaths/min, MAD varies from 1.9 to 8.0 breaths/min and the proportion of the estimated time to the valid time ranged from 40.8% to 90.9%. Large errors are often associated with the recordings in which the subject is positioned facing the video camera. The appearance of the skin and the upper torso can have a large impact on the estimation of respiratory rate.

9.9 Discussion

Having presented the respiratory rate estimation results, we discuss findings in relation to the research objectives and compare the results with previous literature. This section also provides an in-depth analysis of the challenges in non-contact respiratory rate estimation and suggests a suitable condition for the non-contact estimation technique to work.

9.9.1 Respiratory rate estimation process

In recent years, several studies have presented non-contact techniques for estimating respiratory rate based on the principle of the PPGi signal [14, 54, 231, 243], while others have used computer vision techniques to analyse the respiratory-synchronous motion of the subject [213, 234, 244, 245]. The work in this chapter has developed a more robust technique for respiratory rate estimation based on the data fusion of respiratory rate estimates from both the PPGi-based and vision-based principles. This technique has been shown to be able to estimate neonatal respiratory rate in the normal range between 20 and 120 breaths/min and during tachypnoea when respiratory rate is higher than 60 breaths/min. The method is fast and automatic. It can estimate respiratory rate without any human supervision to identify a region of interest.

Although there are challenges caused by the pathological insatiability and spontaneous non-respiratory motion of pre-term infants, our method performs particularly well during quiet and stable periods in a typical recording session. The results in this study indicate that respiratory rate can be estimated for more than 70% of the periods when the reference data are available with an MAE of less than 4.5 breaths/min. Note that the range of respiratory rate in pre-term infants (20 - 120 breaths/min) is much higher and wider than that for adults (12 - 20 breaths/min).

9.9.2 Comparison to conventional monitoring

Using the data from the Neonatal HDU study, Jorge *et al.* [213] presented a comparison between the manual respiratory rate provided by clinical staff and the respiratory rates estimated by processing the IP, ECG and PPG signals. The MAEs between the gold-standard values and the estimates from contact sensors varied from 3.6 to 7.8 breaths/min in the respiratory rate range from 20 to 80 breaths/min. The errors are found to be higher in the respiratory rate estimates derived from the indirect methods (ECG and PPG).

Compared with the reference respiratory rate, the respiratory rate estimates derived from the camera have an MAE of 4.5 beats/min for the training set and an MAE of 3.6 beats/min for the test set. These results are in line with those observed in earlier work [212].

9.9.3 Comparison with other clinical studies

As there is a large difference in the range of respiratory rates between adults (12 - 20 breaths/min) and pre-term infants (20 - 120 breaths/min), this comparison addresses only the clinical studies associated with the neonatal population.

Jorge *et al.* [213] reported the results of respiratory rate estimation on 165 manual breath-counting sessions taken from the Neonatal HDU study. The MAE between the gold-standard and camera-derived respiratory rates was 6.9 breaths/min. Note that during a manual count session, the presence of clinical staff may awaken a subject. Their study did not use any signal quality metrics to exclude these noisy time windows.

van Gastel *et al.* [231] performed non-contact respiratory rate estimation on two pre-term infants, consisting of 588 breaths in total. With the reference respiratory data derived from the ECG signal, this study reported the MAE between the reference and camera-derived respiratory rates to be 4.7 breaths/min, with a correlation coefficient r of 0.87. The results presented in this chapter are consistent with those in the previous studies.

9.9.4 Data fusion techniques

Respiratory rate derived from one respiratory feature has been observed to have a low proportion of estimates (see figure 9.21). In this work, the data fusion technique [189] is used to mitigate these problems by weighting multiple respiratory rate estimates based on their signal quality values. This method is found to produce more reliable respiratory rate estimates and increase the proportion of the estimated time. Figure 9.26 shows the contribution of respiratory rate estimates from each respiratory signal. The estimates from the major axis length has the highest contribution to the estimation of the fused respiratory rate.

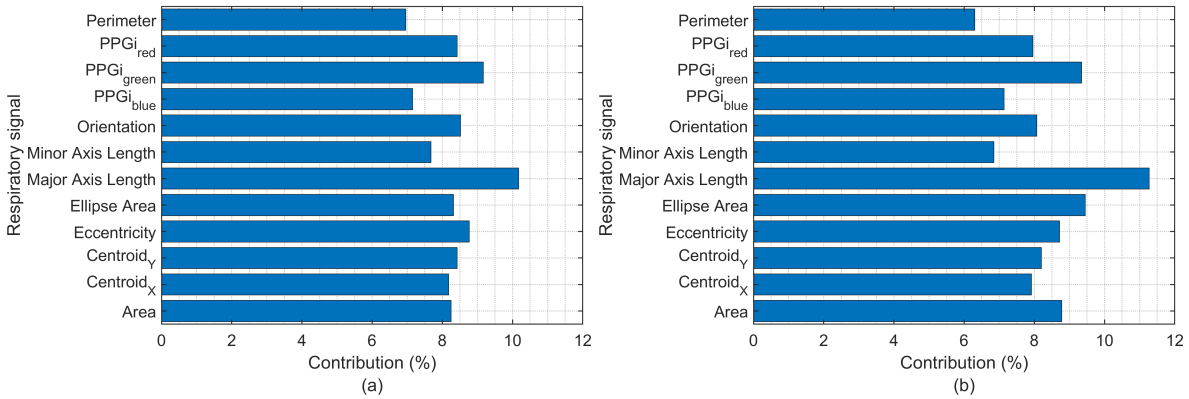


Figure 9.26: Contribution of respiratory rate estimates from each respiratory signal to the fusion respiratory rate RR_{fusion} for (a) the training set and (b) the test set.

9.9.5 Effects of different body postures

Figure 9.27 shows the estimation of respiratory rate for a 2-hour video segment during several challenging scenarios. The estimation performs particularly well when the subject is sleeping on the stomach and the whole of the back is shown to the video camera. Poor-quality respiratory rate values are observed during the periods for which the subject is awake and facing the video camera. Estimation is particularly difficult when the upper body and hands are moving.

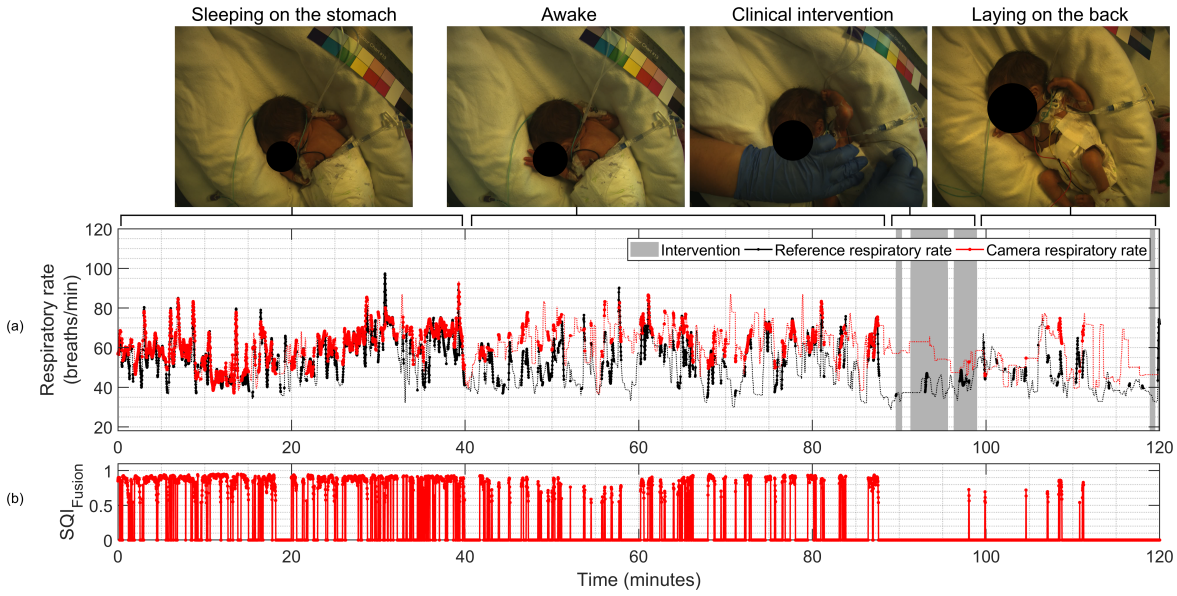


Figure 9.27: Estimation of respiratory rate for a 2-hour video segment during which clinical staff change the subject's body posture from a prone position (the back facing to the camera) to a supine position (the front facing to the camera). (a) Respiratory rate estimates. (b) Signal quality. The algorithm has difficulty estimating respiratory rate when the subject is facing the camera.

9.9.6 Gaps in respiratory rate estimates

Figure 9.28 shows the histogram of the periods for which respiratory rate cannot be estimated for both the training and test sets. The majority of gap durations are concentrated below 30 seconds.

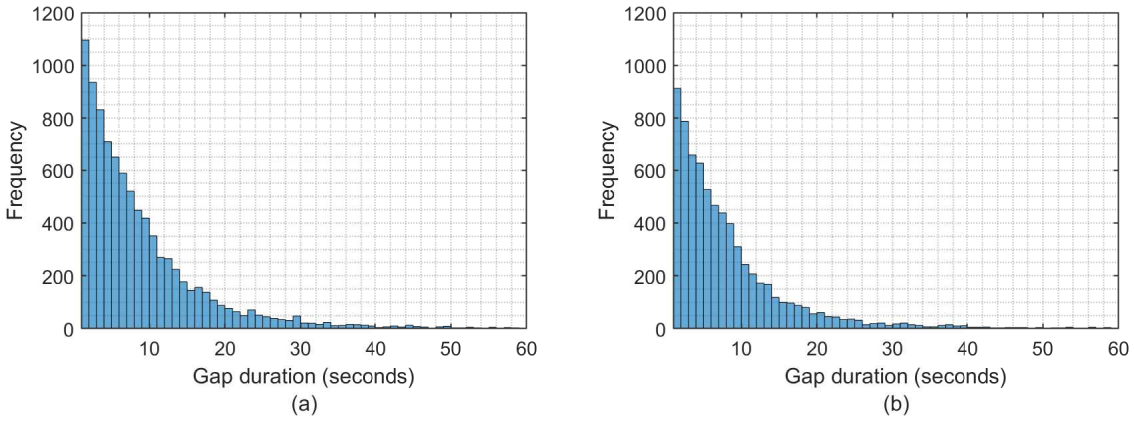


Figure 9.28: Gaps in respiratory rate estimates for (a) the training set and (b) the test set during valid time periods. If gaps of up to 30 seconds are allowed, the percentage of the estimated time increases from 70.2% to 94.5% for the training set and from 78.1% to 96.5% for the test set.

Due to the high prevalence of noise in the respiratory signals, the proportion of the estimated time to the total recording time for respiratory rate is lower than that for heart rate. Hence, the amount of gaps in respiratory rate estimates is also lower in comparison to the gaps for the heart rate estimates (see section 7.7.5). If gaps of up to 30 seconds are allowed, the percentage of the estimated time with respect to the valid time increases from 70.2% to 94.5% for the training set and from 78.1% to 96.5% for the test set. The frequency of measuring respiratory rate in the hospital can vary from second-by-second to every few hours. Clinical staff are mainly interested in the trends of respiratory rate and the interruption of the measurement for several seconds is not considered to be a major problem in a clinical environment.

9.9.7 Heart rate and respiratory rate estimation

Figure 9.29 shows the estimation of heart rate and respiratory rate for a period during which the subject appears to change repeatedly from sleeping to waking states and back again. Vital sign values during this video segment are highly variable, with heart rate varying from 140 to 200 breaths/min and respiratory rate varying from 40 to 100 breaths/min. During the periods for which the subject is sleeping, the estimation of both heart rate and respiratory rate from the video camera agrees well with the reference values. The estimation of both heart rate and respiratory rate cannot be performed when the subject is awake as a result of motion artefacts. As the proportion of respiratory rate estimates is lower than that of heart rate estimates, this result suggests that the estimation of respiratory rate is more affected by motion than the estimation of heart rate.

9.9.8 Challenges in respiratory rate estimation in the clinic

The following challenges are observed during the estimation of respiratory rate in pre-term infants in the Neonatal HDU study. This information could be used to improve the respiratory rate estimation algorithm.

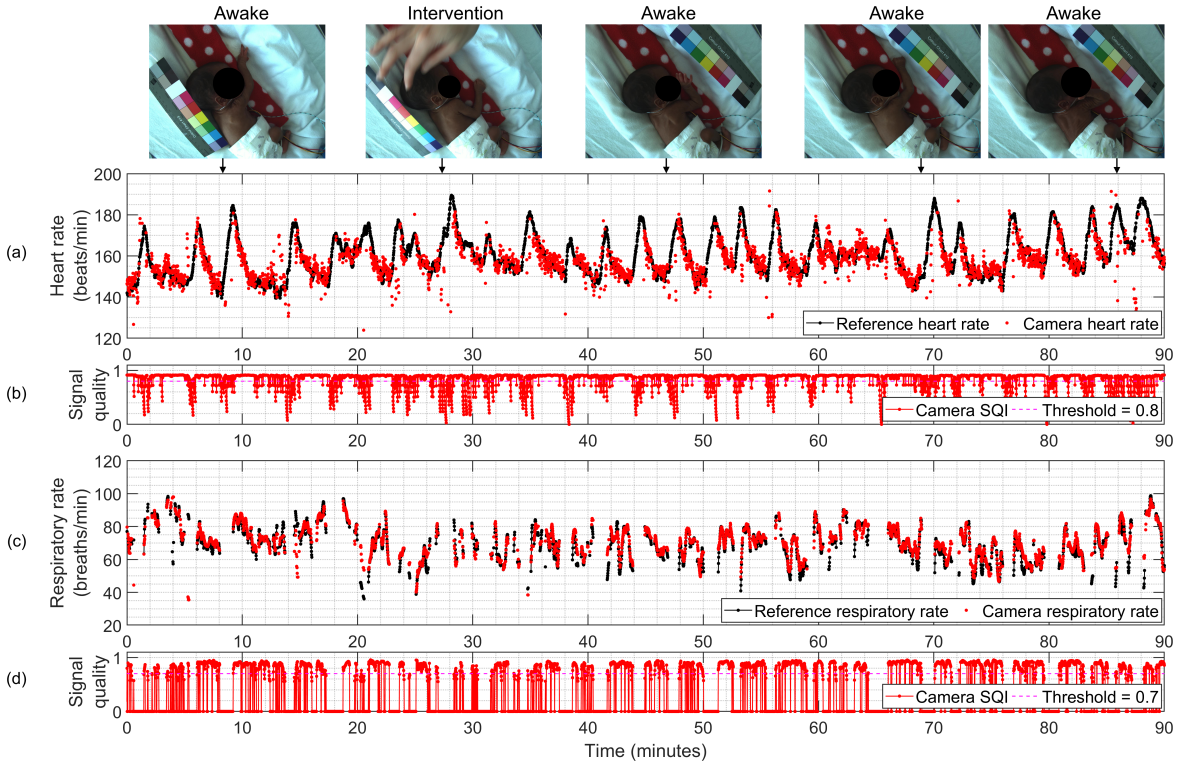


Figure 9.29: Estimation of heart rate and respiratory rate in a challenging scenario. The estimation is performed on a 90-minute video segment during which the subject appears to repeatedly be altering between the sleeping and waking states. During the transition from sleep to wake, the heart rate gradually increases for approximately three minutes. Unfortunately, the movements of the head and hands prevent the estimation of heart rate and respiratory rate. The estimation cannot be performed during sleep for which heart rate gradually decreases. Respiratory rate can be estimated when the respiratory motion is more pronounced, e.g. no head and hand movements. Note that the annotations on the top do not include all awake periods.

Breathing patterns

The respiratory signal is typically non-sinusoidal. Breathing can be slow and shallow or fast and deep. In pre-term infants, the immature mechanisms of the brain controlling the respiratory rhythm can result in irregular breathing patterns. Breathing patterns in pre-term infants are mainly spontaneous. Breathing can stop briefly and the patterns can change very rapidly. Unusual breathing patterns present a challenge to the respiratory rate estimation algorithms presented in this chapter. They may lead to the over- or under-detection of breaths and consequently incorrect breathing rate estimates. Changes in the trends in respiratory rate also present a considerable challenge as they are often associated with changes arousal state (sleeping, awake or alert) and the movement of the upper body and hands (see figure 9.29). Further investigations are needed to assess the robustness of algorithms in different breathing scenarios.

Non-respiratory motion artefacts

The movement of the upper body and hands often causes severe intensity variations in the respiratory signals. These variations are typically stronger than the respiratory-synchronous motion. In this

work, both digital filters and signal quality measures are used to suppress and decrease these variations. As a result, estimation errors are lower than those reported in previous studies but at the cost of a lower proportion of estimated time. The algorithms could be improved by adding a step to reduce non-respiratory related distortions in the respiratory signal.

Clothing

The current method for estimating respiratory rate relies solely on the skin regions extracted by the deep learning algorithm. This method would not work when the majority of the subject's skin is covered by a blanket. Fortunately, this problem affected only a small portion of recording time, since all subjects in the Neonatal HDU study had their upper body naked. In practice, the estimation of respiratory rate could also include features from the movement of the upper body that do not rely on the segmentation of the skin. An algorithm for body part detection could be used to extract local movement related to respiration.

9.10 Conclusion

Respiratory rate is an important indicator of serious adverse events [23]. However, respiratory rate estimation is very difficult as a result of the non-stationary nature of the respiratory signal and the frequent motion artefacts [246]. This chapter has presented a technique for estimating respiratory rate in pre-term infants. The methods work by combining respiratory rate estimates from multiple respiratory signals extracted from video camera data. They can compensate for motion artefacts and missing data, hence providing robust respiratory rate estimation. In the test set, the mean absolute error (MAE) between the reference respiratory rate and the camera-derived respiratory rate is lower than 3.6 breaths/min for over 78% of the time. This chapter also showed sample cases during which respiratory rate could not be estimated and suggested several ideas to improve the estimation of respiratory rate. The findings presented in this chapter are of importance to the development of physiological monitoring algorithms for these patients.

Chapter 10
Towards pain assessment using video cameras

10.1 Introduction

Pre-term infants are exposed to painful clinical procedures several times a day during hospitalisation in a neonatal intensive care unit (NICU). Several examples of painful procedures include intravenous cannulation (inserting a small tube into a peripheral vein to administer fluids or medications), taking blood samples for analysis and lumbar puncture to collect spinal fluids and surgery. Clinically, doctors and nurses assess pain in pre-term infants using multidimensional pain assessment tools which involve measures of physiological changes (heart rate, respiratory rate and oxygen saturation) and behavioural indicators (facial expression and body activities) in response to painful procedures. Many pain assessment tools exist and each has specific conditions for its use (e.g. gestational age and type of pain). The use of pain assessment tools may pose challenges to medical professionals in the clinic. Many tools require a large amount of training before one is qualified to use and then interpret the results. In practice, the assessment of pain is difficult as it is subjective and vulnerable to observer bias [247]. It may lead clinicians to overlook the presence of pain, and later conclude that pain management is not needed.

The purpose of this study is to investigate the use of a video camera to measure the physiological responses of pre-term infants to a heel prick procedure, a routine procedure used to collect blood samples from infants. This procedure appears to cause distress and pain in infants [248, 249].

In this chapter, we first investigate the feasibility of estimating the vital signs of pre-term infants using a video camera during a heel prick procedure. We then evaluate the possibility of assessing the physiological changes and variability of vital signs during each phase of the heel prick procedure (baseline, handling, heel stick, squeeze and recovery phases). The non-contact monitoring algorithms presented in chapters 7 and 9 are modified to estimate heart rate and respiratory rate from the video camera data during the heel prick procedure. Together with computer vision techniques for

detecting pain-related behavioural responses, this would form the basis of a method for automatic pain assessment using entirely video camera data.

