Joana Carlos Mesquita Ferreira

Bachelor of Science

Extraction of Heart Rate from Multimodal Video Streams of Neonates using Methods of Machine Learning

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Biomedical Engineering

Adviser: Christoph Hoog Antink, Assistant Professor,
RWTH Aachen University

Co-adviser: Pedro Manuel Cardoso Vieira, Assistant Professor,
NOVA University of Lisbon

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

December, 2019

Joana Carlos Mesquita Ferreira

Bachelor of Science

Extraction of Heart Rate from Multimodal Video Streams of Neonates using Methods of Machine Learning

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Biomedical Engineering

Adviser: Christoph Hoog Antink, Assistant Professor,
RWTH Aachen University

Co-adviser: Pedro Manuel Cardoso Vieira, Assistant Professor,
NOVA University of Lisbon

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

December, 2019

Extraction of Heart Rate from Multimodal Video Streams of Neonates using Methods of Machine Learning

This document was created using the (pdf)LaTeX processor, based on the "novathesis" template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Acknowledgements

I would first like to thank my supervisor, Dr.-Ing. Christoph Hoog Antink, who gave me the amazing opportunity to conduct my master thesis at MedIT. He always found the time to listen my ideas and doubts and to guide me towards a solution which I am truly proud of. It was a pleasure to work with him.

I would also like to thank my co-supervisor Dr. Pedro Manuel Cardoso Vieira, who is an inspiring teacher, always willing to help.

I want to thank my mother, Helena, who always put me first. The support I received from her through my life and during my stay abroad is immeasurable. I also want to thank Paulo, who always finds the perfect advice, and my sister Luana, who I know is my partner for life. Finally, I want to thank Alexandre for his love and endless encouragement to persue my dreams.

Abstract

The World Health Organization estimates that more than one-tenth of births are premature. Premature births are linked to an increase of the mortality risk, when compared with full-term infants. In fact, preterm birth complications are the leading cause of perinatal mortality. These complications range from respiratory distress to cardiovascular disorders. Vital signs changes are often prior to these major complications, therefore it is crucial to perform continuous monitoring of this signals. Heart rate monitoring is particularly important. Nowadays, the standard method to monitor this vital sign requires adhesive electrodes or sensors that are attached to the infant. This contact-based methods can damage the skin of the infant, possibly leading to infections. Within this context, there is a need to evolve to remote heart rate monitoring methods.

This thesis introduces a new method for region of interest selection to improve remote heart rate monitoring in neonatology through Photoplethysmography Imaging. The heart rate assessment is based on the standard photoplethysmography principle, which makes use of the subtle fluctuations of visible or infrared light that is reflected from the skin surface within the cardiac cycle. A camera is used, instead of the contact-based sensors. Specifically, this thesis presents an alternative method to manual region of interest selection using methods of Machine Learning, aiming to improve the robustness of Photoplethysmography Imaging. This method comprises a highly efficient Fully Convolutional Neural Network to select six different body regions, within each video frame. The developed neural network was built upon a ResNet network and a custom upsampling network. Additionally, a new post-processing method was developed to refine the body segmentation results, using a sequence of morphological operations and centre of mass analysis. The developed region of interest selection method was validated with clinical data, demonstrating a good agreement (78%) between the estimated heart rate and the reference.

Keywords: photoplethysmographic imaging, heart rate monitoring, premature infant, deep learning, convolutional neural network, image semantic segmentation

Resumo

A Organização Mundial de Saúde estima que um décimo dos nascimentos são prematuros. Os nascimentos prematuros estão associados a um aumento do risco de mortalidade. De facto, complicações ligadas a nascimentos prematuros são a principal causa de mortalidade perinatal. Estas complicações abrangem dificuldades respiratórias e complicações cardiovasculares. Estas complicações são frequentemente precedidas por alterações nos sinais vitais. Assim, a monitorização contínua do prematuro é fundamental, particularmente a monitorização do ritmo cardíaco. O método standard para monitorizar este sinal requer eléctrodos adesivos ou sensores que necessitam de ser acoplados à pele. Estes métodos de monitorização de contacto podem provocar danos na pele frágil do prematuro, podendo resultar em infeções. Assim, existe a necessidade de evoluir para métodos de monitorização remota do ritmo cardíaco.

Esta tese introduz um novo método de seleção da região de interesse para melhorar a performance da monitorização do ritmo cardíaco em neonatologia através de fotopletismografia de imagem. A extração do ritmo cardíaco é baseada nos fundamentos do método standard de fotopletismografia: num ciclo cardíaco existem flutuações subtis na quantidade de luz refletida na superfície da pele, na gama do visível e infra-vermelho. Uma câmara de vídeo é utilizada em fotopletismografia de imagem, ao invés de sensores de contacto. Especificamente, a tese apresenta um método alternativo à seleção manual da região de interesse, utilizando métodos de Machine Learning para aumentar a robustez do método de fotopletismografia de imagem. Este método compreende uma rede neuronal altamente eficiente para selecionar seis regiões do corpo, a cada frame de vídeo. A rede neuronal foi desenvolvida com base numa ResNet modificada e uma rede neuronal customizada. Adicionalmente, um novo método de pós-processamento foi desenvolvido, utilizando operações morfológicas e a análise dos centros de massa. O método desenvolvido foi validado, demonstrando uma grande concordância (78%) entre o ritmo cardíaco estimado e a referência.

Palavras-chave: fotopletismografia de imagem, monitorização do ritmo cardíaco, prematuro, deep learning, redes neuronais, segmentação semântica de imagem

Contents

List of Figures

List of Tables

Acronyms

a.u	Arbitrary units
CCD	Charge-coupled device
CM	Center of mass
CMOS	Complementary metal–oxide–semiconductor
CNN	Convolutional neural netwok
CRF	Conditional random field
DFT	Discrete Fourier transform
ECG	Electrocardiogram
ERF	Effective receptive field
FCNN	Fully convolutional neural network
FFT	Fast Fourier transform
FLOPs	Floating point operations
FMAs	floating-point multiply–add operations
FP	False positives
FPS	Frames per second
GPU	Graphics processing unit
GUI	Graphical user interfaces
HR	Heart rate
ILSVRC	ImageNet large scale visual recognition challenge
IoU	Intersection over union
IR	Infra-red

MAE	Mean absolute error
NICU	Newborn intensive care unit
PNG	Portable network graphic
PPG	Photopletismography
PPGI	Photopletismography imaging
RAM	Random-access memory
ReLU	Rectified linear unit
RF	Receptive Field
RGB	Red-Green-Blue
RMSE	Root mean squared error
ROI	Region of interest
RWTH	Rheinisch-Westfälische Technische Hochschule
SD	Standard deviation
SNR	Signal to noise ratio
SST	Wavelet synchrosqueezed transform
UNICEF	United Nations Children's Fund
VGC	Variable gain controls

Symbols

C_i	set of calibration masks
cm_i	set of centres of mass
$\hat{\text{PPGI}}_i(t)$	filtered PPGI signals
$f_{\max}$	dominant frequency in the PPGI signal
$f\text{PPGI}_i(t)$	signal that results from the fusion of $\hat{\text{PPGI}}_i$
$g\text{PPGI}(t)$	PPGI signal with the highest signal quality
i	index
I_{abs}	intensity of light absorbed by the tissue
I_{in}	intensity of light radiated into the tissue
I_{refl}	intensity of light reflected from the tissue
I_{trans}	intensity of light transmitted through the tissue
J	constant, window jump in samples
k	index
L	usually, length of the PPGI signal
$\max(x)$	maximum value of x
$\min(x)$	minimum value of x
N	size of the input frame
QI_i	set of quality indexes
$\text{PPGI}_i(t)$	raw PPGI signals

s_i	set of instances present in the segmentation mask	
W	constant, window length in samples	
$\mathbf{X}$	estimated segmentation mask	*
Y_i^{r}	set of raw segmentation masks, one for each body part	
Y_i^{b}	set of binary segmentation masks, one for each body part	
$\hat{Y}_i$	set of refined segmentation masks, one for each body part	
Y_i^{l}	set of dilated instance segmentation masks	

Chapter One

Introduction

According to the World Health Organization [29], 15 million babies are born premature, each year. This can be translated into: More than one in 10 infants does not complete 37 weeks of gestation. Despite the inhomogeneous incidence between countries, high rates of preterm births are present across different areas of the globe, ranging from 5% to 18% of the total births across the analyzed 184 countries in 2010. Particularly, Portugal, alongside with Germany, holds one of the highest rates of preterm births in Europe (10% in 2017 [71]) , with more than 200 babies born with less than 28 weeks of gestation, i.e. extreme preterm babies. Additionally, the World Health Organization anticipates an increase of the aforementioned numbers. Thus, preterm birth, specially its prevention [7], still poses a dilemma for both obstetricians and neonatologists despite the ongoing research and the general progress of medicine.

Premature infants are born not fully developed. Besides the neurodevelopmental problems highly associated with this type of patients, the functional immaturity of the varied organs and their regularization mechanisms commonly leads to complications such as temperature instability, respiratory distress syndrome, a compromised immune system, sepsis and cardiovascular disorders [7]. Particularly, the latter can result in irregular cardiorespiratory patterns which can lead to clinical complications namely apneas, cardiopulmonary arrest and sudden infant death syndrome [42]. In fact, UNICEF [22] stated that the leading cause of newborn deaths were due to preterm birth complications (35%) followed intrapartum-related events (24%) and severe infections (sepsis or menigitis) (15%). Thus, preterm birth complications are the leading cause of perinatal mortality around the world.

Changes in the vital parameters are often observed prior to the major complications associated with the preterm infants, therefore it is crucial to perform continuous monitoring of this signals. In clinical practice, vital parameters such as body temperature, heart rate, blood pressure, respiratory rate and arterial blood oxygen saturation are extracted using reliable monitoring modalities [23]. However, most of this traditional monitoring modalities are contact based and often invasive. Particularly, heart rate standard monitoring modalities require electrodes or sensors that are placed directly on the infant skin. This kind of monitoring is often linked to discomfort potentially causing infant stress [42]. Additionally,

infants with a gestational age bellow 34 weeks do not have a fully functional skin [69]. The underdeveloped skin structure coupled with the frequent change of electrodes and sensors make preterm infants a patient group prone to the development of infections. As a result, neonates suffering from this complications will be exposed to antibiotics, possibly escalating to last resort antibiotics, thus contributing to the appearance of resistant bacterial strains [9], which could be avoided. Also, the risk of epidermal stripping caused by the infant movements can lead to acute pain, which leads to hemodynamic changes [6]. Besides the aforementioned negative effects of electrodes placement, the associated wires also lead to discomfort and hinder not only the clinical staff activities but also the interaction between parents and infants.

Within this context, it is clear that there is a need to evolve to non-contact heart rate monitoring methods, since many aforementioned clinical complications, associated with the usage of electrodes, could be avoided if the measurement of vital parameters did not rely on mechanical or conductive contact. Recent advances in technology are opening the door for the emergence of a new generation of non-contact and non-invasive vital parameter monitoring modalities.

Photoplethysmography Imaging emerged as a promising heart rate monitoring method for relatively still adult patients. However, it quickly paved its way to address non-contact heart rate monitoring in newborn infants. This topic of research is already a central theme in neonatology. Photoplethysmography Imaging makes use of the subtle fluctuations of visible or infrared light that is reflected from the skin surface within the cardiac cycle. This subtle fluctuations are noticeable in the pixels intensity values of high bit-depth video data, allowing the extraction of the PPGI signal and consecutively the heart rate.

Photoplethysmography Imaging offers diverse advantages over other monitoring modalities. For instance, since the PPGI sensor is a video camera, Photoplethysmography Imaging constitutes a remote non-contact, non-invasive and passive monitoring method. Therefore, it is a painless and stress free method. Additionally, the measured radiation lays within the visible and infrared electromagnetic spectrum, i.e., this monitoring technique does not require the usage of harmful radiation. Also, multiple measuring sites can be used for heart rate extraction given the two-dimensional image provided by the video camera, i.e, each pixel can be considered as an individual sensor. Finally, for heart rate assessment, resorting exclusively in visible light, no dedicated light source may be needed.

Despite the aforementioned advantages, Photoplethysmography Imaging presents some drawbacks. Firstly, despite the recent important contributions to the field, there is still little conclusions about its feasibility in real world use, specially in the scope of neonatology. Second, the accurate extraction of the heart rate is highly dependent on the working conditions, i.e, lightning conditions highly influence PPGI signal quality.

1.1 Aim of the thesis

To access the heart rate of the newborn infant using Photoplethysmography Imaging, a three step process needs to be conducted. Firstly, within the image receptive field, a skin region needs to be selected and continuously tracked along the video frames. This region corresponds to the region of interest. Then, the PPGI signal, corresponding to the average pixel intensity of the previously selected region of interest, needs to be extracted. Finally, the PPGI signal is processed to provide a heart rate estimation.

This thesis aims to develop an innovative region of interest selection method for neonates video recordings, thus replacing the current manual and non robust region of interest tracking methods. This novel region of interest selection method intends to enhance the robustness of heart rate monitoring through photoplethysmography imaging and consequently, contributing to the evolution of this monitoring modality in neonatal care.

In this context, models from the realm of Deep Learning were developed in order to select, for each video frame, six heart rate measuring sites, i.e., regions of interest. The capacity to estimate the predefined regions of interest that are linked to six body parts (head, torso, right arm, left arm, right leg and left leg) during different lightning conditions, skin tones and body positions is explored. Besides the good segmentation results, the developed convolutional neural network also demonstrated to be computationally efficient. Additionally, a post-processing method to further refine the region of interest selection model results was developed.

Algorithms for extraction and processing of the PPGI signal for heart rate assessment were also implemented. Then, the capacity of the developed region of interest selection method to identify successful measuring sites to extract the heart rate was validated using real clinical data, demonstrating a good agreement between the estimated heart rate and the reference.

1.2 Organization of the thesis

The remaining content of this thesis is organized into the following six chapters:

Medical Foundations of the Thesis covers the medical foundations of the thesis including a brief overview of the cardiovascular system and the photoplethysmography imaging method.

Deep Learning covers the fundamentals of Deep Learning.

State-of-the-art briefly presents the standard heart rate monitoring modalities as well as the emerging methods for non-contact heart rate estimation. Then, the state-of-the-art of region of interest selection methods for heart rate estimation through Photoplethysmography Imaging in neonatology is presented. Finally, an overview of the state-of-the-art of methods for image segmentation relying on Deep Learning models is presented.

Clinical Study provides the details of the datasets used to develop and validate the proposed method for region of interest estimation.

Region of Interest Selection describes in detail the developed Deep Learning model for region of interest selection in neonates' recordings. Additionally, a novel post-processing method for region of interest results refinement is presented. The models were validated using data collected at the Neonatology Department of the RWTH Aachen University Hospital as well as at the Saveetha Medical College and Hospital of Chennai, India.

PPGI and Heart Rate Extraction intends to validate the capacity of heart rate assessment when the proposed method for region of interest selection is used to identify the measuring sites along the video frames. In this context, an algorithm to extract and process the PPGI signal for heart rate estimation is described in detail. The performance is evaluated using a subset of the data collected at Saveetha Medical College and Hospital. The agreement between the PPGI estimated heart rate and the reference heart rate is evaluated.

Conclusions concludes this thesis. An overview of the main results and achievements is provided as well as the future perspectives of this topic of research.

Chapter Two

Medical Foundations of the Thesis

This chapter aims to provide the medical background necessary for the following chapters. In Section 2.1, the cardiovascular system is described including its anatomy and physiology. Information about the human physiology presented in this chapter is taken from [32], unless specified otherwise. When relevant, the characteristics of the premature infants cardiovascular system are discussed. Also, the principles of Photoplethysmography Imaging are detailed in Section 2.2.

2.1 Cardiovascular System

The cardiovascular system can be further divided into the heart and circulatory system. While the heart pumps blood through the arteries, the circulatory system carries the blood and is responsible for exchanging nutrients, electrolytes, dissolved gases and waste products between the blood and the surrounding tissues [58]. To accomplish effective and fast exchanges, the circulatory system comprises a vast network of vessels reaching all body tissues, including skin tissues. The two systems work together to ensure that adequate blood flow is delivered to all body tissues [44].

2.1.1 The heart

The heart can be subdivided into the right and left pumps. While the right subdivision pumps blood through the lungs, allowing the exchange of gases between the blood and alveoli, the left subdivision pumps the oxygenated blood to all the body tissues, supplying them with nutrients and oxygen. Thus, the afterload of the left heart subdivision is superior to the afterload of the right heart subdivision.

Each aforementioned heart subdivision is a pulsatile two-chamber pump composed by an atrium, which is a weaker primer pump, and a ventricle which is the main pumping chamber, that pumps the blood through the pulmonary (right ventricle) or peripheral circulation (left ventricle). Therefore, the ventricles supply the main pumping force. Pumping is performed discontinuously by cyclic contraction and relaxation of the chambers that constitute the heart.

2.1.2 Cardiac Cycle

The cardiac cycle comprises the cardiac events that occur from the beginning of one heartbeat to the beginning of the next. Each cycle is initiated by a spontaneous electric triggering impulse, i.e. an action potential, generated in the sinus node. The generated action potential is propagated rapidly through both atria, causing their contraction and consequently the inflow of blood into the respective ventricles. Then, the cardiac impulse travels to the atrioventricular node. Because of this conducting system, there is a delay of more than 0.1 seconds before the impulse reaches the ventricles. Then, the Purkinje fibers distribute the impulse through the ventricles causing their strong contraction and consequently the inflow of blood into the aorta and pulmonary artery.

Figure 2.1 illustrates the cardiac events during a cardiac cycle. The cardiac cycle comprises two periods: a relaxation period, called diastole, during which the heart fills with blood; and a contraction period, called systole.

The diastole period can be further divided into three phases. The first period is called period of rapid filling of the ventricles and it is characterized by the rise of the ventricular volume curve for about the first third of diastole. This fast increase is caused by the opening of the atrioventricular valves due to the large amounts of blood accumulated in the atria

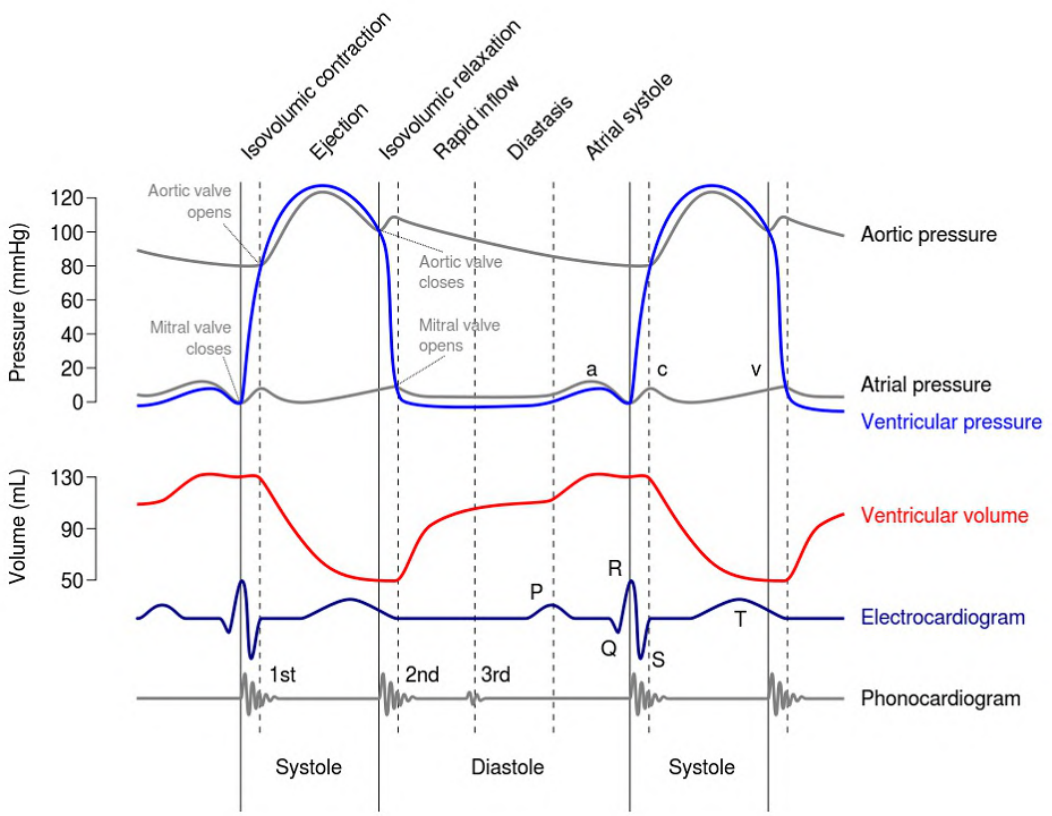


Figure 2.1: Cardiac events within the cardiac cycle. The values both for pressure and volume are linked to the left heart pump (left ventricle and left atria) of an average healthy adult. Figure extracted from [28].

during the ventricular systole period, that caused the development of a moderately increased pressure with respect to the low diastolic ventricular pressure. This pressure difference between the two chambers allow blood to flow rapidly into the ventricles. During the middle third of diastole, only a small amount of blood normally flows into the ventricles. During the last third of diastole, the atria contract and give an additional thrust to the inflow of blood into the ventricles. The latter justifies the term of primer pump attributed to the atria since atrial contraction usually causes an additional 20 % filling of the ventricles, increasing the ventricular pumping effectiveness as much as 20 %. The last diastole period is called atrial systole.

Similarly to diastole, the systole period can be further divided into three phases. Systole starts with a period of isovolumic (Isometric) contraction. Initially, the ventricular pressure rises due to the beginning of ventricular contraction. This rising causes the atrioventricular valves to close. However, the opening of the semilunar valves do not happen until the ventricle builds up enough pressure to equal the pressures in the aorta and pulmonary artery. Thus, during 0.02 to 0.03 seconds, isovolumic or isometric contraction is occurring in the ventricles, but there is no blood flowing from the ventricle, meaning that tension is

increasing in the muscle but the volume inside the chambers is constant. As soon as the pressure inside the ventricles surpass the pressure in the arteries, the semilunar valves open. While in adults, the left ventricular pressure reaches the 120 mmHg, in newborns, it reaches the 70 mmHg [79]. Immediately, blood begins to flow out of the ventricles, initiating the ejection period. About 70 % of the blood emptying occurs during the first third of the period of ejection and the remaining 30 % emptying during the next two thirds. Therefore, the first third is called the period of rapid ejection, and the last two thirds, the period of slow ejection.

Finally, with ventricular relaxation begins the period of isovolumic (Isometric) relaxation. Initially, the ventricular pressure decreases rapidly. Then, when the ventricular pressure equals the pressure in the arteries, the aortic and pulmonary valves close. For another 0.03 to 0.06 seconds, the ventricular muscle continues to relax, even though the ventricular volume does not change, giving rise to the period of isovolumic or isometric relaxation. During this period, the intraventricular pressures decrease rapidly back to their low diastolic levels. Then the atrioventricular valves open to begin a new cycle of ventricular pumping.

2.1.3 Heart Rate

The described periodic activity of the heart can be measured and further analysed in order to access the well-being of the infant. The heart rate corresponds to the number of heart contractions within a minute. A normal heart rate depends on a variety of factors including age. In fact, for a premature or newborn infant, a normal heart rate corresponds to double of the normal heart rate of an adult. Additionally, while in healthy adults diastole comprises the longest time period, in newborn infants, diastole and systole have an equal duration (approximately 0.2 s each). Another difference in the heart rate parameter between the aforementioned age groups is the normal heart rate values range. The heart rate of a healthy newborn infant comprises a wide range of normal heart rate values [39] [79]. Table 2.1 shows the mean heart rate range for each age group.

Table 2.1: Mean heart rate range for each age group [39].

Age	Mean heart rate (range) [bpm]
premature	120-170
0-3 months	100-150
3-6 months	90-120
6-12 months	80-120
1-3 years	70-110
3-6 years	65-110
6-12 years	60-95
> 12 years	55-85

For newborn infants with less than 24 hours of life, the mean heart rate is 120 bpm. Afterwards, the heart rate progressively rises to 160 bpm at the first month. Then, the heart

rate progressively decreases stabilizing at 75 bpm, when reaching 12 years of age.

2.2 Principles of Photoplethysmography

The optical properties of the skin suffer subtle fluctuations within the cardiac cycle. This macroscopic, yet invisible, fluctuations are originated by the periodic pressure changes which cause a rhythmic expansion and contraction of arterial blood vessels: during systole, the arterial blood vessels are expanded; during diastole, the opposite is patent. The expansion of the arterial blood vessels is proportional to its blood volume. Since blood has a different absorption spectrum than the surrounding tissue (see Appendix B), changes in blood volume within a cardiac cycle are detectable through the amount of reflected light from the skin surface: during the low blood pressure period, the tissue contains less blood, leading to a higher light reflection; on the other hand, during the high blood pressure period, the increase in blood volume in the tissue will lead to a decrease in the reflected light [80].

2.2.1 Photoplethysmography and Photoplethysmography Imaging

Photoplethysmography and Photoplethysmography Imaging are monitoring methods that take advantage of the subtle blood volume fluctuations in the skin surface.

The emergence of photoplethysmography dates back to 1938 [37]. The fundamental working principles of this monitoring technique remains unchanged. A light source irradiates light of intensity I_{in} in the visible or near-infrared range into the subject's skin. Then, a portion of this light is reflected, I_{refl}, another portion is absorbed by the tissues, I_{abs}, and the remaining is transmitted, I_{trans}. Figure 2.2 illustrates the photoplethysmography working principle. The periodic changes in the arterial vessels blood volume modulate both the reflected and transmitted light that will reach the sensors [11]. Additionally, the intensity of the reflected, absorbed and transmitted light also depends on a great variety of factors including the skin tone, blood perfusion, blood oxygenation level [1] and the wavelength of the irradiated light. Usually, in neonatoly, PPG sensors are placed, in special socks.

The transition of photoplethysmography into a non-contact monitoring method was not until the year of 2000 [10] [80] with the introduction of photoplethysmography imaging. Based on the same principles of photoplethysmography, photoplethysmography imaging replaces the photodiode, traditionally used as a sensor in contact based photoplethysmography, for a video camera. Thus, instead os having a single measuring site, the video camera provides a two-dimensional array of sensors, i.e. each individual pixel is considered as a small sensor. Each sensor measures the photons that are reflected from the patient's skin surface, I_{refl}. PPGI systems include a CCD or CMOS camera with high SNR, quantum efficiency and frame rate [11]. Besides the video camera, a dedicated light source can be used during the acquisitions [9]. However, PPGI signal can still be obtained using the reflection of existing ambient light, if lightning conditions are favourable [74] [78] [11] [1]. Additionally, an optical bandpass filter can be attached to the camera lens. Figure 2.3 Illustrates the photoplethysmography imaging working principle.

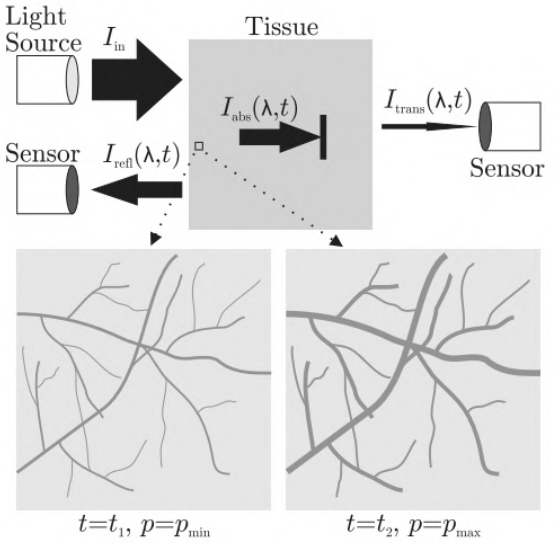


Figure 2.2: Nature of the PPGI signal. Figure extracted from [11].

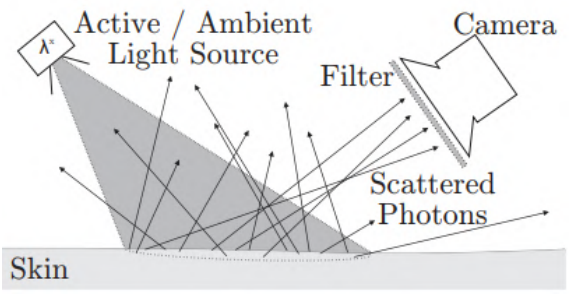


Figure 2.3: PPGI aquisition. Figure extracted from [11].

Chapter Three

Deep Learning

In this chapter, basic concepts regarding Deep Learning for image semantic segmentation are discussed. Section 3.1 introduces the concept of artificial neural networks. Then, a more detailed discussion of Convolutional Neural Networks is provided in Section 3.2. Unless otherwise mentioned, the content presented in the aforementioned sections is taken from [2].

Additionally, this chapter further details the concept of image semantic segmentation (Section 3.3). In Section 3.4, Fully Convolutional Neural Networks are discussed. Finally, this chapter provides an insight into transfer learning (Section 3.5).

3.1 Artificial Neural Networks

Artificial neural networks are inspired in the biological neural system, simulating the learning mechanism of biological organisms. Similarly to the human nervous system, which is based on biological neurons, artificial neural networks are based on basic computational units, also called neurons.

3.1.1 Neuron Model

Figure 3.1 illustrates a biological neuron and the mathematical neuron model. Both neurons receive input signals through the dendrites and produce an output that is sent from the axon. In the mathematical neuron model, the input signals that come from the axons (Figure 3.1 - x_0) interact multiplicatively (Figure 3.1 - $x_0 w_0$) with the dendrite of the receiving neuron, i.e., each artificial synapse has a synaptic strength associated. In other words, the influence of the input signal is controlled by a learned weight (w), also called parameter. All the weighted input signals are then summed in the cell body. The output of a neuron is modelled with a non-linear activation function (f). Commonly, the sigmoid function was employed as the activation function [26]. However, in this thesis, the rectified linear unit function is used instead (more details in Appendix B.1).

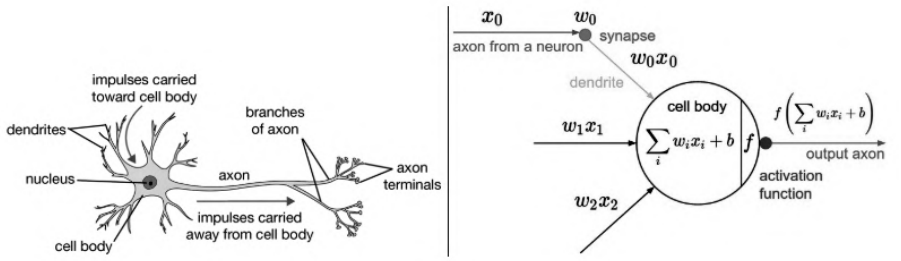


Figure 3.1: Illustration of the biological neuron (left) and mathematical neuron model (right). Figure extracted from [26].

In sum, the neurons response to input signals is modelled as shown in Equation 3.1.

$$a = f(\sum_i w_i x_i + b) \tag{3.1}$$

where a represents the activation value of the neuron, f denotes the activation function, w_i the synaptic strength weight associated with the neuron input, x_i, and b denotes the neuron bias.

3.1.2 Neural Networks

Neural networks are an assembly of neurons that are connected and organized in layers, meaning that the output of the neurons of a layer can become the input of the neurons of the following layer. Note that neurons organized in a cycle are not allowed since it would

originate an infinite loop. Thus, the input is always passed forward. The connections between neurons can be arranged in a fully-connected layer, where neurons between two adjacent layers are fully pairwise connected. This arrangement comprises a high number of operations which is computationally expensive [26]. Figure 3.2 illustrates an example of a neural network structure relying on fully-connected layers.

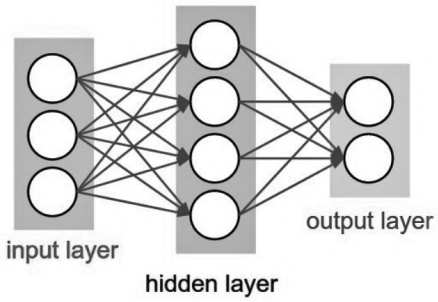


Figure 3.2: Artificial neural network illustration. Figure extracted from [26]

As seen in Figure 3.2, artificial neural networks comprise an input layer, some hidden layers and an output layer. This layer arrangement is able to define a function by using the input and propagating the computed values from the input layer to the output layer using the learned weights as intermediate parameters. The learning procedure comprises the adjustment of the weights using training data, i.e., examples of input-output pairs of the function to be learned. For example, the training data is the pixels of an image (input) and the pixel wise label (output). This adjustments occur by using the predictions obtained from the inputs in the training data and comparing them with the annotated output label in the respective training data pair. Prediction errors cause a weight adjustment in the neural network relying on the backpropagation process (more details in Section 3.1.3). The goal is to provide many different training examples to make the neural network correctly predict an example not seen before (model generalization).

The size of the neural network i.e., the number of hidden layers, is linked to its ability to learn more complex functions from the available training data. Neural networks with multiple hidden layers are frequently referred as deep neural networks.

3.1.3 Backpropagation

Backpropagation comprises the process of updating the weights of a neural network during the training process. By progressively modifying the weights, the function defined by the neural network is adjusted to provide more accurate predictions, reducing the prediction error. In other words, it finds the set of weights that provide the loss function local minima.

3.1.3.1 Loss Function

One can define the training process as an optimization process, where the goal is to find a set of weights that provide the better classification performance. To evaluate the performance

of a specific set of weights a loss function is used. This loss function provides a value that represents the difference between the neural network prediction and the desired annotated output. It is defined as a function of the network predictions. The loss function employed in the current thesis is the Cross-Entropy Loss (Equation 3.2).

$$L = -\sum_{x} \mathrm{p}(x)\log(\mathrm{q}(x)) \tag{3.2}$$

where p refers to the true distribution and q the estimated distribution. Note that the losses are averaged across observations for each minibatch of images. By calculating the loss function over a minibatch of images, instead of computing the loss of one single example, it is possible to have as estimate of the loss over the training set. The quality of the estimate increases with the size of the minibatch. Additionally, computation over a batch of m images can be much more efficient than m computations for individual examples [40].

3.1.3.2 Gradient computation

After the forward pass, where a training image is fed into the network and a forward cascade of computations across the network layers occurs, the network presents an output and the loss function is computed. The gradient computation step comprises the computation of the loss function gradient with respect to the network weights, $\nabla L(x)$ where x is a vector of inputs i.e., the training image and the neural network weights. Note that it is possible to compute the gradient with respect to the input image, however only the gradient for the weights is computed to perform the weights update [26].

The gradient of the loss function, which is the vector of partial derivatives, is calculated using the chain rule of differential calculus. These partial derivatives provide an insight into the weights that have contributed the most to the loss and are used to update the weights. Using the network illustrated in Figure 3.3, an example of gradient computation taken from [73] will be given.

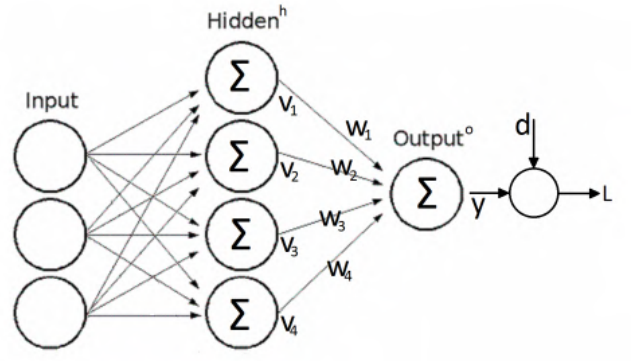


Figure 3.3: Illustration of a simple multilayer network. Figure extracted from [73].

For simplicity reasons, the quadratic loss function will be used in the example. The loss function, L, corresponds to Equation 3.3.

$$L = \frac{1}{2}\sum_{l}(e_l)^2 \tag{3.3}$$

$$e_l = d_l - y_l \tag{3.4}$$

where e_l refers to the error between the desired annotated output, d_l, and the network prediction, y_l. The network prediction, y_l, depends on outputs of the previous layer, v_j and the output layer weights, W_j^o, as seen in Equation 3.5.

$$y_l = \sum_{j} W_j^o v_j \tag{3.5}$$

Using the chain rule of differential calculus, the Jacobian is given by:

$$\frac{\partial L}{\partial W_{jl}^o} = \frac{\partial L}{\partial e_l}\frac{\partial e_l}{\partial y_l}\frac{\partial y_l}{\partial W_{jl}^o} \tag{3.6}$$

Calculating the respective partial derivatives for equations 3.3, 3.4 and 3.5 originates the following Jacobian for the output layer:

$$\frac{\partial L}{\partial W_{jl}^o} = -v_j e_l \tag{3.7}$$

Then, the gradients of the hidden layers are sequentially computed. Using the computed gradients, the weights can be updated using, for example, the Adam optimization algorithm, which will be described in the following Section.

3.1.3.3 Adam optimization algorithm

The Adam optimization algorithm is an optimization algorithm used to update the network weights to minimize the loss function. It is an extension of stochastic gradient descent. As the authors of this algorithm state [43], this method is straightforward to implement, is computationally efficient, has little memory requirements and is well suited for problems that are large in terms of data and/or parameters. Additionally, the Adam optimization algorithm is less likely to stagnate in saddle points of the loss function when compared with the stochastic gradient descent algorithm.

The Adam optimization algorithm combines the idea of introducing a momentum term in the stochastic gradient descent and the RMSProp optimization algorithm. Instead of minimizing the loss function by updating the weighs in the negative gradient direction, the update is made relying on two momentums. The first momentum is computed using the current gradient estimate. The second is computed using the squared gradient estimate. These two momentums are then corrected by introducing the current iteration number, in order to avoid the problem of making large steps at the begin of the optimization process

(since the initial first and second momentum start at zero). The weight updating step uses the aforementioned unbiased momentums. The following pseudo-code summarizes the Adam optimization algorithm.

Algorithm 1: Adam optimization algorithm

first _momentum = 0 ;
second _momentum = 0 ;
for t = *0:num _itearations* **do**
  $\nabla L(x_t)$ = compute _gradient(x) ;
  first _momentum(t+1) = $\beta_1 \times$ first _momentum(t) + (1-β_1) $\times \nabla L(x_t)$;
  second _momentum(t+1) = $\beta_2 \times$ second _momentum(t) + (1-β_2) $\times \nabla L(x_t) \times \nabla L(x_t)$;
  first _unbias(t+1) = first _momentum(t+1) / (1-β_1^t) ;
  second _unbias(t+1) = second _momentum(t+1) / (1-β_2^t) ;
  x -= learning _rate $\times$ first _unbias(t+1) /$\sqrt{\text{second\_unbias}(t+1)+1e-7}$ $\leftarrow$
   update weights ;
end for

where $\nabla L(x_t)$ is the gradient estimates, β_1 and β_2 refers to friction. The β_1 hyperparameter decays the current first momentum and the β_2 hyperparameter decays the current second momentum. They typically assume a high number (0.9 and 0.999 for the first and second momentum respectively).

The addition of these momentums address the saddle points problem since, despite the zero gradient, the momentums will allow the evolution from the local minima. Note that, similarly to stochastic gradient descent, the weights are updated for every minibatch of images by evaluating the loss and respective gradients within the minibatch.

3.1.3.4 Learning Rate

The learning rate is a hyperparameter that determines the step size of each iteration towards the loss function local or global minima. If set too high, it causes suboptimal performance. If set too low, it causes slow convergence. There is a need to formulate the learning rate as a decreasing function over the training process to further improve the network performance. During the initial iterations, when the loss function is still far from its minimum, it is beneficial to have high learning rates, causing bigger weight changes. When the network is closer to its optimal solution, it is beneficial to have a lower learning rate, that will finetune the solution. This process of learning rate scheduling is applied in the current thesis.

3.1.4 Practical issues in Neural Networks training

Despite the potential of a neural network to be an approximation of complex functions, there are some challenges regarding the training process of a neural network, that might compromise its performance.

3.1.4.1 Overfitting

One of the most important challenges refers to the possibility that, by fitting a model to specific training data, the model will not be able to provide a good prediction performance on unseen data. In other words, the possibility of a model to not generalize. Thus, when the model perform extremely good on the training data and poorly in test data, the model suffered overfitting. This problem is common when the training data is insufficient i.e., insufficient number of training examples. It is possible to design a neural network architecture less prone to overfitting.

3.1.4.2 Vanishing and Exploding Gradient Problem

The vanishing gradient problem is linked to neural networks that comprise a high number of layers, making the weight updating process unstable. When this problem occurs, the weight updates in the early layers of the network can either be extremely small (vanishing gradient) or increasingly large (exploding gradient). This phenomenon is primarily caused by the product operation in the gradient computation (chain rule) which can lead to an exponential decrease or increase the gradient across the layers.

The vanishing gradient problem is particularly present when the Sigmoid activation function is used. From Figure 3.4, it is possible to conclude that, for large and small inputs, the derivative is close to zero. When using the chain rule for gradient computation, where the derivatives of each layer are progressively multiplied, the gradient decreases exponentially during backpropagation when the derivatives are close to zero. This causes the gradient to vanish in the early network layers.

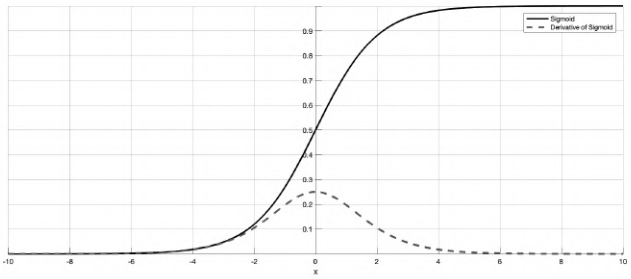


Figure 3.4: Sigmoid function and its derivative. Extracted from [21]

There are several methods to address this problem including the usage of the ReLU activation function, whose derivative is always one for positive values. In addition batch normalization layers, which will be further described in Section 3.2.3, also provide a solution for this problem.

3.1.4.3 Internal Covariance Shift

The internal covariance shift problem arises from the change of the layers' input distributions as the parameters of its previous layer are updated [59]. This makes the network

converge more slowly [40]. Similarly to the vanishing and exploding gradient problem, the internal covariance shift problem scales during propagation across the layers. Thus, this problem is particularly important for deep convolutional networks. The solution lies with the usage of batch normalization layers (described in Section 3.2.3).

3.2 Convolutional Neural Networks

The current thesis introduces a novel Convolutional Neural Network to address the task of image semantic segmentation (details in Section 3.3). Convolutional neural networks extend from the ordinary neural networks. This kind of networks are designed to have an image as the input. Having an image as input makes the use of fully-connected structures not manageable due to the high number of weights to be optimized [26]. For example, an input image of resolution of $576 \times 960 \times 3$ would require, for the first fully-connected layer, 1 658 880 weights. Thus, CNN are tailored to encode certain properties, such as high efficiency and reduced the amount of parameters. In particular, each neuron, instead of being connected to all neurons in the previous layer, will only be connected to a small region of neurons [26].

The CNN neurons are arranged in three dimensions: width, height and depth [26]. For the first layer of a CNN and an input image of resolution of $576 \times 960 \times 3$ the width, height and depth would be 960, 576 and 3, respectively. Thus, for the first layer, the depth refers to the number of input colour channels. However, for the remaining layers, the depth refers to the number of feature maps, also called channels.

As previously described, a CNN comprises a sequence of layers that sequentially transform a three dimensions input into an output volume. The layers used to build the CNN architecture will be detailed forthwith.

3.2.1 Convolutional layer

Convolutional layers are the core building block of a CNN. This layer type comprise sets of three dimensional learnable filters (or kernels). While the width and height of these filters tend to be low, the depth equals the full depth of the input volume, which can reach thousands of channels. Using this filter structure allows each neuron in the convolutional layer to connect with a local region of the input volume [26].

During the forward pass, each filter is convolved across the width and height of the input volume. Meaning that, for each iteration, a dot product between the filter's parameters and the input volume region is computed (see Figure 3.5). At the end, each filter produces a two dimensional feature map that gives the responses of that filter at every spatial position. Thus, by stacking the produced feature maps along the depth dimension, an output volume is produced. The networks early convolutional layers filters commonly activate in response to simple visual features such as an edges. While higher layers usually learn filters that recognize more complex patterns [26].

The size of the convolutional layer output volume depend on three hyperparameters: depth, stride and padding. The first refers to the number of filters. The stride specifies the amount of pixels that the filter shifts at time. Finally, the padding refers to the number of zeros introduced around the input volume borders. All these hyperparameters are set before

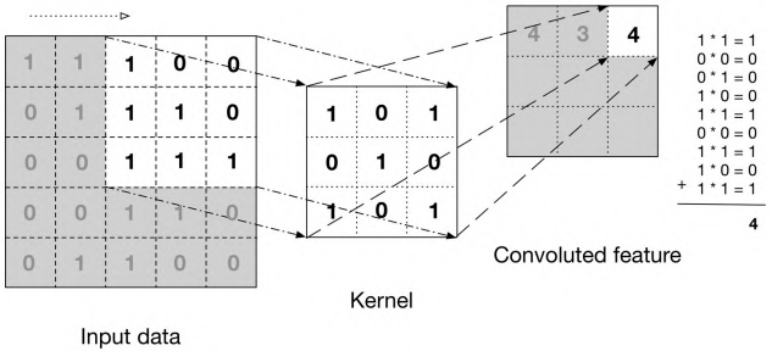


Figure 3.5: Illustration of a convolution operation. Figure extracted from [62].

training. Equation 3.8 shows the relationship between output volume spatial size, O, the input volume spatial size, I, the filter spatial size, K, the stride, S and the padding, P [26].

$$O = \frac{I - K + 2P}{S} + 1 \tag{3.8}$$

Commonly, an activation function proceeds a convolutional layer. In the case of a ReLU activation function, every negative value in the resulting feature map will be replaced with zero.

3.2.2 Pooling Layer

Pooling layers are usually placed in-between successive convolutional layers. This type of layers downsamples the feature maps in order to reduce the number of parameters and the number of operations in the network. The pooling layer also plays a role in decreasing the probability of overfitting [26].

The most common form of this layer is maxpolling (see Figure 3.6), where the output volume is produced by keeping the maximum input value within the filter. Besides maxpolling, average polling is also used in CNN.

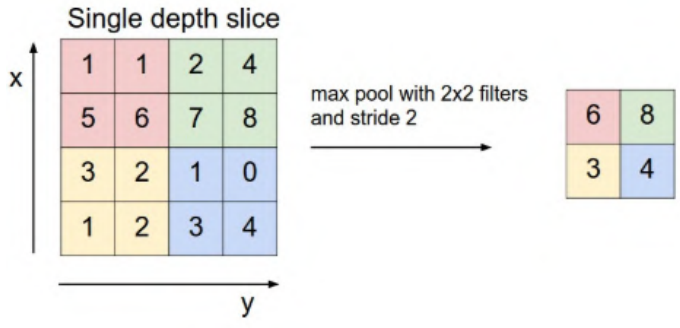


Figure 3.6: Illustration of a maxpolling operation. Figure extracted from [26].

Similarly to the convolutional layer, the output spatial size depend on the filter size (width and height) and the filter's stride. Note that this layer does not have parameters

associated, since it computes a fixed function of the input [26].

3.2.3 Batch Normalization

To address the problem of internal covariate shift, Ioffe et al. [40] propose the Batch normalization layer, in which the input feature map distribution is normalized. Therefore, each batch normalization layer comprises two learnable parameters: the feature map is multiplied by the gama parameter, referring to standard deviation, and then the beta parameter is added, referring to the mean. The normalization is performed for each training minibatch.

Besides providing a solution to the internal covariate shift problem, this layer allows the usage of higher learning rates by reducing the dependence of gradients on the scale of the parameters [40]. Therefore, by incorporating this layer in the CNN structure, it is possible to achieve a higher training speed. Additionally, Batch normalization reduces the probability of overfitting because it has a slight regularization effects. In fact, Ioffe et al. state that the noise added by this layer can eliminate the need for Dropout layers.

3.2.4 Transpose convolution

Transpose convolution is a layer in which a upsampled dense feature map is computed from a downsampled and course input [54]. This layer can also be referred as deconvolutional layer. Contrary to simple interpolation methods, transpose convolutional layers learn to upsample the course inputs in an optimal manner using learnable weights. These weights are bases to reconstruct shape of an input image [54].

In convolutional layers, an activation in the output feature map is connected to a region of the input feature map, the region corresponding to the filter window. In transpose convolutional layers a single input activation will influence multiple activations in the output feature map (see Figure 3.7). In other words, an activation in the input feature map is distributed over a region in the output feature map, by multiplying the input value by the corresponding filter weight. This process is then repeated for every value in the input feature map. Note that there is an overlap of values in the output feature map. Thus, the final activations of the output feature map corresponds to the sum of the overlapped values.

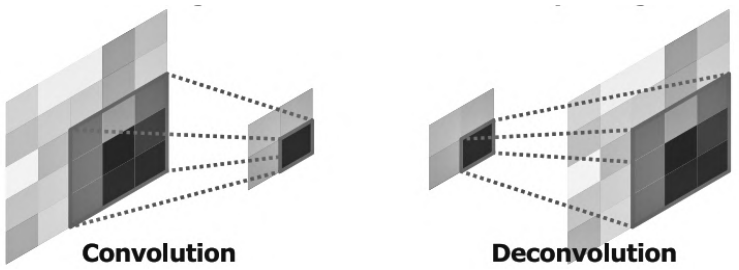


Figure 3.7: Illustration of a transpose convolution operation. Figure extracted from [54].

3.3 Semantic Image Segmentation

Semantic image segmentation aims to assign one of the predefined object classes to every image pixel. This pixelwise labelling task constitutes one of the most challenging problems in computer vision since it comprises the challenge of simultaneous classification and localization: objects in the images are associated to a semantic concept and, its classification label is attributed to the pixels with the appropriate coordinates in the output score map[63].

CNNs have shown an outstanding performance in computer vision tasks given their ability to capture abstract, meaningful and compact feature representations. This success is linked to the CNNs built-in invariance to spatial object transformations namely object rotation, rescaling and translation. While this invariance contributes to the success of the classification tasks, it inherently limits the spatial accuracy in the localization task [16], where abstraction of spatial information is undesired.

Besides the spatial abstraction associated with CNNs structures, there is an additional barrier to accurate object localization prediction: the reduced resolution of the image multidimensional feature representation. The image downsampling arises from the repeated application of pooling and downsampling convolutional layers characteristic of standard CNNs [15].

3.3.1 Receptive Field and Effective Receptive Field analysis

The receptive field is the region of the input feature map space (including the depth) that influences the value of a neuron in the output feature map. The receptive field size of a neuron in the output feature map after a convolutional layer depends on the RF of the input feature map as well as the convolution filter size and stride. A high filter size and stride will generate a large RF.

The ERF is the region of the input image (including the depth) that can possibly modulate the neuron activity in the output feature map. Note that RF and ERF are the same for the first convolutional layer and progressively differ along the FCNN [48].

The RF and ERF analysis constitutes an important step during the design of FCNN based architectures for computer vision tasks. In lower layers of the network, the kernel size should be sufficiently large to originate wide RF capable of capturing the global context from input image. On the other hand, if the context is too wide, it will include an excessively large neighbourhood leading to the presence of noise in the feature maps. In fact, Fakhry et all. [25] proved the above-mentioned statement by applying two networks with extreme kernel sizes to the task of semantic image segmentation. Employing a convolutional layer with a kernel size of 3×3 as the first layer of the network revealed to lead to a very small receptive field to successfully capture global discriminative features from the input image. On the other hand, an initial convolutional layer with a kernel size of 11×11 revealed to be excessively large to lead to a good performance.

3.4 Fully Convolutional Neural Networks

Fully convolutional neural networks are designed to address the task of image semantic segmentation. As previously mentioned in Section 3.3, the task of semantic image segmentation comprises both localization and classification. To achieve a high performance in the image classification task, the models should adopt a deep convolutional structure with large kernel sizes and multiple maxpooling layers to enable the extraction of abstract and global features from the input image. In the other hand, for the task of localization, the models should adopt a fully convolutional structure with lower kernel sizes to retain the object shape and localization information.

Since the model structure requirements for each task are naturally contradictory, the developed FCNN models should adopt a structure that optimizes the trade-off between localization accuracy and classification performance. There are several types of model structures, including image pyramid structures, atrous convolution structures and encoder-decoder structures. The latter will be further discussed since it is the structure used in the current thesis.

3.4.1 Encoder networks, the features extractors

The purpose of the encoder network is to automatically extract meaningful features from the input image. The process of constructing a meaningful multidimensional feature map is achieved by submitting the input image into a sequence of convolution operations, where the complexity of the extracted features gets progressively higher. It is a hierarchical feature engineering process in which the filters in earlier layers capture primitive characteristics of the image (lines, edges, etc.) whereas the filters in deeper layers capture abstract and complex characteristics with semantic meaning [3].

Regular pre-trained image classification networks are usually adapted to address the task of feature extractors. The adaptation comprises the replacement of the fully connected layers for convolutional layers. The ResNet models family is detailed forthwith since it is used in the current thesis as the decoder network, instead of the widely used VGG-16.

3.4.1.1 ResNet

The ResNet models architecture won the first place in the ILSVRC 2015 competition [67]. These models address the challenge of image classification, meaning that the network expects an RGB image and outputs an image label.

The ResNet networks architecture uses a residual learning framework to address the challenge of training deep neural networks. As illustrated in Figure 3.8, the residual block proposed by He et. al. includes shortcut connections that skip one or more layers. These shortcut connections simply perform identity mapping, and their outputs are added to the outputs of the stacked layers. The staked layers in each residual building block fit a residual mapping. The authors of [34] argue that the performance of the network is not compromised

by adding extra layers, using this residual blocks, because if an identity mapping is already optimal, the residual mapping can be pushed to zero.

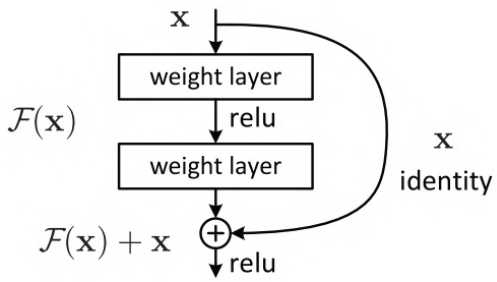


Figure 3.8: Residual learning: a building block. Figure extracted from [34].

ResNet-50 is a 50 layer CNN model that belongs to the ResNet models family [34]. The combination of an initial max pooling operation and the stacked convolutional layers of the four ResNet-50 3-layer bottleneck residual blocks (see Figure 3.9) significantly reduces the spatial resolution of the resulting feature maps by a factor of 32. On the other hand, the number of channels of the feature maps is increased from 3 to 2048 channels.

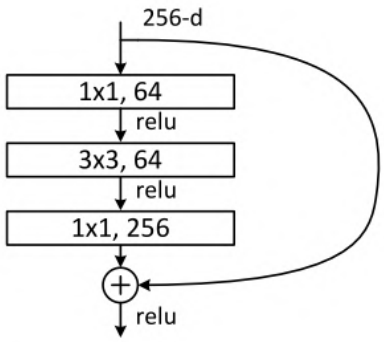


Figure 3.9: Bottleneck building block. Figure extracted from [34].

3.4.1.2 ResNet-50 vs VGG-16

This Section details the ResNet step-change improvements over VGG-16.

Receptive Field: As previously stated in Section 3.3.1, a FCNN model for the task of semantic image segmentation requires the ability to capture global features of its input to be successful. The moderately large kernel size (7×7) of the ResNet-50 first convolutional layer optimizes the trade-off between capturing global discriminative features and noise interference. This kernel size contrasts with the kernel size (3×3) of the VGG-16 first convolutional layer. The larger kernel size of the ResNet-50, leads to an increase of the receptive fields of each neuron in the resulting feature maps, enabling the extraction of more discriminative features which ultimately improves the overall performance. On the other hand, in deeper layers, the ResNet-50 network comprises stacked residual blocks containing

convolutional layers with small kernel sizes to allow the model to grow deeper. Since these successive convolutional layers steadily increase the network receptive field, regardless of the kernel size, the higher layers will still comprise a very large receptive field. However, not all pixels belonging to this large receptive field contribute equally to its corresponding unit in the feature map. In fact, studies [56] show that modern deep CNNs, such as the ResNet-50, tend to gather information mainly from a much smaller region, i.e. the effective receptive field, of the theoretical receptive field. Thus, the effective receptive field does not include an excessively large area of the input image, leaving aside the possibility of undesired noise in the feature maps that can negatively affect the localization performance.

Computation efficiency: Despite being a much deeper network than VGG-16, the ResNet-50 is more efficient with respect to computational complexity. By reducing the features map resolution by a factor of four in the first two layers with respect to the input image, and employing stacked residual blocks containing convolutional layers with small kernel sizes, the ResNet-50 manages to achieve only 3,8 billion FLOPs, which is only 25 % of VGG-16 (15.3 billion FLOPs)[34].

Memory: While the ResNet-50 comprises 25,6 million parameters, the VGG-16 comprises 138 million parameters. This difference reflects on the memory needed to store the network.

3.4.2 Decoder Networks

The feature maps that result from the encoder networks usually have a low-resolution when compared with the input image. To obtain high-resolution predictions, the FCNN architectures include a decoder network, responsible for the feature maps upsampling process.

The existing decoder networks are further discussed in Section 4.2.1.

3.5 Transfer Learning

Transfer learning is a training procedure in which a model is firstly trained for a certain general task, using a large dataset, and then the model is fine-tuned for a specific task, using a small dataset. Note that both tasks need to be related and the used datasets should have similar content.

Training an entire CNN from scratch, using a small dataset, usually leads to overfitting. Instead, it is common to pre-train the model using a very large dataset, and then use the obtained weights either as an initialization of the fine-tuning stage or even fixed for the task of interest. During the fine-tuning process it is possible to fine-tune all the weights of the CNN, or keep some of the earlier layers' weights fixed, and only fine-tune some higher-level portion of the network [26]. For CNNs, the dataset that is commonly used in the pre-training phase is the ImageNet dataset, which comprises 1.2 million images with 1000 categories.

During the fine-tuning process, a smaller learning rate is usually used, in comparison to the learning rate used during the pre-training stage. This is because it is expected that the weights are relatively good after pre-training, thus, it is not desirable to distort them too quickly and too much [26].

All in all, transfer learning is a powerful tool, which allows the development of high performance CNN models using small datasets. Therefore, it reduces the need for large datasets which can be extremely hard to obtain, specially in the field of medicine.

Chapter Four

State-of-the-art

The current project aims to adapt methods from the realm of deep learning, namely methods for image semantic segmentation task, to formulate a model for automatic ROI selection. This chapter aims to provide an insight into the current methods for ROI selection and the current state-of-the-art of image semantic segmentation deep learning models.

Section 4.1 briefly covers the classic contact-based HR monitoring methods, as well as novel unobtrusive HR monitoring methods. Section 4.1.1 focuses on ROI selection methods employed for PPGI signal extraction in the scope of neonatology. Finally, Section 4.2 introduces the state-of-the-art of image semantic segmentation methods, including human body part segmentation models.

4.1 HR Monitoring Methods

In standard clinical environment, the continuous HR assessment commonly depends on reliable and reasonably priced monitoring techniques such as the ECG and/or the PPG [46]. The ECG is the most widely used method to monitor this vital sign. It records the cyclic electrical activity that is generated by the cardiac muscle cells and projected onto the body surface. This is achieved by measuring the voltage between specific measuring sites, requiring the attachment of electrodes to the patients' body. On the other hand, the PPG is based on the fact that fluctuations in the arterial vessels blood volume, within the cardiac cycle, are responsible for the change in the tissue's light abortion, as described in Section 2.2. Besides the HR, the PPG also measures the oxygen saturation given the different absorption spectrum of the oxygenated and deoxygenated hemoglobin [11]. While providing reliable results, both ECG and PPG are contact based methods, which makes them inappropriate for specific types of patients, including patients with skin burns, patients with skin diseases and neonates [46].