This chapter is organised as follows. Section 10.2 provides relevant background regarding pain assessment in pre-term infants. Section 10.3 describes the heel prick sessions investigated in this work. Section 10.4 explains the methods used to extract heart rate and respiratory rate from video camera data in this context. Section 10.5 provides results regarding physiological responses to a heel prick procedure. Section 10.6 discusses the results and gives recommendations for future research. Section 10.7 concludes the chapter.

10.2 Pain assessment in pre-term infants

Pain is an unpleasant sensory and emotional phenomenon. Despite the common response to painful stimuli in human beings, pain experience varies according to individual physiology, personal experience and its context [250]. In infants, biological development causes the experience of pain to change over time, from pre-term to full-term [251] and over the first years of life [250].

For decades, pain management was not provided to pre-term infants during clinical interventions [252]. It was generally believed that pre-term infants did not experience pain because of the immaturity of their brain [253]. In the last two decades, pain management in pre-term infants has improved considerably. A number of pain assessment tools have been developed to assist nurses and doctors in providing effective pain management to neonates [254]. Today, much work is devoted to the study of neonatal pain and the improvement of the quality of pain management for infants [255].

10.2.1 Pain perception in infants

Figure 10.1 summarises the development of pain perception pathways. The neural pathways for pain arise from cutaneous sensory receptors to the cerebral cortex [256, 257]. Cutaneous sensory receptors

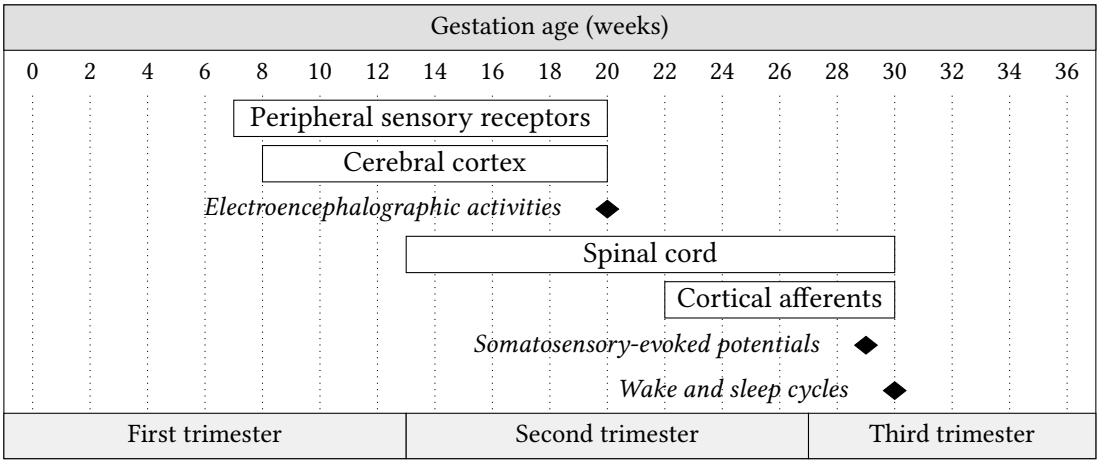


Figure 10.1: Neurodevelopment of pain perception pathways. The anatomical and functional components related to pain perception continuously develop throughout the gestational period. The symbol ◆ indicates when neural activity starts to appear. (Diagram created with the data from [256–258])

start to develop during the 7^{th} week of gestation, and they continue to spread to all cutaneous and mucosal surfaces by the 20^{th} week. The spread of sensory receptors is induced by the development of synapses between the spinal cord and sensory fibres, which begins at the 6^{th} week. The development of various cell bodies in the spinal cord first appears around the 13^{th} week and matures by the 30^{th} week. The connections to the cortex are believed to develop between the 21^{st} and 30^{th} week [258]. Pre-term infants born after 30 completed weeks have the anatomical, physiological and neurological components to perceive pain. The degree of pain that pre-term infants are able to tolerate is related to gestational age [254]. The physiological and behavioural reactivity to pain may reflect the organisation of pain perception pathways and the capacity of the central nervous system [251].

10.2.2 Heel prick procedure

Heel prick is a routine procedure used to obtain blood samples for laboratory examination in infants. The procedure involves piercing the skin of the heel with a lancet and squeezing gently until enough blood is obtained. This procedure is the main way to collect blood from infants in the clinic. In sick pre-term infants, the heel prick procedure is used to carry out daily blood tests (such as blood count and electrolytes) to check the progress of the infants. This procedure has been extensively used in research as a model situation for investigating the pain response in infants, developing pain assessment tools and studying the effectiveness of pharmacological interventions [249].

10.2.3 Physiological and behavioural indicators

Pre-term infants who experience pain may have changes in their vital signs and show changes in facial expression, vocalisation and body posture. Table 10.1 summarises common physiological and behavioural indicators of pain in pre-term infants [254]. The applicability and usefulness of each indicator differs according to the type of pain (e.g. procedural pain and postoperative pain) and the maturity of the infant. These indicators are not assessed individually but as part of multi-dimensional pain scoring system.

Physiological indicators

Physiological responses to pain in pre-term infants include an increase in heart rate [259–261], respiratory rate [262, 263], blood pressure [261] and intracranial pressure (ICP) [260] and a decrease in oxygen saturation [259, 260]. Pre-term infants exhibit similar physiological patterns to those of full-term infants [260]. Although several physiological indicators have been shown to be associated with pain, the following three indicators are commonly used for assessing pain in clinical practice: heart rate, respiratory rate and oxygen saturation. These three indicators are generally the easiest to monitor and the most accessible at the bedside [254].

Table 10.1: Physiological and behavioural indicators of pain in pre-term infants

Indicator	Assessment
Physiological indicators:	
Heart rate, respiratory rate, and blood pressure	Increase from baseline
Peripheral oxygen saturation	Decrease from baseline
Behavioural indicators:	
Infant state	Sleep/wake state and state of arousal
Facial expressions	Brow bulge, eye squeeze, nasolabial furrow, open lips, mouth stretch, lip purse, taut tongue and chin quiver
Movement	Flexion and extension of legs and arms, hand on face and hand guarding
Postures	Fisting, finger splay, arching and limb rigidity
Crying	Increase in frequency, amplitude, or duration; decrease in latency

Stevens *et al.* [248] observed the physiological responses of pre-term infants across different phases of a heel prick procedure (baseline, heel warming, heel stick and squeeze phases). The authors found that heart rate, oxygen saturation and intracranial pressure changed significantly between the baseline and the more invasive phases (heel stick and squeeze). The authors also found that the physiological responses to pain were influenced by behavioural state (quiet sleep, active sleep, quiet awake and active awake) but not the severity of illness. The authors suggested that physiological responses lack specificity to pain perception. They changed not only in response to stimulation, but also according to other physiological and behavioural events (e.g. touch by parents or clinical staff) [248]. The magnitude of physiological changes also differs from individual to individual and from procedure to procedure [248, 254, 259]. Despite these challenges, physiological indicators are the basis of pain assessment in infants hospitalised in the ward. They are generally used in conjunction with other contextual and behavioural indicators to measure pain.

Behavioural indicators

Behavioural indicators provide important contextual information for pain assessment. Facial expressions have been utilised in most pain assessment tools [24, 27, 264]. They have been proven to be the most reliable measures for pain assessment. Facial expressions related to pain include brow bulge, eye squeeze, nasolabial furrow, open lips, mouth stretch, lip purse, taut tongue, and chin quiver [259, 260, 265, 266].

Motor behaviours are regarded as strong indicators of pain. Grunau *et al.* [267–269] evaluated the body activities of infants undergoing painful procedures. The body activities were interpreted using

the Neonatal Individualised Developmental Care and Assessment Program (NIDCAP) [268] which is a standardised behavioural observational tool for infants. The studies reported that the frequencies of occurrence of the following 8 motor behaviours substantially increased in response to heel prick: flex arms, extend arms, flex legs, extend legs, finger splay, fisting, hand on face and frown. On the other hand, the frequencies of occurrence of the following 5 motor behaviours significantly decreased in response to heel prick: twitch extremities, twitch body, twitch face, foot clasping and mouthing.

Crying with high-pitched sound has been proposed as a vocalisation indicator of pain in infants [254]. In very pre-term infants, however, crying may not be a valid measure. They usually require mechanical ventilation to assist their breathing, so they are not able to vocalise.

Modifying factors

Apart from the physiological and behavioural indicators, there exist modifying factors that affect the degree of the pain experienced by infants. The response to a painful procedure varies according to gestational age, behavioural state and severity of illness. Infants born at lower gestational age exhibit greater physiological and less behavioural responses to pain compared to that of infants born at a later stage [254, 270]. The infant's behavioural state is an important contextual variable that influences the ability of an infant to respond to environment and painful stimuli [254]. During sleep or under anaesthetics, infants often exhibit low indicative behaviours to pain. Severity of illness has been reported to influence crying and behavioural responses to pain [271]. Early pre-term and sick infants have been shown to have shorter crying duration and less behavioural responses to pain.

10.2.4 Pain assessment tools

From all indicators of pain presented in the previous section, several indicators have a strong relationship to pain; others may not accurately reflect the pain that infants experience [254]. Several pain assessment tools have been proposed in the literature. These tools were developed by combining a set of suitable physiological and behavioural indicators to produce a single pain index.

Premature Infant Pain Profile (PIPP)

The Premature Infant Pain Profile (PIPP) is most widely adopted in the NICU and research studies [27]. The PIPP employs two infant characteristics (gestational age and behavioural state), two physiological indicators (heart rate and oxygen saturation) and three facial actions (brow burge, eye squeeze and nasolabial furrow). The facial indicators are measured by the amount of time that the infant shows certain facial expressions. PIPP has been extensively validated for assessing acute and procedural pain across gestational age groups from pre-term to term infants [254]. Table 10.2 shows the scoring matrix for the PIPP.

Table 10.2: Premature Infant Pain Profile (PIPP)

Indicator	0 point	1 point	2 points	3 points
Gestation in weeks	≥ 36	$32-35$	$28-31$	< 28
Behavioural state	Active/Awake	Quiet/Awake	Active/Sleep	Quiet/Sleep
Heart rate	$\uparrow 0-4$ bpm	$\uparrow 5-14$ bpm	$\uparrow 15-24$ bpm	$\uparrow \geq 25$ bpm
Oxygen Saturation	$\downarrow 0-2.4\%$	$\downarrow 2.5-4.9\%$	$\downarrow 5-7.4\%$	$\downarrow \geq 7.5\%$
Brow bulge	None	Minimum	Moderate	Maximum
Eye squeeze	None	Minimum	Moderate	Maximum
Nasolabial furrow	None	Minimum	Moderate	Maximum

Scores > 6 indicate moderate pain and > 12 indicate severe pain.

Neonatal Infant Scale (NIPS)

The Neonatal Infant Scale (NIPS) is a multidimensional tool for assessing procedural pain in preterm and full-term infants [264]. It requires an assessment of overall facial expression (relaxed or contracted), breathing patterns, crying, leg movement, arm movement and state of arousal. Each indicator is scored 0 or 1, except for cry which is scored 0, 1 or 2 (see table 10.3). The total score indicates the severity of pain. NIPS has been shown to have high interrater reliability and high internal consistency across all six components [264].

Table 10.3: Neonatal Infant Pain Scale (NIPS)

Indicator	0 point	1 point	2 points
Facial expression	Relaxed	Contracted	–
Cry	Absent	Mumbling	Vigorous
Breathing	Relaxed	Different from baseline	–
Arms	Relaxed	Flexed/Stretched	–
Legs	Relaxed	Flexed/Stretched	–
State of arousal	Sleep/Calm	Uncomfortable	–

Scores ≥ 4 indicate the presence of pain.

10.3 Dataset

In our feasibility study, we focus on four heel pricks in the Neonatal HDU dataset. In these heel pricks, a green-light indicator was placed near the infant to indicate the time when a lancet pierced the skin of the heel to withdraw blood.

For each heel prick, a 20-minute video segment was extracted, comprising of 10 minutes before and after the heel prick. The heel prick time was defined as the time when a lancet pierced the skin of the heel, as indicated by the green-light indicator in the video recording.

10.3.1 Dataset demographics

Each heel prick was administered to a different infant. Table 10.4 shows the patient demographics for each heel prick. Three out of the four infants were female. Two of the infants were White, one

Table 10.4: Patient demographics of each heel prick session

Pain No.	Session	Gender	Ethnicity	Gestational age (weeks)	Weight (grams)
HP1	PN13 Day 2	Female	Mixed White and Asian	28.1	960
HP2	PN14 Day 3	Female	White	29.7	1,430
HP3	PN23 Day 3	Female	Mixed White and Black	30.3	940
HP4	PN26 Day 4	Male	White	34.7	1,210

was mixed White and Asian and one was mixed White and Black. The gestational age ranged from 28.7 to 34.7 weeks. The weight at the first day of the study ranged from 940 to 1,430 grams.

10.3.2 Heel prick phases

The heel prick procedure was divided into five phases (see figure 10.2). The first phase or baseline phase was a one-minute period before the first contact by clinical staff. The second phase or handling phase was a period during which clinical staff were preparing the infant for heel prick. Clinical activities in this period included repositioning the infant, attaching additional sensors and cleaning the heel. The third phase or heel stick phase is a 30-second period after the lancet pierced the skin of the heel. No heel squeezes were performed during this period. The fourth phase or heel squeeze phase was a period during which clinical staff squeezed the heel to draw blood to a blood collection tube. This period started from when the first squeeze was made by clinical staff until when the end of blood collection at which point the infant's heel was wrapped with gauze. Clinical staff normally squeezed the heel several times until enough blood was drawn. Finally, the last phase or recovery phase was a two-minute period after the last contact by clinical staff.

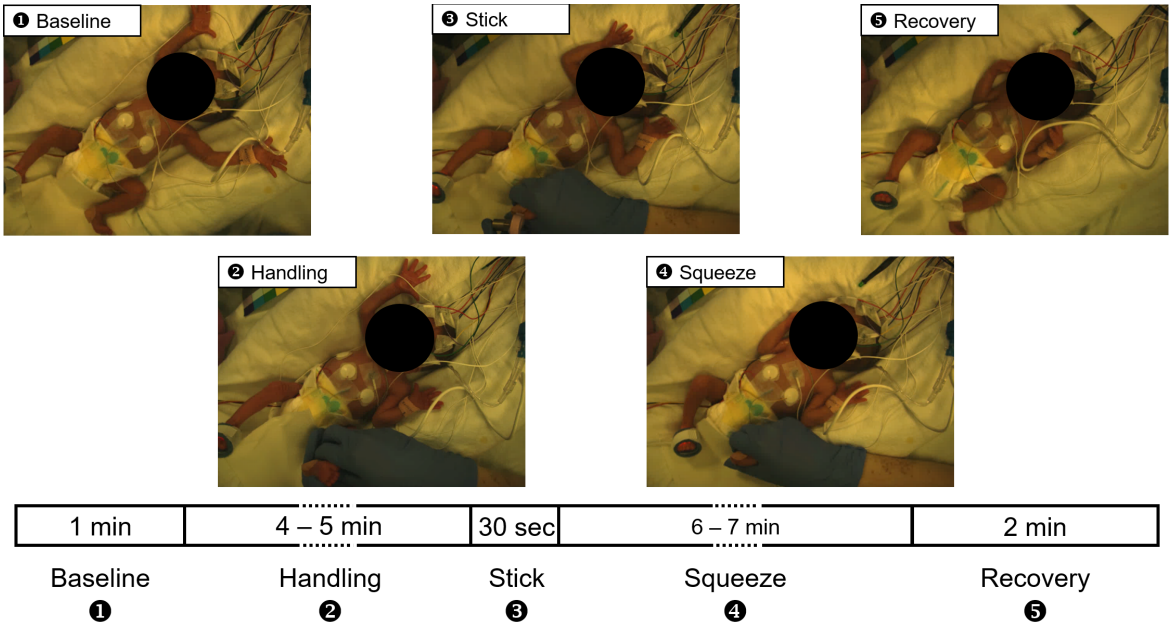


Figure 10.2: The heel prick procedure was divided into 5 phases: baseline (1 minute), preparing (5-6 minutes), heel stick (30 seconds), heel squeeze (6-7 minutes) and recovery (2 minutes).

10.4 Vital sign estimation

This work focuses on the estimation of heart rate and respiratory rate. The estimation of heart rate and respiratory rate is based on the work presented in chapters 6–8 with some modifications to make the algorithms suitable for estimating vital signs under the more challenging conditions.

10.4.1 Reference vital signs

The gold-standard reference heart rate computed in chapter 6 is used to compare with the camera-derived heart rate estimates in this chapter. The reference heart rate is calculated using the mean of two heart rate measurements, one from the electrocardiography (ECG) and the other from the photoplethysmography (PPG), provided by the Philips patient monitor during periods for which the two values do not differ by more than 5 beats/min and the ECG signal is of good quality.

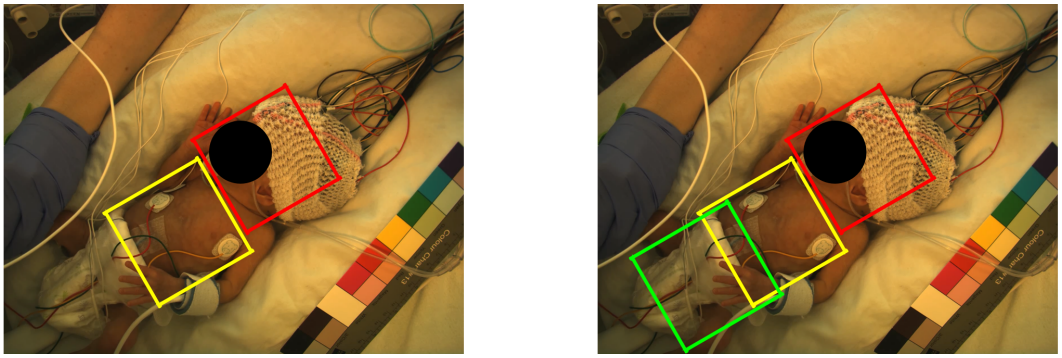
The reference respiratory rate is computed from the impedance pneumography (IP) signal. This differs from the method described in chapter 8 in which the reference respiratory rate was computed from both the IP and ECG signal. In this work, we discard the use of the ECG signal in the computation of the reference respiratory rate as the respiratory component in the ECG signal is not prominent due to the infant's movements during the heel prick procedure. The computation of the reference respiratory rate from the IP signal is performed in the same manner as the description in sections 8.5–8.7 of signal filtering, peak and onset detection and signal quality assessment. The reference respiratory rate is then estimated using the beat counting technique.

10.4.2 Extraction of the PPGi signals

The algorithm presented in chapters 7 and 9 were designed to perform the estimation of vital signs automatically by taking the entire skin regions. The photoplethysmographic imaging (PPGi) signals extracted from the entire skin regions are inherently noisy as a result of infant movement.

In our previous work [272], we demonstrated the benefits of estimating heart rate from multiple regions of interest (ROIs) using data fusion techniques. The heart rate estimates combined from multiple ROIs (e.g. entire skin, head and torso) resulted in lower mean absolute errors and higher estimated time, compared to the heart rate estimates from only the entire skin regions.

In this work, instead of using the entire skin regions, three PPGi signals are extracted from manually fixed ROIs centred at the head, upper torso and lower torso respectively. The size of ROIs is 200×200 pixels. All ROIs are fixed across the 20-minute video segment. The extraction of PPGi signals is similar to what was described in section 7.3.1. The PPGi signal is computed by spatially averaging skin pixels inside the head, upper torso and lower torso ROIs (see figure 10.3b) in the green colour channel. This method prevents the inclusion of legs and arms as well as the skin area of clinical staff to the ROI. Hence, noise in the PPGi signal is reduced.