Given the drawbacks linked to classical HR monitoring methods, several research groups are exploring noninvasive and unobtrusive methods to continuously determinate the HR. New approaches based on different measurement principles have been proposed such as capacitive electrocardiography, ballistocardiography, video-based motion analysis, thermography and photoplethysmographic imaging [11].

Similar to conductive ECG, the capacitive electrocardiogram takes advantage of the voltage signal caused by the de- and repolarization of the cardiac muscle cells. This monitoring method measures the differential bioelectrical signal unobtrusively due to capacitive coupling, not requiring any galvanic contact to acquire the electrocardiogram waveform [11].

The ballistocardiography as well as the video-based motion analysis methods both take advantage of the body surface displacement and vibration effects induced by the contraction of the heart chambers and by the pulse wave that travels trough the vascular system. There are several methods to measure these mechanical effects. Ballistocardiography systems can range from piezoelectric polymer films placed under a bed's mattress to pneumatic and hydraulic methods that detect pressure variations in, for example, air cushions. Additionally, ballistocardiography can also rely on optical methods which include the measurement of optical fibers deformations in a mattress and the scattering of IR light inside the mattress itself [11]. On the other hand, video-based motion analysis methods quantifies subtle displacements of the body surface due to blood pulsation relying on video recordings.

The thermography method is based on the fact that fluctuations in blood flow produce cyclic heat patterns that are synchronized with the HR. The subtle changes in temperature can be visualized and analyzed in the far or mid-infrared range. Note that, similarly to photoplethysmographic imaging, the aforementioned patterns are related to superficial perfusion [11].

Photoplethysmographic imaging, PPGI, captures the blood volume microscopic fluctuations relying on the macroscopic changes in the optical properties of the skin, as detailed

in Section 2.2.1. The photoplethysmographic imaging method state-of-the-art is further detailed forthwith.

4.1.1 HR Estimation through PPGI in Neonatology

Regardless the type of patient, most methods for video-based HR extraction follow the same line of approach. The selection of a ROI and its tracking over time usually constitutes the first step. The ROI comprises skin regions containing pulsatile information. Once ROI is selected, the next step involves the extraction of the signal and the respective HR from the pixels belonging to ROI. Figure 4.1 illustrates the aforementioned steps.

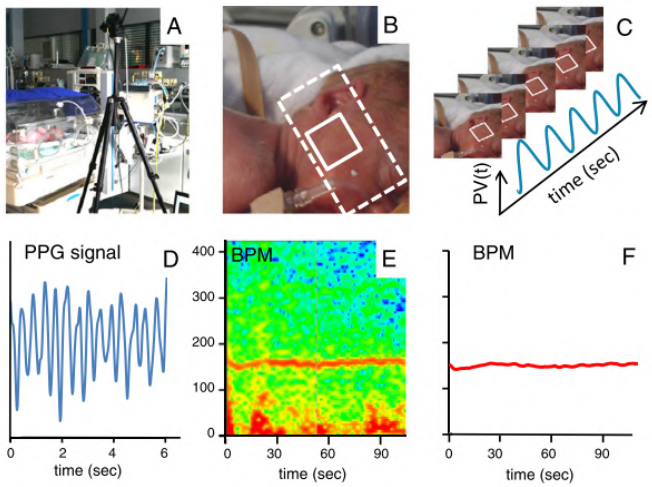


Figure 4.1: Illustration of the steps for HR detection through PPGI. Figure A refers to the experimental setup. Figure B refers to the region of interest selection. Figure C refers to the PPGI signal extraction. Figure D refers to the signal processing step. Figure E and F refer to heart rate computation. Figure extracted from [1].

The majority of studies related to HR extraction from PPGI signals mainly focus on adult subjects restricted to controlled environments where the subject motion is minimum and lightning conditions are optimal [77] [38] [64] [74] [47]. However, the well-established ROI selection methods used in relatively still adults are not suitable in the scope of neonatology given the unpredictability of the neonates' positions and the high level of body motion in the recordings. The latter makes the selection of the ROI and its tracking particularly challenging. Additionally, the incubator restricts the camera angle to the neonate and can cause unwanted refractions. Up to date, few research groups applied the topic of HR estimation through PPGI in neonatology applications. The current section provides an overview of the state-of-art of HR estimation through PPGI in the scope of neonatology, specifying the methods employed in the ROI selection step, the main focus of the thesis.

4.1.1.1 Datasets

The reviewed literature with respect to studies regarding neonates, present results based on datasets that range from seven [68] to 30 [14] different subjects. These small

datasets compromise the application feasibility of the presented works in actual NICUs. The challenge of sharing large amounts of data associated with raw video streams, combined with the rigorous requirements of data acquisition in a NICU environment and the impossibility to properly anonymize the raw data represent a major drawback to the development of large datasets.

The more common type of datasets comprises videos recorded in a controlled environment and dedicated light sources, where the cameras are placed directly above the neonate. The videos were recorded either through the incubator glass [9] [68] [18] [4], directly with open incubators [1], or through a specially-drilled hole in a closed incubator [78] [14]. Some dataset's recordings comprise the majority of the infant's body, whose skin is partially or not covered [78]. While others apply zoom to focus specific uncovered body parts [1]. On the other hand, Sikdar et al. [70] presents HR results based on a dataset that comprises a diverse range of body positions and angles with respect to the video camera.

With the exception of Blanik et al. [9], the aforementioned works rely on RGB cameras. Blanik et al. used an optical filter with a pass band above 720 nm attached to a camera with a sensitive wavelength range of 320-950 nm to filter most of the ambient light, minimizing fluctuating lightning conditions and artifacts. Additionally, the neonates were illuminated by an infrared light emitting diode array (850 nm).

4.1.1.2 ROI selection methods

Despite presenting an important contribution to the field regarding signal processing methods, in [1] [68] [18] [78] [4] [70] the PPGI signal is extracted within a manually chosen region that is sometimes tracked along the frames resorting to rudimentary object tracking methods. Since the selected ROI contains merely skin, it offers only few recognizable image features, making it extremely difficult to track along the video frames. Thus, this leads to a non-continuous HR estimation during motion periods. To address this challenge, researchers are developing new methods for automatic and continuous ROI selection that do not resort to human supervision.

Despite aiming to extract respiration rate, Jorge et al. [42] proposed an approach for ROI selection worth mentioning giving its applicability for HR extraction. In [42], the ROI selection is accomplished through a color-based 2-class classifier based on Gaussian Mixture models. It clusters each pixel from each frame into skin and nonskin classes. In this model the ROI consists of the largest continuous skin region in each frame. However, having a color-based skin classifier, where the ROI is not associated with the anatomical structure of interest, leads to non-robust vital parameter extraction results for continuous monitoring over extended periods of time, as the author states in his conclusions.

Blanik et al. [9] opted to divide the video frames into 30 pixels edge length squares and compute a quality index for each one. This quality index reflects the likelihood of the square to contain HR information. All the squares possessing a quality index above a threshold of 90% of the best quality index value will belong to the ROI and, consequently, will be used

for HR estimation. Despite the fact that the ROI is not directly linked to an anatomical structure, this method guarantees that the selected area is representative of the parameter that will be extracted. However, in periods of intense motion, the technique still yields poor results.

Relying on techniques from the realm of Deep Learning, Chaichuleea et al. [14] [31] managed to detect the presence of the neonate in the incubator, identify the skin region and define two different ROIs for vital-sign estimation. For this purpose, they proposed a CNN with three branches from a shared core network. The patient detection branch was implemented using global average pooling with two outputs containing the prediction of the two classes. The skin segmentation branch was implemented following the FCNN proposed by Long et al. [54]. The body part detection branch locates the neonate's head, torso and diaper relying on bounding boxes using a Faster R-CNN network [66].

The model is capable of producing accurate segmentation results and is robust to changes in different skin tones, pose variations, lighting variations, and routine interaction of clinical staff. The authors reported outstanding results in the HR estimation performance when employing the developed FCNN, proving the method's ability to provide excellent measuring sites. However, as the authors state in the discussion section, the model is unable to achieve real-time performance given its VGG-16 feature extractor (see Section 3.4.1.2) and its region proposal generation network.

The high precision rate obtained in ROI detection using a CNN model prove that this approach may be more appropriate to the ROI detection problem in the scope of neonatology. Therefore, by changing Chaichuleea et al. CNN feature extractor to a model with fewer parameters and by combining the skin segmentation task with the body part identification task, the current thesis aims to develop a semantic image segmentation FCNN model capable of real-time inference and accurate segmentation performance.

4.2 Semantic Image Segmentation using Machine Learning

Computer vision tasks (semantic image segmentation, image classification, object detection, face recognition, etc.) had a substantial progress due to the use of CNNs. With the advantage of extracting compact and meaningful features from images without human supervision, CNNs became the mainstream approach regarding computer vision applications overthrowing classical methods focussing on hand-crafted features to describe images and, for example, nonparametric statistical methods [13] for pixel classification. Although the application of CNNs started as a method to address the image classification problem, it quickly paved his way to address the semantic image segmentation task. The emergence of challenging datasets, such as the PASCAL VOC [24], further encouraged this topic of research.

There are two main families of CNN methods to address the image segmentation task: region-proposal-based methods [36] [33], where multiple region proposals are generated and then each region will be individually processed (image instance segmentation), and FCNN methods, that process the input image at once not differentiating different instances (image semantic segmentation). The latter will be further analyzed, since it is the base of the proposed method for ROI selection, due to its high computational efficiency.

4.2.1 FCNN methods for image semantic segmentation

As detailed in Section 3.4, FCNNs comprise an encoder and a decoder network. Several new FCNN approaches that are proposed for the specific task of image semantic segmentation have contributed to this research topic by presenting new decoder architectures, in other words, new strategies to upsample the low resolution feature maps, outputted by feature extractors. The purpose of the presented decoders is to achieve accurate pixel-wise predictions.

For the encoder network, a modified VGG-16 [72] network is commonly adopted as a feature extractor in CNN models for image semantic segmentation [54] [59] [15] [60] [5] [14] [53] [81]. Typically, the weights of the feature extractor correspond to the weights of the network when trained on the ImageNet object classification dataset [67].

Long et al. [54] introduced the first FCNN for image semantic segmentation by presenting the first decoder network with a novel architecture feature, the skip connections. The novel skip architecture comprises the combination of information encoded in early, high-resolution intermediate feature maps with information encoded in deeper feature maps. The latter leads to an increase in the segmentation performance, particularly along the object boundaries. The model accepts the whole image as input. Despite the promising results, this architecture comprises a high number of parameters hindering the application of an end-to-end training style. Thus, the authors use a stage wise training process where additional decoder layers are progressively added to the previous trained network until no progress in the performance is observed.

Several authors used the core FCNN structure proposed by Long et al.. For example, Oliveira et al. [60] further refined Long et al. decoder architecture for the particular task of segmenting human body parts. The latter architecture is further discussed in Section 4.2.2 since it will be used as inspiration for the decoder network presented in the current thesis.

Additionally, some other decoder variations have been proposed. The authors of Segnet [5] introduced the unpooling operation. Instead of using skip connections, the decoder network proposed in [5] uses the unpooling operation, where the max-pooling indices of the encoder network are reused to perform non-linear upsampling. Noh et al. [59] proposed a new image semantic segmentation model by learning a deconvolutional network. This deconvolutional network comprises a series of deconvolution and unpooling layers to learn the upsampling of low-resolution feature maps. The latter upsampling process, conducted by the deconvolutional layers, is mediated by learned filters that constitute the bases to reconstruct the shape of the input object from the multidimensional feature map outputted by the encoder network. Additionally, some models [16] [15] [81] [51] further refine the object segmentation with the dense CRF post-processing method [45], achieving outstanding results.

Some approaches address the challenge of upsampling low resolution feature maps by modifying the feature extractor architecture to provide feature maps at a higher spacial resolution. In [17] [15] [16] , Chen et al. employs the atrous convolution operation which allows the prediction of mid-resolution feature maps that can be upsampled to match the input image resolution using bilinear interpolation.

After the release of ResNet networks [34], semantic image segmentation has made a new breakthrough: by replacing the VGG-16 layers with a ResNet network a significant improvement both in the prediction performance and computational cost was observed [66] [16] [65] [17] [52] [63]. Particularly, Peng et al. [63], following the skip architecture feature, uses the output of each residual block of the ResNet-101 in the upsampling process. The intermediate ResNet-101 feature maps are fed into a sequence of convolutional layers before being combined with the previous decoder layer. This sequence, named Global Convolutional Network, intends to enlarge the effective receptive field to make up for the typical ResNet-101 small effective receptive field. With this effective receptive field adjustment, the proposed model addresses both localization and classification task, characteristic of image semantic segmentation. Following this trend, the proposed method for ROI selection employs a ResNet-50 as the feature extractor.

4.2.2 FCNN model for human body parts segmentation

Some authors [16] [41] [52] report the results of its proposed models on the PASCAL human parts dataset (details of the dataset in Section 6.2.2). Despite the promising results, the architectures are complex. The FCNN model proposed by Oliveira et al. [60] is designed to address the human body parts segmentation problem relying on a less complex decoder network but still presenting outstanding results.

Similarly to [54], the encoder network of Oliveira et al. corresponds to a modified VGG-16 image classification network. The modification tailors the VGG-16 network for the task of image semantic segmentation and comprises the replacement of its fully connected layers by convolutional layers. This alteration allows the encoder network to produce coarse feature maps.

The novelty of the model proposed by Oliveira el al. relies on the upsampling process, in other words, in the decoder network. Despite following a skip connection architecture, where each layer of the decoder network combines the upsampled output of its previous layer with the pooled features of the corresponding layer of the encoder network, the combination procedure includes an additional layer, a spacial dropout layer [75]. The combination of the feature maps corresponds to the following procedure: the intermediate encoder network feature map is fed into a convolutional layer followed by dropout and is then element wise summed to the output of the corresponding layer in the decoder network (see Figure 4.2). This output is then submitted into a bilinear interpolation operation followed by a convolutional operation to expand the feature map resolution by a factor 2. The described upsampling process is repeated until the feature map reaches the input image resolution.

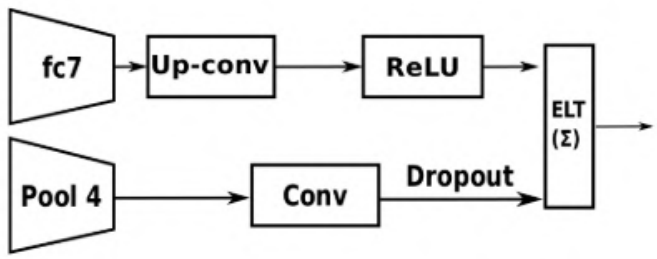


Figure 4.2: Illustration of the first decoder layer of Oliveira et. al. model. Figure extracted from [60].

The addition of the dropout layer in the decoder network improved the robustness to overfitting resulting in excellent segmentation results (details of the results in Appendix D.2.4). However, the model proposed by Oliveira et al. requires excessive computational cost due to the high number of parameters and high number of floating point operations. Therefore, it is not suitable for real time performance applications. The model proposed in this work, for ROI selection, addresses the aforementioned problem using the Oliveira et al. decoder as a reference and a different feature extractor as the encoder network.

Chapter Five

Clinical study

This chapter details the characteristics of the Neonaten and Navpani datasets used to develop and validate the proposed method for continuous non-contact heart rate monitoring of preterm infants. Section 5.1 details the characteristics of the studies. Section 5.2 describes the acquirement hardware.

5.1 Description of the studies

In this thesis, a dataset recorded in two different hospital settings was used.

The Neonaten dataset was collected at University Hospital Aachen (UKA), Department of Neonatology (Aachen subset) and was approved by the ethics committee of the UKA, Aachen, Germany (EK 327/16). Both video recordings (IR and RGB) and reference data (PPG and ECG) were acquired from nine neonates that were placed in incubators or in warming beds / cribs. Except for one infant, two measurements per infant were performed giving a total of 17 recordings. The clinical study involved five males and four females with a mean gestational age at birth of 30 weeks and a mean biological age of 49 days (minimum:8 days. maximum:241 days).

The Navpani dataset (Chennais subset) was recorded at Saveetha Medical College and Hospital and was approved by the institutional ethics committee of Saveetha University (SMC/IEC/2018/03/067). Neonates were recorded either under an infant radiant warmer or in a transport incubator. The clinical study involved seven males and 13 females with a mean gestational age at birth of 35 weeks; a mean biological age of 17 days (minimum:1 day. maximum:77 days).

In both datasets, the majority of the infants were awake during the whole measurements. Consequently, the recordings comprise a high level of bodily activity. Also, no constrains were imposed regarding the neonates' position and clinical staff activity which proceeded normally with the patient care routine. The environment of the NICU is not modified thus, there is a presence of both natural and artificial lightning. The latter can negatively affect the PPGI signal.

The combination of the Neonaten and Navpani datasets generates a datset that comprises a great variety of body positions and orientations, camera angles and skin colours. The patient demographics of the 29 neonates that comprise the dataset are listed in Appendix C.

5.2 Experimental Setup

The Neonaten RGB data was recorded using the CMOS colour camera GS3-U3-23S6C-C (FLIR, USA). Simultaneously, IR data was recorded using the monochrome CMOS camera GS3-U3-23S6M-C (FLIR, USA) equipped with a 940 nm filter (BN940, Midwest Optical Systems, Inc., USA). The cameras acquired 16-bit images with a resolution of 1200×1920 pixels at 25 frames per second. In addition to ambient light, a S75-WHI lighting module (STEMMER IMAGING AG, Germany) was used. The reference HR was provided through the PPG waveform filmed from the patient monitor.

For the Navpani data acquisition, the same cameras as in the Aachen subset was used. While illumination in the visual domain was ambient, active IR illumination was provided using a matching LED lamp (S75-940-W, Smart Vision Lights, USA) and an additional diffusion filter (LEE Filters, UK). The reference HR was directly extracted from the Philips patient monitor.

The basic setup for PPGI monitoring and its components are shown in Figure 5.1.

Figure 5.1: Experimental setup of the Neonaten dataset acquisitions. The cameras and dedicated light are installed in a tripod pointing at the neonate at a distance of approximately 1m.

Chapter Six

Region of Interest Selection

Deep Learning for Skin and Body Part Semantic Segmentation

In this chapter, a new method for region of interest selection is discussed. The proposed method comprises a Convolutional Neural Network for the task of simultaneous skin and body part semantic segmentation and a refinement algorithm to further refine both classification and localization performance.

Section 6.1 describes the proposed Convolutional Neural Network. Section 6.2 details the process of Dataset construction. Section 6.3 describes the training procedure. In Section 6.4 a new post-processing method is described. Finally, Section 6.6 presents the performance evaluation of the new method for region of interest selection. Finally, Section 6.7 discusses the results of the proposed method for ROI selection.

6.1 Skin and Body Part Segmentation Network

Similar to [5] [54] [59] [60], the proposed model follows an encoder-decoder architecture. The model is an optimized version of Oliveira's et al. FCNN to further improve efficiency in terms of both memory and computation time during inference.

6.1.1 Reimplementing an Encoder-Decoder Architecture

Figure 6.1 illustrates the detailed configuration of the proposed FCNN. The network is composed of two parts: encoder and decoder networks. The encoder network (Figure 6.1 A-F) takes an image as input and outputs a rich multidimensional feature representation, whereas the decoder network (Figure 6.1 G-S) gradually recovers the object shape and detail information from the coarse feature representation extracted from the encoder network. The output of the decoder network provides a prediction mask in the same resolution as the input image. Subsequently, the softmax layer (Figure 6.1 T) outputs a probability map for each class that contains the probability of each pixel to belong to the respective class.

For the encoder network, a modified version of the ResNet-50 [34] pretrained on ImageNet dataset [20] is used. For the decoder network, the architecture proposed by Olivera et all. [60] is used as a base.

6.1.2 Encoder Network

Given the outstanding results of employing very deep CNNs for the semantic image segmentation task [63] [16] [52] [41] [65] [17], the proposed encoder network is a re-purposed ResNet-50 [34], a state-of-art deep CNN designed for image classification. Section 3.4.1.1 details the architecture of the ResNet-50 network.

The original ResNet-50 takes a fixed-size image and estimates a probability for each one of the predefined classes. To tailor the ResNet-50 to the semantic image segmentation task, the average pooling layer and the final fully-connected layer are eliminated. This alteration prevents the complete loss of localization information and promotes feature maps with higher resolution. Thus, the encoder network consists of one initial convolutional layer followed by one max pooling layer and four bottleneck residual blocks.

6.1.3 Decoder Network

The decoder network was implemented using the FCNN proposed by Oliveira et all. [60] as a base which, in turn, is a refinement of the architecture of Long et al. [54]. This FCNN performs a series of spatial upsampling to progressively enlarge the feature maps, while incorporating intermediate feature maps from the encoder network to recover the original shape of the input object with high localization accuracy. Table 6.1 provides a full description of the employed decoder architecture.

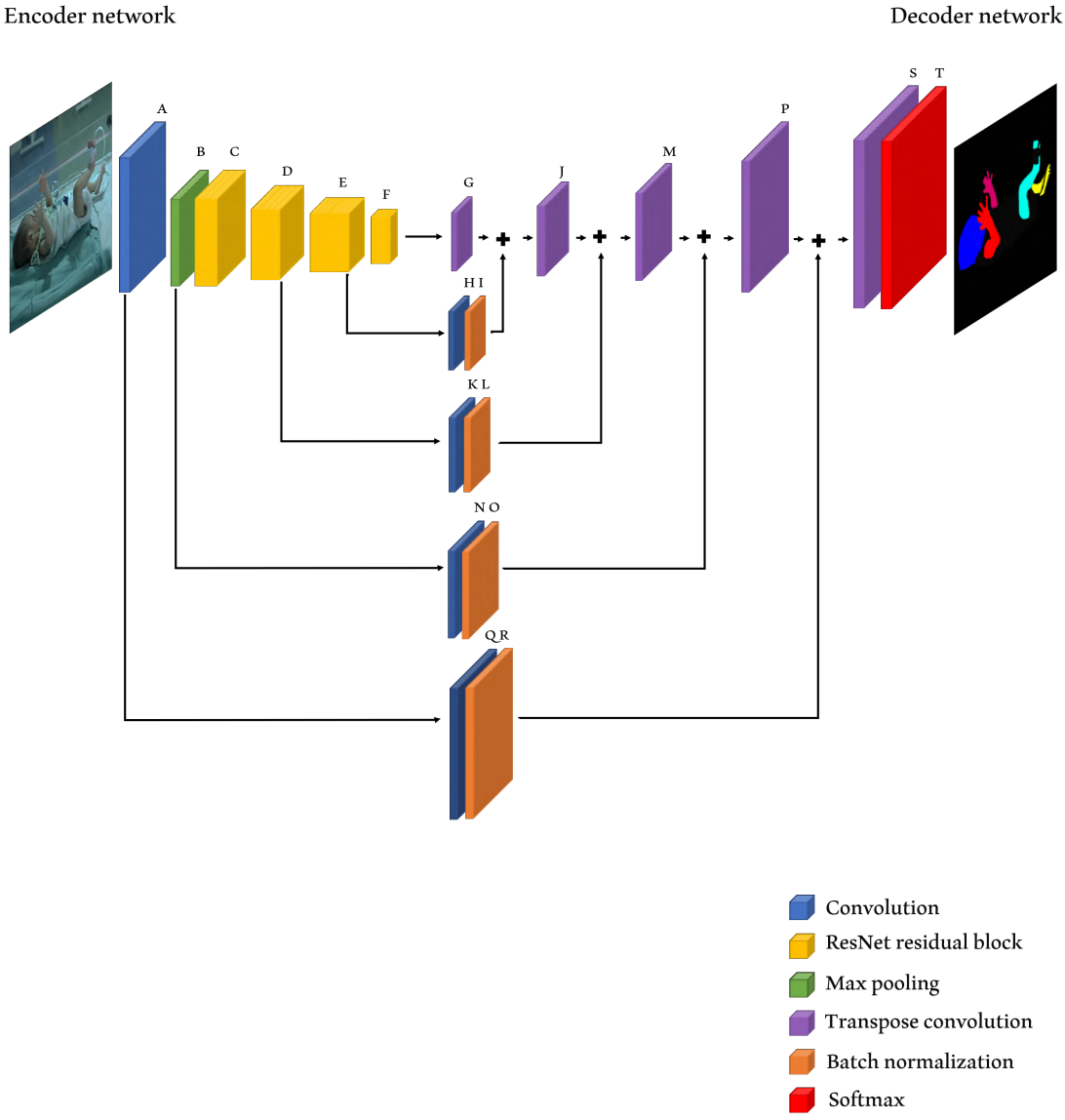


Figure 6.1: Overview of the proposed FCNN architecture. The encoder network (A-F) outputs a coarse multidimensional feature representation of the input. The decoder network (G-S) progressively increases the feature maps resolution to generate a dense pixel-wise class prediction. Intermediate feature maps from the encoder are used during the upsampling process to refine the prediction result. For brevity reasons ReLUs are omitted.

In contrast to methods relying on simple bilinear interpolation [60] [16], where no learning is involved, the proposed decoder network generates high-resolution segmentation masks using five transposed convolution operations (Figure 6.1 G, J M, P and S). Each transposed convolution upsamples the feature map by a factor of two, using multi-channel upsampling kernels. Section 3.2.4 details the transpose convolution operation. The first transposed convolutional layer takes the output of the fourth residual block of the ResNet-50 as an input. The following four transpose convolutional layers take, as input, the element-wise sum of the previous transposed convolutional layer output and an intermediate feature map from the encoder network, followed by a convolutional and batch normalization layer. The previous transposed convolutional layer output provides a preliminary feature map at a coarse resolution, whereas the ResNet-50 intermediate feature map contributes with information to refine the low-resolution preliminary feature map.

The intermediate feature maps selected to incorporate the decoder network have the same subsampling factor as the output of each transposed convolutional layer, allowing the direct concatenation operation. Thus, the outputs of the first convolutional layer, the following max-pooling layer and second and third residual blocks of the ResNet-50 (Figure 6.1 A, B, D and E, respectively) constitute the intermediate layers that will be fused during the up-sampling process (subsampling factor of 2, 4, 8 and 16 respectively). For example, the input of the second transposed convolutional layer (Figure 6.1 J) is the concatenation result of the first transposed convolutional layer output (Figure 6.1 G) and the output of the third residual block of the ResNet-50 (Figure 6.1 E), both with a sixteenth of the input image resolution.

The key design step to allow the element wise sum of the feature maps is the dimensionality reduction step conducted by the convolutional layers of the decoder-network (Figure 6.1 H, K, N and Q). These convolutional layers compress the intermediate encoder feature maps that will be used during the upsampling process. Thus, convolving the intermediate feature maps with $1 \times 1 \times K \times 7$ learnable parameters, where K denotes the number of channels of the intermediate feature maps, ensures that the intermediate feature maps have the same number of channels as the transposed convolutional layer output. The batch normalization layer that follows the convolutional layer improves the robustness to over-fitting and reduces the internal covariate shift (mode details in Section 3.2.3). The employed batch normalization layer replaces the spacial dropout layer used in the Oliveira et al. encoder-decoder architecture given its proven effectiveness [40]. Note that batch normalization and dropout layers can not be employed simultaneously because of their an incompatibility, which causes a decrease in performance [50].

Each transposed convolutional layer is followed by a ReLU activation function to better deal with the vanishing gradient problem [30]. The final high-resolution feature maps are fed to a softmax layer (Figure 6.1 T) which generates a dense score map at the same size as the input image.

Table 6.1: Detailed configuration of the proposed Decoder network. For brevity reasons ReLUs are omitted from the table. The terms "conv", "transconv"and "batchnorm"denote convolution, transpose convolution and batch normalization, respectively. The letters next to each layer name correspond to the respective layer in the Encoder-Decoder network illustration 6.1. An image with a resolution of 3 × 576 × 960 is assumed as input (channels × height × width).