(a) (b)

Figure 10.3: Regions of interest (ROIs) for estimating heart rate and respiratory rate. (a) Heart rate estimates are obtained from the data fusion of the head ROI (red square) and upper torso ROI (yellow square). (b) Respiratory rate estimates are obtained from the data fusion of the head ROI (red square), upper torso ROI (yellow square) and lower torso ROI (blue square).

10.4.3 Estimation of vital signs

Heart rate estimation is performed using data fusion from the head and torso ROIs (see figure 10.3a). Each raw PPGi signal is first detrended and filtered using a 60th-order band-pass Finite Impulse Response (FIR) filter with cut-off frequencies of 2.3 Hz (140 beats/min) and 3.5 Hz (210 beats/min). The choice of the filter is determined by the range of heart rate in all heel prick sessions. The signal quality assessment is performed using the algorithms described in section 7.4. Heart rate is estimated using an autoregressive model with model order 10 (see section 7.5.3). The estimation is performed over a window length of 10 seconds with a step size of 1 second. Finally, heart rate estimates from the head and upper torso ROIs are combined using data fusion with multiple Kalman filters to produce the final heart rate estimates (see section 8.8.3).

Respiratory rate estimation is performed using data fusion from the head, upper torso and lower torso ROIs (see figure 10.3b). The latter two ROIs tend to have a strong respiratory component since the breathing of pre-term infants is more prominent in the stomach area [241]. The respiratory rate estimation is performed in a similar manner to the heart rate estimation. Each PPGi signal is first detrended and filtered using a cascade of a 120th-order low-pass FIR filter with a cut-off frequency of 2.0 Hz (120 breaths/min) and a 200th-order high-pass FIR filter with a cut-off frequency of 0.3 Hz (20 breaths/min). The assessment of signal quality is performed using the algorithms described in section 9.6. Respiratory rate is estimated using an autoregressive model with model order 8 (see section 7.5.3) over a window length of 10 seconds with a step size of 1 second. Finally, once respiratory rate has been computed for every ROI (head, upper torso and lower torso ROIs), the data fusion technique with multiple Kalman filters (see section 8.8.3) is applied to produce the final respiratory rate estimates.

10.5 Results

The estimation of heart rate and respiratory rate is performed for each video segment (20 minutes). The camera-derived estimates are plotted against the reference vital signs. The results are reported both over the course of the video segment and over the five phases of the heel prick procedure.

10.5.1 Vital sign estimation during the heel prick procedure

Figure 10.4 shows the heart rate and respiratory rate estimation results for heel prick session HP1. In this session, the infant is actively awake before the heel prick procedure, so heart rate is particularly high during the baseline phase. During the handling phase, the heart rate continuously decreases. In comparison to the moment before the heel prick, the heart rate suddenly increases right after the lancet pierces the skin of the heel. The heart rate and respiratory rate also increases after the first squeeze. After the blood is drawn and clinical staff wrap the heel with gauze, the heart rate and respiratory rate decrease during in the recovery phase.

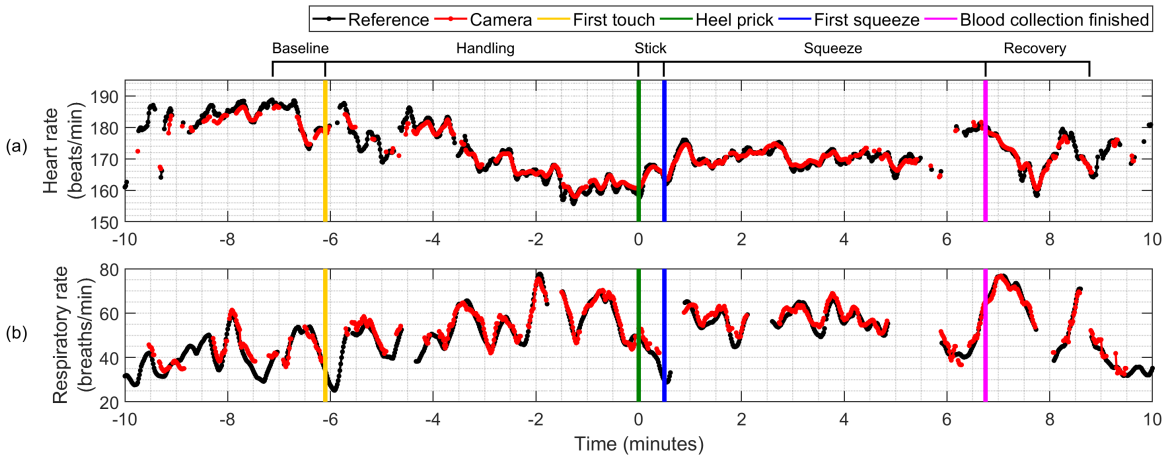


Figure 10.4: Vital sign estimation results for the heel prick session HP1. (a) Heart rate estimates with each heel prick phase indicated at the top of the plot. (b) Respiratory rate estimates.

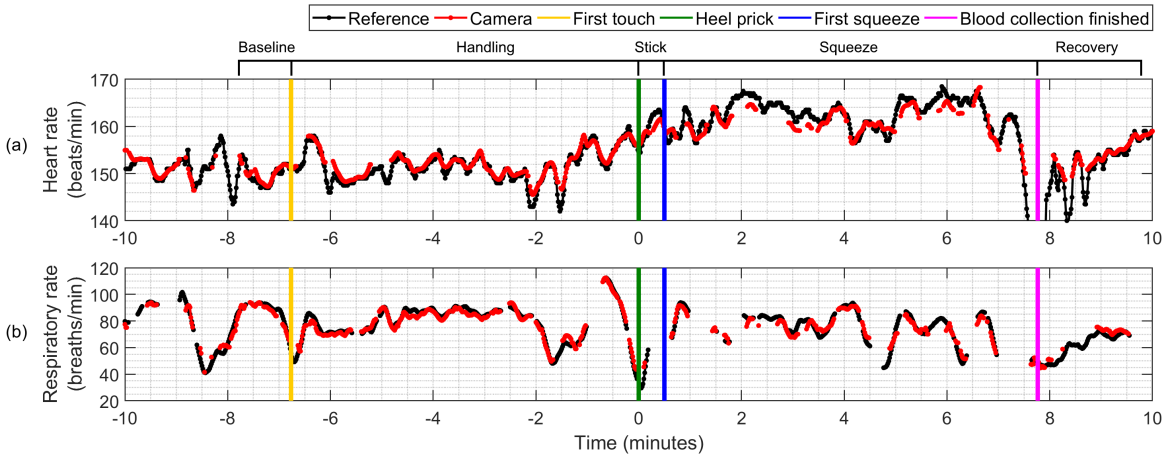


Figure 10.5: Vital sign estimation results for the heel prick session HP2. (a) Heart rate estimates with each heel prick phase indicated at the top of the plot. (b) Respiratory rate estimates.

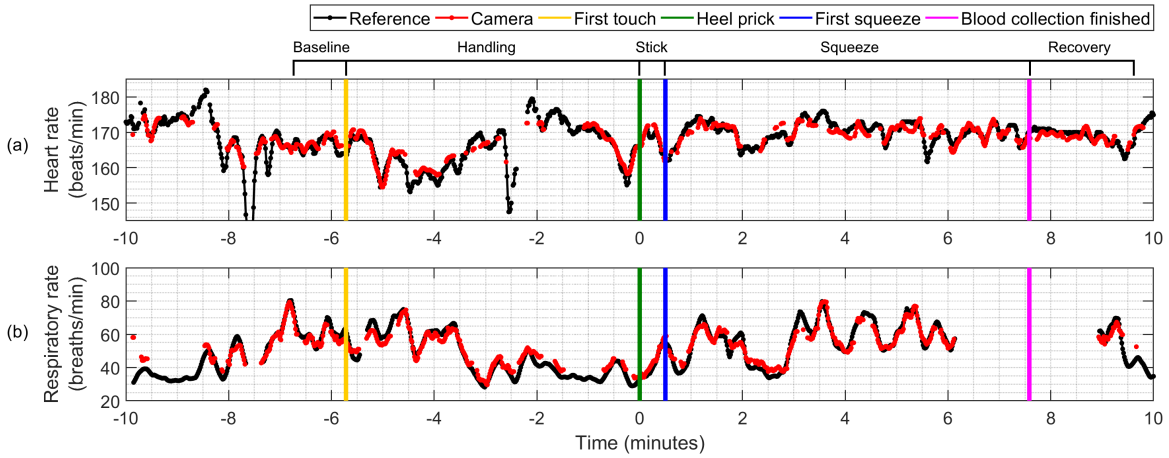


Figure 10.6: Vital sign estimation results for the heel prick session HP3. (a) Heart rate estimates with each heel prick phase indicated at the top of the plot. (b) Respiratory rate estimates.

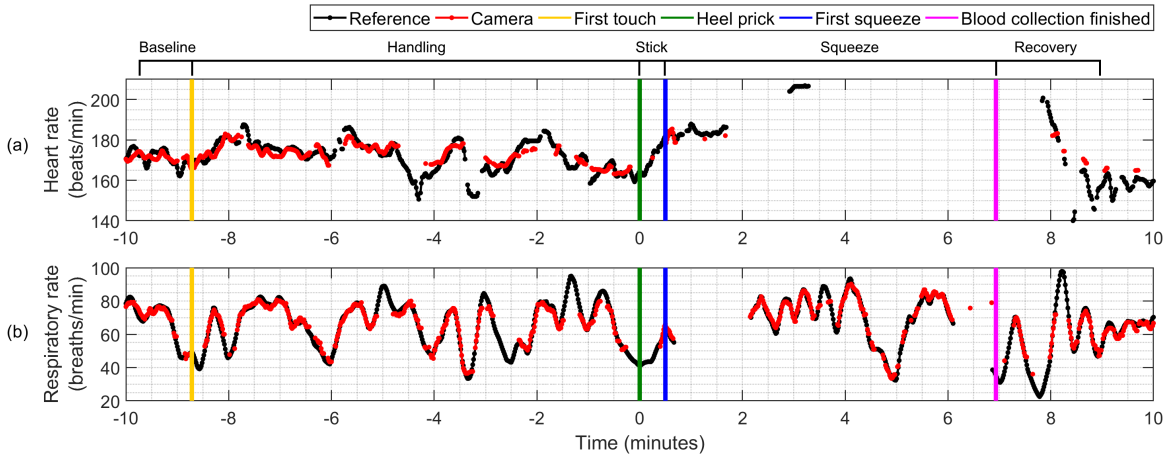


Figure 10.7: Vital sign estimation results for the heel prick session HP4 with each heel prick phase indicated at the top of the plot. (a) Heart rate estimates. (b) Respiratory rate estimates.

Figure 10.5 shows the heart rate and respiratory rate estimation results for heel prick session HP2. In this session, the infant is asleep throughout the heel prick procedure. During the handling phase, heart rate and respiratory rate increase slightly with respect to the baseline phase. Sudden increases in heart rate from the baseline and handling phases occur during the heel stick and squeeze phases. The respiratory rate also increases, in comparison to its value right before the heel prick. The heart rate and respiratory rate decrease back to baseline values after the blood collection has finished.

Figure 10.6 shows the heart rate and respiratory rate estimation results for heel prick session HP3. The infant is actively awake and crying at the beginning of the session (from minute -10 to minute -8). During the handling phase, the heart rate fluctuates and the respiratory rate decreases with respect to the baseline phase. Increases in both heart rate and respiratory rate are observed during the heel stick and squeeze phases. The infant is active and awake during the recovery phase.

Figure 10.7 shows the heart rate and respiratory rate estimation results for heel prick session HP4. The infant is awake during the heel prick procedure, and crying during the squeeze phase.

The pulse oximetry became loose as a result of his crying, and so the reference heart rate is not available for some parts of the heel squeeze and recovery phases. Increases in heart rate and higher fluctuations in respiratory rate are observed during the heel stick and squeeze phases, compared to those during the baseline and handling phases. Similar to the other sessions, a decrease in heart rate and respiratory rate were observed in the recovery phase.

10.5.2 Physiological responses across heel prick phases

The maximum, mean and standard deviation (SD) are calculated for both heart rate and respiratory rate over the five phases of the heel prick procedures (baseline, handling, heel stick, heel squeeze and recovery) for all four heel prick sessions. Table 10.5 summarises the physiological changes for each phase of the heel prick procedure.

Figure 10.8 shows the maximum heart rate and the maximum respiratory rate for the different phases of the heel prick procedure for each heel prick session. Figure 10.9 shows mean heart rate and respiratory rate for the different phases of the heel prick procedure for each heel prick session. In both figures, the camera-derived vital-sign estimates generally agree with the reference vital-sign values. There are mixed responses during the handling phase. The vital signs can increase or decrease depending on the state of arousal (sleep or awake) and the effects of handling by clinical staff. The maximum and mean of both heart rate and respiratory rate in the squeeze phase are found to be higher than those in the stick phase. Figure 10.10 shows the standard deviation of

Table 10.5: Physiological changes for the different phases of the heel prick procedure

Session	Baseline phase		Handling phase		Stick phase		Squeeze phase		Recovery phase	
	Mean (SD)	Max	Mean (SD)	Max	Mean (SD)	Max	Mean (SD)	Max	Mean (SD)	Max
Heart rate (beats/min)										
HP1 Ref	182.7 (4.9)	188.5	170.3 (8.3)	187.7	165.1 (3.4)	168.1	170.9 (3.6)	181.3	171.3 (5.5)	180.6
Cam	179.6 (4.0)	186.8	169.4 (7.5)	184.2	165.3 (2.2)	167.5	170.5 (3.2)	181.7	171.0 (5.1)	178.6
HP2 Ref	149.8 (2.0)	154.0	151.5 (3.3)	160.0	160.1 (3.0)	163.5	161.9 (4.9)	168.5	148.8 (13.9)	159.0
Cam	150.0 (1.7)	153.8	152.2 (2.9)	158.4	159.2 (1.7)	161.5	160.9 (3.0)	168.3	153.8 (2.3)	158.3
HP3 Ref	166.3 (1.8)	169.0	165.2 (6.5)	179.5	168.3 (1.5)	170.9	170.0 (3.0)	176.0	169.0 (1.9)	171.5
Cam	166.9 (1.7)	170.1	164.9 (5.0)	172.8	168.7 (2.4)	172.2	170.0 (2.4)	173.9	168.9 (1.1)	171.7
HP4 Ref	170.3 (3.7)	176.0	171.8 (7.5)	187.5	170.5 (6.2)	181.3	189.3 (10.0)	206.9	169.6 (19.1)	200.8
Cam	171.1 (1.7)	174.6	172.8 (4.8)	183.0	171.1 (0.0)	171.1	182.0 (2.3)	185.4	176.9 (5.7)	183.2
Respiratory rate (breaths/min)										
HP1 Ref	42.7 (10.4)	53.8	53.7 (9.1)	77.6	37.5 (7.5)	56.0	54.4 (8.3)	76.6	55.8 (13.0)	76.8
Cam	45.9 (4.9)	53.8	56.4 (8.0)	75.5	48.5 (5.2)	55.8	57.5 (6.7)	76.4	57.4 (11.6)	76.7
HP2 Ref	88.2 (8.0)	95.0	79.1 (13.7)	112.4	62.2 (20.9)	88.6	71.2 (14.5)	93.8	61.7 (8.9)	72.8
Cam	87.2 (8.2)	94.6	80.1 (11.9)	112.4	73.1 (11.1)	83.8	72.4 (11.3)	91.7	66.0 (10.0)	75.5
HP3 Ref	59.7 (4.4)	70.1	46.2 (12.9)	75.0	45.2 (8.1)	55.7	57.9 (11.3)	82.8	61.9 (9.0)	74.3
Cam	57.0 (3.6)	66.2	49.4 (10.2)	74.6	46.3 (7.8)	59.0	56.2 (9.4)	79.7	62.4 (5.1)	75.5
HP4 Ref	64.8 (11.9)	77.2	65.6 (13.9)	95.0	48.0 (6.4)	62.4	64.4 (17.3)	93.4	54.3 (20.1)	97.9
Cam	68.6 (9.3)	76.6	67.0 (10.7)	80.9	58.9 (4.6)	64.9	69.4 (13.7)	90.2	60.2 (9.2)	74.0

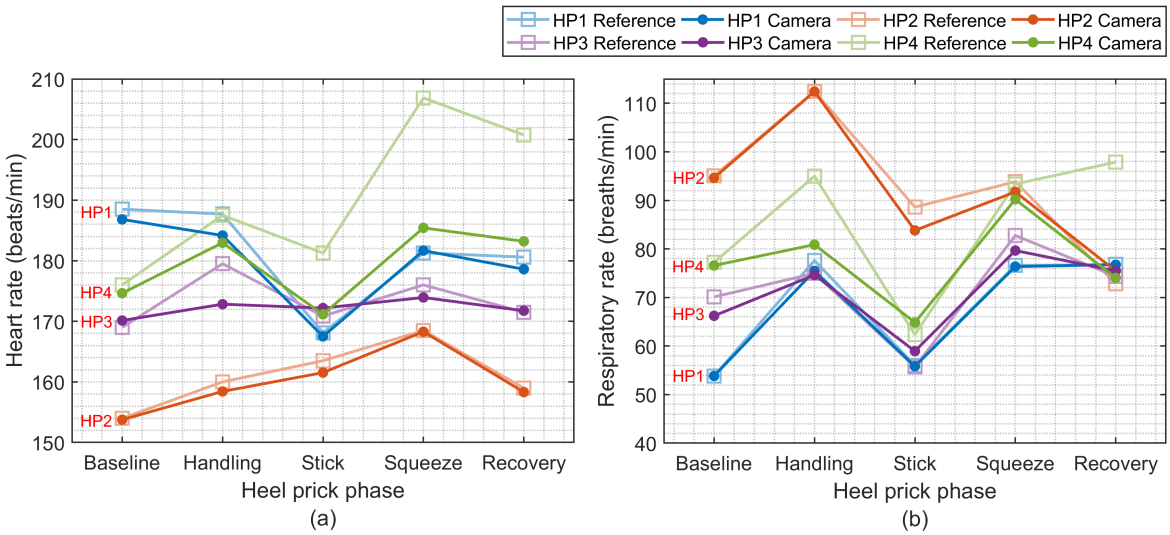


Figure 10.8: Maximum (a) heart rate and (b) respiratory rate for the different heel prick phases

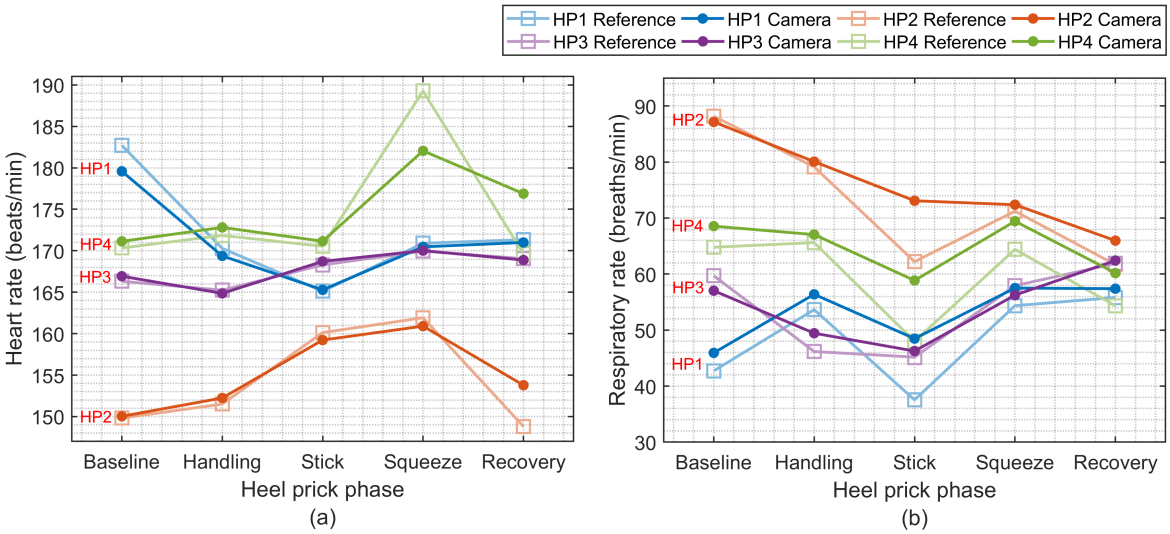


Figure 10.9: Mean (a) heart rate and (b) respiratory rate for the different phases of the heel prick procedure

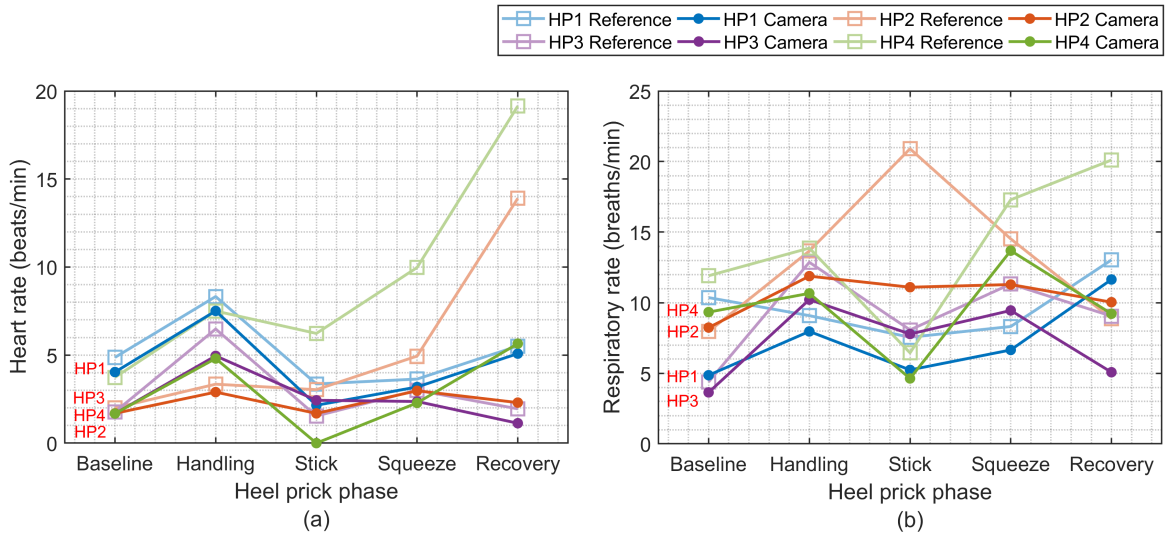


Figure 10.10: Standard deviation of (a) heart rate and (b) respiratory rate for the different heel prick phases.

the heart rate and the respiratory rate for the different phases of the heel prick procedure for each heel prick session. The standard deviation of the camera-derive estimates is observed to be lower than that from the reference vital signs.

10.6 Discussion

10.6.1 Non-contact vital sign estimation

Non-contact algorithms provide accurate estimates of vital signs that generally agree with the reference values over the course of each heel prick session. Even though the non-contact estimation could not be performed for certain time periods as a result of the infant's movements or shadow cast by the clinical staff, statistical variables (maximum, mean and standard deviation) can be calculated for each heel prick phase. The resulting values are also in agreement with the reference vital signs. An important step in the analysis is the suppression of unwanted low-frequency content by using prior knowledge of the suitable range of vital signs from the reference data. The manual selection of ROIs is also of importance. It allows the selection of regions that give the best vital sign estimates. For example, a cardiac component is prominent in a smooth skin area (e.g. the face), whereas respiration is prominent in the torso regions since pre-term infants normally have abdominal breathing [241]. The use of data fusion is found to improve the proportion of the time that vital signs can be estimated.

10.6.2 Physiological responses

Baseline and handling phases

Pre-term infants exhibit physiological responses over the course of the heel prick procedure. The infant state and the clinical procedures during handling (such as infant positioning, attaching sensors, heel warming and alcohol swab cleansing) affect the physiological responses of the infant during the handling phase [262, 273]. The handling by clinical staff is found to decrease the values of the vital signs for the infants who are actively awake and to increase the values of the vital signs for the infants who are asleep.

For session HP1, the infant was actively awake before the heel prick session. The procedures performed during the handling phase can be seen to reduce heart rate and increase respiratory rate. For session HP2, the infant was asleep throughout the session. Heart rate and respiratory rate during the handling phase increase from their values during the baseline phase. This pattern has also been observed in previous studies [248, 259, 260]. For session HP3, the infant was crying from the beginning of the session. The handling of the infant by clinical staff is found to calm the infant down and reduce both the heart rate and respiratory rate. For session HP4, the infant was awake. Fluctuations of both heart rate and respiratory rate are found during the handling period.

Heel stick and squeeze phases

Increases in heart rate and respiratory rate are present during both the heel stick and squeeze phases. These responses are consistent with previous studies [248, 259, 260, 262, 273]. The values of heart rate and respiratory rate during the heel squeeze phase are found to be higher than those during the heel prick phase. Non-contact techniques are able to capture vital-sign changes in most heel prick sessions.

In all four sessions, heart rate and respiratory rate increase during the heel stick phase, compared to the values at the moment before heel lancing. Both heart rate and respiratory rate increase further during the heel squeeze phase. Changes in heart rate and respiratory rate are obvious in some sessions. For session HP2, heart rate increases by more than 15 beats/min compared to baseline values (see figure 10.5). The increase in heart rate continues for more than seven minutes. For session HP4, heart rate increases by more than 25 beats/min, with the infant crying as a result of heel squeezing (see figure 10.7). Figures 10.8 and 10.9 show large errors when comparing camera-estimated heart rate with reference heart rate as a result of gaps caused by unavailable reference data and motion artefacts.

Recovery phase

After the squeeze phase, heart rate and respiratory rate decrease in all four sessions. However, for about 1 or 2 minutes after the squeeze phase, both values begin to increase again as a result of clinical staff removing sensors from the infant.

Impact of data fusion

When estimating the maximum and mean changes of vital sign values for the different heel prick phases (see figures 10.8 and 10.9 respectively), the camera-derived physiological values generally agree with the reference physiological values. Data fusion smooths the estimates of heart rate and respiratory rate with rapid changes being eliminated. The standard deviations of heart rate and respiratory rate measured by the video camera have the same trends as those measured by conventional technology, but their amplitudes are reduced. Maximum and mean changes in heart rate and respiratory rate are more discriminatory than standard variation changes in heart rate and respiratory rate when measured using the video camera.

10.6.3 Integration with behavioural responses

Behavioural responses play an important role in pain assessment. Infant state (asleep or awake) is known to influence the physiological response to pain [248]. The assessment of behavioural responses using video cameras could be performed using deep learning techniques, which have been shown to be capable of identifying facial features, recognising actions and estimating human body pose in images and videos [274]. These techniques could be used to estimate infant state, recognise facial expressions (e.g. brow bulge, eye squeeze and crying) and body movements (e.g.

arm and leg movements). Deep learning techniques could be used to interpret the motor behaviours of pre-term infants according to the NIDCAP tool [268]. The integration of non-contact vital sign algorithms and camera-based behavioural assessment tools would enable the assessment of pain in pre-term infants using video cameras.

10.6.4 Study limitations

The major limitation of this pilot study was the sample size, just four heel prick sessions, which may limit the generalisability of the results to a larger population. The design of the clinical study was not specifically to study the physiological and behavioural responses to a heel prick. It involved very long infant handling phase (5-6 minutes) and long heel squeezing phase (6-7 minutes). The physiological responses may become blurred during these lengthy phases.

In two out of four heel prick sessions, video recordings ended a few minutes after the heel prick procedure, so the study was limited to only 10 minutes of data after the heel prick. This prevented the estimation of the time required for vital-sign values to return to baseline values.

10.6.5 Recommendations for future research

The goal of this study was to investigate whether non-contact technologies could be used to estimate the vital signs of pre-term infants during a heel prick procedure. No attempt was made to assess pain or develop a new pain assessment tool using video camera data. Future studies should ensure even distribution across important variables such as gestational age, infant state, severity of illness and pain score. The effects of camera resolution and field of view should also be taken into account, so that the infant's face can be captured clearly within the field or view.

10.7 Conclusion

This chapter has presented a preliminary investigation of the use of a video camera for the estimation of heart rate and respiratory rate during a heel prick procedure. The camera-derived vital sign estimates were shown to be in agreement with the reference vital signs. The results from this exploratory study lay the foundation for the development of algorithms to derive objective pain scores in pre-term infants entirely from video camera data.

Chapter 11
Conclusion

11.1 Introduction

Non-contact vital-sign monitoring using video cameras enables the measurement of vital signs without sensors attached to the skin. Camera-based technologies are ubiquitous and provide high performance at low cost. They could be integrated into a patient ward, an in-home health monitoring system or a telemedicine system. In the clinic, non-contact technologies present numerous advantages in patient comfort and the management of skin irritation and infection. However, there are several challenges associated with the use of camera-based non-contact monitoring technologies in the clinic such as variable lighting conditions, patient movement, the automatic detection and segmentation of a patient in the video frame and the selection of suitable regions of interests (ROIs) from which vital signs can be extracted.

Over the past decade, most research has been performed in controlled environments where healthy subjects remained relatively still during video recording. The research conducted in this thesis introduced a framework that delivered the accurate and continuous measurement of heart rate and respiratory rate for pre-term infants in the clinic. The framework can cope with the challenges presented in a real hospital setting, to provide robust vital-sign estimates. It has been developed and validated on clinical data recorded from thirty pre-term infants hospitalised in the neonatal intensive care unit (NICU) of the John Radcliffe Hospital in Oxford, comprising a total video recording of 426.6 hours.

11.2 Thesis review

11.2.1 Clinical study

Chapter 3 introduced the Neonatal High-Dependency Unit (Neonatal HDU) study involving the double monitoring of 30 pre-term infants with a video camera and conventional patient monitoring devices in regular clinical care. This study was designed with the aim of developing non-contact

monitoring algorithms in an unconstrained clinical setting. Pre-term infants provide a wide spectrum of physiological values for validating the proposed algorithms.

The Neonatal HDU study data were acquired during more than 90 recording sessions (over 400 hours in total and nearly 400 terabytes of data storage). The study was approved by a Medical Research Ethics Committee (MREC), compliant with the Data Protection Act and adhered to the National Health Service (NHS) code of confidentiality.

It produced a large dataset of video data and reference signals collected from participants from a variety of backgrounds. The dataset can be used not only for the development and calibration of non-contact estimation algorithms but also for the development of algorithms for detecting important clinical events, e.g. neonatal apnoea and pain.

11.2.2 Patient detection and skin segmentation

Chapter 4 answered two important questions: is the patient in the scene? and if so, which pixels belong to skin? Skin pixels are used to extract the pulsatile information required for the estimation of vital signs.

The chapter presented a multi-task convolutional neural network (CNN) for the detection of the presence of the infant in the video frame and the subsequent segmentation of skin regions [275]. To generate ground truth data for training the network, a custom-built semi-automatic skin annotation tool was developed using a graph cuts algorithm with geodesic star convexity [153]. The network was trained using images labelled by three annotators. It demonstrated robustness with different skin colours and changes in the lighting conditions of the hospital ward. The network achieved an accuracy of 98.8% for patient detection and an average intersection-over-union (IOU) score of 88.6% for skin segmentation.

11.2.3 Intervention detection

Chapter 5 answered the final question before estimating vital signs in pre-term infants: is the infant undergoing a clinical intervention? Clinical staff perform routine clinical interventions several times a day, depending on the state of the infant. Clinical interventions can cause motion artefacts which may prevent the estimation of vital signs from video camera data.

The CNN model for patient detection and skin segmentation in chapter 4 was extended for intervention detection, requiring the processing of temporal information in the video data. Different network architectures and fusion strategies were explored. It was shown that the information already processed in the patient detection and skin segmentation network could be effectively used to assist the new intervention detection task. As part of the implementation, custom software was developed for labelling intervention periods in video recordings. The network was trained using

ground truth data provided by three annotators. The network achieved an accuracy of 94.5% in classifying intervention and non-intervention events.

The CNN models proposed in chapter 4 and 5 can be used to efficiently process a video recording to detect a patient, segment skin regions and identify time periods during which clinical intervention was undertaken.

11.2.4 Reference heart rate

Chapter 6 presented the analysis of the quality and accuracy of heart rate measurements provided by the Philips IntelliVue patient monitoring system and described the computation of a gold-standard reference heart rate. In the Neonatal HDU study, the Philips patient monitor provided two heart rate measurements: one derived from the electrocardiogram (ECG) and the other derived from the photoplethysmogram (PPG). This study revealed discrepancies between the two measurements possibly due to different device clocks, different proprietary algorithms, the intrinsic differences between ECG and PPG signals, physiological events, clinical interventions and motion artefacts.

The reference heart rate was computed from the mean of two heart rate measurements (from the ECG and PPG signals) during periods for which the two values did not differ by more than 5 beats/min (as recommended by the "ANSI/AAMI EC13:2002 Cardiac monitors, heart rate meters, and alarms" standard [176]) and the ECG signal was of good quality. The reference heart rate was available for 94.5% of the total recording time (403.1 hours out of 426.6 hours). The reference heart rate was later used to compare against the heart rate estimate from the video camera.

11.2.5 Heart rate estimation from a video camera

Chapter 7 proposed algorithms for the continuous estimation of heart rate from video-camera data recorded in the NICU. The robust framework presented in chapter 4 provided an accurate per-frame skin segmentation which could be used as a ROI for estimating physiological parameters.

The chapter described the extraction of the PPGi signal, the assessment of the quality of the signal and the estimation of heart rate. Four different techniques for heart rate estimation were compared. The autoregressive model technique with the optimal model order achieved the best results. The comparison between the reference heart rate and the camera-derived heart rate was performed on valid data during which (1) the reference heart rate was available; (2) the infant was present in front of the camera as identified by algorithms in chapter 4; and (3) the infant was not undergoing any clinical intervention as identified by the algorithms of chapter 5. The results showed a strong positive correlation and good agreement between the reference heart rate and the camera-derived heart rate, with a mean absolute error of less than 2.3 beats/min for over 76% of the valid recording time in the test set. Heart rate estimation was also possible across different skin tones. Estimation errors mostly occurred because of artefacts caused by subject motion. The errors

between the reference heart rate and the camera-derived heart rate were similar to the errors found when comparing two different conventional heart rate monitoring devices such as the ECG and PPG.

11.2.6 Reference respiratory rate

Chapter 8 described the computation of a gold-standard reference respiratory rate. The Neonatal HDU study employed impedance pneumography (IP) to record respiratory signals in pre-term infants. The IP signal is a standard technology for respiratory rate measurement in the clinic; however, it is susceptible to noise and motion artefacts [21, 210] arising from the movement of the upper body and irregular breathing patterns.

In this chapter, two new respiratory rates were calculated from the analysis of both IP and ECG signals using two different techniques: breath counting and autoregressive model. The reference respiratory rate was calculated using data agreement between the two respiratory rate estimates by computing the mean of the two estimates when both values did not vary by more than 5 breaths/min and their data were both of good quality. The reference respiratory rate was available for 44.3% of the total recording time (189.0 hours out of 426.6 hours). This finding is consistent with the previous studies in pre-term infants [77, 213].

11.2.7 Respiratory rate estimation from a video camera

Chapter 9 proposed robust algorithms for estimating respiratory rate continuously from video cameras. Respiratory rate estimation is very challenging due to the non-stationary nature of the respiratory signal in pre-term infants and frequent motion artefacts.

During each respiratory cycle, the chest and abdomen of the infant expands and then contracts. This phenomenon causes movement of the body and changes light reflection from the skin. Multiple respiratory signals were extracted from the colour changes of the skin area, the shape properties of the skin area and morphology analysis of the skin area. Respiratory rate was estimated for each respiratory signal. All respiratory rate estimates were then combined together using data fusion with multiple Kalman filters. The data fusion strategy was able partially to compensate for the problems of motion artefacts and missing data, which occur when respiratory rate is estimated using only one respiratory signal source. The results showed good agreement between the reference respiratory rate and the camera-derived respiratory rate, with a mean absolute error lower than 3.6 beats/min for over 78% of the valid recording time in the test set. The results are in line with previous studies in pre-term infants [212, 213].

11.2.8 Preliminary study of pain assessment

Chapter 10 presented a preliminary investigation of the feasibility of using camera-based non-contact monitoring algorithms to measure the physiological response of pre-term infants to a painful heel

prick procedure. By using prior knowledge of the range of vital-sign values from the reference data, it was possible to derive both heart rate and respiratory rate during the heel prick procedure, for each heel prick phase. The camera-derived vital sign estimates were consistent with the reference vital signs, showing that it would be possible to use a video camera to quantify the physiological response of pre-term infants in each phase of a heel prick procedure.

11.3 Future work

Video cameras are an attractive choice for non-contact monitoring due to their ubiquity, high performance and low cost. Non-contact vital-sign estimation algorithms could be improved by developing algorithms to search for an optimal and localised skin area from which to estimate vital-sign values. Such a strategy should allow the estimation algorithms to run during periods of small movements. Camera-based monitoring technologies could be also applied to provide other assessments (e.g. pain scores, motor development and neurological status [276]) from the analysis of facial expressions and patient activity.

11.3.1 Improving the selection of ROIs using deep learning techniques

Chapter 4 presented the CNN model for patient detection and skin segmentation. Chapters 7 and 9 showed that heart rate and respiratory rate can be estimated from the entire segmented skin area, as provided by the algorithms. Modifications to the original algorithms to include body part detection and body pose estimation would enable the vital-sign estimation algorithms to be more robust and accurate. ROIs could be selected from the most stable parts of the skin, avoiding the movement of arms and legs. Respiratory rate could be estimated from the torso area, where breathing is more prominent than in other skin areas. Algorithms could also be developed to search for optimal ROIs based on the visual clues obtained from the deep learning algorithms, leading not only to an improvement in the results but also in the proportion of time that vital signs could be estimated.

11.3.2 Clinical activity detection using deep learning techniques

Chapter 5 proposed a CNN model for detecting intervention and non-intervention events in videos. This could be extended to classify the type of clinical interventions in videos (e.g. checking the normal functionality of equipment, changing a nappy, taking temperature readings, administering medications or withdrawing blood from the heel). The automatic classification of clinical intervention events could then be included in an electronic patient record (EPR) to keep track of clinical care activities without clinical staff manually recording them.

11.3.3 Pain assessment

The research conducted in chapter 10 showed that it was possible to use non-contact algorithms to estimate vital signs and derive physiological indicators in pre-term infants during a heel prick

procedure. The next step is to develop algorithms for deriving behavioural indicators related to pain. Deep learning techniques could be used to implement algorithms to estimate infant states (sleep or awake), recognise facial expressions (e.g. brow bulge, eye squeeze and crying) and body movements (e.g. arm and leg movements).

The development of non-contact pain assessment could be performed by using a rule-based classifier based on the changes in vital signs and the presence of behavioural indicators that are associated with pain. This is similar to how a nurse observes the infant's pain as part of routine care [24, 27, 264].

11.3.4 Monitoring of neonatal resuscitation

A newly-born infant must undergo a transition from intrauterine to extrauterine life. Around 5-10% of all newborns need help to support the establishment of their airflow, breathing or circulation [277]. Respiratory distress syndrome (RDS) usually occurs when the infant's lungs are not fully developed and have not produced enough surfactant (fluid in the lungs), with an incidence rate around 50% in the infants born at 26–28 weeks [278]. RDS could lead to pneumothorax (collapsed lung) and other breathing problems. Camera-based monitoring technology could be used to monitor respiratory functions (e.g. tidal volume, respiratory effort or apnoea) in infants during delivery room resuscitation.