Name	Kernel size	Stride	Pad	Input size	Output size
encoder network (A-F)	-	-	-	3 × 576 × 960	2048 × 18 × 30
transconv-G	2	2	0	2048 × 18 × 30	7 × 36 × 60
conv-H	1	1	0	1024 × 36 × 60	7 × 36 × 60
batchnorm-I	-	-	-	7 × 36 × 60	7 × 36 × 60
transconv-J	2	2	0	7 × 36 × 60	7 × 72 × 120
conv-K	1	1	0	512 × 72 × 120	7 × 72 × 120
batchnorm-L	-	-	-	7 × 72 × 120	7 × 72 × 120
transconv-M	2	2	0	7 × 72 × 120	7 × 144 × 240
conv-N	1	1	0	64 × 144 × 240	7 × 144 × 240
batchnorm-O	-	-	-	7 × 144 × 240	7 × 144 × 240
transconv-P	2	2	0	7 × 144 × 240	7 × 288 × 480
conv-Q	1	1	0	64 × 288 × 480	7 × 288 × 480
batchnorm-R	-	-	-	7 × 288 × 480	7 × 288 × 480
transconv-S	2	2	0	7 × 288 × 480	3 × 576 × 960

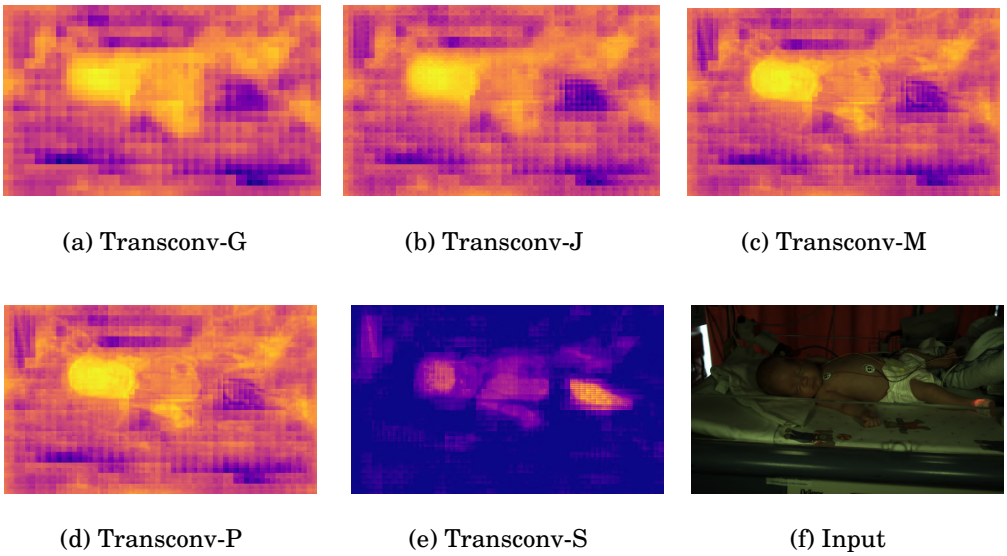
6.1.3.1 Analysis of the Decoder Network

The hierarchical structure of the filters in the encoder network is also applied to the filters' structure of the transposed convolutional layers: the complexity of the object's characteristics encoded in the filters is progressively higher. Thus, the details of the object shape are recovered, as the feature maps are propagated through the layers in the decoder network. Figure 6.2 emphasize the progressive increment of the object's reconstruction complexity through the decoder network.

6.1.4 Decoder Variants

Decoder variants were developed to allow the comparison between different upsampling methods and decoding techniques. Note that all the following decoders variants share the same ResNet-50 encoder network.

In order to analyse the impact of an upsampling process mediated by learned parameters, a decoder variant was developed, where the transposed convolutional layer is replaced by a bilinear interpolation layer where no learning is involved - Encoder-decoder-bilinear. Following the bilinear interpolation layer is a convolutional layer, with trainable parameters, that will densify the sparse upsampled maps. Note that, by setting the sampling factor to two, no other alterations to the decoder network are necessary besides the inclusion of an initial convolutional layer. The main purpose of the initial convolutional layer is the dimensionality reduction of the encoder output feature maps, to match the number of the



(a) Transconv-G (b) Transconv-J (c) Transconv-M

(d) Transconv-P (e) Transconv-S (f) Input

Figure 6.2: Activation maps of each transpose convolution in the decoder network. Each body part class has a separate activation map however, for effective visualization, the illustrated activation maps are the sum of the individual activation maps from all the body part classes. A progression from coarse to detailed body parts shape can be seen from top left to bottom right. Lower transpose convolutional layers capture a coarse object configuration and localization, whereas the finer details are encoded in the deeper layers. Noisy activations are suppressed in the final transpose convolution because of the colour information that is introduced by the output score map generated by the last batch normalization layer (Figure 6.1 R).

decoder output classes (6 body part classes plus a background class).

A decoder variant that does not include the concatenation of intermediate encoder network feature maps in the decoder network is also created - Encoder-decoder-unconnected. In this decoder variant no structure information will be harnessed from the encoder network meaning that the upsampling process will rely exclusively in learned multi-dimensional upsampling kernels.

In addition to the above variants, a new decoder variance is included where the batch normalization layers of the encoder networks are replaced with dropout layers - Endoder-decoder-dropout. The latter approach is similar to the decoder network proposed by Oliveira et al. [60].

6.2 Datasets

6.2.1 Neonaten-Navpani Dataset

To train the FCNN to identify skin pixels and simultaneously categorize them into one of the six predefined human body parts classes, a dataset of frames from the neonates' recordings and its ground truth masks need to be created. To this end, a set of frames was selected from the available recordings and manually annotated.

The designed dataset comprises frames from both the Neonaten and Navapani recordings. Therefore, apart from including a great variety of body positions, this dataset contains neonates with a wide range of skin tones. This variety, coupled with data augmentation methods (subsection 6.3.2.3), will keep the model from overfitting.

The developed method for continuous non-contact HR monitoring comprises the PPGI signal extraction from RGB and IR recordings. Thus, a Neonaten-Navpani-RGB dataset containing RGB frames and a Neonaten-Navpani-IR dataset containing IR frames needs to be constructed.

6.2.1.1 Frame Selection

The high frame rate associated with the neonates' recordings leads to consecutive frames containing similar body positions and illumination. To ease the process of selection of meaningful and distinct frames within a recording, a MATLAB algorithm was developed.

The algorithm sequentially selects a pair of frames $(k, k+2)$ where $k \in [0, N-2]$ being N the recording's length. The motion between the pair of frames is computed relying on the Block-matching algorithm.

Frames with a motion value above a defined threshold are displayed in a GUI to allow the final manual selection of the frames that will be saved to be part of the training and validation dataset.

To reduce computational effort, the user can manually crop the first frame of the recording to reduce the background (incubator walls, empty mattress, etc.). The selected crop will be propagated along all the recording's frames.

A total of 563 RGB PNG images and 81 IR PNG images were saved at their original resolution, ($3 \times 1200 \times 1920$ and 1200×1920, respectively). Additionally, 48 RGB images of a dummy in several body positions positions were included in the dataset.

6.2.1.2 Ground Truth Annotation Protocol

The ground truth annotation is the process of manually labelling the neonate's body regions where the skin is exposed. The selected frames are labelled using the polygon and smart polygon tool in the MATLAB Image Labeler application. The MATLAB image labeller application generates a PNG ground truth image and a mat file containing the label

definitions. The label definitions were kept the same for all frames and are displayed in table 6.2.

Table 6.2: Label definitions.

Label name	Pixel label ID	Description
Background	0	background
Head	1	skin pixels belonging to the head
Torso	2	skin pixels belonging to the torso
Right arm	3	skin pixels belonging to right arm
Left arm	4	skin pixels belonging to left arm
Right leg	5	skin pixels belonging to right leg
Left leg	6	skin pixels belonging to left leg

6.2.1.3 Infrared Dataset

Although both IR and RGB recordings were acquired simultaneously, the camera angles were slightly different. Thus, the ground truth annotations of the RGB frames can not be linked to the IR respective frames. To construct the Neonaten-Navpani-IR dataset, RGB frames of the Neonaten-Navpani-RGB dataset were manipulated to resemble IR frames. Two Neonaten-Navpani-IR datasets were constructed with two levels of image manipulation. The selected 81 true IR frames will allow the testing of the manipulation quality.

The wavelength of the IR radiation is closer to the wavelength captured in the red channel when compared with the remaining colour channels. Thus, one Neonaten-Navpani-IR dataset is constructed upon the red channel of the RGB frames - Neonaten-Navpani-IR-redchannel.

The second Neonaten-Navpani-IR dataset is generated by manipulating the histogram of the red channel of the RGB frames. A MATLAB algorithm will range over the Neonaten-Navpani-RGB dataset and, for each RGB frame, it will select the respective frame in the IR recording. The histogram of the red channel of the RGB frame is then adjusted to match the histogram of the IR frame, using the imhistmatch MATLAB function. An example of this histogram transformation is illustrated in Figure 6.3. The second Neonaten-Navpani-IR dataset, Neonaten-Navpani-IR-manipulated, comprises the histogram adjusted red channel frames.

6.2.2 PASCAL Human Parts Dataset

The PASCAL human parts dataset is a subset of the general PASCAL VOC 2010 dataset [24], which contains extra detailed annotations of human body parts (eyes, nose, upper arm, etc.). The annotations were merged to form six body part classes and one background class, similarly to the Neonaten-Navpani dataset annotations. Only images containing human subjects were used, giving a total of 3584 images.

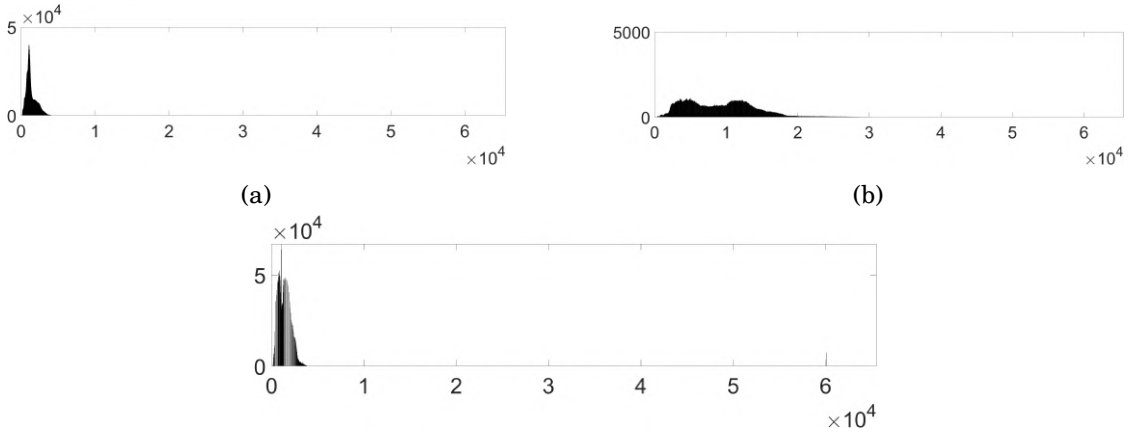


Figure 6.3: Illustration of the histogram manipulation process. The histogram of the red channel of the RGB frame (6.3b) is adjusted to match the histogram of the true IR corresponding frame (6.3a). The result of the histogram adjustment is displayed in Figure 6.3c.

6.2.3 Freiburg Sitting People Dataset

The Freiburg sitting people dataset was created by Oliveira et al. [60] and provides high-resolution annotations of images of people. The dataset comprises 200 images of sitting people from multiple viewing angles and orientations. Similarly to PASCAL human parts dataset, the dataset contains detailed ground truth annotations for multiple body parts. Once again, the annotations were merged to generate six body part classes.

6.3 Network Training

The proposed FCNN is very deep and, despite the effort to reduce the number of parameters of the encoder network, the whole architecture contains a great amount of parameters to tune. In addition, the Neonaten-Navpani dataset is extremely small to successfully train the model without overfitting. Thus, to train the model for the specific task of segmenting neonates' body parts, a two-stage training procedure is employed: pre-training stage and fine-tuning stage. This stage wise training system falls into the concept of transfer learning.

The designed model can be trained end-to-end thanks to the long range encoder-decoder connections that efficiently propagate the gradient to early low level layers during the backward pass.

6.3.1 Pre-training Stage

Firstly, the network is trained using the combination of the PASCAL human parts and the Freiburg sitting dataset. These datasets contain a great variety of body scales and poses, allowing the generation of a flexible and generalized base model. The pre-training stage is the first stage of the transfer learning process where the model is trained to perform the general task of body part segmentation.

For training, 90 % of the combined dataset images is used, while the remaining will be employed for validation.

6.3.1.1 Weight Initialization

Despite the ResNet-50 network alterations (see Section 6.1.2), the initial architecture remains unchanged enabling the initialization of the encoder network with the publicity available pre-trained weights for the classification task on the large ImageNet dataset. Note that by using the pre-trained weights as a starting point, the knowledge learned for the task of image classification is transferred for the task of semantic image segmentation.

The decoder network needs to be trained from scratch. The weights of the convolutional and transpose convolutional layers are initialized using the initialization scheme proposed by He et al. specially designed for CNNs that rely on asymmetric, non linear activations [35]. The weights of the batch normalization layers are initialized with ones and the biases with zeros.

6.3.1.2 Data Preprocessing

Before each epoch, the training dataset is shuffled and each batch is then selected orderly to ensure that all images are used and that each image is only selected once within an epoch. The batch size was set to six due to GPU memory limitations (more details see Section 6.5).

The selected batch of images and their respective ground truth masks are resized to a resolution of 320×320. The images are normalized using the mean and standard deviation

of the ResNet-50 pretrained on ImageNet. Before being fed into the network, the training images are submitted into a process of data augmentation that is described in Section 6.3.2.3.

For the IR model training, both the training and validation images are transformed into greyscale images by substituting the RGB channels for their mean. Note that, an image with three channels is still fed into the network.

6.3.2 Fine-tune Stage

The network is fine-tuned using the Neonaten-Navpani dataset. Given the small number of neonates of the clinical study, a balanced 5-fold cross validation technique is employed.

For two neonates, the two measurements were not performed within the same conditions: one measurement was performed when the neonate was inside the incubator and the other when the neonate was in the infant radiant warmer . Thus, for training purposes, the latter measurements are treated like measurements associated to different subjects. Therefore, the clinical study size is assumed to be 31 neonates. The 31 neonates are splitted into five groups having a fairly equal representation of skin colour and body positions. The patient demographics and body positions present in each fold are listed in table C.1. The distribution of the dataset frames for each fold is listed in Appendix D.1.

Table 6.3: Summary of the patient demographics in the five folds. M = Male, F = Female, W = White, B = Black, WB = Mixed White and Black, Su = Supine, P = Prone, Si = Side.

Fold	**Subjects**	**Gender**		**Skin colour**			**Inside the incubator**		**Lying position**		
		M	F	B	W	BW	Yes	No	Su	P	Si
1	7	1	6	2	1	4	1	6	6	1	0
2	7	2	5	2	2	3	2	5	5	1	1
3	5	4	1	1	2	2	1	4	4	1	0
4	5	2	3	1	2	2	1	4	5	0	0
5	7	3	4	2	1	4	0	7	7	0	0
Total	31	12	19	8	8	15	5	26	27	3	1

The model is trained on four folds and the remaining fold is employed for validation. The process is repeated five times so that each fold is employed for validation once. Thus, from the fine-tuning stage, five sets of weights are generated. The validation results from the five trained models are combined to produce an overall class prediction performance metric.

6.3.2.1 Weight Initialization

The parameters of the model pre-trained on the PASCAL human parts and Freiburg sitting dataset are used to initialize all the layers of the encoder-decoder network. This procedure falls into the second stage of the transfer learning process, where the knowledge learned for the general task of human body parts segmentation is transferred for the specific task of simultaneous skin and body part segmentation.

6.3.2.2 Data Preprocessing

Depending on the selected fold and the type of model (RGB or IR), the training and validation datasets are constructed from the Neonaten-Navpani-RGB or Neonaten-Navpani-IR dataset. The batch size was set to three.

Similarly to the pre-training stage, each batch is selected orderly from the training dataset. The patch of images and their respective ground truth mask are resized to a resolution of 960×600 and then cropped equally from both sides to a resolution of 960×576. The cropping step make both the width and height divisible by 32 (the downsampling factor of ResNest-50) avoiding dimensions errors in the element wise sum step. Before being fed into the network, the images are also normalized using the mean and standard deviation of the Neonaten-Navpani-RGB or Neonaten-Navpani-IR dataset, depending on the model type (RGB or IR). The training images are then submitted into the process of data augmentation described in Section 6.3.2.3.

6.3.2.3 Data Augmentation

The relative small size of the Neonaten-Navpani dataset, compared to popular datasets for computer vision applications [24] [67], may induce overfitting of the CNN. To improve the generalization of the network's parameters, the training data was synthetically modified. At each batch, a set of image transformations were applied.

- **Scaling:** since cameras can be positioned at different distances from the neonate, it is important for the model to be invariant to different body scales. Each training image was randomly resized by a scale factor between 0,7 and 1,4.
- **Rotation:** to increase the robustness of the network to different rotations each training image was randomly rotated by an angle of up to 30 degrees.
- **Flipping:** since human body is symmetric, it is acceptable to randomly horizontally flip the training image.
- **Colour variations:** to increase robustness to variations in illumination and skin colour, the brightness, contrast and saturation were slightly modified within an interval of 0,9 to 1,1.

6.3.3 Optimization and Loss Function

The standard Adam optimization algorithm is employed for optimization, where the initial learning rate, β_1, β_2 and ϵ are set to 0.0001, 0.9, 0.999 and 1×10^{-8}, respectively. For the pre-training stage, the learning rate remained constant throughout the training. While for the fine-tuning stage a "step"learning rate policy was employed, where the learning rate was progressively reduced by a factor of 0,5 every 30 epochs. The training procedure is stooped after training and testing loss convergence.

The cross-entropy loss is used as the loss function to train the network.

6.4 Refinement Algorithm

The proposed deep FCNN was designed to perform image semantic segmentation. Inherently to this concept is the inability to distinguish two instances of the same class. This has an negative impact on the PPG signal, when a subject besides the neonate enters the field of view of the recording. To overcome this limitation and further improve the body parts segmentation overall performance, a post-processing method was developed. A high-level illustration of the whole model pipeline is shown in Figure 6.4.

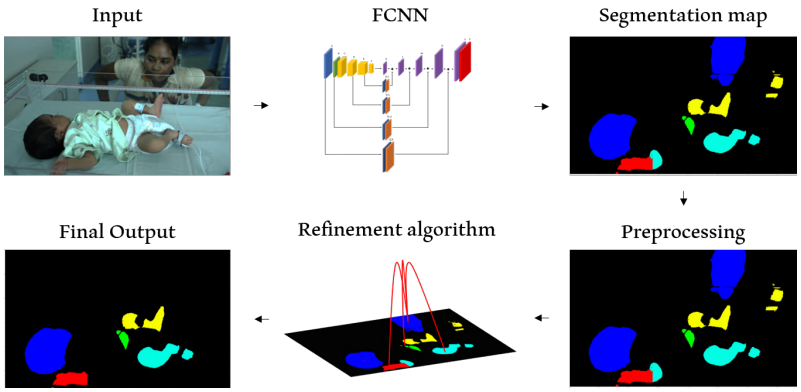


Figure 6.4: Model illustration. The proposed FCNN provides a semantic segmentation mask of the input frame. The semantic segmentation mask is preprocessed and then fed into the refinement algorithm to refine the segmentation results and eliminate instances besides the neonate.

The refinement algorithm assigns a score to each instance, favouring larger instances that are spatially closer to the global CM and anatomically correct positioned with respect to the remaining body parts. The instance with the higher score is kept and the remainder will form a calibration mask that will be propagated to the following frames to erase mislabelled instances.

6.4.1 Notation

Consider a segmentation mask $\boldsymbol{X}$ defined over a set of variables $\{X_1,...,X_N\}$. The domain of each variable is a set of labels $L = \{l_1,...,l_p\}$. In the proposed FCNN, X ranges over seven possible pixel-level image labellings, $p \in \{0,...,6\}$, given an input frame of size N. X_q is the label assigned to pixel q.

The segmentation mask $\boldsymbol{X}$ is divided into six segmentation masks, $Y_i^r = \{Y_{i_1}^r,...,Y_{i_N}^r\}$ for $i \in \{1,...,6\}$, containing the segmentation of each body part. The domain of each of Y_i^r is a pair of labels, being one of the elements of the pair the background label and the other the label l_i.

The refinement model generates six refined segmentation masks with only one instance, $\hat{Y}_i = \{\hat{Y}_{i_1},...,\hat{Y}_{i_N}\}$, by computing a score associated to each instance present in Y_i, s_i =

$\{s_{i_1},...,s_{i_{k_i}}\}$ where k_i denotes the number of instances of Y_i. These scores arise from operations over the CMs of each body part instance in $\mathbf{Y}$, $cm_i = \{cm_{i_1},...,cm_{i_{k_i}}\}$, and the global CM, cm_g.

At inference time, an additional set of calibration masks will be generated, $C_i = \{C_{i_1},...,C_{i_N}\}$, where $C_i \in [0,1]$. These calibration masks will propagate the segmentation refinement and instance elimination to the following frames.

6.4.2 Segmentation Mask Preprocessing

The primary goal of the preprocessing step is to transform the raw semantic segmentation masks provided by the FCNN, $\boldsymbol{X}$, into instance segmentation masks, $\boldsymbol{Y}$. To ease the score computation process and make the refinement algorithm computationally efficient, residual isolated segments will be eliminated from the raw semantic segmentation masks to avoid unnecessary computation of CMs. The preprocessing relies on morphological operations.

The semantic segmentation mask $\boldsymbol{X}$ is divided into six semantic segmentation masks, $Y_i^{\mathrm{r}} = \{Y_{i_1}^{\mathrm{r}},...,Y_{i_N}^{\mathrm{r}}\}$ for $i \in \{1,...,6\}$, containing the raw segmentation of each body part. Each raw segmentation mask, Y_i^{r}, is converted into a binary mask, Y_i^{b}:

$$Y_i^{\mathrm{b}} = \begin{cases} 0 & \text{if} \quad Y_i^{\mathrm{r}} = 0, \\ 1 & \text{if} \quad Y_i^{\mathrm{r}} = l_i, \end{cases}$$

The binary segmentation mask Y_i^{b} is then submitted into a erosion process where the erosion operation is repeated six times with the structuring element illustrated in Figure 6.5a. Small segments in the mask will be removed in this erosion process. To approximately restore the shrinked regions and connect close segments that probably belong to the same instance, the binary segmentation mask Y_i^{b} is then submitted into a dilation process, where the dilation operation is repeated 100 times with the structuring element illustrated in Figure 6.5a. Note that the number of both erosion and dilation operations was manually optimised for the neonates recordings' field of view.

Isolated segments in the processed Y_i^{b} are considered individual instances. The application of the connected-component labelling algorithm, with the structuring element illustrated in Figure 6.5b and a squared connectivity equal to one, a instance segmentation mask of dilated instances is generated, Y_i^{l}. The element wise multiplication of the labelled dilated instance mask with the unprocessed binary semantic segmentation mask, Y_i^{b}, generates an instance segmentation mask, Y_i. The process is repeated for $i \in \{1,...,6\}$ generating a set of instance segmentation masks $\boldsymbol{Y}$.

The different segmentation masks that result from the preprocessing process are illustrated in Figure 6.6.

1	1	1
1	1	1
1	1	1

a

0	1	0
1	1	1
0	1	0

b

Figure 6.5: Structuring elements employed during the image preprocessing. 6.5a is the structuring element used in the erosion and dilation process and 6.5b is the structuring element employed in the connected-elements algorithm

6.4.3 Score Computation

A score is computed for each instance in Y_i when the number of instances is superior to one, $k_i > 1$. The score function (equation 6.1) is denoted by $s_m(j)$, where $j \in \{0,...,k_m\}$, and models the compatibility of the instance m_j, with label l_m and CM cm_{m_j}, with the remaining body parts instances n_h, with label l_n and CMs cm_{n_h}. Note that $n \cap m = 0$ and $n \cup m = i$.

$$s_m(j) = \sum_n \sum_h f(l_m, l_n, cm_{m_j}, cm_{n_h}) + g(l_m, cm_{m_j}, cm_g) \tag{6.1}$$

where $f(l_m, l_n, cm_{m_j}, cm_{n_h})$ is the pairwise score function (equation 6.2) and $g(l_m, cm_{m_j}, cm_g)$ is the proximity score function (equation 6.3).

$$f(l_m, l_n, cm_{m_j}, cm_{n_h}) = \mu(l_m, l_n) \times d(cm_{m_j}, cm_{n_h}) \tag{6.2}$$

where d denotes the function that computes the distance between the cm_{m_j} and cm_{n_h} and $\mu(l_m, l_n)$ is a weight matrix that encodes the weights that will penalize CMs distances that are not in line with the anatomically correct position of the human body. The following example illustrates the intuition behind the pairwise score: an instance j labelled as right leg, $l_m = 5$, and with a CM, cm_{m_j}, close to the CM, cm_{n_h}, of an instance classified as head, $l_n = 1$, will generate a lower pairwise score when compared with a different instance, $j+1$, with the same label, $l_m = 5$, with a CM ,$cm_{m_{j+1}}$, distant to the the same CM cm_{n_h} with $l_n = 1$,

$$g(l_m, cm_{m_j}, cm_g) = \theta(l_m) \times d(cm_{m_j}, cm_g) \tag{6.3}$$

where θ is a weight matrix that encodes the weights that will penalize instances which are distant from the global CM, cm_g, depending on the label l_m. Both weight matrices, θ and μ are displayed in Figure 6.7b and Figure 6.7a, respectively. Note that both θ and μ were manually optimised.

(a) Segmentation mask illustration.

$$\begin{matrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 & 6 & 6 & 0 & 6 \\ 0 & 1 & 1 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 2 & 5 & 5 & 0 \\ 0 & 0 & 3 & 5 & 0 & 0 & 0 & 0 \end{matrix}$$

(b) Segmentation mask, $\boldsymbol{X}$.

$$\begin{matrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \quad \cdots \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 & 5 & 0 \\ 0 & 0 & 0 & 5 & 0 & 0 & 0 & 0 \end{matrix} \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 & 6 & 6 & 0 & 6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

(c) Individual body part raw segmentation masks, Y_i^{r}.

$$\begin{matrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \quad \cdots \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{matrix} \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

(d) Individual body part binary segmentation masks, Y_i^{b}.

$$\begin{matrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \end{matrix} \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{matrix} \quad \cdots \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 & 0 & 1 & 1 & 0 \end{matrix} \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 2 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 2 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 2 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

(e) Individual body part instance segmentation masks of dilated instances, Y_i^{l}.

$$\begin{matrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \quad \cdots \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \end{matrix} \quad \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 2 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

(f) Individual body part instance segmentation masks, Y_i.

Figure 6.6: Illustration of the semantic segmentation mask ($\boldsymbol{X}$) evolution across the preprocessing process.

Finally, a score that reflects the relative size of the instance, m_j, with respect to the size of all the remaining instances classified with the same label, l_m, is added to $s_m(j)$ to generate the final score, $S_m(j)$, of the instance m_j (equation 6.4).

$$S_m(j) = \beta \times (\max(s_m) - \min(s_m)) \times \frac{N_{m_j}}{N_m} + s_m(j) \tag{6.4}$$

where β is a parameter that regulates the influence of the size score in the final instance score. N_{m_j} denotes the number of pixels of the instance m_j and N_m denotes the total number of pixels of the instances with the same label as m_j, l_m.

	head	torso	right arm	left arm	right leg	left leg
head	0	-1	-2.1	-2.1	1	1
torso	-1	0	-1	-1	-1	-1
right arm	-2.1	-1	0	-1	0.5	0.5
left arm	-2.1	-1	-1	0	0.5	0.5
right leg	1	-1	0.5	0.5	0	-1.7
left leg	1	-1	0.5	0.5	-1.7	0

a

	head	torso	right arm	left arm	right leg	left leg
CMg	-2	-4	-4	-4	0.1	0.1

b

Figure 6.7: Weight matrices. 6.7a corresponds to the μ weight matrix. 6.7b corresponds to the θ weight matrix

6.4.4 Calibration Masks

The segmentation mask preprocessing and the score computation is too computationally expensive (approximately one second per image) to be applied in every frame of the recording for real time applications. In addition, since multiple consecutive frames contain similar information, the application of the refinement algorithm to every frames is redundant.

Thus, for each application of the refinement algorithm, a set of calibration masks will be generated, $C_i = \{C_{i_1}, ..., C_{i_N}\}$, where $C_i \in [0,1]$ and $i \in \{1, ..., 6\}$. These calibration masks are matrices with the same size, N, as the recording frame.

The calibration mask C_i is initialized with ones, then, it will be filled with zeros on the coordinates of the mislabelled instances i.e instances with the lower scores. The element wise multiplication of the calibration mask C_i with the raw semantic segmentation masks Y_i^r, generates refined segmentation masks with only one instance, $\hat{Y}_i$. Figure 6.8 illustrates the calibration masks operation.

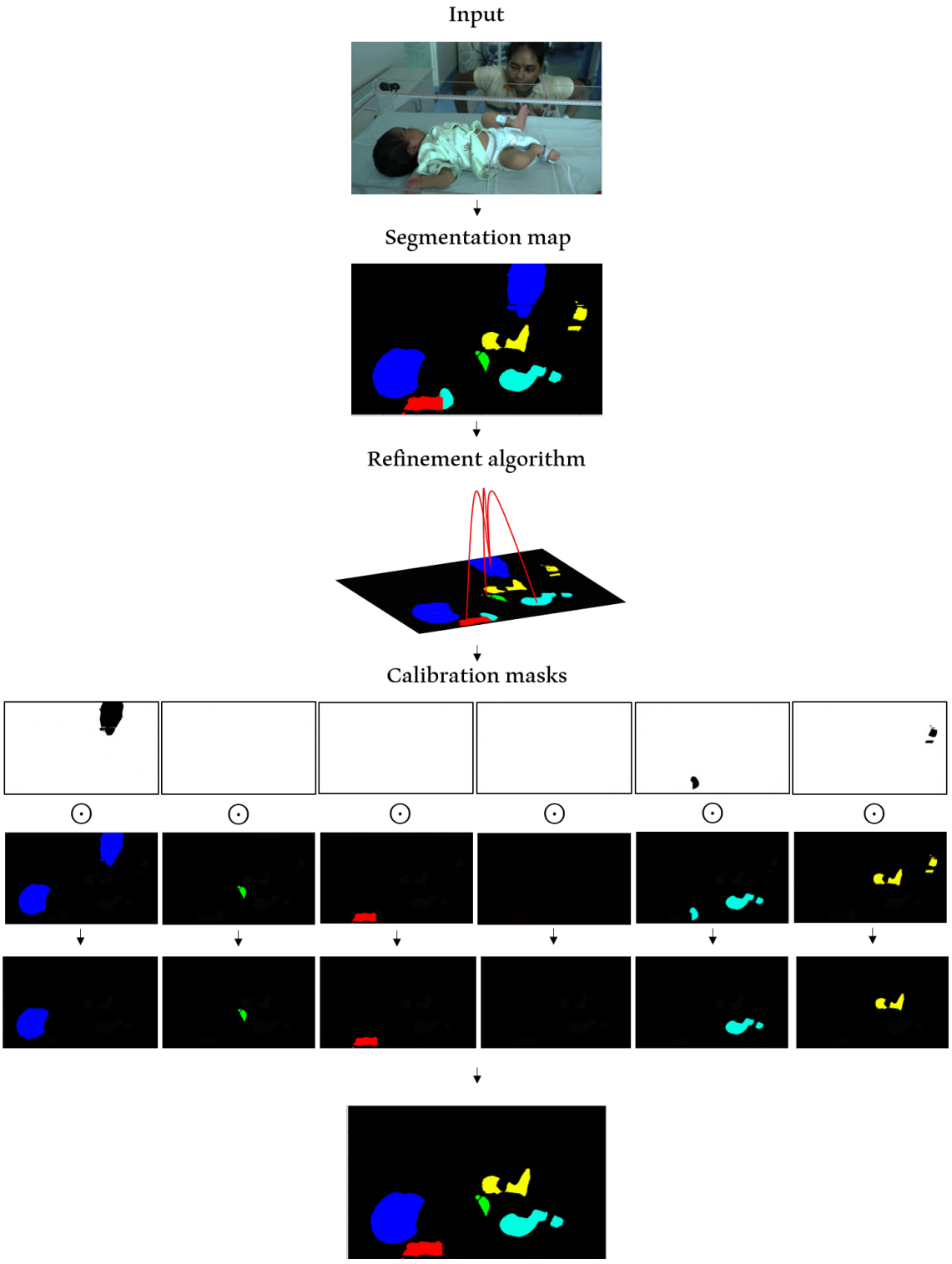


Figure 6.8: Illustration of the calibration masks operation. The refinement algorithm generates one calibration mask per body part. The element wise multiplication of the calibration mask with the respective body part segmentation mask will erase the mislabelled instances.

6.5 Hardware and Software

The training and testing framework was implemented in Python, specifically the PyTorch package [61] was used to implement the proposed FCNN model and all its different decoder variants.

For the PASCAL human parts and Freiburg sitting dataset, the training and testing procedure of the models was conducted in the Google Colaboratory research tool, whereas, for the Neonaten-Navpani dataset, the training and testing procedure of the models was conducted on a workstation:

1. The Google Colaboratory research tool is a cloud service based on Jupyter Notebooks that allows the execution of python code in a virtual machine fully configured for deep learning applications. The virtual machine is equipped with an Intel Xeon Processor CPU @2.3 Ghz, 12.6 GB RAM and a NVIDIA Tesla K80 GPU with 2496 CUDA cores and 12 GB of GDDR5 memory.

2. The workstation is equipped with an Intel Xeon Gold CPU @ 3.4 Ghz, 128 GB RAM and a NVIDIA Quadro P4000 glsGPU with 1792 CUDA cores and 8 GB of GDDR5 memory.

6.6 Evaluation

To perform a complete analysis of the proposed FCNN, its decoder variants (see section 6.1.4) and the model proposed by Oliveira et al. [60], multiple performance indices are considered to evaluate both the computational effort and class prediction performance. Specifically, to analyse the computational effort, the model complexity, the computational complexity, the peak GPU memory usage and inference time are reported. To quantitatively compare the class prediction performance of the considered FCNN models in the PASCAL human parts and Freiburg sitting people dataset, two evaluation metrics were employed: class average accuracy and class average IoU. To evaluate the performance of the proposed FCNN model in the Neonaten-Navpani dataset, a new evaluation metric was introduced: class average precision. The inclusion of the latter metric, that is not commonly used in the evaluation of image semantic segmentation models, is specially important for the task of ROI selection for PPGI extraction, where the most important factor to consider is the number of FP that can negatively interfere with the PPGI signal quality.

In order to perform a direct and fair comparison in the computational effort evaluation, the considered models are implemented using the same PyTorch framework, meaning that the publicity available model proposed by Oliveira et al. [60] is convert in PyTorch. The considered models are designed to predict seven classes (background, head, torso, right and left arm and right and left leg) and to expect batches of images with shape $3 \times H \times W$, where H and W are the height and with of the image. For the class prediction performance analysis the accuracy and IoU values reported for the model proposed by Oliveira et al. are extracted from the original paper [60].

Unless otherwise specified, the following results are originated from models that follow the training procedure specified in Section 6.3.

6.6.1 Computational Effort

6.6.1.1 Model Complexity

To analyse the complexity of the different considered FCNN models, the total amount of learnable parameters is taken into account. Additionally, following the approach in [8], the size of the file that contains the learned parameters of each considered FCNN model is compared in Table 6.4, having a direct relation with the total amount of learnable parameters. The file size also provides an insight into the minimum amount of GPU memory required for each model during training and inference.

The proposed FCNN model and its decoder variants have 18 % of the learnable parameters of the encoder-decoder architecture proposed by Oliveira et al. [60]. This substantial difference in the model's complexity mainly derives from the employed encoder network: the proposed model and its decoder variants rely on a modified ResNet-50, which is significantly smaller in the number of learnable parameters when compared with the modified VGG-16 in the Oliveira et al. architecture. The characteristic high number of learnable parameters,

Table 6.4: Total amount of learnable parameters, number of learnable parameters in the encoder and decoder network and size of the file containing the learnable parameters values for each considered model.

Method	Total nº of parameters	Nº of encoder parameters	Nº of decoder parameters	File size (KB)
Encoder-decoder-bilinear	23 592 480	23 508 032	84 448	276 341
Encoder-decoder-unconnected	23 567 140	23 508 032	59 108	276 480
Encoder-decoder-dropout	23 577 836	23 508 032	69 804	276 617
Encoder-decoder-batchnorm	23 577 892	23 508 032	69 860	276 624
Encoder-decoder-Oliveira	134 729 180	134 260 544	468 636	-

which exceeds the hundreds of millions, of the encoder-decoder architectures that rely on the VGG-16 network [60] [54] often hinder the adoption of a end-to-end training style thus leading to a multi-stage training style. On the other hand, as mentioned in Section 6.3, the proposed FCNN model is able to perform an end-to-end training procedure.

The proposed FCNN model and its decoder variants have a large number of learnable parameters in the encoder network (23M) which contrasts with the relatively small number of learnable parameters in the decoder network. Thus, the relatively small fluctuations of the considered decoder variants learnable parameters number are not meaningful in terms of both model complexity and memory. However, it is worth noticing the lower number of learnable parameters of the Encoder-decoder-unconnected model, which mainly derives from the elimination of the convolutional layers that allow the concatenation of the intermediate feature maps from the encoder network and the feature maps outputted by the transpose convolutional layers.

6.6.1.2 Peak Memory Usage

The maximum occupied GPU memory during one forward pass, which includes the memory required to process the batch of images besides the memory required to store the network learnable parameters and feature maps, is also evaluated. For the sake of simplicity, only values for the Encoder-decoder-batchnorm, Encoder-decoder-unconnected and Encoder-decoder-Oliveira models are reported in Figure 6.9, since the remaining decoder variants reveal similar results to those for the Encoder-decoder-batchnorm model.

From Figure 6.9, it can be seen that, for all the considered batch sizes, the proposed FCNN model and its variants are able to perform a forward pass, for an image with a resolution of 576×960, without surpassing the maximum available GPU memory (8GB) of the NVIDIA quadro p4000. The memory efficiency of the proposed FCNN model was one of the primary motivations behind the its design. On the other hand, the model proposed by Oliveira et al. is unable to perform a single forward pass for batch sizes superior to nine images on the same GPU.

As expected, the encoder-decoder-unconnected model reveals a slightly reduction on the maximum occupied GPU memory across different batch sizes when compared to the remaining FCNN models. The latter mainly derives from the fact that the proposed FCNN and the models with decoder variants where information from different intermediate encoder

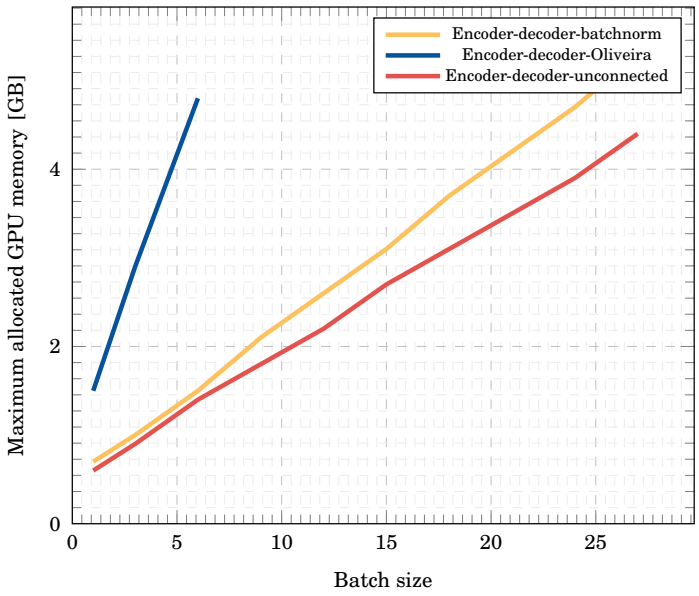


Figure 6.9: Maximum occupied GPU memory *vs*. batch size. This graph shows the maximum occupied GPU memory during one forward pass across different batch sizes of images with a resolution of 576×960.

layers are combined during the upsampling process, imply the storage of intermediate feature maps from the encoder network which requires the extensive usage of memory during inference and training time.

6.6.1.3 Computational Complexity

The computational complexity of each considered model is measured relying on the number of floating-point multiplication and adds operations, FMAs, required for a single forward pass. Note that the number of FLOPs is approximately twice the number of FMAs, since multiply-add operations are counted as single operations.

The proposed FCNN model and its decoder variants have approximately 6.5 % of the total number of FMAs of the encoder-decoder architecture proposed by Oliveira et al. (see Appendix D.2.1). This major difference between the number of FMAs will be reflected on the inference time since the number of operations required to perform a forward pass has a linear relationship with the time necessary for a image to be processed [8]. Thus, the limited number of operations of the proposed FCNN model keeps the processing speed in a range suitable for real-time applications, as discussed in Section 6.6.1.4.

6.6.1.4 Inference Time

Figure 6.10 reports the per image inference time for a image of size 576×960 for each considered model, as a function of batch size. For statistical validation, the reported times corresponds to the average over 10 runs on the NVIDIA Quadro p4000 GPU. For the sake of simplicity, only values for the proposed FCNN model and for Oliveira et al. model

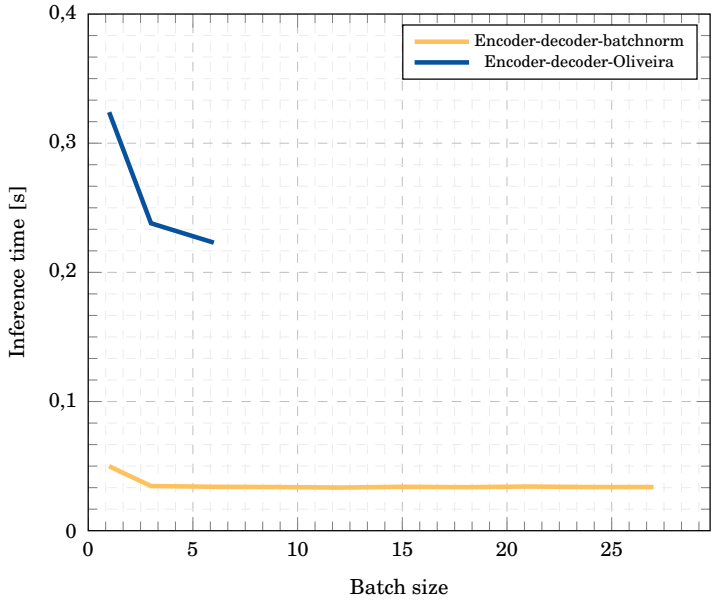


Figure 6.10: Inference time *vs*. batch size. This graph shows the inference time during one forward pass across different batch sizes of images with a resolution of 576×960. Missing data points for the Encoder-decoder-Oliveira model plot are due to lack of enough system memory required to process larger batches.

are reported, since the remaining decoder variants reveal similar results to those for the proposed FCNN model.

Not surprisingly, the Oliveira et al. model yields the higher inference time per image proving once more the linear relationship between computation complexity and inference time [12]. As previously stated, the Oliveira et al. model is unable to perform a forward pass for batch sizes superior to nine, for images with a resolution of 576×960, given the GPU memory restrictions of the Quadro p4000 GPU. Thus, no data points for batch sizes superior to nine can be reported. It is worth noticing the substantial decrease in inference time as the batch size increases, for the Oliveira et al. model. This inference time shortening derives mainly from batch processing optimisation.

The proposed FCNN and its decoder variants can process 30 images per second on the NVIDIA Quadro p4000 GPU for images of size 576×960. Thus, the proposed FCNN model is able to achieve real-time performances for recordings captured at 25 FPS. On the other hand, the Oliveira et al. model does not exceed the 4 FPS frame rate, making it an impossible contender for real-time applications on an NVIDIA Quadro p4000 GPU.

6.6.2 Class Prediction Performance

6.6.2.1 PASCAL human parts dataset and Freiburg sitting RGB dataset

In table 6.5, the results of the proposed FCNN model and its decoder variants on the combined PASCAL human parts and Freiburg sitting validation RGB dataset are displayed.

Table 6.5: Quantitative results on the PASCAL human parts and Freiburg sitting validation RGB dataset. For each method, the IoU and accuracy for each class and the mean IoU and accuracy are reported. The proposed encoder-decoder architecture (Encoder-decoder-batchnorm) outperforms all the other methods in all the body part classes. Particularly noteworthy are the significant improvements in both accuracy and IoU for thinner classes such as the right and left arm and right and left leg.

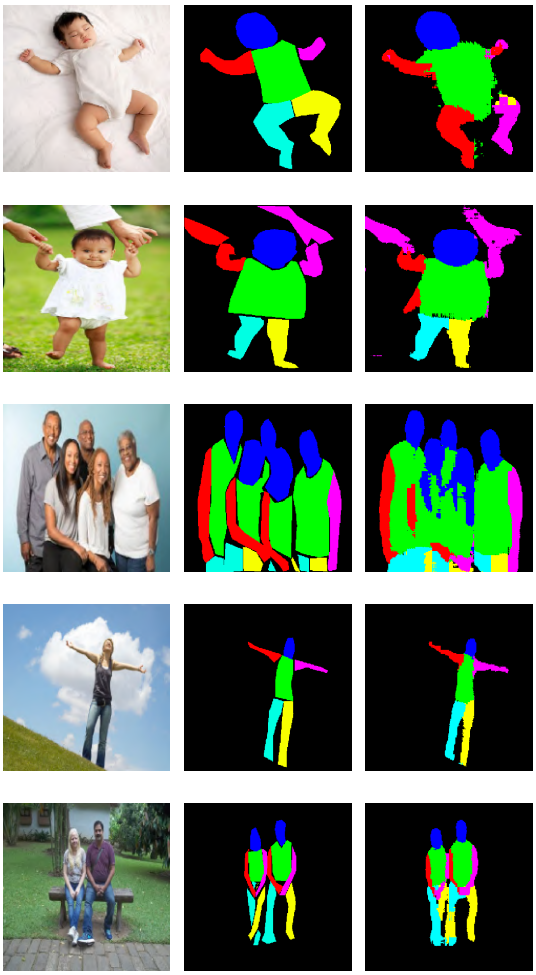
	IoU (%)						Accuracy (%)							
Method	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	**Mean IoU (%)**	**Mean Accuracy (%)**
Encoder-decoder-bilinear	66	55	18	12	29	33	73	71	24	12	31	41	36	50
Encoder-decoder-unconnected	47	52	10	12	30	32	51	72	10	12	30	39	31	43
Encoder-decoder-dropout	65	53	33	32	44	43	73	67	39	40	48	46	45	63
Encoder-decoder-batchnorm	**67**	**56**	**35**	**36**	**46**	**45**	**72**	**69**	**40**	**43**	**52**	**52**	**48**	**66**

The proposed FCNN achieves a better performance with 48 % mean IoU and 66 % mean accuracy, when compared to other decoder variants. As expected, the decoder variant where skip connections are not present yields the worst class prediction performance. This result proves the positive impact of incorporating the object shape information present on intermediate encoder feature maps during the upsampling process. The poor performance of the decoder variant where the upsampling process was conducted by bilinear interpolation proves the importance of an upsampling process mediated by learned parameters, specifically to recover the spacial information of thinner body part classes such as the right and left arm and right and left leg.

Qualitative visual results of the proposed FCNN model on new testing images are illustrated in Figure 6.11.

The proposed model is also compared with the encoder-decoder model proposed by Oliveira et al.. It is worth noting that the model proposed by Oliveira et al. and other state-of-art semantic segmentation models have achieved outstanding results on the PASCAL humans parts dataset, since their ultimate goal is to obtain the highest accuracy, regardless of the computational effort. Also, while the proposed encoder-decoder model is trained end-to-end, the model proposed by Oliveira et al. adopts a stage wise training method.

To perform a fair comparison, the proposed model is trained following the same dataset division as Oliveira et al. for a coarse body part prediction. Thus, the model is trained to segment the body into four parts, meaning that the model will predict 5 classes: head, torso, arms, legs and background. The model will be trained exclusively on 70% of the PASCAL human parts dataset and the remaining 30% will be employed for testing. The class prediction performance results on the PASCAL human parts testing dataset are displayed in Appendix D.2.2.



(a) input frame (b) ground truth (c) prediction

Figure 6.11: Qualitative results of the proposed FCNN model trained on the PASCAL human parts and Freiburg sitting people dataset. The model shows great segmentation results on new testing images. However, the model has some difficulties in segmenting infants. Note that the testing images have the same image resolution as the training and validation dataset.

Although results provided by the proposed FCNN are surpassed significantly by the results provided by Oliveira et al. architecture, its poor performance on the remaining performance indices constitute a hard constraint in its practical application, specifically for the ROI selection task where low computational effort is mandatory. Additionally, the learning efficiency [12] of Oliveira et al. model is low compared with the learning efficiency of the proposed FCNN model, which reveals its poor ability to utilise its parametric space (see Appendix D.2.2).

Only the proposed FCNN model and the decoder variant model where dropout is employed will be further tested on the Neonaten-Navpani dataset since they yield the best trade-off relation between class prediction performance and computational effort.

6.6.2.2 Neonaten-Navpani-RGB dataset

Table 6.6 summarises the class prediction performance of the proposed FCNN model and of the encoder-decoder-dropout model. Note that the reported validation results are the combination of the model's validation results for each fold. The class prediction performance of the proposed FCNN model for each fold is reported in Appendix D.2.3.

Table 6.6: Quantitative results on the Neonaten-Navpani-RGB dataset for images of size 576×960. For the considered methods, the mean IoU, accuracy and precision across the five folds are reported for each class. Additional, the overall mean IoU, accuracy and precision are reported.

	IoU (%)						Accuracy (%)						Precision (%)								
Method	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	**Mean IoU (%)**	**Mean Accuracy (%)**	**Mean Precision (%)**
Encoder-decoder-dropout	79	35	46	34	50	34	84	45	62	51	58	41	90	50	60	45	70	54	46	57	62
Encoder-decoder-batchnorm	80	41	49	35	56	40	85	49	62	55	64	49	89	58	61	46	72	56	50	61	64

The proposed FCNN outperforms the encoder-decoder-dropout model by 3.5% for mean accuracy, 3.7% for mean IoU and 2.1% for mean precision. The higher performance achieved by the proposed FCNN model proves, once more, the superiority of batch normalization layers over dropout layers [40].

Despite the relatively low mean IoU, reported for the proposed FCNN model, for the majority of the frames, the model did not produce labels outside the the neonate's skin, meaning that the low IoU mainly derives from mislabelled body parts and not from poor skin identification performance. Also, a substantial difference between the segmentation performance of the right and left side can be noted. This difference is caused by the unbalanced nature of the Neonaten-Navpani dataset, where the majority of the neonates are mostly lying with their head on the left side of the recording, making the left arm and leg often farthest away from the camera and often hidden behind the remaining body parts, thus hindering the proper training of the left arm and leg classes.

Qualitative visual results of the proposed FCNN model are illustrated in Figure 6.12 and Figure 6.13. As shown in Figure 6.12, the model is able to successfully identify the skin pixels belonging to the neonate and discard the skin pixels belonging to the hands and arms of the clinical staff. Thus, the proposed model reveals a certain independence on colour information. Additionally, the model shows a good performance in body part segmentation, correctly labelling the skin pixels into one of the predefined six body part classes even for challenging body positions and lightening conditions. Thinner and smaller body part regions (e.g. foot that is separated from the continuous skin region of the leg by sensors and cables) are also precisely labelled. The proposed model is also able to affectively distinguish the right and the left when the neonate is in a supine position, despite the similar visual appearance of the categories. Also, the body region boundaries are well preserved, proving the effectiveness of using the shape information from early layers of the encoder-network to recover fine spatial details (e.g. segmentation around cables, sensors).

On the other hand, as shown in Figure 6.13, the model is unable to correctly identify the right and left arm if the neonate is in a prone position. Also, the incubator and the breathing mask pose a challenge to accurate segmentation, which mainly derives from the lack of training examples where these two components are present.

6.6.2.3 PASCAL human parts dataset and Freiburg sitting grey scale images dataset

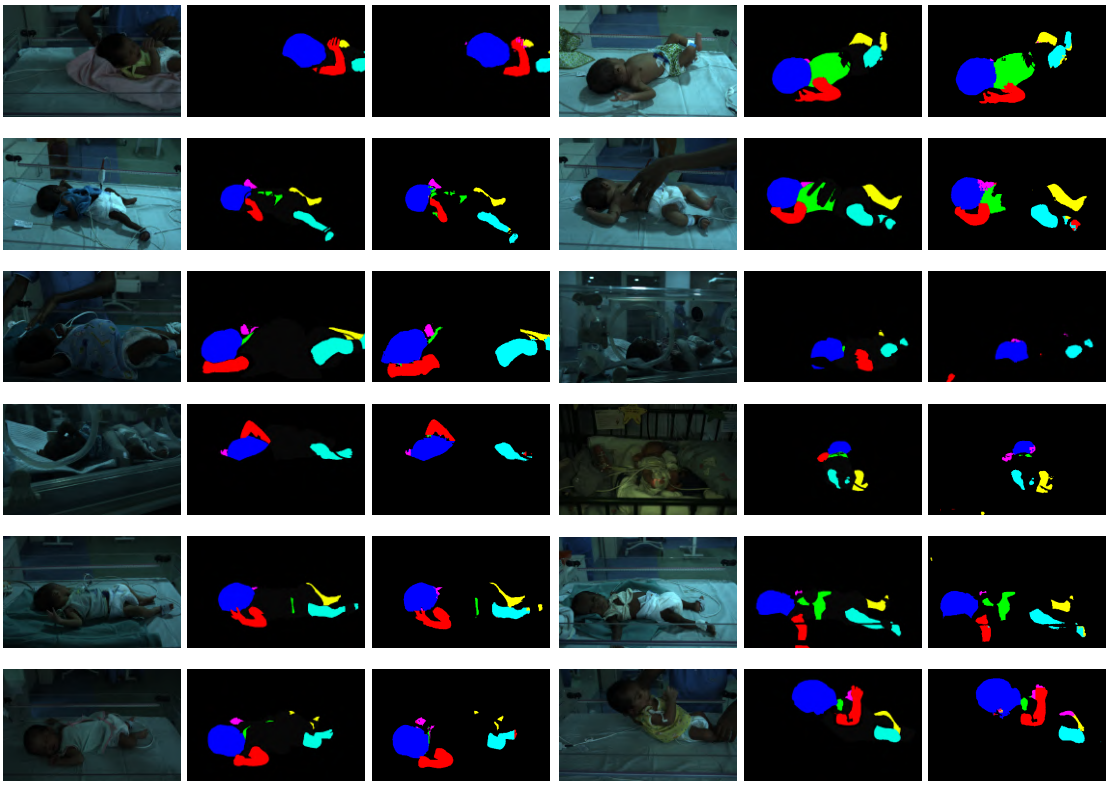
The evaluation of the model trained to select the ROI in IR recordings, will be performed exclusively on the proposed FCNN model and on the model where a dropout layer is employed instead of the batch normalization layer.

Contrary to RGB data, the proposed model achieves a worse class prediction performance on the PASCAL human parts dataset and Freiburg sitting grey scale images dataset, when compared with the encoder-decoder-dropout model. The latter outperforms the proposed model by 8.5% and 14.4% in the overall mean IoU and accuracy, respectively (see Appendix D.2.4). Thus, for a dataset with a reduced amount of information per image, the dropout layer reveals a stronger regularization action, creating a more dependent and robust set of parameters. Those parameters will lead to a better class prediction performance in a diverse dataset, such as the PASCAL human parts dataset.

The class prediction performance of the models for grey scale images is lower when compared with the class prediction performance for RGB images, given the lower number of degrees of freedom in the input information. Note that the proposed FCNN model yields a higher decrease in class prediction performance, specially for thinner classes such as the right and left leg and right and left arm.

6.6.2.4 Neonaten-Navpani-IR dataset

In this Section, the success of the image manipulation techniques employed in the construction of the two Neonaten-Navpani-IR datasets is analysed. Note that both datasets



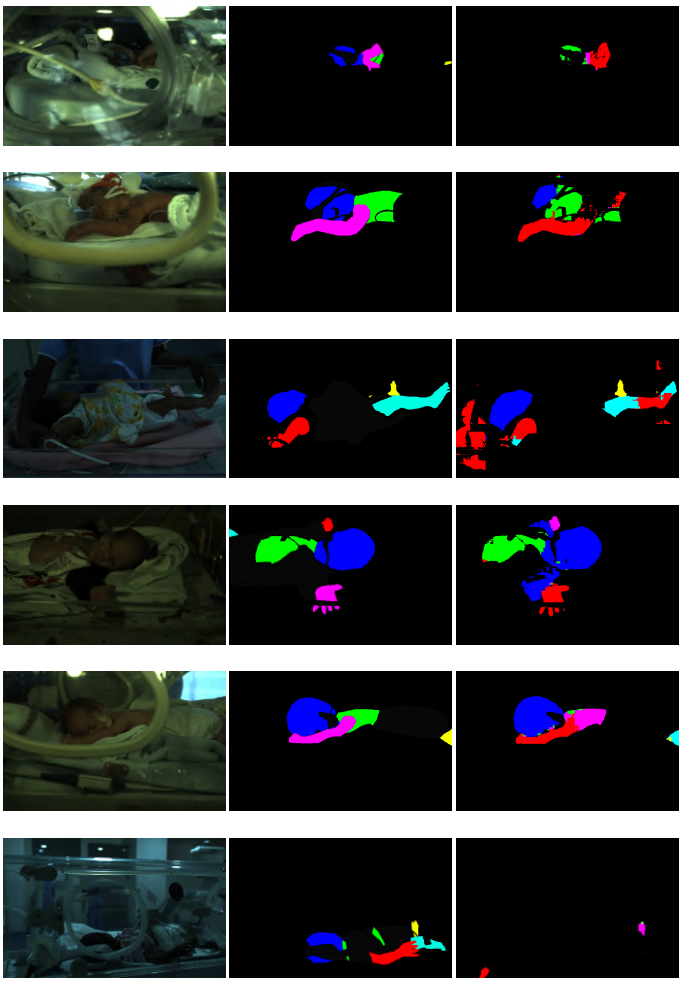
(a) input frame (b) ground truth (c) prediction (d) input frame (e) ground truth (f) prediction

Figure 6.12: Qualitative results of the proposed FCNN model for simultaneous skin and body part semantic segmentation. All the input images have the same image resolution as the training dataset. The model shows a good performance, particularly with its ability to identify the neonate and its body parts within a variety of positions. Additionally, the model successfully ignores the hands and arms of the clinical staff which highlights a certain independence on colour information.

comprise true IR images and manipulated RGB images. For testing, only true IR images are employed. The experiments are conducted relying on the proposed FCNN model and on the encoder-decoder-dropout model.

The class prediction performance of both models is higher when they are trained on the Neonaten-Navpani-IR dataset where the red channel of the RGB images is further manipulated (see Appendix D.2.5). Thus, the histogram manipulation of the RGB images red channel is a better approximation to true IR images than the raw red channel. An experiment where only IR images are employed during training needs to be conducted to extract further conclusions.