11.3.5 Monitoring of newborns in the first 72 hours of life

The first few hours of a newborn's life is the most critical moment. Post-delivery assessment and examination are performed to confirm normality and rule out any congenital abnormalities. If needed, emergency medical intervention may be required to prevent lifelong consequences. Initial post-delivery assessment typically includes an assessment of the APGAR score (measured from Appearance, Pulse rate, Grimace, Activity and Respiratory effort) [279] and physical examination (e.g. observing patterns of movement and respiration of the infant) [280]. Video cameras could be utilised to provide the automated and continuous assessment over the first 72 hours of life.

11.4 Concluding remarks

The research conducted in this thesis has proposed a framework for the monitoring of heart rate and respiratory rate using video cameras in a real-world clinical setting. The proposed methods have been shown to be robust with respect to different skin tones, different body postures and variable lighting conditions.

Camera-based non-contact monitoring technologies have the potential to expand beyond the monitoring of vital signs. With recent advances in computer vision and deep learning techniques, new research avenues have opened up the field of non-contact monitoring, indicating its integration into patient care in the NICU and telehealth applications.

Appendices

Appendix A
Convolutional neural network

A convolutional neural network (CNN) is a computational model formed by multiple processing layers that can learn the representation of a problem [134]. The network comprises a series of convolution operations which allow it to extract features at multiple levels of abstraction. The network is learned to discover a complex structure in the dataset by adjusting weights of each learnable layer through the backpropagation algorithm [281]. The CNN has brought great success in various computer vision tasks such as image classification, image segmentation, video classification, object detection and action recognition.

A.1 Building blocks

There are variants of CNN architectures designed to suit the task that one wants to tackle, but their basic building blocks are similar [163]. Given that each layer has the same functional interface:

$$f : \Re^{H \times W \times D} \rightarrow \Re^{H' \times W' \times D'} \tag{A.1}$$

where H, W and D are the height, width and depth respectively. $\mathbf{x} \mapsto \mathbf{y}$ that transforms an *input* $\mathbf{x}$ into an *output* $\mathbf{y}$ and the output shape depends on the function f. It may have weights and biases which are learned from the training data.

Convolutional layer

The convolutional layer is the core of any convolutional network. This layer is made of a set of small *filters* $\mathbf{w}$ which learn the representation of the input. Each filter is convolved across the input by computing the dot product between the filter and local regions of the input resulting a *feature map* $\mathbf{y}$. Given that each filter of size $H'' \times W''$ receives an input $\mathbf{x}$ with D channels (depth) and generates an output $\mathbf{y}$ with D' channels. The convolution operation is formulated as [163]:

$$y_{i'j'k'} = \sum_{p=1}^{H''} \sum_{q=1}^{W''} \sum_{r=1}^{D} w_{pqrk'} x_{i'+p,j'+q,r} \tag{A.2}$$

where the i and j are spatial indices (in vertical and horizontal directions) and k is a channel index (depth).

Non-linear activation layer

A non-linear activation layer is usually placed after the convolutional layer to introduce nonlinearities to the network. The Rectifier Linear Unit (ReLU) activation function is usually employed in his layer. The ReLU is defined as:

$$y_{ijk} = \max\{0, x_{ijk}\} \tag{A.3}$$

which applies max operation across each element of the input producing the output whose negative elements are zeroed out [282].

Pooling layer

A pooling layer is periodically added between successive convolutional layers to progressively down-sample the spatial size of the input in order to introduce spatial invariant and obtain more abstract representations. Given that $H'' \times W''$ is the spatial size of subregion. There are two main variations of the pooling layer. The *max*-pooling layer takes the maximum value over each subregion of the input:

$$y_{i'j'k} = \max_{1 \le p \le H'', 1 \le q \le W''} x_{i'+p, j'+q, k}. \tag{A.4}$$

The *average*-pooling compute an average over each subregion of the input:

$$y_{i'j'k} = \frac{1}{W''H''} \sum_{p=1}^{H''} \sum_{q=1}^{W''} x_{i'+p, j'+q, k}. \tag{A.5}$$

Fully-connected layer

The fully-connected layer is added after several convolutional and pooling layers to provide high-level reasoning such as image classification to the network. The fully-connected layer connects to every neuron in the proceeding layer. No spatial information is preserved. It can be thought as convolutional layers with 1×1 filters. Its operation is formulated as:

$$y_{k'} = \sum_{r=1}^{D} w_{rk'} x_r. \tag{A.6}$$

Convolutional transpose layer

The convolutional transpose layer computes the reverse of the convolution operation. It generally used to up-sample the input with the learned filters for the task required pixel-wise prediction such as image segmentation [140]. The layer is formulated as:

$$y_{i'j'k'} = \sum_{r=1}^{D} \sum_{p=1}^{u(H'', S_h)} \sum_{q=1}^{u(W'', S_w)} w_{a(i',p), b(j',q), k', r} \cdot x_{c(i',p), d(j',q), r} \tag{A.7}$$

where

$$a(i', p) = S_h p + m(i', S_h) \tag{A.8}$$

$$b(j', q) = S_w q + m(j', S_w) \tag{A.9}$$

$$c(i', p) = u(i', S_h) - p \tag{A.10}$$

$$d(j', q) = u(j', S_w) - q \tag{A.11}$$

$$m(k, S) = (k - 1) \bmod S \tag{A.12}$$

$$u(k, S) = \left\lfloor \frac{k-1}{S} \right\rfloor \tag{A.13}$$

and (S_h, S_w) is an upsampling factor. A combination of the convolutional transpose layer followed by the ReLU layer can produce a nonlinear upsampling output.

Batch normalisation layer

The batch normalisation layer [160] is added between convoluational and ReLU layers during training. The layer normalises the input at each *batch* by shifting the input to zero mean and unit variance. Let T be the batch size and the input and output are 4D tensors where the fourth dimension refers to the different data (e.g. images) in the same batch, the operation is defined as:

$$y_{ijkt} = w_k \frac{x_{ijkt} - \mu_k}{\sqrt{\sigma_k^2}} \tag{A.14}$$

where

$$\mu_k = \frac{1}{HWT} \sum_{p=1}^{H} \sum_{q=1}^{W} \sum_{s=1}^{T} x_{pqkt} \tag{A.15}$$

$$\sigma_k^2 = \frac{1}{HWT} \sum_{p=1}^{H} \sum_{q=1}^{W} \sum_{s=1}^{T} (x_{pqkt} - \mu_k)^2 \,. \tag{A.16}$$

Summation layer

The summation layer receives two inputs with the identical shape and produces the element-wise summation of two input layers for joining the data from two different sources:

$$y_{ijk} = x_{ijk}^1 + x_{ijk}^2 \tag{A.17}$$

Softmax layer

The softmax layer is an output layer of the network which performs multi-class classification to generate well-form probabilities over classes. The input is transformed using the softmax function

$$y_{i'j'k'} = \frac{e^{x_i jk}}{\sum_{t=1}^{D} e^{x_{ijt}}} \,. \tag{A.18}$$

The function squashes every element of the input to the values between 0 and 1, with all elements are added up to 1.

Loss layer

The loss layer is employed to penalise the network during training with a given cost function. Given the input-output pairs $(x_1, z_1), \ldots, (x_n, z_n)$ for n samples and a cost $C(\mathbf{z}, \bar{\mathbf{z}})$ representing the penalty if the prediction $\bar{\mathbf{z}}$ is taken instead of $\mathbf{z}$, the loss layer is defined as

$$y = \frac{1}{n} \sum_{i=1}^{n} C(z_i, f(x_i; \mathbf{w})) \, . \tag{A.19}$$

The learning is driven by this loss layer and a good set of parameters yield a minimum loss.

CNNs are commonly made up of a stack of convolutional layers with non-linear activation layers, and periodically followed by pooling layers. The structure repeats until the output has spatially small size and a decision is made upon that output by fully-connected layers and a softmax. A CNN architecture is designed depending on the task one wants to tackle as sequential or graph-based models. In some cases, the sequential topology may restrict the implementation of a complex network. A directed acyclic graph (DAG) may be utilised to represent a complex network connection. It provides more flexibility, compared to the linear topology, by allowing any layer or parameter to communicate freely in the network [163].

A.2 Network training

The training operation of CNNs involves learning all layers jointly on an end-to-end manner using *gradient descent* and *backpropagation* [281]. During training, the network is annexed by a loss layer, so the output of the function f is the loss with respect to the parameter $\mathbf{w}$. The parameter $\mathbf{w}$ is iteratively learned using gradient descent. The update of the parameter $\mathbf{w}$ is proportional to the negative gradient of the function f at current iteration t. The parameter $\mathbf{w}$ gradually moves in the direction of steepest descent defined by the loss. The update is commonly performed in batch in order to reduce the parameter variance during update and improve the learning stability. Let γ_t be the learning rate at iteration t, the update is defined as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma_t \frac{\partial f}{\partial \mathbf{w}}(\mathbf{w}_t) \, . \tag{A.20}$$

Since the network is composed of several building blocks, the partial derivative of each building block can be derived using the *chain rule* to propagate the loss backwards through the network. This approach is called *backpropagation* [281]. Let x^l be the input of layer l, the gradient with respect to the parameter w^l at layer l can be determined using:

$$\frac{\partial f}{\partial w^l} = \frac{\partial f}{\partial x^L} \frac{\partial x^L}{\partial x^{L-1}} \cdots \frac{\partial x^{l+1}}{\partial x^{l+1}} \frac{\partial x^l}{\partial w^l} \, . \tag{A.21}$$

The partial derivative of f with respect to w^l determines whether the parameter w^l should be increased or decreased. It can be calculated recursively by taking the network composition that relate w^l to f into consideration.

Appendix B
Interactive image segmentation

Image segmentation is a method to assign a label to each pixel of the image. In this work, a graph cut segmentation with geodesic star shape priors was used for annotating skin areas in order to generate a skin dataset and for post-processing the result of colour-based skin detection. The technique was employed to partition an image between skin and non-skin pixels. Gaussian mixture models were utilised for propagating a skin label throughout subsequent images in a video recording.

B.1 Graph cuts with geodesic star convexity

Interactive segmentation allows a user to interact with the segmentation framework by specifying certain pixels as object or background. The marked pixels put hard constraints and provide appearance information to the algorithm [283]. Recently, graph cut segmentation with geodesic star convexity has demonstrated to be a powerful technique for the segmentation of an arbitrary shape. It optimises an energy function which integrates region and boundary properties as well as shape constraints. Hence, accurate image segmentation can be achieved with minimal user effort [153].

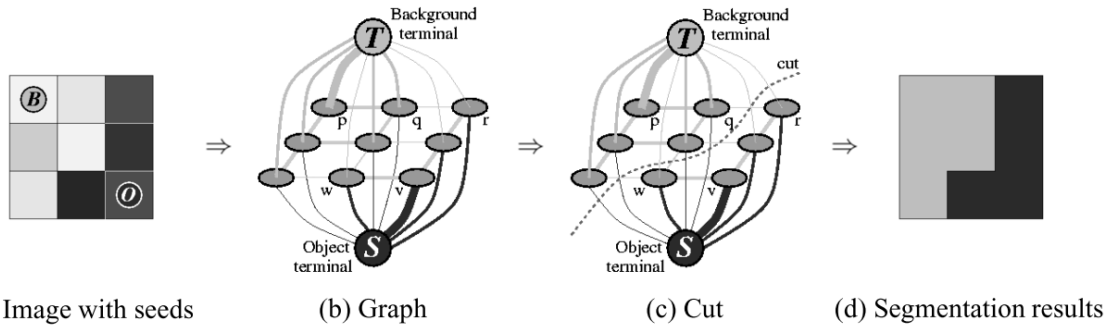


Figure B.1: Example of a segmentation on a 3×3 image. Given $B = \{p\}$ and $O = \{s\}$ are user provided seeds for background and object respectively. The thickness of each edge indicates the edge's weight. The minimum cut is a set of edges with lowest cost whose disconnects a graph. (Image adapted from [284])

Graph cuts

Let $G = (V, E)$ be a directed weighted graph defined as a set of nodes V and a set of edges E connecting two nodes in V. Each edge $e \in E$ is assigned to a nonnegative weight w_e. In image segmentation, a set of nodes V corresponds to a set of pixels P in the image . The graph G has two special terminal nodes called terminals. A cut $C \subset E$ on the graph G separates the graph into two disjoint sets, S and T where $V = S \cup T$. The cost of the cut C to $\{S, T\}$ is defined as the sum of all edge weights [284]:

$$|C| = \sum_{e \in C} w_e. \tag{B.1}$$

The minimum cut problem is to find a cut C which has the minimum cost among all possible cuts. The max-flow/min-cut theorem proves that the minimum cut is equivalent to the maximum flow [284]. The minimum cut on the graph G can be efficiently computed in polynomial time by a max-flow algorithm [285], which finds a maximum network flow from S to T as shown in figure B.1.

Geodesic stars

In image segmentation, a shape prior is utilised to bias the segmentation towards a certain shape. A star shape prior [286] is used to represent a generic shape not specific to any object. The star shape has a centre point c which is generally provided by the user. For any point p inside an object, all the points on the straight line formed between c and p are also inside the object (see figure B.2a). In other words, if p is regarded as the object, all points on the path between c and p must be regarded as the object.

As oppose to a single star, geodesic stars have multiple stars and employ geodesic paths which can adapt and bend to image gradients [153]. Therefore, the geodesic stars has a greater visibility of an object (see figure B.2b). In geodesic stars, the length of a discrete path is defined as

$$L(\Gamma) = \sum_{i=1}^{n-1} \sqrt{(1 - \gamma_g) d(\Gamma^i, \Gamma^{i+1})^2 + \gamma_g \parallel \Delta I(\Gamma^i, \Gamma^{i+1}) \parallel^2} \tag{B.2}$$

where Γ is a discrete path with n pixels comprised of $\{\Gamma^i, \ldots, \Gamma^{i+1}\}$, $d(\Gamma^i, \Gamma^{i+1})$ is the Euclidean

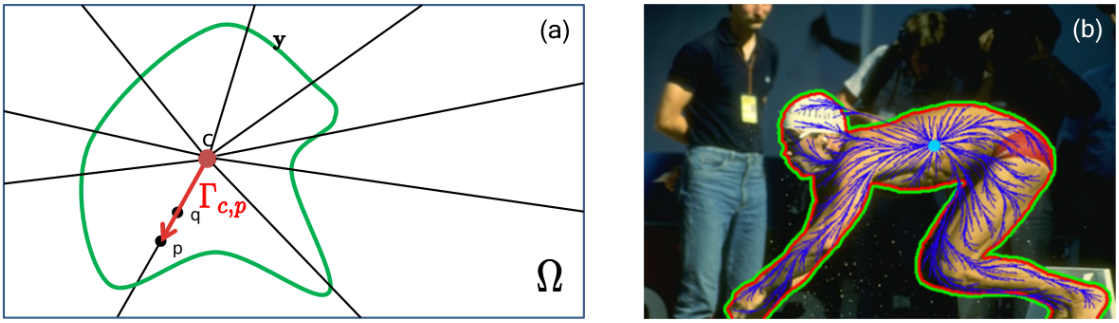


Figure B.2: (a) Single star convexity. For any point p inside an object $\mathbf{y}$, $\Gamma_{c,p}$ is the straight line formed between the star center c and point p. All points along the line are inside the object. (b) In Geodesic star convexity, geodesic paths can bend and adapt the appearance of object. (Image adapted from [153])

distance between the points Γ^i and Γ^{i+1}, $\Delta I(\Gamma^i, \Gamma^{i+1})$ is the difference of the image gradient between successive pixels and γ_g is the weighting parameter between Euclidean and geodesic.

Thus, the geodesic path is defined as the shortest discrete path between the closest star centre of point p, $c(p)$, and point p

$$\Gamma_{c(p),p} = \arg\min_{\Gamma \in \mathcal{Q}_{c(p),p}} L(\Gamma) \tag{B.3}$$

where $\mathcal{Q}_{c(p),p}$ is the set of all discrete paths between points $c(p)$ and p. Given the image Ω, if every point $p \in \Omega$ connects to its nearest star centre $c(p)$ using the geodesic path $\Gamma_{c(p),p}$, a *Geodesic forest* is obtained (see figure B.2b). Geodesic stars can capture the large deviations and fine details of an object.

Energy function

Many computer vision problems can be solved using energy minimisation [287]. Geodesic star constraints are imposed into the energy minimisation of Boykov and Jolly [284]. Let Ω represents the image, $\mathcal{P}$ denotes the collection of all pixels and $\mathcal{N}$ is the collection of neighbour pairs. Each pixel must be assigned to the label set $\mathcal{L} = \{0, 1\}$ where 0 and 1 denote background and object pixels respectively. Let $y_p \in \mathcal{L}$ be the label assigned to pixel p and $\mathbf{y} = \{f_p | p \in \mathcal{P}\}$ be the set of all label assignments. Let $\mathbf{c}$ be the set of all star centres. The energy function is defined as

$$E(\mathbf{y}|\Omega, \mathbf{c}) = \sum_{p \in \mathcal{P}} U(y_p) + \gamma \sum_{(p,q) \in \mathcal{N}} V(y_p, y_q) + \sum_{p \in \mathcal{P}} \sum_{q \in \Gamma_{c(p),p}} S_{pq}(y_p, y_q) \tag{B.4}$$

where λ denotes a weighting parameter of the second term. The first term called the *data* term. It calculates how well each pixel is fitted to background or object models. The data term is computed as the negative log likelihood of the models learned from star centres or, more specifically, user-provided brush strokes. The second term called the *contrast dependent* term [288] which is defined as $V(y_p, y_q) = exp(-\beta \parallel \Omega_p - \Omega_q \parallel^2)$ where β is a smoothness parameter. The third term is the *star-convexity* which is defined as

$$\forall q \in \Gamma_{c(p),p}, \quad S_{pq}(y_p, y_q) = \begin{cases} \infty, & \text{if } y_p = 1 \text{ and } y_q = 0 \\ 0, & \text{otherwise} \end{cases} \tag{B.5}$$

The constraint is set to an ∞ energy to restrict the shape-space. Thus, the energy function is minimised

$$\bar{\mathbf{y}} = \arg\min_{\mathbf{y} \in \mathcal{Y}} E(\mathbf{y}|\mathbf{x}, \mathbf{c}) \tag{B.6}$$

where $\mathcal{Y}$ is the set of all possible assignments. The star convexity reduces the output space significantly [153]; hence, the solution can be computed efficiently.

Energy minimization using graph cuts

The graph cuts are a powerful technique. It can be modified to fit a specific knowledge of the object in order to direct a partition between the object and background. In order to minimise the energy function, a graph is constructed whose nodes represent pixels and whose edges represent the weights which reflect the parameters in the data, contrast-dependent and shape prior terms of the energy function [284]. The optimal segmentation is obtained by computing the minimum cost of cutting of this graph.

B.2 Gaussian mixture models

Gaussian mixture models (GMM) are a probabilistic density model comprised a finite number of Gaussian distributions in order that the model is able to describe the distribution of a complex shape. The density function of GMM is defined as the sum of individual Gaussian density functions: [159]

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) \tag{B.7}$$

where $\mathbf{x}$ is a input vector, K is the number of Gaussians, π_k is the mixing coefficient of he k^{th} Gaussian, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)$ is the k^{th} Gaussian density function given by its mean $\boldsymbol{\mu}_\mathbf{k}$ and covariance matrices $\boldsymbol{\sigma}_\mathbf{k}$.

An effective method for fitting the parameters of GMM to the training data is the expectation-maximization (EM) technique [159]. Given a good initialisation of model parameters, the EM algorithm iteratively estimates the expectation function of the negative log-likelihood from the current parameters and tunes the model parameters by maximizing the likelihood found earlier.