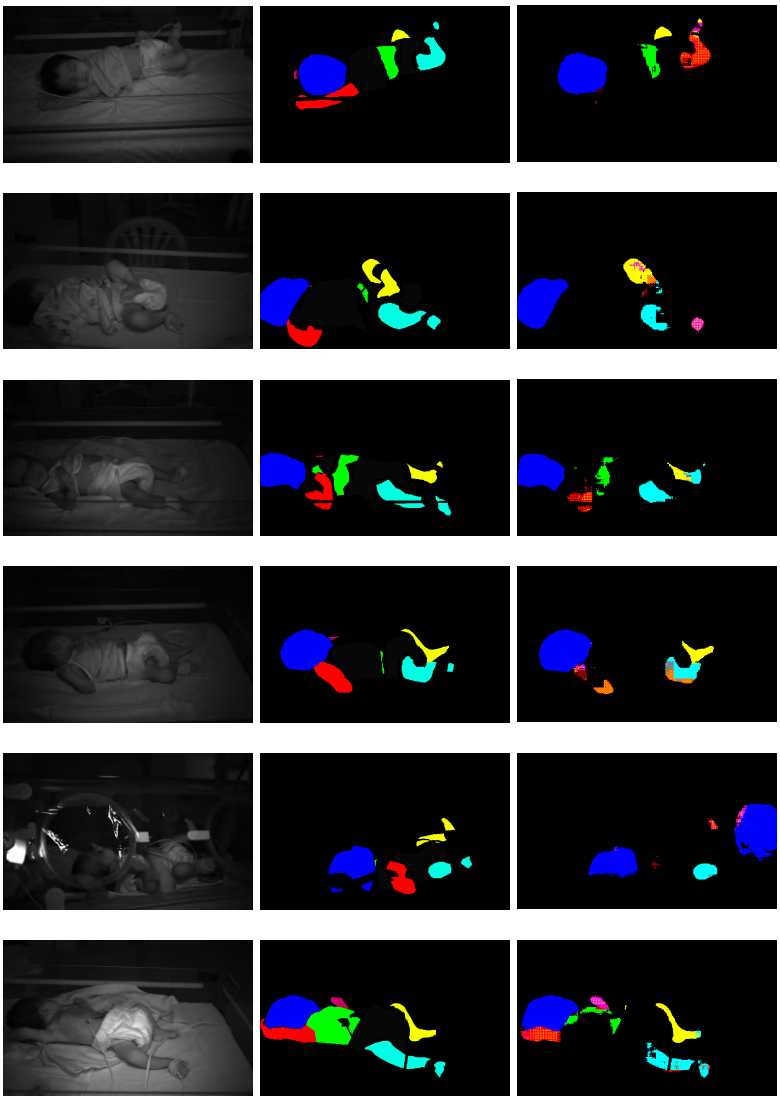
The stronger regularization action of the dropout layer, when there is less information per image available, is further emphasized when comparing the difference in class prediction performance of the considered models between the two Neonaten-Navpani-IR datasets. The difference for the encoder-decoder-dropout model does not surpass the 2% for all the evaluation metrics. Whereas, for the proposed FCNN model, the difference reaches the 13% for overall mean precision. Thus, the encoder-decoder-dropout model is able to better



(a) input frame (b) ground truth (c) prediction

Figure 6.13: Qualitative failure results of the proposed CNN model for simultaneous skin and body part semantic segmentation. All the input images have the same image resolution as the training dataset. The model shows an inability to correctly label the left arm if the neonate is in a prone position, labelling it as right arm. Additionally, the breathing mask hinder the correct labelling of the head.

generalize its weights and therefore have a better class prediction performance when using a dataset where grey scale images comprise very distinct pixel values from those of testing images (true IR frames).



(a) input frame (b) ground truth (c) prediction

Figure 6.14: Qualitative results of the proposed CNN model for simultaneous skin and body part semantic segmentation on testing IR frames.

Despite the higher encoder-decoder-dropout model's ability to better generalize its learnable parameters, the proposed FCNN model still outperforms the encoder-decoder-dropout model by 3.5% on the overall mean precision for the Neonaten-Navpani-IR-manipulated dataset where histogram manipulation is applied to the red channel of Neonaten-Navpani-RGB images. Thus, and despite the reduced number of testing images per fold, which limits the formulation of a generalized conclusion, the proposed FCNN seems to be more indicated

for ROI selection, where precision is one of the key elements for successful PPGI extraction. The class prediction performance of the proposed FCNN model for each fold on the aforementioned Neonaten-Navpani-IR is reported in Appendix D.2.5.

Despite the efforts to adapt the Neonaten-Navpani-RGB dataset to resemble IR images, the class prediction performance of the proposed model trained on the manipulated Neonaten-Navpani-IR dataset is still low compared with the class prediction performance of the model trained on the Neonaten-Navpani-RGB dataset. A higher class prediction performance would be achieved if the model was only trained with IR images. However, similarly to what happened with the model trained on the PASCAL human parts and Freiburg sitting people grey images dataset, the class prediction performance is expected to be lower given the lack of colour information.

Qualitative visual results of the proposed FCNN model trained on the Neonaten-Navpani-IR-manipulated on IR images are illustrated in Figure 6.14. As shown in the figure, the model is able to identify the skin pixels for despite the lack of colour information. However, the model reveals some difficulties in labelling the skin pixels into the correct body part classes.

6.6.3 Impact of Transfer Learning

In this Section the effect of transfer learning is studied, relying exclusively on the proposed FCNN model. Experiments are carried out where the training on the Neonaten-Navpani-RGB dataset starts from the weights of the model trained on the PASCAL human parts and Freiburg sitting RGB dataset and where the training on the Neonaten-Navpani-RGB dataset starts from the pre-trained weights of ResNet-50, for the image classification task on the ImageNet dataset.

When no transfer learning is employed, the individual and overall class prediction performance is much worse, specially on thinner classes such as arms and legs, due to poor generalization (see Appendix D.2.6). When the weights of the model trained on the PASCAL human parts and Freiburg sitting RGB dataset are used as a starting point, the mean IoU improves from 19% to 50 %, showing that transfer learning is essential when training deep learning models on small and challenging datasets.

Besides the substantial improvement in the final class prediction performance, leveraging knowledge from the general human body part segmentation task revealed to improve the learning process of the proposed model in two more aspects. Firstly, the initial prediction performance achieved on the test Neonaten-Navpani dataset is significantly higher when compared with the non pre-trained model. The latter indicates that the human body parts segmentation task is highly related to the simultaneous skin and body parts segmentation task. Secondly, despite the unchanged learning rate, the pre-trained model is able to achieve a higher prediction performance in a lower number of epochs. These three aspects are pointed in [76] as indicators of positive transfer learning. Figure 6.15 illustrates the aforementioned indicators.

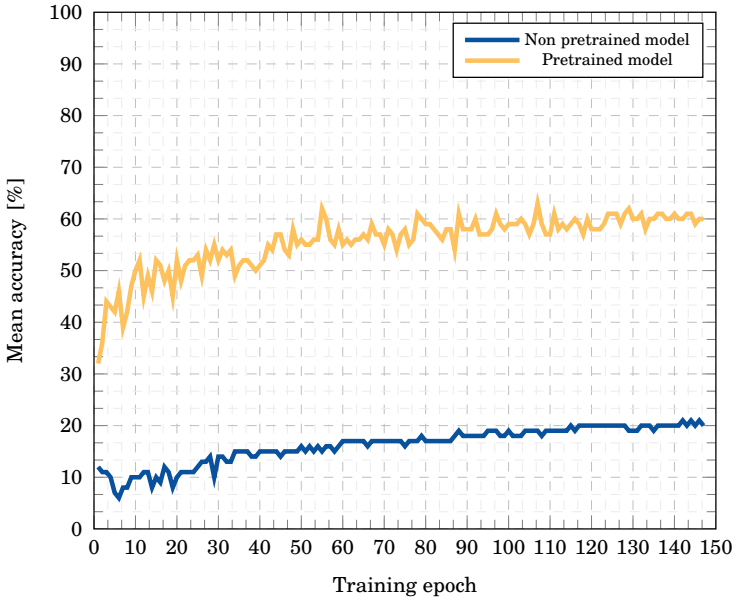


Figure 6.15: Mean accuracy *vs*. training epoch. This graph shows the evolution of the mean accuracy for the pre-trained and non pre-trained encoder-decoder-batchnorm model during the training on second fold of the Neonaten-Navpani dataset. The graph emphasizes the three benefits of transfer learning: the initial mean accuracy of the pre-trained model is higher than the mean accuracy of the non pre-trained model, the pre-trained model mean accuracy evolution reveals a higher slope indicating a higher rate of performance improvement and the higher asymptote of the pre-trained model mean accuracy indicates a higher performance after reaching the global optimal solution.

6.6.4 Effect of Data Augmentation

To examine the impact of data augmentation, an additional experiment is conducted where the network is fine-tuned with the Neonaten-Navpani-RGB dataset without resorting to data augmentation methods.

An overall performance of 54.6% mean IoU is obtained which is 5% more than the performance achieved when the model is fine-tuned using data augmentation methods. The latter contradicts the intuition that data augmentation always leads to an increase in the class prediction performance. This contradiction mainly derives from the balanced nature of the five folds, meaning that, despite the diversity of the Neonaten-Navpani dataset, the balanced dataset distribution between the five folds allows the training of the model with subjects similar to ones in which the model will be validated. Thus, the overfitting of the model to the training examples leads to an improvement in the class prediction performance of the model since validation and training examples are similar. On the other hand, the data augmentation generalizes the learnable parameters of the model, leading to a poorer class prediction performance in the validation examples. However, this generalization is desirable to enable the model to accurately identify the ROIs in distinct subjects.

6.6.5 Receptive Field Analysis

The proposed model exhibits the characteristic limitation of semantic image segmentation algorithms based on FCNNs [59]. The network fixed receptive field size makes it unable to handle multiple object scales. Thus, subjects that are too small with respect to the size of the receptive field are often classified as background. The latter limitation is illustrated in Figure 6.16. However, for the Neonaten-Navpani dataset, the described limitation is not relevant since subjects have a favourable and fairly constant scale.

(a) input frame (b) ground truth (c) prediction

Figure 6.16: Qualitative results of the proposed FCNN model, trained on the PASCAL human parts and Freiburg sitting people dataset. The model shows an inability to identify subjects that are too small with respect to the size of the receptive field, labelling them as background. The first line illustrates the latter limitation: the model is not able to correctly identify the body parts of the subjects that have a small scale. The second line is the prediction of the first image after zooming, to prove that the model is able to correctly label and identify the body parts when the same subjects have a favourable scale. Note that the illustrated input images have the same resolution as the training images, 320×320.

Despite the fixed receptive field size, the context information that each neuron capture from the input image can be changed by varying the input image resolution. Experiments were carried out where the proposed FCNN model is trained on the Neonaten-Navpani-RGB dataset using three different input resolutions: 288×480, 576×960 and 1184×1920. The model yields substantially better class prediction performance for the intermediate resolution (see Appendix D.2.8) outperforming the higher and lower resolution by 12.5% and 4.4% for the mean IoU, respectively. For the higher resolution, the receptive field of the model is not sufficiently large to capture the global context of the input image, whereas, for the lower resolution, the receptive field of the model is to wide, including a excessively large neighbourhood. Particularly noteworthy is that, despite the scales being equal in absolute, the results of having a excessively small receptive field are far more worse than having a excessively large receptive field. The described receptive field analysis is valid for the Neonaten-Navpani dataset, where all the subjects have a relative constant size in the

images.

6.6.6 Refinement Algorithm

As mentioned in Section 6.4, the refinement algorithm aims to improve the body part segmentation overall performance. In this section, the effect of this new post-processing method is quantitatively measured. Note that the proposed refinement algorithm will not improve the class prediction accuracy. Instead, it will improve the IoU and precision metrics by converting the mislabelled pixels into background, thus reducing the number of FP. For the sake of simplicity, only the mean IoU and the mean precision across the 31 subjects are reported in Table 6.7, the box plot comparing the precision percentage obtained before and after the refinement algorithm is presented in the appendix.

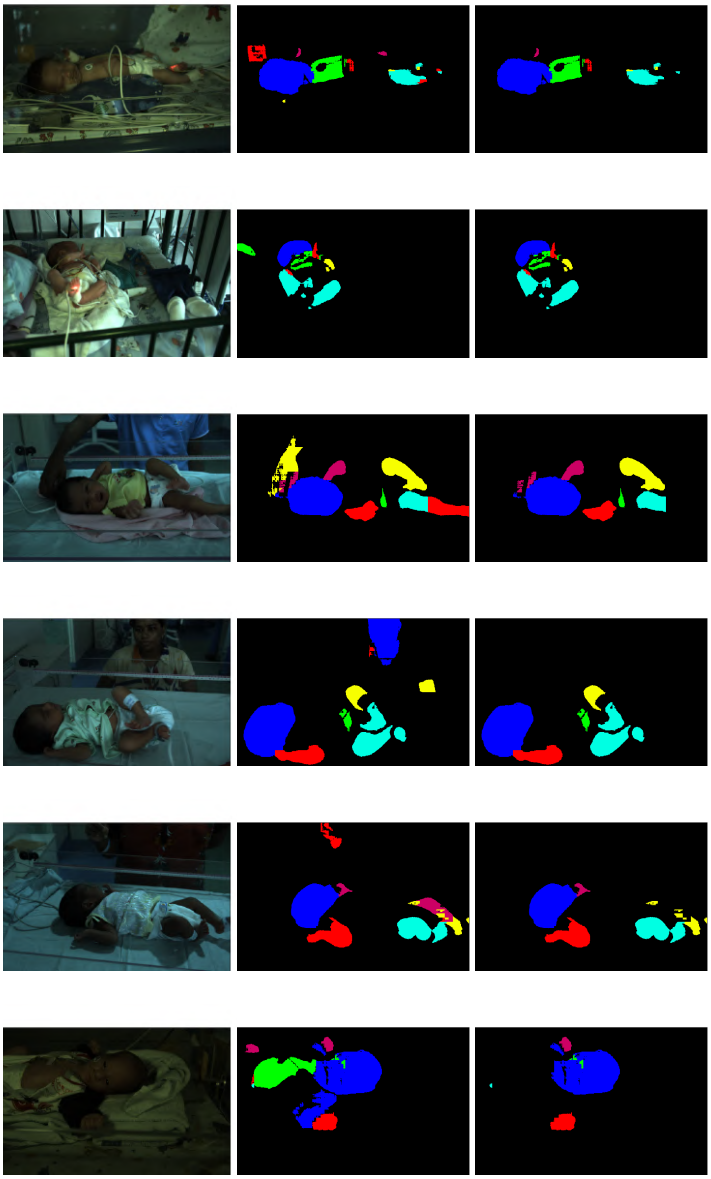
Table 6.7: Quantitative results on the Neonaten-Navpani-RGB dataset before and after the application of the refinement algorithm. The reported results for each class correspond to the mean IoU and mean precision across the 31 subjects. Note that the results before the refinement are collected after the pre-processing procedure, where residual isolated segments are eliminated. The latter justifies the difference between the results presented in Table D.4 and the results in the present table.

	IoU (%)						Precision (%)							
	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	**Mean IoU (%)**	**Mean Precision (%)**
Before refinement	89	38	52	39	60	42	97	53	66	48	82	57	53	67
After refinement	89	38	57	44	60	42	98	53	73	55	82	57	55	70

Table 6.7 shows that the refinement algorithm substantially boosts the class prediction performance of the proposed FCNN, offering a 5% absolute increase in the mean IoU and a 7% absolute increase in the mean precision of the right and left arm class. For the remaining four classes no substantial difference was attained. Besides the head class, where the mean accuracy across all the subjects suffered a decrease of 1%, the mean accuracy for the remaining body part classes remains unchanged after the application of the refinement algorithm.

The performance increase does not occur for all subjects: for recordings where only one instance per class is present, the application of the refinement algorithm does not affect the class prediction performance. On the other hand, for the remaining subjects, the refinement algorithm yields consistent improvements over the baseline model. For example, for subject S009_S006 (Figure 6.17 fourth line), the refinement algorithm yields about 8% improvement

in both IoU and precision for the head class, and a 8% and 10% improvement in the IoU and precision, respectively, for the right arm class.



(a) input frame (b) before refinement (c) after refinement

Figure 6.17: Qualitative results of the refinement algorithm.

Qualitative visual comparisons of the proposed FCNN model's results before and after the refinement algorithm are illustrated in Figure 6.17. The segmentation results of the proposed FCNN before the application of the refinement algorithm include the segmentation of body parts that do not belong to the neonate. Employing the refinement algorithm further improves the performance by removing the false positive body parts and isolated segmentations.

Despite the great results of the refinement algorithm as a post-processing method, its

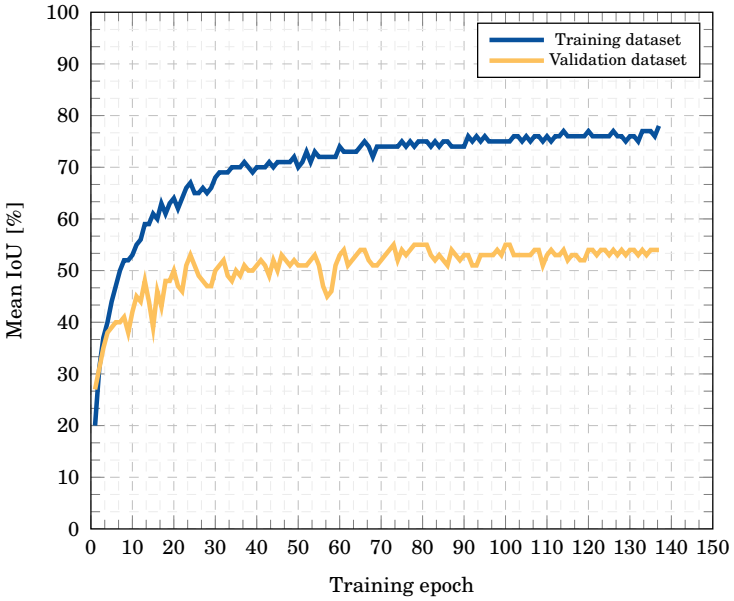


Figure 6.18: Mean IOU *vs*. training epoch. This graph shows the evolution of the mean IOU of the proposed FCNN model on the Neonaten-Navpani-RGB training and validation dataset during the training on the first fold.

parameters were manually optimized to fit the Neonaten-Navpani dataset, where the size of the subjects in the field of view are approximately constant across the different recordings. Thus, the refinement algorithm results can not be generalized to larger and more diverge dataset, such as the PASCAL human parts dataset, where the subjects appear at different scales.

6.6.7 Network Training

6.6.7.1 RGB model

In the pre-training stage, where the proposed FCNN model is trained on the PASCAL human parts and Freiburg sitting people RGB dataset, the model converges to global optimal solution after approximately epoch 300, using a batch size of six. The training takes approximately 50 hours on Google Colab. The amount of time required for the proposed model to converge is substantially low, when compared with the 240 hours required to train the model proposed by Oliveira et al. on the same dataset.

In the fine-tuning stage, where the proposed FCNN model is trained on the Neonaten-Navpani-RGB dataset, the model converges to global optimal solution after approximately epoch 100, using a batch size of three. The training takes approximately 9 hours on the workstation. Note that the fine-tuning stage is repeated five times, one for each fold. Thus, the five models take around 95 hours to train.

The evolution of the overall mean IoU of the Neonaten-Navpani-RGB training and validation dataset during the training process is shown in Figure 6.18.

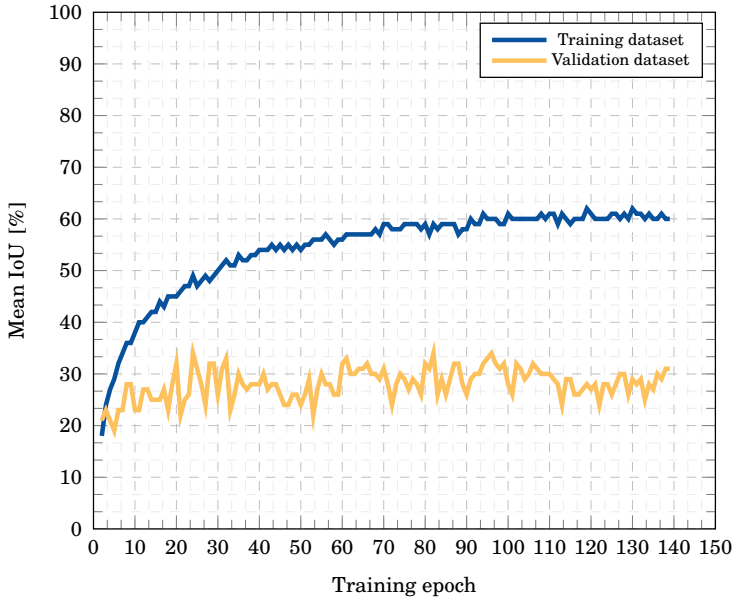


Figure 6.19: Mean IoU *vs.* training epoch. This graph shows the evolution of the mean IoU of the proposed FCNN model on the Neonaten-Navpani-IR-manipulated training and validation dataset during the training on the first fold..

6.6.7.2 IR model

In the pre-training stage, where the proposed FCNN model is trained on the PASCAL human parts and Freiburg sitting people grey images dataset, the model converges to global optimal solution after approximately epoch 450, using a batch size of three. The training takes approximately 75 hours on Google Colab.

In the fine-tuning stage, where the proposed FCNN model is trained on the Neonaten-Navpani-IR-manipulated dataset, the model converges to global optimal solution after approximately epoch 90, using a batch size of three. The training takes approximately 7.5 hours on the workstation. Note that the fine-tuning stage is repeated five times, one for each fold. Thus, the five models take around 112.5 hours to train.

The evolution of the overall mean IoU of the Neonaten-Navpani-IR-manipulated training and validation dataset during the training process is shown in Figure 6.19.

6.7 Discussion

The proposed FCNN model produces semantic segmentation masks that are progressively refined across the decoder network in a coarse-to-fine manner. During this process, intermediate feature maps from the encoder network are incorporated to progressively recover the fine details in the produced segmentation mask. A new architecture feature was introduced, the batch normalization layer, that replaces the dropout layer. Additionally, the ResNet-50 was introduced to replace the traditional VGG-16. Experimental results on several computational effort indicators proved the computational efficiency of the proposed model both in terms of computational efficiency but also in terms of memory usage. Thus, making the model appropriate to address the ROI selection task at real time, which was one of the primary motivations behind its design. However, contrary to previous works, there was a class prediction performance decrease when replacing the VGG-16 for the ResNet-50. To justify the latter statement new tests need to be conducted, however, the end-to-end training procedure could be the main factor behind the class prediction performance discrepancy since, as pointed in [5], the multi-stage training procedure employed by Oliveira et al. enables the model to achieve higher accuracy.

Despite the discouraging results on the PASCAL human parts and Feiburg sitting dataset, the proposed model revealed good segmentation results on the Neonaten-Navpani-RGB dataset being robust to various skin tones, body position and locations (e.g. incubator or infant radiant warmer) and routine interactions of clinical staff. Despite the lower class prediction performance for the Neonaten-Navpani-IR dataset the model also revealed good results given the fact that colour information is not present. Applying the refinement algorithm leads to a class prediction improvement.

Worth noticing is the high precision of the head class both for the RGB and IR model (80% and 62%, respectively), which is the most common area to extract the PPGI signal in adults [77] [64] [47].

In future work, more subjects should be included to further generalize the parameters of the model. Also, the downsampling of the ground truth masks during the training process, to match the size of the input frame, was not a good practice. As pointed in [17], the downsampling process removes the fine annotations which leads to no back-propagation of fine structure details. Instead, the segmentation masks provided by the FCNN should be upsampled to match the original resolution of the ground truth masks which, in turn, is the resolution of the recordings' frames.

Overall, this new method for ROI selection is capable of tracking multiple possible measurement areas for continuous extraction of the PPGI signal. The PPGI extraction will be conducted by the proposed FCNN model, pre-trained on the PASCAL human parts and Freiburg sitting people dataset and further fine-tuned on the Neonaten-Navpani dataset with images with a quarter-resolution of the neonates' recordings frames.

Chapter Seven

PPGI and Heart Rate Extraction

The current chapter's aim is to prove that the new developed model for ROI selection provides the necessary measuring sites for HR assessment. Section 7.1.1 presents the algorithm to extract the PPGI signals. The proposed algorithm to compute the HR is described in Section 7.1.2. The results are presented in Section 7.2 and discussed in Section 7.3.

7.1 Methodology

The current approach to address the HR estimation through RGB and IR video recordings is based on the fact that subtle changes in the skin's reflected light accompany the blood volume fluctuations of superficial blood vessels due to the cardiac cycle. The approach takes a sequence of frames of a neonate as input and returns a HR value. The first step of the proposed method comprises the automatic selection of the ROIs for each frame, i.e the identification of the neonate's skin pixels and their simultaneous classification into one of the predefined body parts. To this end, the method proposed in Chapter 6 is employed. From the tracked individual measurement areas, individual PPGI signals are extracted. For each PPGI signal, a quality index based on the signal's frequency spectrum is calculated, reflecting the likelihood of containing HR information. A new PPGI signal is computed based on the weighted fusion of the extracted PPGI signals. The HR value will be computed from the PPGI signal with the highest quality index. The selected PPGI is filtered to remove signal fluctuations originated by external sources. To identify the HR, the wavelet synchrosqueezed transform is analysed. The main steps necessary to compute the HR through video recordings are further detailed forthwith.

7.1.1 PPGI Signal Extraction

The PPGI signal extraction comprises the computation of the average intensity for the pixels within the identified ROIs. The average intensity of each ROI is computed for each frame of the neonate's recording according to Equation 7.1.

$$\mathrm{PPGI}_i(t) = \frac{\sum_{\mathrm{ROI}} I_{x,y}(t)}{N} \tag{7.1}$$

where N corresponds to the number of pixels inside the ROI and $I_{x,y}(t)$ corresponds to the value of the pixel positioned in the coordinates x and y in the two dimensional grid $I(t)$. $I(t)$ corresponds to the difference between the green and red channel of the frame at time t. The first contains the strongest PPGI signal amongst the remaining channels, that better correlates with the HR. The second, whose amplitude is poorly modulated by the changes in the blood volume in superficial blood vessels [19], cancels the illumination intensity changes due to external factors, that are present in both green and red channel [1] [77] (see Figure 7.1). The average pixels' intensity within each ROI for all frames forms the PPGI signals, $PPGI_i(t)$, where $i \in \{1,...,8\}$ indexes the eight identified ROIs: six ROIs identified by the FCNN and the two extra ROIs. The notation used for i is explained in Appendix E.1.1. Thus, for each ROI, a PPGI signal is computed.

The mean intensity fluctuations that accompany blood volume changes within the cardiac cycle are more visible in raw uncompressed video data. Thus, the green and red channel used to compute I are the raw uncompressed channels after the demosaicing process at the original resolution of 1200×1920 and a frame rate of 25 fps.

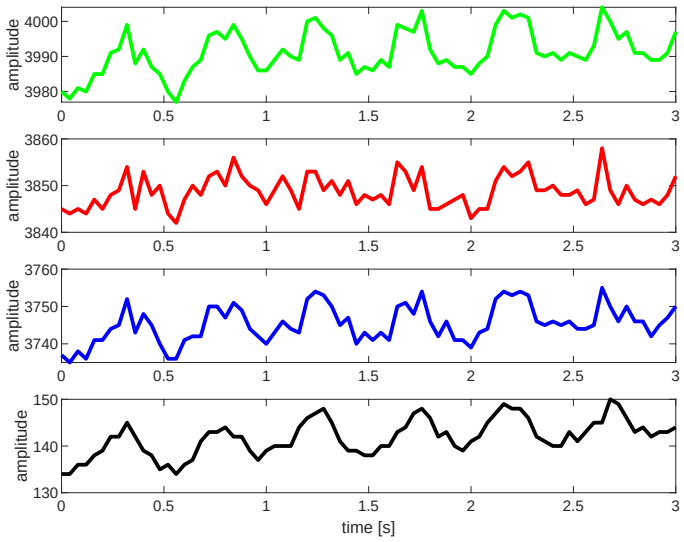


Figure 7.1: From top to bottom: average of the head's pixels' intensity from the green, red and blue channels. At the bottom: average of the head's pixels' intensity from the two dimensional grid originated by the difference between the green and red channel, $I(t)$. The signals were extracted from the subject S009_S012.

The PPGI extraction was implemented in Python, specifically using the Pytorch library [61].

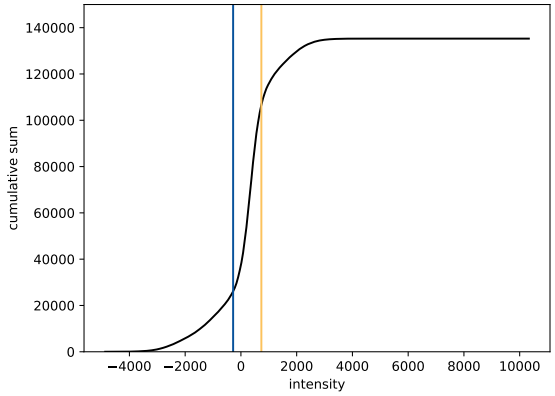
7.1.1.1 Extra ROI

The proposed method for ROI selection identifies the head of the neonate as a measuring site. However, structures that do not contain HR information (e.g. eyes and hair) are included in the identified head region. Additionally, there is a uneven distribution of the blood carrying capillaries density in the head's skin region, which means that the total amount of blood volume change within a cardiac cycle will depend on the skin region. Thus, the strength of the PPGI signal component that is modulated by the blood volume fluctuations will vary according to the head's skin region. For instance, the forehead region has a high blood perfusion when compared with other skin regions [27] and therefore usually yields a strong PPGI signal [47] [55].

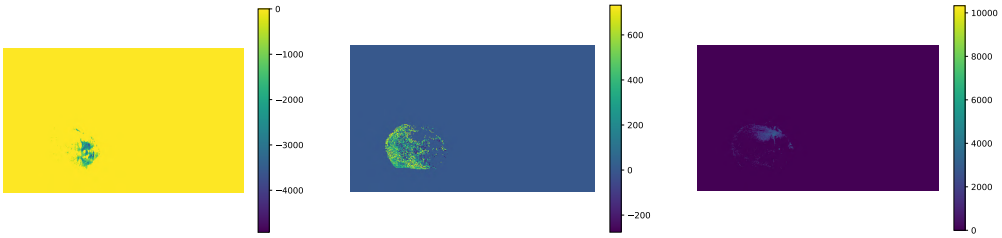
When extracting the PPGI signal from the whole identified head region, regions with low or absent blood perfusion are included, introducing noise into the signal. An algorithm to further divide the identified head region into two segments is developed in order to include two extra ROIs to address the aforementioned problem.

The method to define two extra ROIs is based on the head's pixels' intensity distribution and creates two sets of pixels with similar reflected light values. The method starts by computing the cumulative sum of the values of the head's pixels' intensity histogram. Then, two limits are created: the lower limit which corresponds to the intensity value which is higher than the intensity value of 20% of the total number of the head's pixels and the upper

limit which is higher than the intensity value of 80% of the total number of the head's pixels. Using the latter limits, two sets of pixels are created: the head middle section comprising pixels whose intensity values are higher than the lower limit and lower than the higher limit and the head lower section comprising pixels whose intensity values are lower than the lower limit. Figure 7.2 illustrates the pixels' head division.



(a) Cumulative distribution of the head's pixels' intensity. The two vertical lines separate the three sets of pixels.



(b) Head lower section. (c) Head middle section. (d) Head upper section.

Figure 7.2: Illustration of the head's division into three sets of pixels according to the head's pixels' intensity distribution. The head lower section comprises the skin area around the cheeks and nose of the neonate which are normally associated to a better PPGI signal quality when compared with the signal provided by the remaining sets of pixels. The head upper section will not be used given its poor correlation with the HR. Note that the intensity values are originated from the two dimensional grid $I(t)$

.

7.1.1.2 Implementation Details

To minimize the processing time, the algorithm iteratively processes a batch of 15 frames at the time. Thus, the algorithm starts by loading 15 frames from the raw uncompressed video file. After demosaicing, the algorithm creates 2 groups of images: a RGB 8-bit image group, H_n, and a one channel 16-bit image group, I_n, corresponding to the difference between the green and red channel of each considered frame. Note that $n \in \{1, ..., 15\}$.

Both image groups are converted into Pytorch tensors and transferred to the GPU memory. The frames belonging to H are resized to a quarter of the original frames' resolution

and cropped equally from both sides to match the resolution of 576×960. Then, the H tensor, of size $15 \times 3 \times 576 \times 960$, is fed into the proposed FCNN model for ROI selection. The first prediction mask is processed by the refinement algorithm proposed in Section 6.4 to originate segmentation masks, C_i where $i \in \{1,...,6\}$. The output prediction masks, M, of size $15 \times 576 \times 960$, and the calibrations masks are then upsampled using nearest-neighbour interpolation to a resolution of 1200×1920. Note that the domain of M ranges over seven possible labels, $p \in \{0,...,6\}$.

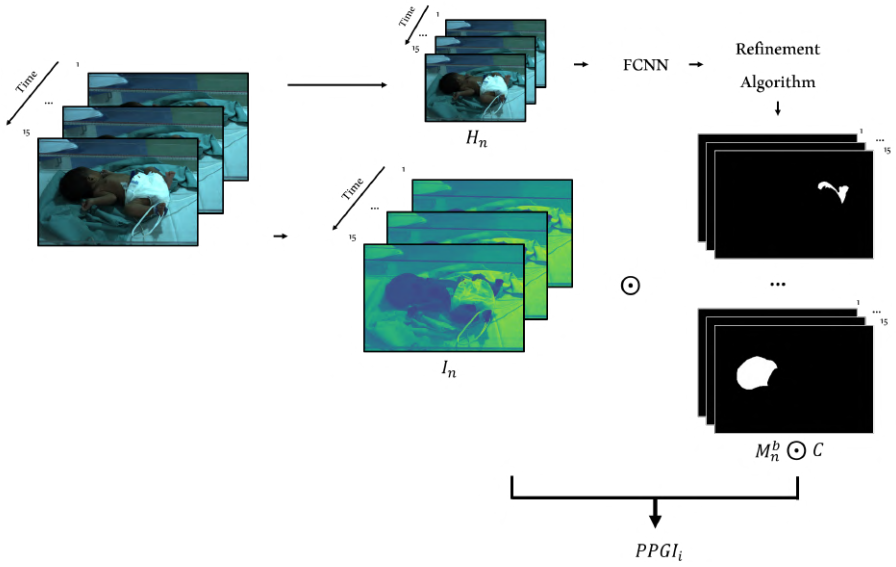


Figure 7.3: Illustration of the PPGI extraction steps.

For each M_n and their correspondent I_n, six PPGI values are computed, each corresponding to a PPGI value from a body part. Note that if the body part class was not identified in the frame H_n, its correspondent PPGI value is zero. The PPGI values are extracted following Equation 7.1. Firstly, the mask M_n is divided into six binary segmentation masks, $M^{\mathrm{b}}_{n_k}$, containing the individual segmentation of each body part. The element-wise multiplication of I_n, C_i and $M^{\mathrm{b}}_{n_i}$ followed by the sum of the result's elements results in the numerator of the Equation 7.1. The sum of the elements of $M^{\mathrm{b}}_{n_i}$ originates the denominator of the Equation 7.1. Thus, the division of the aforementioned values results in the average intensity value at the moment t from the ROI i. The process is repeated for all the ROIs in the frame and for all the frames of the loaded batch of images. Additionally, from the element wise multiplication of $M^{\mathrm{b}}_{n_1}$, C_i and I_n the two extra $PPGI_i$ values are computed relying on the process described in Section 7.1.1.1. All the PPGI values are saved. Then, a new batch of frames is loaded and the process of PPGI extraction is repeated.

The process of PPGI extraction takes, on average, 0.46 s per image on an NVIDIA Quadro p4000 GPU. The latter processing time includes the FCNN model inference time and the time of the computations mentioned in the current section. When the refinement algorithm step is excluded, the average processing time decreases to 0.26 s per image. The reported processing time can be further optimized.

7.1.2 HR Estimation

The current section describes the procedures employed in the HR estimation process.

7.1.2.1 Quality Index

The strength of the PPGI signal is not homogenous throughout the skin pixels. As previously observed [47] [49] [57], the signal quality varies depending on the body part and skin region within the body part, according to the density of blood carriyng capillaries. Figure 7.4 illustrates the difference in the PPGI signal quality for different ROIs.

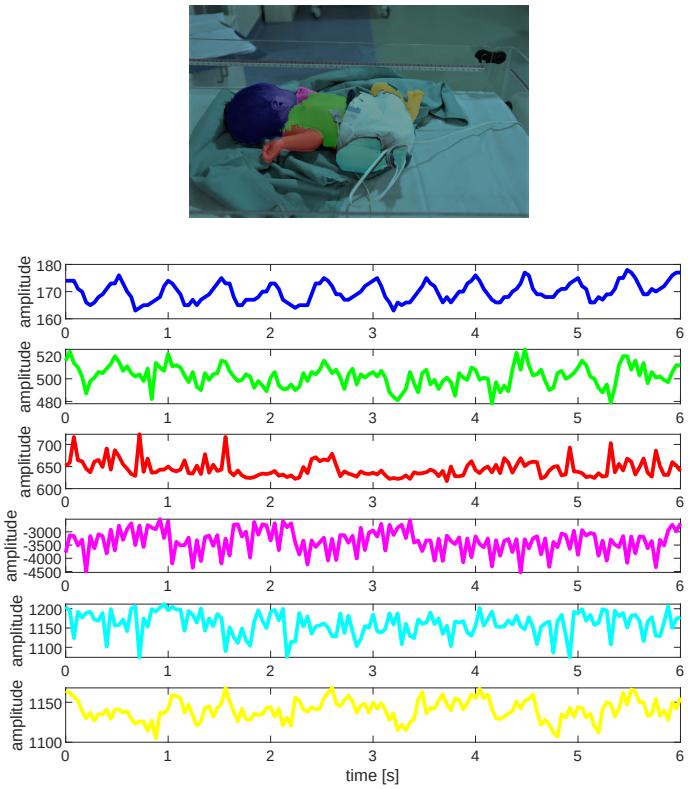


Figure 7.4: From top to bottom: average of the pixels' intensity for the head, torso, right arm, lef arm, right leg and left leg regions. The strenght of the PPGI signal varies according to the body region: the signal extracted from the head yields a better quality when compared with the signals originated from the remaining body regions.

Thus, there is a need to compute a quality index for each extracted PPGI signal that will reflect the likelyhood of containing HR information and the quality of the signal itself. The quality index will determine which ROIs will allow a better HR estimation and which ROIs should be rejected. The quality index computation follows the method employed in [9].

For the quality index, the PPGI signals are bandpass filtered in order to remove the signal's components related to breathing-synchronous motion.

By knowing that blood volume changes within the cardiac cycle causes low amplitude fluctuations in the average pixels' intensity, when compared with the artefacts introduced

by sudden changes in illumination (e.g. shadows caused by clinical staff movement), PPGI signals can be rejected for HR estimation, within the studied time window, if large amplitude changes are present. Thus, within the studied time window, the quality index is set to zero if the difference between the minimum and maximum signal's amplitude is greater than 100. The latter threshold was optimized manually and it is based on the fact that the observed PPGI component that is modulated by blood volume fluctuations does not surpass the ± 50 amplitude value, after filtering.

For the remaining PPGI signals, a quality index is computed relying on the fact that the PPGI component that is modulated by the blood volume changes has a dominant frequency corresponding to the HR. Therefore, the frequency components of a good PPGI signal should be concentrated in a small frequency band centred in the HR, in the signal's frequency spectrum. Frequency components outside the aforementioned small frequency band correspond to noise in the PPGI signal. Based on this assumption, a quality index can be estimated by computing the ratio of the power of the PPGI signal in turn of the HR to the power of the noise in the bandpass filter range. Thus, the quality index can be defined as:

$$QI_i(t) = \begin{cases} \frac{\int_{f_{\max}-b}^{f_{\max}+b} \mathrm{DFT}_i(f)df}{\int_{B1}^{B2} \mathrm{DFT}_i(f)df - \int_{f_{\max}-b}^{f_{\max}+b} \mathrm{DFT}_i(f)df} & f_{\max} \in [60\text{bpm}, 170\text{bpm}] \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

where $\mathrm{DFT}_i(f)$ denotes the DFT of the filtered PPGI_i signal over the considered time window. The interval $[B1, B2]$ correspond to the passband frequency of the bandpass filter and the interval $[f_{\max}-b, f_{\max}+b]$ denotes a small frequency band centred in the HR. Note that the computed maximum frequency needs to be inside the HR's physiological range to be considered the HR. Otherwise, the quality index is set to zero. The parameter b was manually optimized and set to 0.25 Hz.

Figure 7.5 shows eight seconds of the filtered PPGI signals from the head, torso and left leg regions and their respective DFT and quality index. The DFTs illustrate how the quality index is defined: the blue area corresponds to the numerator and the yellow area corresponds to the denominator of Equation 7.2. By comparing the PPGI signals and their respective quality indexes, it is evident that the quality index correctly reflects the quality of the PPGI signal.

7.1.2.2 PPGI signal fusion

Following [74] [31] and [78], that state that the HR estimation is more reliable when fusing the signals obtained from different skin regions, the PPGI signals extracted from the different tracked measuring sites are combined through a weighted average, following the signal fusion method proposed in [47]. The weights are determined by the computed quality indexes.

Given the fact that blood reaches different skin regions at slightly different times, it is expected to observe varying delays amongst the extracted PPGI signals from the six different body parts. However, when calculating the cross-correlation between the PPGI

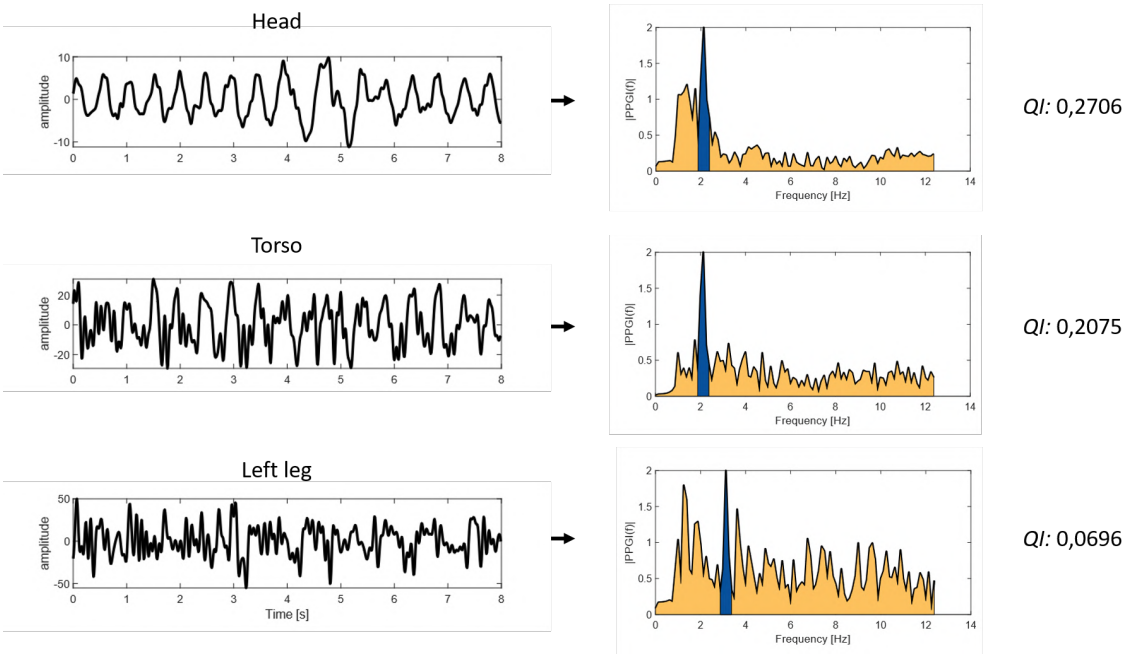


Figure 7.5: Illustration of the quality index definition. The yellow area represents the denominator and the blue area corresponds to the numerator of Equation 7.2. The PPGI signal extracted from the head detains the higher quality index due to the fact that its frequency spectrum consists of a clear peak at the HR physiological range and no significant peaks at the remaining bandpass filter frequencies.

extracted from the head and the PPGI extracted from the torso and arms, no significant time lag between the PPGI signals was found. The small delays between the PPGI signals are lower than 20 ms and, for the frame rate of 25 Hz, can be neglected. Therefore, no preprocessing methods are needed for the PPGI signal fusion step. Given the poor quality of the PPGI signal extracted from the legs, no assumptions on the time lag between the PPGI signals can be made. The cross-correlation plot of the aforementioned PPGI signals can be consulted in Appendix E.2.1.

Thus, the PPGI signal fusion can be simply defined as:

$$\text{fPPGI}(t) = \sum_{i=1}^{n} GI_i \times \hat{\text{PPGI}}_i(t) \tag{7.3}$$

where GI_i denotes the quality index for the ROI_i, $\hat{PPGI}_i(t)$ denotes the filtered PPGI signal from ROI_i and n corresponds to eight, the maximum number of ROIs. The signal fusion operation is illustrated in Figure 7.6.

7.1.2.3 Temporal Filtering

Not all frequencies present in the extracted signal reflect the blood volume fluctuations due to cardiac activity: the camera sensors captures the PPGI signal mixed with light fluctuations coming from different sources. For instance, frequencies bellow 1Hz, are caused by breathing-synchronous motion (see Figure 7.7).

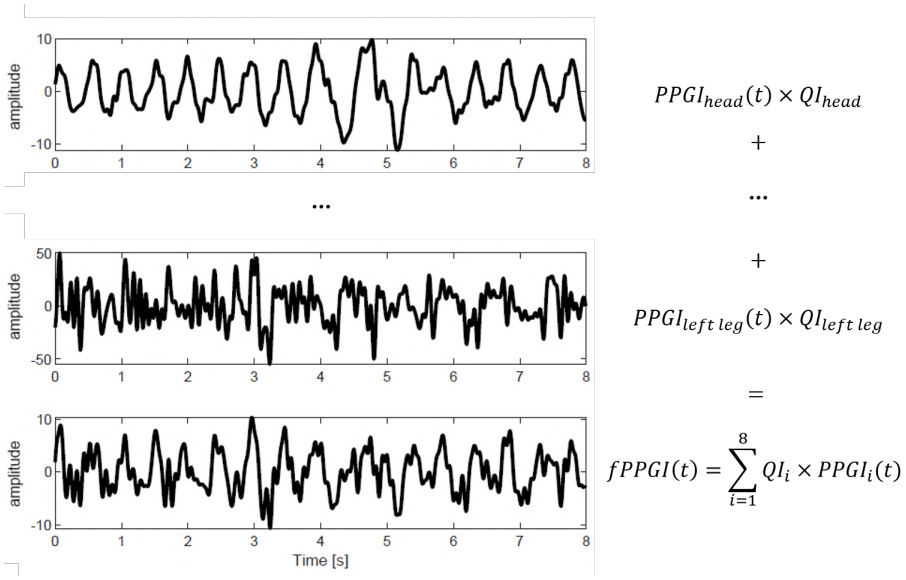


Figure 7.6: From top to bottom: filtered PPGI signals from the head and left leg. At the bottom: PPGI's weighted average. Note that the PPGI's weighted average reveals to be worse than the PPGI signal from the head region. The latter can be attributed to the fact that the PPGI signal from the legs mainly contain noise.

The normal heart rate of a neonate ranges from 1.7Hz to 2.8Hz or from 100 to 170 beats per minute. Therefore, the extracted PPGI signals are preprocessed by applying a 2^{th} order Butterworth bandpass filter with lower and upper cutoff frequency of 1.5Hz and 5Hz, respectively. Note that the upper cutoff frequency corresponds to 300bpm which is considerably higher than the physiological HR range of the neonates. The wide passband of the aforementioned filter allows the preservation of the PPGI waveform shape maintaining possible systolic and diastolic peaks.

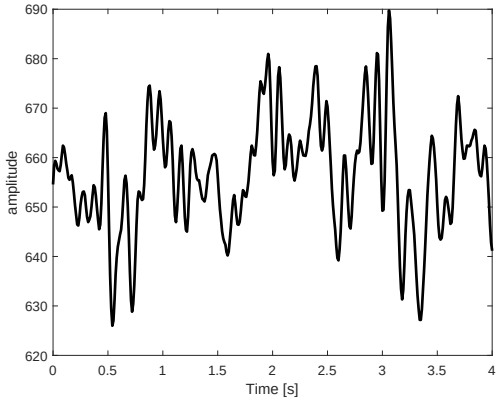


Figure 7.7: Raw PPGI signal extracted from the torso region of the neonate S009_S012 where PPGI fluctuations due to breathing-synchronous motion is visible.

7.1.2.4 HR computation

The extraction of the HR from the PPGI signal relies on the analysis of the signal's time-frequency plane. To this end, the wavelet synchrosqueezed transform is employed using the analytic Morlet wavelet. From the time-frequency plane, the maximum-energy time-frequency ridge is computed providing the HR estimation. This method computes a HR value for each signal sample.

7.1.2.5 Implementation details

All the steps necessary for the HR estimation are implemented in MATLAB (MATLAB 2019a, The MathWorks, Natick, 2019) and are summarized in the Appendix E.2.2.

The raw extracted $\text{PPGI}_i(t)$ signals are resampled to a sampling frequency of 100 and are then bandpass filtered relying on a 12^{th} order Butterworth filter with cutoff frequencies set to 1Hz and 20Hz. The filtered signals, $\hat{\text{PPGI}}_i(t)$, will be used to compute a sequence of quality indexes. A quality index is computed for each $\hat{\text{PPGI}}_i(t)$ at each second, using a eight seconds window, with a seven seconds overlap.

The quality index, QI_i, is computed relying on Equation 7.2 and following the description in Section 7.1.2.1. After computing the quality indexes of the eight ROIs, the $\hat{\text{PPGI}}_i(t)$ signals are combined using the weighted average (see Equation 7.6), forming the fused PPGI, $f\text{PPGI}(t)$. Then, a new quality index is computed from the eight seconds $f\text{PPGI}(t)$.

The PPGI signal that yields the greater quality index amongst the $\hat{PPGI}_i(t)$ and the $f\text{PPGI}(t)$ will be used for HR estimation. The raw $\text{PPGI}_i(t)$ signals are filtered, $\hat{\text{PPGI}}_i^2(t)$, following the specifications reported in Section 7.1.2.3. To extract the HR, the wavelet synchrosqueezed transform is computed for the selected PPGI signal over the eight seconds window. The HR estimation, for each signal sample, corresponds to the frequency with maximum energy in the estimated spectrum.

Once HR sequence is computed, a median filter of 6 seconds was applied to the HR estimates to reduce artefacts introduced by noise and motion.

7.2 Results

In the current section, the performance of the proposed algorithm is described relying on five RGB recordings. The remaining RGB and IR recordings are excluded due to bad VGC settings or lack of reference HR.

The HR estimation performance was examined for two different scenarios. Firstly, the HR estimation performance is analysed assuming that only the PPGI signal from the whole head region is available. Then, the same analysis is performed using the nine computed PPGI signals and the methods described in Section 7.1.2. Additionally, for each scenario, an analysis where high intensity motion periods are excluded from evaluation is performed, using the automatic movement evaluation method from Section 6.2.1.1.

7.2.1 Performance Metrics

To perform a complete analysis of the HR estimation, three performance indices are considered: the RMSE, the MAE and the percentage of time in which there is an agreement between the reference HR and the HR estimate provided by the extracted PPGI signal, with a tolerance of ± 5 bpm. Within the latter percentage of time, the estimates are deemed to be accurate. Note that the HR of a neonate is typical higher than 100 bpm thus, an error of 5 bpm corresponds to a 5 % difference which is not clinically significant. During bradycardia episodes, the difference can rise to 7 %, considering a HR of 75 bpm.

7.2.2 Head

The performance results for each considered subject are listed in Table 7.1. When analysing the complete recording, the RMSE averaged 18 ± 10 bpm. In addition, the MAE was 10 ± 5 bpm and the mean percentage of accurate estimates was 71 %. If high motion intensity periods are excluded from the evaluation, the RMSE improves to 13 ± 7 bpm; the MAE improves to 6 ± 3 bpm and the mean percentage of accurate estimates improves to 78%.

Table 7.1: Performance of the proposed method for HR estimation relying on the PPGI extracted from the head. HMI stands for high motion intensity. The first columns of the RMSE, MAE and prediction accuracy columns are the results when the complete recording is considered, the second columns refer to the results when HMI periods are excluded.

Patient ID	**Measurement**	**Recording time** [s]	**HMI time** [s]	**RMSE** [bpm]		**MAE** [bpm]		**Prediction accuracy** % of time	
S009_S009	1	587	105	31	18	17	8	61	67
S009_S010	1	587	55	17	16	8	7	68	70
S009_S012	1	587	21	3	2	2	1	97	98
S009_S014	1	587	522	19	18	12	8	59	77
S009_S016	2	587	79	18	12	9	5	70	78
Mean				18	13	10	6	71	78
SD				10	7	5	3	15	12

Particularly, the recording of the neonate S009_S012 originated a PPGI signal that provides a HR estimation that is perfectly correlated with the reference HR, having an agreement with the reference HR 97% of the recording time. Figure 7.8 depicts a Bland-Altman plot which compares the proposed method and the gold standard method for HR estimation, for subject S009_S012. According to the results, the estimated mean difference was 0.76 bpm and the limits of agreement ranged from -5.2 bpm to 6.7 bpm. Note that the majority of the outliers are coloured in red, meaning that they correspond to HR estimates during high motion intensity periods.

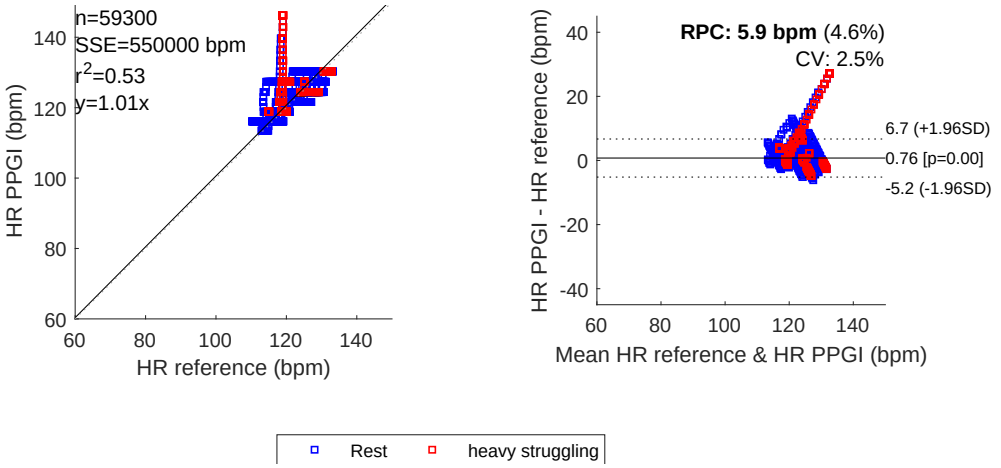


Figure 7.8: Bland-Altman plot comparing the proposed method (HR PPGI) and the gold standard method ($HR_{\text{reference}}$) for HR estimation. The plot comprises the results from the subject S009_S012. The graph shows a bias of 0.76 and the 95% limits of agreement range from -5.2 bpm to 6.7 bpm.

Lastly, Figure 7.9 shows the HR estimated using the PPGI signal extracted from the head region (blue line) as well as the reference HR (yellow line), and the respective time-resolved motion intensity profile. Both signals are extracted from the recording of the subject S009_S012. When analysing the figure, it is evident that the major discrepancy between the HR estimated by the proposed method and the reference, at approximately t = 430 s, coincides with a high motion intensity period. The latter visually empathises the improvements in the HR estimation performance indices when excluding high motion intensity intervals from the evaluation.

7.2.3 Multiple Regions of Interest

The performance results for each considered subject are listed in Appendix E.3.1. The comparison between both monitoring modalities for the complete recording showed a RMSE of 16 ± 9 bpm. In addition, the MAE was 10 ± 5 bpm and the mean percentage of accurate estimates was 62 %. If high motion intensity periods are excluded from the evaluation, the RMSE improves to 13 ± 7 bpm; the MAE improves to 8 ± 5 bpm and the mean percentage of accurate estimates improves to 68%.

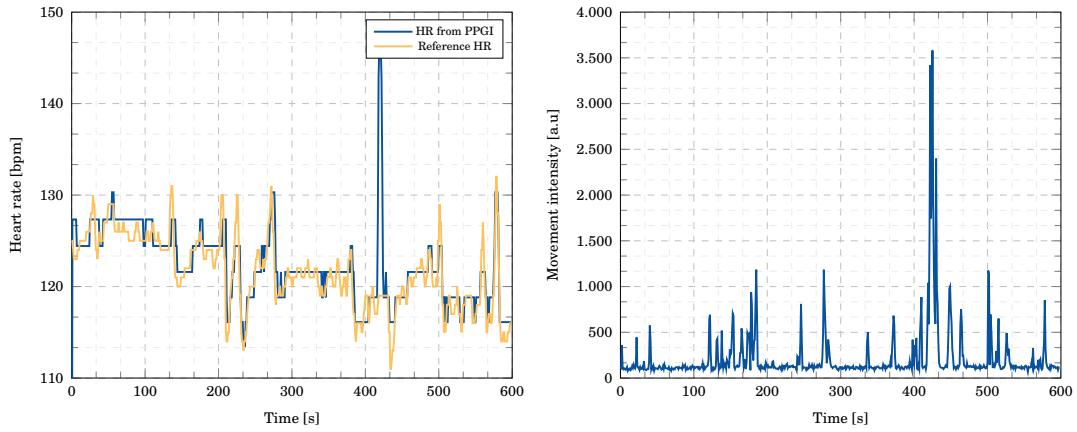


Figure 7.9: Left graph: illustrative example showing the HR obtained from the head's PPGI signal (blue line) and the reference HR (yellow line). Right graph: Movement intensity *vs*. time. The signals correspond to subject S009_S012.

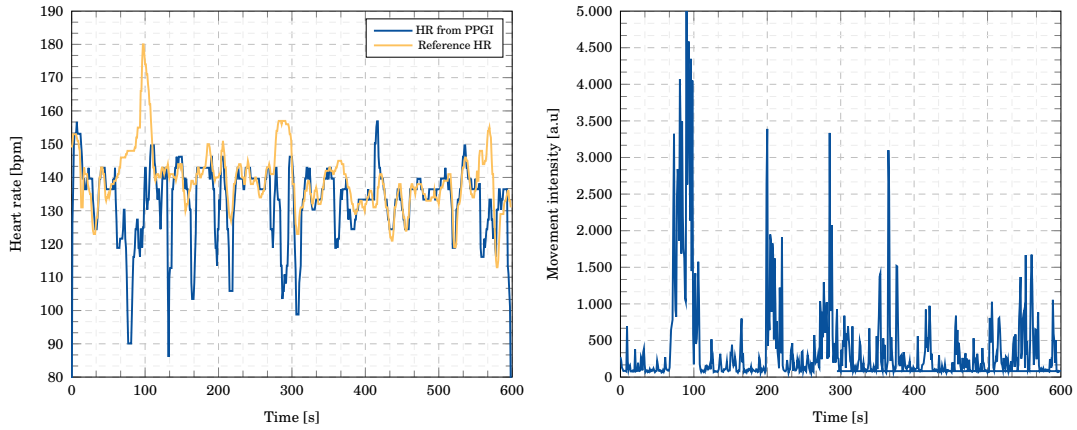


Figure 7.10: Left graph: illustrative example showing the HR obtained with the proposed method (blue line) and the reference HR (yellow line). Right graph: Movement intensity *vs*. time. The signals correspond to subject S009_S016.

Figure 7.10 shows the HR estimated with the proposed method (blue line) as well as the reference HR (yellow line), and the respective time-resolved motion intensity profile. Both signals are extracted from the recording of the subject S009_S016. Similarly to Figure 7.9, it is evident that the major discrepancies between the HR estimated by the proposed method and the reference coincides with a high motion intensity period.

7.2.4 Performance for Different Skin Tones

Darker skin usually poses a challenge to HR estimation due to an expected decrease in the PPGI signal strength [1], when compared with the PPGI signals extracted from lighter skin tones. This decrease results from the decrease in the amount of light that reaches the pulsatile vessels due to higher light absorption by the epidermal melanin, that is present in a higher concentration in subjects with a darker skin tone.

From the evaluated subjects, three had dark skin and the remaining two had light brown skin. Particularly, Figure 7.11 displays the filtered PPGI signals extracted from neonate S009_S009 and neonate S009_S012. The first has a dark skin and the second has a light brown skin. Despite the signals' amplitude difference, the PPGI signal extracted from the infant with a darker skin tone still reveals a high correlation with the HR. Thus, there is an indication that the extraction of the HR relying on the proposed method still works well with subjects with darker skin, when using the appropriate light conditions.