GMM is often employed in colour analysis in order to model the colours of an object of interest which often form a small cluster in the colour space. In addition, GMM provides a greater flexibility to perform online adaptation in order to deal with slight variations in an image sequence.

Appendix C
Optical flow estimation

Optical flow refers to the pattern of apparent motion in a visual scene. It describes the point correspondences of moving objects in images. Optical flow can be used in, for example, video classification, object tracking, motion segmentation and velocity estimation. The estimation of optical flow is performed by computing the displacement vector between two images.

Modern optical flow estimation algorithms are based on energy optimisation techniques Brox *et al.* [172] proposed an algorithm to estimate optical flow based on a set of hypotheses (brightness, gradient and smoothness) and energy optimisation. This algorithm was shown to produce accurate optical flow vectors.

C.1 Variational model

According to Brox *et al.* [172], a variational formulation for computing optical flow is based on three important assumptions: a brightness constancy, a gradient constancy and a smoothness constraint.

Let I_1 and I_2 be an image at time t and another image at time $t+1$ respectively, $\mathbf{x} := (x, y)$ be a point in the image and $\mathbf{w} := (u, v)^\intercal$ be the sought displacement vector between the two images. For optical flow estimation, the first assumption made is that the brightness of a pixel should not be varied by the displacement. This assumption can measured by the following energy function:

$$E_{\text{Brightness}}(\mathbf{w}) = \int \Psi\left(|I_2(\mathbf{x}+\mathbf{w}) - I_1(\mathbf{x})|^2\right) d\mathbf{x} \tag{C.1}$$

where $\Psi(s^2) = \sqrt{s^2 + \epsilon^2}$ results in L^1 minimisation and ϵ is set to a fixed value of 0.001.

The second assumption is that the gradient of the image can be assumed not to change by the displacement. This assumption yields to the following energy function:

$$E_{\text{Gradient}}(\mathbf{w}) = \int \Psi\left(|\nabla I_2(\mathbf{x}+\mathbf{w}) - \nabla I_1(\mathbf{x})|^2\right) d\mathbf{x} \tag{C.2}$$

The third assumption takes the interaction between nearby pixels into account. This leads to the assumption of a smooth optical flow by penalising the total variation of the optical flow vector:

$$E_{\text{Smoothness}}(\mathbf{w}) = \int \Psi\left(|\nabla u(\mathbf{x})|^2 + |\nabla v(\mathbf{x})|^2\right) d\mathbf{x} \tag{C.3}$$

The total energy is the weighted sum of the brightness, gradient and smoothness terms:

$$E(\mathbf{w}) = E_{\text{Brightness}} + \gamma E_{\text{Gradient}} + \alpha E_{\text{Smoothness}} \tag{C.4}$$

where γ and α are weighing parameters ($\gamma, \alpha > 0$) steering the importance of the gradient and smoothness terms respectively. The goal is to find a displacement vector $\mathbf{w}$ that minimizes this total energy function.

C.2 Energy minimisation framework

The total energy function $E(\mathbf{w})$ is highly non-linear, and the optimisation is not trivial [172]. The minimisation can be performed through the Euler–Lagrange equations. Given that

$$\begin{aligned}
I_x &:= \delta_x I_2(\mathbf{x} + \mathbf{w}), \\
I_y &:= \delta_x I_2(\mathbf{x} + \mathbf{w}), \\
I_z &:= I_2(\mathbf{x} + \mathbf{w}) - I_1(\mathbf{x}), \\
I_{xx} &:= \delta_{xx} I_2(\mathbf{x} + \mathbf{w}), \\
I_{xy} &:= \delta_{xy} I_2(\mathbf{x} + \mathbf{w}), \\
I_{yy} &:= \delta_{yy} I_2(\mathbf{x} + \mathbf{w}), \\
I_{xz} &:= \delta_x I_2(\mathbf{x} + \mathbf{w}) - \delta_x I_1(\mathbf{x}), \\
I_{yz} &:= \delta_y I_2(\mathbf{x} + \mathbf{w}) - \delta_y I_1(\mathbf{x}).
\end{aligned} \tag{C.5}$$

An optimiser can find a solution by minimising the following Euler-Lagrange equations, which include all necessary constraints:

$$\Psi'(I_z^2) I_x I_z + \gamma \Psi'(I_{xz}^2 + I_{yz}^2)(I_{xx} I_{xz} + I_{xy} I_{yz}) - \alpha \text{div}(\Psi'(|\nabla u|^2 + |\nabla v|^2)\nabla u) = 0 \tag{C.6}$$

$$\Psi'(I_z^2) I_y I_z + \gamma \Psi'(I_{xz}^2 + I_{yz}^2)(I_{yy} I_{yz} + I_{xy} I_{xz}) - \alpha \text{div}(\Psi'(|\nabla u|^2 + |\nabla v|^2)\nabla u) = 0 \tag{C.7}$$

where $\Psi'(s^2)$ is the first derivative of $\Psi(s^2)$ with respect to s^2.

References

[1] R. E. Behrman and A. S. Butler, “Mortality and acute complications in preterm infants”, in *Preterm Birth: Causes, Consequences, and Prevention*, Washington, DC: National Academies Press, 2006.

[2] M. M. Baharestani, “An overview of neonatal and pediatric wound care knowledge and considerations”, *Ostomy Wound Management*, vol. 53, no. 6, pp. 34–55, 2007.

[3] R. O. Lloyd, R. M. Goulding, P. M. Filan, and G. B. Boylan, “Overcoming the practical challenges of electroencephalography for very preterm infants in the neonatal intensive care unit”, *Acta Paediatrica*, vol. 104, no. 2, pp. 152–157, 2015.

[4] V. Blazek, T. Wu, and D. Hoelscher, “Near-infrared CCD imaging: Possibilities for noninvasive and contactless 2D mapping of dermal venous hemodynamics”, in *SPIE Proceedings Volume 3923: Optical Diagnostics of Biological Fluids V*, 2000, pp. 2–9.

[5] F. P. Wieringa, F. Mastik, and A. F. W. Van Der Steen, “Contactless multiple wavelength photoplethysmographic imaging: A first step toward "SpO2 camera"technology”, *Annals of Biomedical Engineering*, vol. 33, no. 8, pp. 1034–1041, 2005.

[6] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, “Remote plethysmographic imaging using ambient light”, *Optics Express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[7] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Non-contact, automated cardiac pulse measurements using video imaging and blind source separation”, *Optics Express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[8] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Advancements in noncontact, multiparameter physiological measurements using a webcam”, *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 7–11, 2011.

[9] Y. Sun, “Imaging photoplethysmography: Towards effective physiological measurements”, PhD Thesis, Loughborough University, 2011.

[10] L. Scalise, N. Bernacchia, I. Ercoli, and P. Marchionni, “Heart rate measurement in neonatal patients using a webcamera”, in *2012 IEEE International Symposium on Medical Measurements and Applications Proceedings*, 2012, pp. 6–9.

[11] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, “Eulerian video magnification for revealing subtle changes in the world”, *ACM Transactions on Graphics*, vol. 31, no. 4, 2012.

[12] L. A. M. Aarts, V. Jeanne, J. P. Cleary, C. Lieber, J. S. Nelson, S. Bambang Oetomo, and W. Verkruysse, “Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit”, *Early Human Development*, vol. 89, no. 12, pp. 943–48, 2013.

[13] L. Kong, Y. Zhao, L. Dong, Y. Jian, X. Jin, B. Li, Y. Feng, M. Liu, X. Liu, and H. Wu, “Non-contact detection of oxygen saturation based on visible light imaging device using ambient light”, *Optics Express*, vol. 21, no. 15, pp. 17 464–17 471, 2013.

[14] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh, “Non-contact video-based vital sign monitoring using ambient light and auto-regressive models”, *Physiological Measurement*, vol. 35, no. 5, pp. 807–31, 2014.

[15] W. Hu, D. Y. C. Lie, M. U. Kakade, R. Ichapurapu, S. Mane, J. Lopez, Y. Li, C. Li, R. E. Banister, A. Dentino, T. Nguyen, S. Zupancic, and J. Griswold, “An intelligent non-contact wireless monitoring system for vital signs and motion detection”, *2010 International Conference on System Science and Engineering*, pp. 190–194, 2010.

[16] S. Suzuki, T. Matsui, M. Kagawa, T. Asao, and K. Kotani, “An approach to a non-contact vital sign monitoring using dual-frequency microwave radars for elderly care”, *Journal of Biomedical Science and Engineering*, vol. 6, no. 7, pp. 704–711, 2013.

[17] J. Kuutti, M. Paukkunen, M. Aalto, P. Eskelinen, and R. E. Sepponen, “Evaluation of a Doppler radar sensor system for vital signs detection and activity monitoring in a radio-frequency shielded room”, *Measurement*, vol. 68, pp. 135–142, 2015.

[18] K. Wardell, A. Jakobsson, and G. Nilsson, "Laser Doppler perfusion imaging by dynamic light scattering", *IEEE Transactions on Biomedical Engineering*, vol. 40, no. 4, pp. 309–316, 1993.

[19] H. Lu, Y. Li, H. Li, L. Yuan, Q. Liu, Y. Sun, and S. Tong, "Single-trial estimation of the cerebral metabolic rate of oxygen with imaging photoplethysmography and laser speckle contrast imaging", *Optics Letters*, vol. 40, no. 7, p. 1193, 2015.

[20] S. S. Ricci and T. Kyle, *Maternity and pediatric nursing*. Philadelphia, PA: Lippincott Williams & Wilkins, 2009, p. 1902.

[21] V. H. Larsen, P. Christensen, H. Oxhøj, and T. Brask, "Impedance pneumography for long-term monitoring of respiration during sleep in adult males", *Clinical Physiology*, vol. 4, no. 4, pp. 333–342, 1984.

[22] J. M. Ernst, D. A. Litvack, D. L. Lozano, J. T. Cacioppo, and G. G. Berntson, "Impedance pneumography: Noise as signal in impedance cardiography", *Psychophysiology*, vol. 36, no. 3, pp. 333–338, 1999.

[23] M. A. Cretikos, R. Bellomo, K. Hillman, J. Chen, S. Finfer, and A. Flabouris, "Respiratory rate: the neglected vital sign", *Medical Journal of Australia*, vol. 188, no. 11, pp. 657–659, 2008.

[24] L. Holsti and R. E. Grunau, "Initial validation of the Behavioral Indicators of Infant Pain (BIIP)", *Pain*, vol. 132, no. 3, pp. 264–272, 2007.

[25] P. J. Mathew and J. L. Mathew, "Assessment and management of pain in infants.", *Postgraduate Medical Journal*, vol. 79, no. 934, pp. 438–443, 2003.

[26] J. W. B. Peters, H. M. Koot, R. E. Grunau, J. de Boer, M. J. van Druenen, D. Tibboel, and H. J. Duivenvoorden, "Neonatal Facial Coding System for assessing postoperative pain in infants: item reduction is valid and feasible", *The Clinical Journal of Pain*, vol. 19, no. 6, pp. 353–363, 2003.

[27] B. Stevens, C. Johnston, P. Petryshen, and A. Taddio, "Premature Infant Pain Profile: Development and initial validation", *The Clinical Journal of Pain*, vol. 12, no. 1, pp. 13–22, 1996.

[28] J. Allen, "Photoplethysmography and its application in clinical physiological measurement", *Physiological Measurement*, vol. 28, no. 3, R1–R39, 2007.

[29] A. A. Alian and K. H. Shelley, "Photoplethysmography", *Best Practice & Research Clinical Anaesthesiology*, vol. 28, no. 4, pp. 395–406, 2014.

[30] W. Wang, "Robust and automatic remote photoplethysmography", PhD Thesis, Technische Universiteit Eindhoven, 2017.

[31] D. Shier, J. Butler, and R. Lewis, *Holes's essentials of human anatomy & physiology*, 13th. New York, NY: McGraw-Hill Education, 2012.

[32] A. B. Lumb, *Nunn's applied respiratory physiology*, 7th. London: Churchill Livingstone/Elsevier, 2010.

[33] J. A. Lemons, C. R. Bauer, W. Oh, S. B. Korones, L.-A. Papile, B. J. Stoll, J. Verter, M. Temprosa, L. L. Wright, R. A. Ehrenkranz, A. A. Fanaroff, A. Stark, W. Carlo, J. E. Tyson, E. F. Donovan, S. Shankaran, and D. K. Stevenson, "Very low birth weight outcomes of the national institute of child health and human development neonatal research network, January 1995 through December 1996", *Pediatrics*, vol. 107, no. 1, 2001.

[34] J. Ersch and T. Stallmach, "Assessing gestational age from histology of fetal skin: an autopsy study of 379 fetuses", *Obstetrics and Gynecology*, vol. 94, no. 5, pp. 753–757, 1999.

[35] Y. N. Kalia, L. B. Nonato, C. H. Lund, and R. H. Guy, "Development of skin barrier function in premature infants", *Journal of Investigative Dermatology*, vol. 111, no. 2, pp. 320–326, 1998.

[36] T. Oranges, V. Dini, and M. Romanelli, "Skin physiology of the neonate and infant: Clinical implications", *Advances in wound care*, vol. 4, no. 10, pp. 587–595, 2015.

[37] P. W. Post, A. N. Krauss, S. Waldman, and P. A. M. Auld, "Skin reflectance of newborn infants from 25 to 44 weeks gestational age", *Human Biology*, vol. 48, pp. 541–557, 1976.

[38] L. M. Dubowitz, V. Dubowitz, and C. Goldberg, "Clinical assessment of gestational age in the newborn infant", *The Journal of pediatrics*, vol. 77, no. 1, pp. 1–10, 1970.

[39] C. Lockwood, T. Conroy-Hiller, and T. Page, "Vital signs", *JBI Reports*, vol. 2, pp. 207–230, 2004.

[40] D. Evans, B. Hodgkinson, and J. Berry, "Vital signs in hospital patients: A systematic review", *IJNS*, vol. 38, no. 6, pp. 643–650, 2001.

[41] D. L. Schriger, "Approach to the patient with abnormal vital signs", in *Goldman's Cecil Medicine*, L. G. I. Schafer, Ed., 24th Editi, Philadelphia, PA: W.B. Saunders, 2012, ch. 7, pp. 27–30.

[42] S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Plüddemann, I. MacOnochie, L. Tarassenko, and D. Mant, "Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: A systematic review of observational studies", *The Lancet*, vol. 377, no. 9770, pp. 1011–1018, 2011.

[43] R. M. Kliegman, B. M. Stanton, J. S. Geme, and N. F. Schor, "Nelson Textbook of Pediatrics", in *Nelson Textbook of Pediatrics*, 20th ed., Philadelphia, PA: Elsevier, 2015.

[44] E. Phillipos, A. L. Solevåg, K. Aziz, S. van Os, G. Pichler, M. O'Reilly, P.-Y. Cheung, and G. M. Schmölzer, "Oxygen Saturation and Heart Rate Ranges in Very Preterm Infants Requiring Respiratory Support at Birth", *The Journal of Pediatrics*, vol. 182, 41–46.e2, 2017.

[45] L. M. Schnapp and N. H. Cohen, "Pulse oximetry: Uses and abuses", *Chest*, vol. 98, no. 5, pp. 1244–1250, 1990.

[46] G. Pearson and H. Duncan, "Early warning systems for identifying sick children", *Paediatrics and Child Health*, vol. 21, no. 5, pp. 230–233, 2011.

[47] K. W. Cross and T. E. Oppe, "The respiratory rate and volume in the premature infant", *The Journal of Physiology*, vol. 116, no. 2, pp. 168–174, 1952.

[48] H. E. Thomas, *Handbook of biomedical instrumentation and measurement*. Reston, VA: Reston Publishing Company, 1974.

[49] M. Folke, L. Cernerud, M. Ekström, and B. Hök, "Critical review of non-invasive respiratory monitoring in medical care", *Medical & Biological Engineering & Computing*, vol. 41, no. 4, pp. 377–83, 2003.

[50] A. F. Pacela, "Impedance pneumography — a survey of instrumentation techniques", *Medical & Biological Engineering*, vol. 4, no. 1, pp. 1–15, 1966.

[51] O. Mimoz, T. Benard, A. Gaucher, D. Frasca, and B. Debaene, "Accuracy of respiratory rate monitoring using a non-invasive acoustic method after general anaesthesia", *British Journal of Anaesthesia*, vol. 108, no. 5, pp. 872–875, 2012.

[52] T. M. Baird and M. R. Neuman, "Effect of infant position on breath amplitude measured by transthoracic impedance and strain gauges.", *Pediatric pulmonology*, vol. 10, no. 1, pp. 52–6, 1991.

[53] E. Pasquali, "Problems in impedance pneumography: Electrical characteristics of skin and lung tissue", *Medical and Biological Engineering*, vol. 5, no. 3, pp. 249–258, 1967.

[54] M. Villarroel, "Non-contact vital sign monitoring in the clinic", DPhil Thesis, University of Oxford, 2017.

[55] B. J. Semmes, M. J. Tobin, J. V. Snyder, and A. Grenvik, "Subjective and objective measurement of tidal volume in critically ill patients", *Chest*, vol. 87, no. 5, pp. 577–9, 1985.

[56] P. B. Lovett, J. M. Buchwald, K. Stürmann, and P. Bijur, "The vexatious vital: neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage", *Annals of Emergency Medicine*, vol. 45, no. 1, pp. 68–76, 2005.

[57] M. J. Mayotte, J. G. Webster, and W. J. Tompkins, "A comparison of electrodes for potential use in paediatric/infant apnoea monitoring", *Physiological Measurement*, vol. 15, no. 4, pp. 459–467, 1994.

[58] M. P. Dyke, R. Kohan, and S. Evans, "Morphine increases synchronous ventilation in preterm infants", *Journal of Paediatrics and Child Health*, vol. 31, no. 3, pp. 176–9, 1995.

[59] J. Flammer, M. Mozaffarieh, and H. Bebie, *Basic sciences in ophthalmology: Physics and chemistry*. New York, NY: Springer, 2013.

[60] L. O. Björn, "The nature of light and its interaction with matter", in *Photobiology*, New York, NY: Springer, 2008, pp. 1–39.

[61] J. G. Webster, *Design of pulse oximeters*. Bristol, UK: Institute of Physics Publishing, 1997.

[62] S. Hu, V. Azorin-Peris, and J. Zheng, "Opto-physiological modeling applied to photoplethysmographic cardiovascular assessment", *Journal of Healthcare Engineering*, vol. 4, no. 4, pp. 505–528, 2013.

[63] G. Baranoski and A. Krishnaswamy, "An introduction to light interaction with human skin", *Revista de Informática Teórica*, vol. 11, no. 1, pp. 33–62, 2004.

[64] A. J. Thody, E. M. Higgins, K. Wakamatsu, S. Ito, S. A. Burchill, and J. M. Marks, "Pheomelanin as well as Eumelanin is present in human epidermis", *Journal of Investigative Dermatology*, vol. 97, no. 2, pp. 340–344, 1991.

[65] M. J. Van Gemert, S. L. Jacques, H. J. Sterenborg, and W. M. Star, "Skin optics", *IEEE Transactions on Biomedical Engineering*, vol. 36, no. 12, pp. 1146–1154, 1989.

[66] M. R. Hamblin, P. Avci, and G. K. Gupta, *Imaging in dermatology*, 1st ed. Cambridge, MA: Academic Press, 2016.

[67] A. N. Krauss, P. W. Post, S. Waldman, and P. A. M. Auld, "Skin reflectance in the newborn infant", *Pediatric Research*, vol. 10, no. 9, pp. 776–778, 1976.

[68] D. P. West, J. M. Halket, D. R. Harvey, J. Hadgraft, L. M. Solomon, and J. I. Harper, "Percutaneous absorption in preterm infants", *Pediatric Dermatology*, vol. 4, no. 3, pp. 234–237, 1987.

[69] M. Nitzan, A. Romem, and R. Koppel, "Pulse oximetry: Fundamentals and technology update", *Medical Devices: Evidence and Research*, vol. 7, no. 1, pp. 231–239, 2014.

[70] S.-H. Tseng, P. Bargo, A. Durkin, and N. Kollias, "Chromophore concentrations, absorption and scattering properties of human skin in-vivo", *Optics Express*, vol. 17, no. 17, pp. 14 599–14 617, 2009.

[71] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida, "Wearable photoplethysmographic sensors — past and present", *Electronics*, vol. 3, no. 2, pp. 282–302, 2014.

[72] J. L. Reuss, "Multi-layer modeling of reflectance pulse oximetry", *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 2, pp. 153–159, 2005.

[73] E. D. Chan, M. M. Chan, and M. M. Chan, "Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations", *Respiratory Medicine*, vol. 107, no. 6, pp. 789–799, 2013.

[74] A. C. Dassel, R. Graaff, M. Aardema, W. G. Zijlstra, and J. G. Aarnoudse, "Effect of location of the sensor on reflectance pulse oximetry", *International Journal of Obstetrics and Gynaecology*, vol. 104, no. 8, pp. 910–916, 1997.

[75] Z. Zhu, T. Liu, G. Li, T. Li, and Y. Inoue, "Wearable sensor systems for infants", *Sensors*, vol. 15, no. 2, pp. 3721–3749, 2015.

[76] M. R. Grubb, J. Carpenter, J. A. Crowe, J. Teoh, N. Marlow, C. Ward, C. Mann, D. Sharkey, and B. R. Hayes-Gill, "Forehead reflectance photoplethysmography to monitor heart rate: Preliminary results from neonatal patients", *Physiological Measurement*, vol. 35, no. 5, pp. 881–893, 2014.

[77] A. Johansson, P. A. Oberg, and G. Sedin, "Monitoring of heart and respiratory rates in newborn infants using a new photoplethysmographic technique", *Journal of Clinical Monitoring and Computing*, vol. 15, no. 7-8, pp. 461–467, 1999.

[78] P. S. Addison, J. N. Watson, M. L. Mestek, and R. S. Mecca, "Developing an algorithm for pulse oximetry derived respiratory rate (RRoxi): a healthy volunteer study", *Journal of Clinical Monitoring and Computing*, vol. 26, no. 1, pp. 45–51, 2012.

[79] M. A. Pimentel, A. E. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton, "Towards a robust estimation of respiratory rate from pulse oximeters", *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2016.

[80] B. P. Imholz, W. Wieling, G. A. van Montfrans, and K. H. Wesseling, "Fifteen years experience with finger arterial pressure monitoring: assessment of the technology", *Cardiovascular research*, vol. 38, no. 3, pp. 605–616, 1998.

[81] M. Chang, "Optimal oxygen saturation in premature infants", *Korean Journal of Pediatrics*, vol. 54, no. 9, pp. 359–362, 2011.

[82] S. Fouzas, K. N. Priftis, and M. B. Anthracopoulos, "Pulse oximetry in pediatric practice", *Pediatrics*, vol. 128, no. 4, pp. 740–752, 2011.

[83] A. T. Boutin, A. Weidemann, Z. Fu, L. Mesropian, K. Gradin, C. Jamora, M. Wiesener, K. U. Eckardt, C. J. Koch, L. G. Ellies, G. Haddad, V. H. Haase, M. C. Simon, L. Poellinger, F. L. Powell, and R. S. Johnson, "Epidermal sensing of oxygen is essential for systemic hypoxic response", *Cell*, vol. 133, no. 2, pp. 223–234, 2008.

[84] W. T. Cecil, "A clinical evaluation of the accuracy of the Nellcor N-100 and Ohmenda 3700 pulse oximeters", *Journal of Clinical Monitoring*, vol. 4, no. 3, pp. 31–36, 1988.

[85] A. L. Ries, L. M. Prewitt, and J. J. Johnson, "Skin color and ear oximetry", *Chest*, vol. 96, no. 2, pp. 287–290, 1989.

[86] J. T. B. Moyle, "Uses and abuses of pulse oximetry", *Archives of Disease in Childhood*, vol. 74, no. 1, pp. 77–80, 1996.

[87] M. T. Petterson, V. L. Begnoche, and J. M. Graybeal, "The effect of motion on pulse oximetry and its clinical significance", *Anesthesia and Analgesia*, vol. 105, no. 6, pp. 78–84, 2007.

[88] F. A. Selig, E. R. Tonolli, E. V. C. M. D. Silva, and M. F. D. Godoy, "Heart rate variability in preterm and term neonates", *Arquivos Brasileiros de Cardiologia*, vol. 96, no. 6, pp. 443–449, 2011.

[89] D. L. J. Bunker, R. Kumar, A. Martin, and S. P. Pegg, "Thermal injuries caused by medical instruments", *Journal of Burn Care & Research*, vol. 35, no. 2, pp. 132–134, 2014.

[90] P. D. Mannheimer, "The light-tissue interaction of pulse oximetry", *Anesthesia and Analgesia*, vol. 105, no. 6, pp. 10–17, 2007.

[91] W. Verkruysse, M. Bartula, E. Bresch, M. Rocque, M. Meftah, and I. Kirenko, "Calibration of contactless pulse oximetry", *Anesthesia & Analgesia*, vol. 124, no. 1, pp. 136–145, 2017.

[92] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video", in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3430–3437.

[93] A. A. Kamshilin, E. Nippolainen, I. S. Sidorov, P. V. Vasilev, N. P. Erofeev, N. P. Podolian, and R. V. Romashko, "A new look at the essence of the imaging photoplethysmography", *Scientific Reports*, vol. 5, 2015.

[94] G. De Haan and A. Van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature", *Physiological Measurement*, vol. 35, no. 9, pp. 1913–1926, 2014.

[95] W. Wang, A. C. Den Brinker, S. Stuijk, and G. De Haan, "Algorithmic principles of remote PPG", *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2017.

[96] D. J. McDuff, E. B. Blackford, and J. R. Estepp, "The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography", in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 63–70.

[97] C. Takano and Y. Ohta, "Heart rate measurement based on a time-lapse image", *Medical Engineering and Physics*, vol. 29, no. 8, pp. 853–857, 2007.

[98] H. Nisar, M. B. Khan, W. T. Yi, Y. V. Voon, and T. S. Khang, "Contactless heart rate monitor for multiple persons in a video", in *2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2016.

[99] T. Tamura, T. Togawa, M. Ogawa, and M. Yoda, "Fully automated health monitoring system in the home", *Medical Engineering and Physics*, vol. 20, no. 8, pp. 573–579, 1998.

[100] Y. G. Lim, K. H. Hong, K. K. Kim, J. H. Shin, S. M. Lee, G. S. Chung, H. J. Baek, D.-U. Jeong, and K. S. Park, "Monitoring physiological signals using nonintrusive sensors installed in daily life equipment", *Biomedical Engineering Letters*, vol. 1, no. 1, pp. 11–20, 2011.

[101] D. Finn, G. B. Boylan, C. A. Ryan, and E. M. Dempsey, "Enhanced monitoring of the preterm infant during stabilization in the delivery room", *Frontiers in Pediatrics*, vol. 4, p. 30, 2016.

[102] S. Ohta, H. Nakamoto, Y. Shinagawa, and T. Tanikawa, "A health monitoring system for elderly people living alone", *Journal of Telemedicine and Telecare*, vol. 8, no. 3, pp. 151–156, 2002.

[103] E. Stone and M. Skubic, "Evaluation of an inexpensive depth camera for in-home gait assessment", *Journal of Ambient Intelligence and Smart Environments*, vol. 3, no. 4, pp. 349–361, 2011.

[104] E. Stone and M. Skubic, "Evaluation of an inexpensive depth camera for passive in-home fall risk assessment", in *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, 2011.

[105] Y. Sun, S. Hu, V. Azorin-Peris, S. Greenwald, J. Chambers, and Y. Zhu, "Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise", *Journal of Biomedical Optics*, vol. 16, no. 7, 2011.

[106] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera", *Biomedical Optics Express*, vol. 6, no. 5, pp. 1565–1588, 2015.

[107] M. Villarroel, A. Guazzi, J. Jorge, S. Davis, G. Green, A. Shenvi, P. Watkinson, K. McCormick, and L. Tarassenko, "Continuous non-contact vital sign monitoring in neonatal intensive care unit", *Healthcare Technology Letters*, vol. 1, no. 3, pp. 87–91, 2014.

[108] A. R. Guazzi, M. Villarroel, J. Jorge, J. Daly, M. C. Frise, P. A. Robbins, and L. Tarassenko, "Non-contact measurement of oxygen saturation with an RGB camera", *Biomedical Optics Express*, vol. 6, no. 9, pp. 3320–3338, 2015.

[109] S. Kwon, J. Kim, D. Lee, and K. Park, "ROI analysis for remote photoplethysmography on facial video", in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015.

[110] K. Ishihara, "Region-of-interest setting apparatus for respiration monitoring and a respiration monitoring system", US Patent Number: US6110123A, 2000.

[111] G. Cennini and V. Jeanne, "Method of controlling a function of a device and system for detecting the presence of a living being", US Patent Number: US8542878B2, 2013.

[112] V. Jeanne, F. J. D. Bruijn, R. Vlutters, G. Cennini, and D. Chestakov, "Processing images of at least one living being", US Patent Number: US8542877B2, 2013.

[113] J. Muehlsteff, M. J. Bartula, E. Breshch, S. W. Kaestle, and C. Shan, "Device and method for obtaining vital sign information of a subject", US Patent Application Number: US20140275832A1, 2014.

[114] M. J. Rocque, "Differentiating among a plurality of subjects in a field-of-view based on vital signs", US Patent Number: US9218668B2, 2015.

[115] M. J. Rocque, M. J. Bartula, and G. M. Kersten, "Apparatus and method for determining vital sign information from a subject", US Patent Application Number: US20140243649A1, 2014.

[116] M. J. Rocque, "Apparatus and method for detecting subjects on the basis of vital signs", US Patent Number: US9218668B2, 2014.

[117] T. C. Chen and Y. T. Chen, "Living body determination devices and methods", US Patent Application Number: US20160150986A1, 2015.

[118] C. Kenner, J. Wright, and J. W. Lott, *Comprehensive neonatal care: An interdisciplinary approach*. St. Louis, MO: Saunders Elsevier, 2007.

[119] Neonatal Data Analysis Unit (NDAU), "Neonatal Data Analysis Unit 2015 Report", Imperial College London, London, Tech. Rep., 2016.

[120] Neonatal Network Northern Ireland (NNNI), "Neonatal Care in Northern Ireland 2014 Report", Queen's University Belfast, Belfast, Tech. Rep., 2015.

[121] Neonatal Data Analysis Unit (NDAU), "Neonatal Data Analysis Unit 2014 Report", Imperial College London, London, Tech. Rep., 2015.

[122] H. C. Glass, A. T. Costarino, S. A. Stayer, C. M. Brett, F. Cladis, and P. J. Davis, "Outcomes for Extremely Premature Infants", *Anesthesia & Analgesia*, vol. 120, no. 6, pp. 1337–1351, 2015.

[123] M. Villarroel, J. Jorge, C. Pugh, and L. Tarassenko, "Non-contact vital sign monitoring in the clinic", in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, 2017, pp. 278–285.

[124] A. K. Abbas and S. Leonhardt, "Intelligent neonatal monitoring based on a virtual thermal sensor", *BMC Medical Imaging*, vol. 14, no. 1, 2014.

[125] A. K. Abbas, K. Heimann, K. Jergus, T. Orlikowsky, and S. Leonhardt, "Neonatal non-contact respiratory monitoring based on real-time infrared thermography", *BioMedical Engineering OnLine*, vol. 10, no. 1, 2011.

[126] J. H. Klaessens, M. van den Born, A. van der Veen, J. Sikkens-van de Kraats, F. A. van den Dungen, and R. M. Verdaasdonk, "Development of a baby friendly non-contact method for measuring vital signs: First results of clinical measurements in an open incubator at a neonatal intensive care unit", in *Proceedings Volume 8935, Advanced Biomedical and Clinical Diagnostic Systems XII*, 2014.

[127] The British Association of Perinatal Medicine (BAPM), *Categories of Care 2011*. London, 2011.

[128] D. Clifton, M. Villaroel, and L. Tarassenko, "Remote monitoring of vital signs", US Patent Number: US9615749B2, 2013.

[129] A. Karpathy, "Connecting Images and Natural Languag", PhD thesis, Stanford University, 2016, p. 95.

[130] S. L. Hijazi, R. Kumar, and C. Rowen, "Using Convolutional Neural Networks for Image Recognition", Cadence Design Systems, San Josem, CA, Tech. Rep., 2015.

[131] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[132] M. T. Jones, "A beginner's guide to artificial intelligence, machine learning, and cognitive computing", Armonk, NY, Tech. Rep., 2017.

[133] I. Goodfellow, A. Courville, and Y. Bengio, *Deep learning*. Cambridge, MA: MIT Press, 2016.

[134] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[135] A. Krizhevsky, I. Sulskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", in *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[136] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: delving deep into convolutional nets", *Computing Research Repository (CoRR)*, 2014. arXiv: 1405.3531.

[137] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", in *2015 International Conference on Learning Representations (ICLR)*, 2015.

[138] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[139] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 4, 2016, pp. 770–778.

[140] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation", in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[141] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation", in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1520–1528.

[142] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[143] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks", in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[144] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos", in *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014, pp. 568–576.

[145] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal residual networks for video action recognition", in *NIPS'16 Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2016. arXiv: 1611.02155.

[146] Y. A. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, "Efficient backprop", *Lecture Notes in Computer Science*, vol. 7700, pp. 9–48, 2012.

[147] M. A. Mielsen, *Neural Networks and Deep Learning*. Determination Press, 2015.

[148] M. Lin, Q. Chen, and S. Yan, "Network in network", in *2014 International Conference on Learning Representations (ICLR)*, 2014.

[149] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs", *Computing Research Repository (CoRR)*, 2014. arXiv: 1412.7062.

[150] G. Lin, C. Shen, I. Reid, and A. van dan Hengel, "Efficient piecewise training of deep structured models for semantic segmentation", *Computing Research Repository (CoRR)*, 2015. arXiv: 1504.01013.

[151] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks", *Computing Research Repository (CoRR)*, 2015. arXiv: 1502.03240.

[152] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning", *Computing Research Repository (CoRR)*, 2016. arXiv: 1603.07285.

[153] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman, "Geodesic star convexity for interactive image segmentation", in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3129–3136.

[154] D. P. Mitchell, "Spectrally optimal sampling for distribution ray tracing", *ACM SIGGRAPH Computer Graphics*, vol. 25, no. 4, pp. 157–164, 1991.

[155] M. Paulin, J. Revaud, Z. Harchaoui, F. Perronnin, and C. Schmid, "Transformation Pursuit for Image Classification", in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2014, pp. 3646–3653.

[156] Q. V. Le, J. Ngiam, Z. Chen, D. Chia, P. W. Koh, and A. Y. Ng, "Tiled convolutional neural networks", in *NIPS'10 Proceedings of the 23rd International Conference on Neural Information Processing Systems*, 2010, pp. 1279–1287.

[157] M. K. Agoston, *Computer graphics and geometric modeling*. London: Springer-Verlag, 2005.

[158] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "R-CNNs for pose estimation and action detection", *Computing Research Repository (CoRR)*, 2014. arXiv: 1406.5212.

[159] C. M. Bishop, *Pattern recognition and machine learning*, 6th. New York, USA: Springer-Verlag, 2006.

[160] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift", in *ICML'15 Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, 2015, pp. 448–456.

[161] P. Baldi and P. J. Sadowski, "Understanding dropout", in *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013, pp. 2814–2822.

[162] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", in *Machine Learning Research Volume 9*, 2010, pp. 249–56.

[163] A. Vedaldi and K. Lenc, "MatConvNet – Convolutional Neural Networks for MATLAB", in *Proceedings of the 23rd ACM international conference on Multimedia - MM '15*, New York, New York, USA: ACM Press, 2015, pp. 689–692.

[164] K.-Z. Lee, P.-C. Hung, and L.-W. Tsai, "Method and system for contact-free heart rate measurement", US Patent Application Number: US20130096439A1, 2013.

[165] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection", *International Journal of Computer Vision*, vol. 46, pp. 81–96, 2002.

[166] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[167] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods", *Pattern Recognition*, vol. 40, no. 3, pp. 1106–22, 2007.

[168] S. Bianco and R. Schettini, "Two new von Kries based chromatic adaptation transforms found by numerical optimization", *Color Research and Application*, vol. 35, no. 3, pp. 184–192, 2010.

[169] S. Bianco and R. Schettini, "Computational color constancy", in *3rd European Workshop on Visual Information Processing*, 2011, pp. 1–7.

[170] M. D. Fairchild, *Color appearance models*. Chichester: John Wiley & Sons, Ltd, 2013.

[171] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability", *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.

[172] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping", in *Computer Vision - ECCV 2004*, vol. 4, 2004, pp. 25–36.

[173] X. Zeng, W. Ouyang, and X. Wang, "Multi-stage contextual deep learning for pedestrian detection", in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 121–128.

[174] R. Kliegman, R. Behrman, and H. Jenson, *Nelson textbook of pediatrics*, 2nd. Philadelphia, PA: W.B. Saunders Company, 2015.

[175] P. Schwartz, "Guidelines for the interpretation of the neonatal electrocardiogram", *European Heart Journal*, vol. 23, no. 17, pp. 1329–1344, 2002.

[176] Association for the Advancement of Medical Instrumentation, *ANSI/AAMI EC13:2002 Cardiac monitors, heart rate meters, and alarms*. Arlington, VA: Association for the Advancement of Medical Instrumentation, 2002.

[177] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement", *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[178] D. Giavarina, "Understanding Bland-Altman analysis.", *Biochemia Medica*, vol. 25, no. 2, pp. 141–151, 2015.

[179] R. E. Klabunde, *Cardiovascular Physiology Concepts*, 2nd. Philadelphia, PA: Lippincott Williams & Wilkins, 2011.

[180] G. Rose, "Automatic analysis of electrocardiograms", *Mathematics and Computer Science in Biology and Medicine*, pp. 177–183, 1965.

[181] M. Tipple, "Interpretation of electrocardiograms in infants and children", *Images in Paediatric Cardiology*, vol. 1, no. 1, pp. 3–13, 1999.

[182] J. Charles, T. Pfister, M. Everingham, and A. Zisserman, "Automatic and efficient human pose estimation for sign language videos", *International Journal of Computer Vision*, vol. 110, no. 1, pp. 70–90, 2014.

[183] G. Friesen, T. Jannett, M. Jadallah, S. Yates, S. Quint, and H. Nagle, "A comparison of the noise sensitivity of nine QRS detection algorithms", *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 1, pp. 85–98, 1990.

[184] J. A. Van Alste and T. S. Schilder, "Removal of base-line wander and power-line interference from the ECG by an efficient FIR filter with a reduced number of taps", *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 12, pp. 1052–1060, 1985.

[185] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm", *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.

[186] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database", *IEEE Transactions on Biomedical Engineering*, vol. BME-33, no. 12, pp. 1157–1165, 1986.

[187] W. Zong, G. G. Moody, and D. Jiang, "A robust open-source algorithm to detect onset and duration of QRS complexes", in *Computing in Cardiology*, vol. 30, 2003, pp. 737–740.

[188] M. Schiavenato, C. Oliu, E. Bello, J. Bohorquez, and N. Claure, "Development of a system for the assessment of heart rate variability in the NICU", in *2013 29th Southern Biomedical Engineering Conference*, 2013, pp. 25–26.

[189] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter", *Physiological Measurement*, vol. 29, no. 1, pp. 15–32, 2008.

[190] T. He, G. Clifford, and L. Tarassenko, "Application of independent component analysis in removing artefacts from the electrocardiogram", *Neural Computing and Applications*, vol. 15, no. 2, pp. 105–116, 2006.

[191] T. Riggs, B. Isenstein, and C. Thomas, "Spectral analysis of the normal electrocardiogram in children and adults", *Journal of Electrocardiology*, vol. 12, no. 4, pp. 377–379, 1979.

[192] B. Mulgrew, P. Grant, and J. Thompson, *Digital signal processing: Concepts and applications*, 2nd. Hamshire: MacMillan Press, 2002.

[193] J. A. Crowe and D. Damianou, "The wavelength dependence of the photoplethysmogram and its implication to pulse oximetry", in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1992, pp. 2423–2424.

[194] Y. Maeda, M. Sekine, and T. Tamura, "The advantages of wearable green reflected photoplethysmography", *Journal of Medical Systems*, vol. 35, no. 5, pp. 829–834, 2011.

[195] G. C. Millman, "Fanaroff and Martin's neonatal-perinatal medicine diseases of the fetus and infant, 8th edn, Vols I and II", *Archives of Disease in Childhood: Fetal and Neonatal Edition*, vol. 91, no. 6, 2006.

[196] J. N. Jensen and M. Hannemose, "Camera-based heart rate monitoring", BSc Thesis, Technical University of Denmark, 2014.

[197] S. Bobbia, Y. Benezeth, and J. Dubois, "Remote photoplethysmography based on implicit living skin tissue segmentation", in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 361–365.

[198] W. Zong, T. Heldt, G. Moody, and R. Mark, "An open-source algorithm to detect onset of arterial blood pressure pulses", in *Computers in Cardiology*, 2003, pp. 259–262.

[199] J. J. K. Ó Ruanaidh and W. J. Fitzgerald, "Retrospective changepoint detection", in *Numerical Bayesian Methods Applied to Signal Processing*, New York, NY: Springer, 1996, pp. 96–121.

[200] T. Riemer, M. Weiss, and M. Losh, "Discrete clipping detection by use of a signal matched exponentially weighted differentiator", in *IEEE Proceedings on SouthEastCon*, 1990, pp. 245–248.

[201] M. Müller, *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer, 2007.

[202] Fitriani, "Multiscale Dynamic Time and Space Warping", PhD Thesis, Massachusetts Institute of Technology, 2008.

[203] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space", *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[204] K. Ohtsu, H. Peng, and G. Kitagawa, *Time series modeling for analysis and control*, ser. SpringerBriefs in Statistics. Tokyo: Springer Japan, 2015.

[205] J. P. Burg, "Maximum entropy spectral analysis", in *Proceedings of the 37th Ann. Int. SEG Meeting*, vol. 6, 1975.

[206] A. V. Oppenheim and R. W. Schafer, *Discrete time signal processing*, 3rd. Essex: Pearson Education, 2009, p. 1120.

[207] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion", *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.

[208] L. Tarassenko, L. Mason, and N. Townsend, "Multi-sensor fusion for robust computation of breathing rate", *Electronics Letters*, vol. 38, no. 22, p. 1314, 2002.

[209] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian process regression in vital-sign early warning systems", in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2012, 2012, pp. 6161–6164.

[210] J. M. Richards, J. R. Alexander, E. A. Shinebourne, M. de Swiet, A. J. Wilson, and D. P. Southall, "Sequential 22-hour profiles of breathing patterns and heart rate in 110 full-term infants during their first 6 months of life", *Pediatrics*, vol. 74, no. 5, pp. 763–77, 1984.

[211] C. J. Morley, A. J. Thornton, M. A. Fowler, T. J. Cole, and P. H. Hewson, "Respiratory rate and severity of illness in babies under 6 months old", *Archives of Disease in Childhood*, vol. 65, pp. 834–837, 1990.

[212] J. Jorge, M. Villarroel, S. Chaichulee, A. Guazzi, S. Davis, G. Green, K. McCormick, and L. Tarassenko, "Non-contact monitoring of respiration in the neonatal intensive care unit", in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, 2017, pp. 286–293.

[213] J. Jorge, M. Villarroel, S. Chaichulee, K. McCormick, and L. Tarassenko, "Data fusion for improved camera-based detection of respiration in neonates", in *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics*, G. L. Coté, Ed., vol. 10501, SPIE, 2018, p. 36.

[214] A. K. Gupta, "Respiration rate measurement based on impedance pneumography", Texas Instruments, Dallas, TX, Tech. Rep., 2011.

[215] B. G. Goudra, L. C. Penugonda, R. M. Speck, and A. C. Sinha, "Comparison of acoustic respiration rate, impedance pneumography and capnometry monitors for respiration rate accuracy and apnea detection during gi endoscopy anesthesia", *Open Journal of Anesthesiology*, vol. 03, no. 02, pp. 74–79, 2013.

[216] G. B. Moody, R. G. Mark, M. A. Bump, J. S. Weinstein, A. D. Berman, J. E. Mietus, and A. L. Goldberger, "Clinical validation of the ecg-derived respiration (EDR) technique", in *Computers in Cardiology*, vol. 13, 1986, pp. 507–510.

[217] F. Yasuma and J. I. Hayano, "Respiratory sinus arrhythmia: Why does the heartbeat synchronize with respiratory rhythm?", *Chest*, vol. 125, no. 2, pp. 683–690, 2004.

[218] M. K. Hathorn, "Respiratory sinus arrhythmia in new-born infants", *The Journal of Physiology*, vol. 385, no. 1, pp. 1–12, 1987.

[219] C. Orphanidou, S. Fleming, S. A. Shah, and L. Tarassenko, "Data fusion for estimating respiratory rate from a single-lead ECG", *Biomedical Signal Processing and Control*, vol. 8, no. 1, pp. 98–105, 2013.

[220] D. Cysarz, R. Zerm, H. Bettermann, M. Frühwirth, M. Moser, and M. Kröz, "Comparison of respiratory rates derived from heart rate variability, ECG amplitude, and nasal/oral airflow", *Annals of Biomedical Engineering*, vol. 36, no. 12, pp. 2085–2094, 2008.

[221] R. Ruangsuwana, G. Velikic, and M. Bocko, "Methods to extract respiration information from ECG signals", in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010, pp. 570–573.

[222] W. Lu, M. M. Nystrom, P. J. Parikh, D. R. Fooshee, J. P. Hubenschmidt, J. D. Bradley, and D. A. Low, "A semi-automatic method for peak and valley detection in free-breathing respiratory waveforms", *Medical Physics*, vol. 33, no. 10, pp. 3634–3636, 2006.

[223] Q. Li, R. G. Mark, and G. D. Clifford, "Artificial arterial blood pressure artifact models and an evaluation of a robust blood pressure and heart rate estimator", *BioMedical Engineering OnLine*, vol. 8, 2009.

[224] L. Sörnmo and P. Laguna, *Bioelectrical signal processing in cardiac and neurological applications*. Cambridge, MA: Academic Press, 2005.

[225] R. J. Huntsman, N. J. Lowry, and K. Sankaran, "Nonepileptic motor phenomena in the neonate", *Paediatrics & Child Health*, vol. 13, no. 8, pp. 680–684, 2008.

[226] J. A. Hirsch and B. Bishop, "Respiratory Sinus Arrhythmia in Humans: How Breathing Pattern Modulates Heart Rate", *The American Journal of Physiology*, vol. 241, no. 4, pp. 620–629, 1981.

[227] G. B. Moody, R. G. Mark, A. Zoccola, and S. Mantero, "Derivation of respiratory signals from multi-lead ECGs", *Computers in Cardiology 1985*, no. January 1985, pp. 113–116, 1985.

[228] D. J. Meredith, D. Clifton, P. Charlton, J. Brooks, C. W. Pugh, and L. Tarassenko, "Photoplethysmographic derivation of respiratory rate: a review of relevant physiology", *Journal of Medical Engineering & Technology*, vol. 36, no. 1, pp. 1–7, 2012.

[229] L. Dougherty and S. Lister, *The Royal Marsdden Hospital Manual of Clinical Nursing Procedures*, 9th. Hoboken, NJ: John Wiley & Sons, Ltd, 2015.

[230] A. Lundström and M. Nylund, "Clinical Skills in Children's Nursing", *Nurse Education in Practice*, vol. 11, no. 5, 2011.

[231] M. van Gastel, S. Stuijk, and G. de Haan, "Robust respiration detection from remote photoplethysmography", *Biomedical Optics Express*, vol. 7, no. 12, p. 4941, 2016.

[232] V. Makkapati, P. Raman, and G. Pai, "Camera based respiration rate of neonates by modeling movement of chest and abdomen region", *2016 International Conference on Signal Processing and Communications, SPCOM 2016*, no. 2, 2016.

[233] L. Cattani, D. Alinovi, G. Ferrari, R. Raheli, E. Pavlidis, C. Spagnoli, and F. Pisani, "Monitoring infants by automatic video processing: A unified approach to motion analysis", *Computers in Biology and Medicine*, vol. 80, pp. 158–165, 2017.

[234] F. Benetazzo, A. Freddi, A. Monteriù, and S. Longhi, "Respiratory rate detection algorithm based on RGB-D camera: theoretical background and experimental results.", *Healthcare Technology Letters*, vol. 1, no. 3, pp. 81–86, 2014.

[235] R. L. Eisenberg and N. M. Johnson, *Comprehensive Radiographic Pathology*. New York, NY: Elsevier, 2013.

[236] P. Sharma, "Edge detection using moore neighborhood", *International Journal of Computer Applications*, vol. 61, no. 3, pp. 975–8887, 2013.

[237] D. Kalman, "The most marvelous theorem in mathematics", *Math Horizons*, vol. 15, no. 4, pp. 16–17, 2008.

[238] M. Stojmenovic and A. Nayak, "Direct ellipse fitting and measuring based on shape boundaries", in *Advances in Image and Video Technology*, vol. 4872, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 221–235.

[239] P. L. Rosin, "Further five-point fit ellipse fitting", *Graphical Models and Image Processing*, vol. 61, pp. 245–259, 1999.

[240] R. M. Haralick and L. G. Shapiro, *Computer and robot vision*. Boston, MA: Addison-Wesley Longman Publishing Company, 1992.

[241] A. B. Te Pas, C. Wong, C. O. F. Kamlin, J. A. Dawson, C. J. Morley, and P. G. Davis, "Breathing patterns in preterm and term infants immediately after birth", *Pediatric Research*, vol. 65, no. 3, pp. 352–356, 2009.

[242] G. Cioni and H. F. Prechtl, "Preterm and early postterm motor behaviour in low-risk premature infants", *Early Human Development*, vol. 23, no. 3, pp. 159–191, 1990.

[243] E. B. Thoman, "Sleeping and waking states in infants: A functional perspective", *Neuroscience & Biobehavioral Reviews*, vol. 14, pp. 93–107, 1990.

[244] K. Tan, R. Saatchi, H. Elphick, and D. Burke, "Real-time vision based respiration monitoring system", *7th International Symposium on Communication Systems Networks and Digital Signal Processing (CSNDSP)*, pp. 770–774, 2010.

[245] N. Koolen, O. Decroupet, A. Dereymaeker, K. Jansen, J. Vervisch, V. Matic, B. Vanrumste, G. Naulaers, S. V. Huffel, and M. D. Vos, "Automated respiration detection from neonatal video data", *4th International Conference on Pattern Recognition Applications and Methods*, pp. 164–169, 2015.

[246] S. Nemati, A. Malhotra, and G. D. Clifford, "Data fusion for improved respiration rate estimation", *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 1, 2010.

[247] L. J. Duhn and J. M. Medves, "A systematic integrative review of infant pain assessment tools", *Advances in Neonatal Care*, vol. 4, no. 3, pp. 126–40, 2004.

[248] B. J. Stevens and C. C. Johnston, "Physiological responses of premature infants to a painful stimulus.", *Nursing research*, vol. 43, no. 4, pp. 226–31, 1994.

[249] V. Lindh, U. Wiklund, and S. Hakansson, "Heel lancing in term new-born infants: an evaluation of pain by frequency domain analysis of heart rate variability", *Pain*, 1999.

[250] R. P. Riddell, M. Fitzgerald, R. Slater, B. Stevens, C. Johnston, and M. Campbell-Yeo, "Using only behaviours to assess infant pain", *Pain*, vol. 157, no. 8, pp. 1579–1580, 2016.

[251] B. O. Valeri and M. B. M. Linhares, "Pain in preterm infants: Effects of sex, gestational age, and neonatal illness severity.", *Psychology & Neuroscience*, vol. 5, no. 1, pp. 11–19, 2012.

[252] J. de Lima, A. R. Lloyd-Thomas, R. F. Howard, E. Sumner, and T. M. Quinn, "Infant and neonatal pain: anaesthetists' perceptions and prescribing patterns.", *British Medical Journal (Clinical Research Ed.)*, vol. 313, no. 7060, p. 787, 1996.

[253] A. M. Wagner, "Pain control in the pediatric patient.", *Dermatologic Clinics*, vol. 16, no. 3, pp. 609–17, 1998.

[254] P. S. Beacham, "Behavioral and physiological indicators of procedural and postoperative pain in high-risk infants.", *Journal of obstetric, gynecologic, and neonatal nursing : JOGNN*, vol. 33, no. 2, pp. 246–55, 2004.

[255] X. Cong, J. M. McGrath, R. M. Cusson, and D. Zhang, "Pain assessment and measurement in neonates: an updated review.", *Advances in neonatal care*, vol. 13, no. 6, pp. 379–95, 2013.

[256] K. J. Anand, "Pain, plasticity, and premature birth: a prescription for permanent suffering?", *Nature Medicine*, vol. 6, no. 9, pp. 971–973, 2000.

[257] C. L. Lowery, M. P. Hardman, N. Manning, R. W. Hall, and K. J. S. Anand, "Neurodevelopmental Changes of Fetal Pain", *Seminars in Perinatology*, vol. 31, no. 5, pp. 275–282, 2007.

[258] R. Slater, M. Fitzgerald, and J. Meek, "Can Cortical Responses Following Noxious Stimulation Inform Us About Pain Processing in Neonates?", *Seminars in Perinatology*, vol. 31, no. 5, pp. 298–302, 2007.

[259] K. D. Craig, M. F. Whitfield, R. V. Grunau, J. Linton, and H. D. Hadjistavropoulos, "Pain in the preterm neonate: behavioural and physiological indices", *Pain*, vol. 52, no. 3, pp. 287–99, 1993.

[260] B. J. Stevens, C. C. Johnston, and L. Horton, "Multidimensional Pain Assessment in Premature Neonates: A Pilot Study", *Journal of Obstetric, Gynecologic & Neonatal Nursing*, vol. 22, no. 6, pp. 531–541, 1993.

[261] C. C. Johnston and M. E. Strada, "Acute pain response in infants: a multidimensional description.", *Pain*, vol. 24, no. 3, pp. 373–82, 1986.

[262] N. McIntosh, L. Van Veen, and H. Brameyer, "The pain of heel prick and its measurement in preterm infants", *Pain*, vol. 52, no. 1, pp. 71–74, 1993.

[263] A. M. Taksande, K. Y. Vilhekar, M. Jain, and D. Chitre, "Pain response of neonates to venipuncture.", *Indian journal of pediatrics*, vol. 72, no. 9, pp. 751–3, 2005.

[264] J. Lawrence, D. Alcock, J. Kay, and P. McGrath, "The development of a tool to assess neonatal pain", *Journal of Pain and Symptom Management*, vol. 6, no. 3, p. 194, 1991.

[265] R. V. Grunau and K. D. Craig, "Pain expression in neonates: facial action and cry", *Pain*, vol. 28, no. 3, pp. 395–410, 1987.

[266] C. C. Johnston, B. Stevens, F. Yang, and L. Horton, "Developmental changes in response to heelstick in preterm infants: a prospective cohort study.", *Developmental Medicine & Child Neurology*, vol. 38, no. 5, pp. 438–45, 1996.

[267] R. E. Grunau, L. Holsti, M. F. Whitfield, and E. Ling, "Are twitches, startles, and body movements pain indicators in extremely low birth weight infants?", *The Clinical Journal of Pain*, vol. 16, no. 1, pp. 37–45, 2000.

[268] L. Holsti, R. E. Grunau, T. F. Oberlander, and M. F. Whitfield, "Specific Newborn Individualized Developmental Care and Assessment Program movements are associated with acute pain in preterm infants in the neonatal intensive care unit.", *Pediatrics*, vol. 114, no. 1, pp. 65–72, 2004.

[269] L. Holsti, R. E. Grunau, T. F. Oberlander, and M. F. Whitfield, "Prior pain induces heightened motor responses during clustered care in preterm infants in the NICU", *Early Human Development*, vol. 81, no. 3, pp. 293–302, 2005.

[270] K. Andrews and M. Fitzgerald, "Cutaneous flexion reflex in human neonates: a quantitative study of threshold and stimulus-response characteristics after single and repeated stimuli.", *Developmental medicine and child neurology*, vol. 41, no. 10, pp. 696–703, 1999.

[271] K. Michelsson, P. Sirviö, and O. Wasz-Höckert, "Pain cry in full-term asphyxiated newborn infants correlated with late findings", *Acta Pædiatrica*, 1977.

[272] S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, G. Green, K. McCormick, A. Zisserman, and L. Tarassenko, "Localised photoplethysmography imaging for heart rate estimation of pre-term infants in the clinic", in *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics*, 2018.

[273] N. McIntosh, L. van Veen, and H. Brameyer, "Alleviation of the pain of heel prick in preterm infants", *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 70, no. 3, F177–F181, 1994.

[274] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[275] S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, G. Green, K. McCormick, A. Zisserman, and L. Tarassenko, "Multi-task convolutional neural network for patient detection and skin segmentation in continuous non-contact vital sign monitoring", in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, 2017, pp. 266–272.

[276] L. Wallin and M. Eriksson, "Newborn Individual Development Care and Assessment Program (NIDCAP): a systematic review of the literature", *Worldviews on Evidence-Based Nursing*, vol. 6, no. 2, pp. 54–69, 2009.

[277] A. C. C. Lee, S. Cousens, S. N. Wall, S. Niermeyer, G. L. Darmstadt, W. A. Carlo, W. J. Keenan, Z. A. Bhutta, C. Gill, and J. E. Lawn, "Neonatal resuscitation and immediate newborn assessment and stimulation for the prevention of neonatal deaths: a systematic review, meta-analysis and Delphi estimation of mortality effect.", *BMC public health*, vol. 11 Suppl 3, S12, 2011.

[278] A. A. Fanaroff and R. J. Martin, *Neonatal-perinatal medicine: diseases of the fetus and infant, volume 1*, 7th ed. Mosby, 2002, p. 1732.

[279] V. Apgar, "A proposal for a new method of evaluation of the newborn infant", *Anesthesia & Analgesia*, vol. 120, no. 5, pp. 1056–1059, 2015.

[280] H. Baston and H. Durward, *Examination of the newborn: a practical guide.* Abingdon: Taylor & Francis, 2016.

[281] R. Hecht-Nielsen, "Theory of the backpropagation neural network", *Proceedings Of The International Joint Conference On Neural Networks*, vol. 1, pp. 593–605, 1989.

[282] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines", *Proceedings of the 27th International Conference on Machine Learning*, no. 3, pp. 807–814, 2010.

[283] L. Gorelick, O. Veksler, Y. Boykov, and C. Nieuwenhuis, "Convexity shape prior for binary segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 258–271, 2017.

[284] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images", in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1, IEEE Comput. Soc, 2001, pp. 105–112.

[285] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, 2004, pp. 1124–1137.

[286] O. Veksler, "Star shape prior for graph-cut image segmentation", *Lecture Notes in Computer Science*, vol. 5304, pp. 454–467, 2008.

[287] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147–159, 2004.

[288] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts", *ACM Transactions on Graphics*, vol. 23, no. 3, 2004.