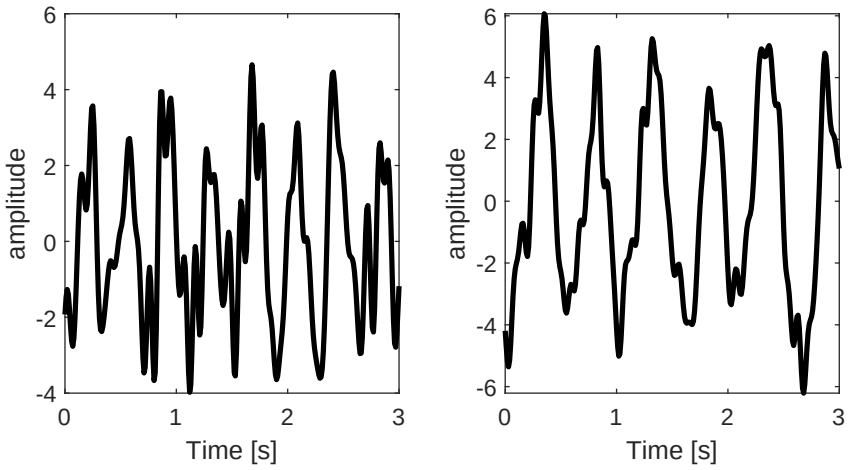


Figure 7.11: Left: PPGI signal extracted from neonate S009_S009 which has a dark skin tone. Right: PPGI signal extracted from neonate S009_S012 which has a light brown skin tone. Both signals are extracted from the head region.

7.3 Discussion

The findings presented in the current chapter validate the proposed FCNN model for ROI selection by demonstrating that it is possible to accurately and continuously monitor the HR of neonates relying on the automatically selected ROIs from RGB recordings. The proposed method also demonstrates the feasibility of continuous non-contact HR monitoring in a NICU environment without interfering in the patient clinical routines, using a video camera. The method is based on the fact that the cyclical blood volume fluctuations causes periodic changes in the light reflected from the skin. Note that, no conclusions upon IR recordings can be made.

To test the viability of the proposed method for ROI selection, a pilot study involving five neonates was conducted. The considered RGB recordings include periods of high motion intensity as well as periods of patient care routine, where there is an interaction between the neonate and the clinical staff. The HR assessment was tested for two scenarios. A scenario where only the PPGI signal from the head region is used, and a scenario where the nine PPGI signals and the method described in Section 7.1.2 are used. For both scenarios, outstanding results were obtained. Both Table 7.1 and Table E.2 report an excellent agreement between the HR estimated using the proposed method and the reference HR. However, for the first scenario, best agreements between the HR estimations and the reference HR were obtained for all considered neonates when compared with the results of the second scenario. The latter indicates that including multiple measuring sites, besides the head region, do not benefits the HR assessment and that the head region contains the strongest PPGI signal. Additionally, the employed quality index reveals to be a good metric for PPGI signal quality assessment, given the small discrepancy between the results of the first and second scenario.

Regarding the HR estimation scenario, the best results were achieved for neonate S009_S012. Figure 7.8 shows the Bland-Altman plot for the aforementioned neonate RGB recording HR measurements. The plot demonstrates not only a small bias (0.76 bpm) but also the high precision that can be achieved by this HR monitoring technique. The latter reinforces again an excellent agreement between the studied monitoring modalities.

As previously mentioned, the employed signal fusion method was based on the work of Kumar et al. [47]. Kumar et al. fused PPGI signals from small ROIs placed on the head region. On the other hand, in the present method, the fused PPGI signals are extracted from multiple body parts and, despite the negligible delay, the originated fused PPGI signal reveals a low signal quality when compared to the signal quality of the PPGI signal extracted from the head region. Therefore, fusing PPGI signals from multiple body parts do not benefits the HR assessment.

The addition of the two extra ROIs by further dividing the $\mathrm{ROI}_{\mathrm{head}}$ did not improve the HR estimation performance. Instead, the PPGI signals extracted from the extra ROIs contained more artefacts and the PPGI component modulated by the arterial blood volume fluctuations was almost extinguished. Thus, a better PPGI signal is obtained by averaging all the pixels belonging to $\mathrm{ROI}_{\mathrm{head}}$, which leads to the cancellation of the asynchronous

reflection component changes across the pixels belonging to ROI_{head}, and to the addiction of the synchronous reflection component changes across the pixels originated from the blood volume fluctuations.

The low lightning conditions in the NICU coupled with the frequent interactions of the clinical staff, creating pronounced shadows, negatively affects the PPGI signal. Additionally, despite the FCNN's ability to successfully track the ROI across the frames during high motion intensity movement periods, the neonate's movement where the orientation of the tracked skin with respect to the illumination sources changes (e.g. rotational movement) causes the average intensity value to change dramatically due to changes in the dominant light source that incidents the tracked skin region. As expected, significant improvements in the HR estimation metrics are verified when high motion intensity periods are excluded from the evaluation (see Table 7.1 and Table E.2). To address the challenge of accurate HR extraction during high motion intensity periods, a more robust algorithm for HR extraction needs to be developed combining, more efficiently, the PPGI signals from the static ROI. Note that standard contact based monitoring techniques, such as PPG or ECG, also contain signal artefacts during high motion intensity movement periods [46] and they are still the primary method to access the HR in the NICU.

In general, this chapter demonstrates that the HR assessment is possible for RGB recordings relying on the ROIs proposed by the developed FCNN, as corroborated by the reported results. Thus proving that HR monitoring through video cameras might be a proper alternative to adhesive electrodes in a NICU environment.

Chapter Eight

Conclusions

Continuous monitoring of heart rate is fundamental in the routine care of a premature infant, since changes in the vital parameters are often observed prior to major complications. Despite the advances in neonatal monitoring, the standard methods for heart rate monitoring rely on adhesive electrodes and sensors that are attached to the skin. In addition to causing stress, these contact-based methods can damage the fragile skin of the premature causing pain and possible infections. Therefore, the development of a robust remote heart rate monitoring technique is a great contribute to neonatal monitoring. In this thesis, the heart rate assessment is based on the photoplethysmography imaging technique.

The goal of this master thesis was to develop an automatic method for region of interest selection in order to improve the robustness of heart rate estimation, under the difficult NICU scenario. In this context, novel region of interest selection algorithms were developed, whose feasibility was tested in two different datasets. In addition, signal processing methods for heart rate estimation were implemented. The performance of heart rate estimation was tested on a small set of subjects.

In Chapter 6 a Fully Convolutional Neural Network model was developed. Besides being computationally efficient, the segmentation results demonstrated that this approach is capable of accurately identify the six predefined body parts within the RGB or IR frame, even in challenging conditions. Particularly, the head yielded the best segmentation performance with 89 % of precision for RGB data and 79 % of precision for IR data. Pre-training the network on a general large dataset revealed to be extremely important to achieve a good segmentation performance. The latter is a major finding as it can be broadly applied in other networks' training in the field of medicine, where large datasets are scarce. In addition, a novel post-processing method was developed, aiming to eliminate body part instances besides those from the premature. This method further improved the segmentation performance (up to 40 % in some body parts) for recordings where the care taker or parents are present in the field of view. The combination of the latter tools provide an evolution from manual to automatic robust selection of the region of interest.

Chapter 7 presents two approaches for heart rate assessment through the measuring sites provided by the developed region of interest selection methods. For both approaches,

the heart rate prediction performance was evaluated on five RGB recordings. For IR recordings, the heart rate assessment performance was not evaluated. The first algorithm extracts the heart rate using exclusively the photoplethysmography imaging signal from the head region. In this single region of interest approach, simple signal processing methods were used. The second algorithm extends form the first by considering not only the photoplethysmography imaging signal from the head region, but also photoplethysmography imaging signal from the remaining five body parts. In this multiple region of interest approach the heart rate is extracted from the measuring site with the highest signal quality. Despite the fact that, in both approaches, there was a high agreement between the reference heart rate and the estimated heart, the single region of interest approach had a slightly better performance. Therefore, the head region, the body part which yields better segmentation performance, is the best to estimate the heart rate.

Despite the small number of subjects in the heart rate estimation study, the main goal of the experiments reported in Chapter 7 is to prove that the proposed method for automatic region of interest selection provides the necessary measuring sites for heart rate assessment. Additionally, the results prove the feasibility of photoplethysmography imaging, making it an alternative for monitoring the heart rate remotely in newborn intensive care units. The major challenge in such scenario is the constant body movements which lead to changes in the light that incidents on the successfully tracked skin leading to amplitude fluctuations on the photoplethysmography imaging signal. Thus, the recording settings have a major impact on the signal quality.

In the future, more subjects should be included in the study and new datasets should be created following the same experimental setup, meaning that the camera position would be constant throughout the datasets. Ideally, the camera should be placed from an angle from above, similarly to the setup adopted by Chaichulee et al. [14]. By improving the camera setup, illumination, and location of the neonate within the field of view the impact of the variability of illumination during motion should be diminished. Additionally, the next step would comprise the development of a robust and computationally efficient signal processing algorithm to assess the heart rate.

This thesis also plays a role in heart rate monitoring using IR recordings. Despite the impossibility to evaluate the heart rate estimation performance through this kind of recordings, the segmentation results revealed a great potential.

All in all, this thesis contributes with a robust algorithm for region of interest selection, bringing photoplethysmography imaging one step closer to be as accurate as standard heart rate monitoring methods. Thus, in the future, photoplethysmography imaging may play a major role in neonatal remote monitoring, being a good alternative to adhesive electrodes. In sum, the future incubator should be a cable-free incubator with an incorporated imaging system. Despite the good progress towards remote heart rate assessment in neonatology, there is still a long way to go.

Bibliography

[1] L. A. 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 — A pilot study." In: *Early Human Development* 89.12 (2013), pp. 943–948. URL: http://dx.doi.org/10.1016/j.earlhumdev.2013.09.016https://linkinghub.elsevier.com/retrieve/pii/S0378378213002375.

[2] C. C. Aggarwal. "An Introduction to Neural Networks." In: *Neural Networks and Deep Learning*. Cham: Springer International Publishing, 2018, pp. 1–52. URL: http://link.springer.com/10.1007/978-3-319-94463-0{_}1.

[3] C. C. Aggarwal. "Convolutional Neural Networks." In: *Neural Networks and Deep Learning*. Cham: Springer International Publishing, 2018, pp. 315–371. URL: http://link.springer.com/10.1007/978-3-319-94463-0{_}8.

[4] L. Antognoli, P. Marchionni, S. Nobile, V. Carnielli, and L. Scalise. "Assessment of cardio-respiratory rates by non-invasive measurement methods in hospitalized preterm neonates." In: *MeMeA 2018 - 2018 IEEE International Symposium on Medical Measurements and Applications, Proceedings* (2018).

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495. arXiv: 1511.00561v3. URL: http://www.ncbi.nlm.nih.gov/pubmed/28060704.

[6] M. Bartocci, L. L. Bergqvist, H. Lagercrantz, and K. J. S. Anand. "Pain activates cortical areas in the preterm newborn brain." In: *Pain* 122.1 (2006), pp. 109–117. URL: http://search.ebscohost.com/login.aspx?direct=true{\&}db=edselp{\&}AN=S0304395906000431{\&}site=eds-live.

[7] R. E. Behrman and A. S. Butler. *Preterm birth: causes, consequences and prevention*. 1st ed. National Academies Press (US), 2007.

[8] S. Bianco, R. Cadene, L. Celona, and P. Napoletano. "Benchmark Analysis of Representative Deep Neural Network Architectures." In: *CoRR* abs/1810.0 (2018). arXiv: 1810.00736v2. URL: https://github.com/CeLuigi/models-comparison.pytorch.

[9] N. Blanik, K. Heimann, C. Pereira, M. Paul, V. Blazek, B. Venema, T. Orlikowsky, and S. Leonhardt. "Remote vital parameter monitoring in neonatology - robust, unobtrusive heart rate detection in a realistic clinical scenario." In: *Biomedizinische Technik. Biomedical engineering* 61.6 (2016), pp. 631–643.

[10] V. Blazek, T. Wu, and D. Hoelscher. "Near-infrared CCD imaging: possibilities for non-invasive and contactless 2D mapping of dermal venous hemodynamics." In: *Optical Diagnostics of Biological Fluids V*. Ed. by A. V. Priezzhev and T. Asakura. 2000, pp. 2–9. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=917217.

[11] C. Brüser, C. H. Antink, T. Wartzek, M. Walter, and S. Leonhardt. "Ambient and unobtrusive cardiorespiratory monitoring techniques." In: *IEEE Reviews in Biomedical Engineering* 8 (2015), pp. 30–43.

[12] A. Canziani, E. Culurciello, and A. Paszke. "An analysis of deep neural network models for practical applications." In: *CoRR* abs/1605.0 (2016). arXiv: 1605.07678v4. URL: https://arxiv.org/pdf/1605.07678.pdf.

[13] Ce Liu, J. Yuen, and A. Torralba. "Nonparametric Scene Parsing via Label Transfer." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.12 (2011), pp. 2368–2382. URL: http://ieeexplore.ieee.org/document/5936073/.

[14] 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)*. IEEE, 2017, pp. 266–272. URL: http://ieeexplore.ieee.org/document/7961751/.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs." In: *CoRR* abs/1412.7 (2014). arXiv: 1412.7062. URL: https://arxiv.org/pdf/1412.7062.pdfhttp://arxiv.org/abs/1412.7062.

[16] L.-C. Chen, G. Papandreou, S. Member, I. Kokkinos, K. Murphy, and A. L. Yuille. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." In: *CoRR* abs/1606.0 (2016). arXiv: 1606.00915v2. URL: http://liangchiehchen.com/projects/.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. "Rethinking Atrous Convolution for Semantic Image Segmentation." In: *CoRR* abs/1706.0 (2017). arXiv: 1706.05587v3. URL: https://arxiv.org/pdf/1706.05587.pdf.

[18] J. C. Cobos-Torres, M. Abderrahim, and J. Martínez-Orgado. "Non-contact, simple neonatal monitoring by photoplethysmography." In: *Sensors (Switzerland)* 18.12 (2018).

[19] L. F. Corral Martinez, G. Paez, and M. Strojnik. "Optimal wavelength selection for non-contact reflection photoplethysmography." In: ed. by R. Rodríguez-Vera and R. Díaz-Uribe. Vol. 8011. International Society for Optics and Photonics, 2011, p. 801191. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.903190.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. *ImageNet: A Large-Scale Hierarchical Image Database*. 2009, pp. 248–255. URL: http://www.image-net.org..

[21] *Derivative of the Sigmoid function - Towards Data Science*. URL: https://towardsdatascience.com/derivative-of-the-sigmoid-function-536880cf918e (visited on 12/23/2019).

[22] S. Devine and G. Taylor. *EVERY CHILD ALIVE*. Tech. rep. United Nations Children's Fund, 2018.

[23] M. Elliott and A. Coventry. "Critical care: the eight vital signs of patient monitoring." In: *British journal of nursing (Mark Allen Publishing)* 21.10 (), pp. 621–5. URL: http://www.ncbi.nlm.nih.gov/pubmed/22875303.

[24] M. Everingham, L. Van Gool, C. K. I Williams, J. Winn, A. Zisserman, M Everingham, L. K. Van Gool Leuven, B. CKI Williams, J Winn, and A Zisserman. "The PASCAL Visual Object Classes (VOC) Challenge." In: *Int J Comput Vis* 88 (2010), pp. 303–338. URL: http://www.flickr.com/.

[25] A. Fakhry, H. Peng, and S. Ji. "Deep models for brain EM image segmentation: novel insights and improved performance." In: *Bioinformatics* 32.15 (2016), pp. 2352–2358. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw165.

[26] L. Fei-Fei. *Stanford university computer science class cs231n: Convolutional neural networks for visual recognition.* 2017. URL: http://cs231n.github.io/convolutional-networks/.

[27] M. Fernandez, K. Burns, B. Calhoun, S. George, B. Martin, and C. Weaver. "Evaluation of a new pulse oximeter sensor." In: *American journal of critical care : an official publication, American Association of Critical-Care Nurses* 16.2 (2007), pp. 146–52. URL: http://www.ncbi.nlm.nih.gov/pubmed/17322015.

[28] *File:Wiggers Diagram.svg - Wikimedia Commons*. URL: https://commons.wikimedia.org/wiki/File:Wiggers{_}Diagram.svg (visited on 12/23/2019).

[29] Global Health Organization. *Preterm birth*. 2018. URL: https://www.who.int/news-room/fact-sheets/detail/preterm-birth (visited on 08/06/2019).

[30] X. Glorot, A. Bordes, and Y. Bengio. "Deep Sparse Rectifier Neural Networks." In: *Journal of Machine Learning Research* 15 (2010). URL: https://www.utc.fr/{~}bordesan/dokuwiki/{_}media/en/glorot10nipsworkshop.pdf.

[31] G. Green, S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, A. Zisserman, L. Tarassenko, and K. McCormick. "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*. Ed. by G. L. Coté. SPIE, 2018, p. 26. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10501/2289759/Localised-photoplethysmography-imaging-for-heart-rate-estimation-of-pre-term/10.1117/12.2289759.full.

[32] A. C. Guyton and J. E. Hall. *Textbook of Medical Physiology*. 11th. Elsevier Health Sciences, 2006.

[33] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. "Simultaneous Detection and Segmentation." In: *CoRR* abs/1407.1 (2014). arXiv: 1407.1808v1. URL: http://www.eecs.berkeley.edu/Research/Projects/CS/vision/shape/sds..

[34] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." In: *CoRR* abs/1512.0 (2015). arXiv: 1512.03385. URL: http://image-net.org/challenges/LSVRC/2015/http://arxiv.org/abs/1512.03385.

[35] K. He, X. Zhang, S. Ren, and J. Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015). arXiv: 1502.01852v1. URL: https://arxiv.org/pdf/1502.01852.pdf.

[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick. "Mask R-CNN." In: *CoRR* abs/1703.0 (2017). arXiv: 1703.06870v3. URL: http://arxiv.org/abs/1703.06870.

[37] A. B. Hertzman. "The blood supply of various skin areas as estimated by photoelectric plethysmograph." In: *American Journal of Physiology-Legacy Content* 124.2 (1938), pp. 328–340. URL: http://www.physiology.org/doi/10.1152/ajplegacy.1938.124.2.328.

[38] M. Huelsbusch and V. Blazek. "Contactless mapping of rhythmical phenomena in tissue perfusion using PPGI." In: ed. by A. V. Clough and C.-T. Chen. Vol. 4683. International Society for Optics and Photonics, 2002, p. 110. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.463573.

[39] R. M. Insoft and I. D. Todres. "Growth and Development." In: *A Practice of Anesthesia for Infants and Children*. 6th. Elsevier, 2009. Chap. 2, pp. 7–24. URL: http://www.crossref.org/deleted{_}DOI.html.

[40] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." In: *CoRR* abs/1502.0 (2015). arXiv: 1502.03167v3. URL: https://arxiv.org/pdf/1502.03167.pdf.

[41] M. A. Islam, M. Rochan, S. Naha, N. D. B. Bruce, and Y. Wang. "Gated Feedback Refinement Network for Coarse-to-Fine Dense Semantic Image Labeling." In: *CoRR* abs/1806.1 (2018), pp. 1–14. arXiv: 1806.11266. URL: http://arxiv.org/abs/1806.11266.

[42] 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: *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge* (2017), pp. 286–293.

[43] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization." In: *International Conference on Learning Representations* (2014). arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980.

[44] R. E. Klabunde. *Cardiovascular Physiology Concepts*. 2nd. Lippincott Williams & Wilkins, 2012.

[45] P. Krähenbühl and V. Koltun. "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials." In: *CoRR* abs/1210.5 (2012). arXiv: 1210.5644. URL: http://arxiv.org/abs/1210.5644.

[46] J. Kranjec, S. Begus, J. Drnovsek, and G. Gersak. "Novel methods for noncontact heart rate measurement: A feasibility study." In: *IEEE Transactions on Instrumentation and Measurement* 63.4 (2014), pp. 838–847.

[47] M. Kumar, A. Veeraraghavan, and A. Sabharval. "DistancePPG: Robust non-contact vital signs monitoring using a camera." In: 6.5 (2015), pp. 200–215. arXiv: 1502.08040. URL: http://arxiv.org/abs/1502.08040.

[48] H. Le and A. Borji. "What are the Receptive, Effective Receptive, and Projective Fields of Neurons in Convolutional Neural Networks?" In: *CoRR* abs/1705.0 (2017). arXiv: 1705.07049. URL: http://www.scholarpedia.org/article/Projectivehttp://arxiv.org/abs/1705.07049.

[49] G. Lempe, S. Zaunseder, T. Wirthgen, S. Zipser, and H. Malberg. "ROI Selection for Remote Photoplethysmography." In: *Bildverarbeitung f{\"u}r die Medizin 2013*. Ed. by H.-P. Meinzer, T. M. Deserno, H. Handels, and T. Tolxdorff. Springer, Berlin, Heidelberg, 2013, pp. 99–103. URL: http://link.springer.com/10.1007/978-3-642-36480-8{_}19.

[50] X. Li, S. Chen, X. Hu, and J. Yang. "Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift." In: *CoRR* abs/1801.0 (2018). arXiv: 1801.05134v1. URL: https://arxiv.org/pdf/1801.05134.pdf.

[51] G. Lin, C. Shen, A. van dan Hengel, and I. Reid. "Efficient piecewise training of deep structured models for semantic segmentation." In: *CoRR* abs/1504.0 (2015). arXiv: 1504.01013. URL: http://arxiv.org/abs/1504.01013.

[52] G. Lin, A. Milan, C. Shen, and I. Reid. "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation." In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua (2017), pp. 5168–5177. arXiv: arXiv:1611.06612v3.

[53] Z. Liu, X. Li, P. Luo, C. Change, and L. X. Tang. "Semantic Image Segmentation via Deep Parsing Network *." In: *CoRR* abs/1509.0 (2015). arXiv: 1509.02634v2. URL: https://arxiv.org/pdf/1509.02634.pdf.

[54] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 3431–3440. URL: https://people.eecs.berkeley.edu/{~}jonlong/long{_}shelhamer{_}fcn.pdfhttp://ieeexplore.ieee.org/document/7298965/.

[55] S. K. Longmore, G. Y. Lui, G. Naik, P. P. Breen, B. Jalaludin, and G. D. Gargiulo. "A Comparison of Reflective Photoplethysmography for Detection of Heart Rate, Blood Oxygen Saturation, and Respiration Rate at Various Anatomical Locations." In: *Sensors* 19.8 (2019), p. 1874. URL: https://www.mdpi.com/1424-8220/19/8/1874.

[56] W. Luo, Y. Li, R. Urtasun, and R. Zemel. "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks." In: *CoRR* abs/1701.0 (2017). arXiv: 1701.04128. URL: https://arxiv.org/pdf/1701.04128.pdfhttp://arxiv.org/abs/1701.04128.

[57] A. V. Moço, S. Stuijk, and G. de Haan. "Skin inhomogeneity as a source of error in remote PPG-imaging." In: *Biomedical optics express* 7.11 (2016), pp. 4718–4733. URL: http://www.ncbi.nlm.nih.gov/pubmed/27896011http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5119611.

[58] J. Moini. "Anatomy and Physiology of the Cardiovascular System." In: *Phlebotomy: Principles and Practice*. 1st. 2013. Chap. 5.

[59] H. Noh, S. Hong, and B. Han. "Learning Deconvolution Network for Semantic Segmentation." In: *CoRR* abs/1505.0 (2015). arXiv: 1505.04366. URL: https://arxiv.org/pdf/1505.04366.pdfhttp://arxiv.org/abs/1505.04366.

[60] G. L. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox. "Deep learning for human part discovery in images." In: *Proceedings - IEEE International Conference on Robotics and Automation* 2016-June (2016), pp. 1634–1641.

[61] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. D. Facebook, A. I. Research, Z. Lin, A. Desmaison, L. Antiga, O. Srl, and A. Lerer. *Automatic differentiation in PyTorch*. Tech. rep. 2017. URL: https://openreview.net/pdf?id=BJJsrmfCZ.

[62] J. Patterson and A. Gibson. *Deep Learning. A Practitioner's Approach*. Ed. by M. Loukides and T. McGovern. First edit. O'Reilly, 2017. URL: http://oreilly.com/safari.

[63] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. "Large Kernel Matters – Improve Semantic Segmentation by Global Convolutional Network." In: *CoRR* abs/1703.0 (2017). arXiv: 1703.02719. URL: https://arxiv.org/pdf/1703.02719.pdfhttp://arxiv.org/abs/1703.02719.

[64] M.-Z. Poh, D. J. Mcduff, and R. W. Picard. "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam." In: *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING* 58.1 (2011). URL: http://ieeexplore.ieee.org..

[65] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe. "Full-resolution residual networks for semantic segmentation in street scenes." In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua (2017), pp. 3309–3318. arXiv: arXiv:1611.08323v2.

[66] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." In: *CoRR* abs/1506.0 (2016). arXiv: 1506.01497v3. URL: http://image-net.org/challenges/LSVRC/2015/results.

[67] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252. URL: http://link.springer.com/10.1007/s11263-015-0816-y.

[68] L. Scalise, N. Bernacchia, I. Ercoli, and P. Marchionni. "Heart rate measurement in neonatal patients using a webcamera." In: *MeMeA 2012 - 2012 IEEE Symposium on Medical Measurements and Applications, Proceedings* (2012), pp. 6–9.

[69] T. Shwayder, Tor, Akland. "Neonatal skin barrier : structure , function , and disorders." In: *Dermatologic Therapy* 18 (2005), pp. 87–103.

[70] A. Sikdar, S. K. Behera, D. P. Dogra, and H. Bhaskar. "Contactless vision-based pulse rate detection of Infants under Neurological Examinations." In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* 2015-Novem (2015), pp. 650–653.

[71] C. Silva. *Nascem mais bebés prematuros e de mães mais velhas*. 2018. URL: https://www.jn.pt/nacional/interior/nascem-mais-prematuros-e-de-maes-mais-velhas-10191208.html.

[72] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In: *ICLR* (2014). arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556.

[73] I. Snuverink. "Deep Learning for Pixelwise Classification of Hyperspectral Images." Doctoral dissertation. Delft University of Technology, 2017.

[74] 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." In: *Institute of Physics and Engineering in Medicine Physiological Measurement Physiol. Meas* 35 (2014), p. 807. URL: http://iopscience.iop.org/0967-3334/35/5/807.

[75] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. "Efficient Object Localization Using Convolutional Networks." In: *CoRR* abs/1411.4 (2014). arXiv: 1411.4280. URL: http://arxiv.org/abs/1411.4280.

[76] L. Torrey and J. Shavlik. "Transfer Learning." In: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. 2010. Chap. Transfer L, pp. 242–264. URL: www.igi-global.com/chapter/perturbation-size-independent-analysis-.

[77] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. "Remote plethysmographic imaging using ambient light." In: *Optics express* 16.26 (2008), pp. 21434–45. URL: http://www.ncbi.nlm.nih.gov/pubmed/19104573http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2717852.

[78] M. Villarroel, A. Guazzi, S. Davis, P. Watkinson, A. Guazzi, K. McCormick, L. Tarassenko, J. Jorge, M. Villarroel, A. Shenvi, and G. Green. "Continuous non-contact vital sign monitoring in neonatal intensive care unit." In: *Healthcare Technology Letters* 1.3 (2014), pp. 87–91. URL: https://digital-library.theiet.org/content/journals/10.1049/htl.2014.0077.

[79] J. Wright Lott. "Cardiovascular System." In: *Comprehensive Neonatal Care: An Interdisciplinary Approach*. Elsevier Health Sciences, 2007. Chap. 3rd.

[80] T. Wu, V. Blazek, and H. J. Schmitt. "Photoplethysmography imaging: a new non-invasive and noncontact method for mapping of the dermal perfusion changes." In: *Optical Techniques and Instrumentation for the Measurement of Blood Composition, Structure, and Dynamics*. Ed. by A. V. Priezzhev and P. A. Oberg. 2000, p. 62. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.407646.

[81] 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." In: *CoRR* abs/1502.0 (2015). URL: https://www.robots.ox.ac.uk/{~}szheng/papers/CRFasRNN.pdf.

Appendix A

Medical Foundations of the thesis

A.1 Visible and Near-infrared Spectrum

The contrast between the reflection coefficient of blood and bloodless tissue attributes mainly to the fact that all forms of the haemoglobin molecule present in the blood absorb light more strongly than the remaining tissues [74].

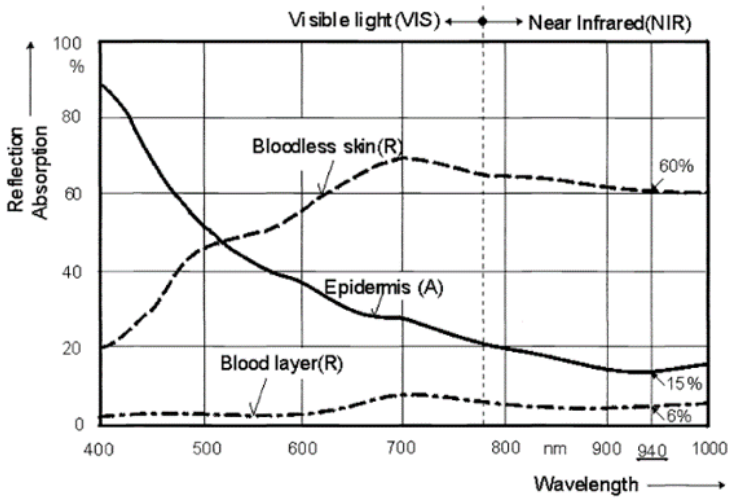


Figure A.1: Optical properties of the skin in visible and near-infrared spectrum [80].

Appendix B

Deep Learning

B.1 Rectified Linear Unit Function

The rectified linear unit function, or ReLU, is commonly used as the activation function nowadays. This activation function computes the function $f(x) = max(0, x)$, meaning that the activation is zero if $x < 0$ otherwise, the function is linear with slope 1. When compared with the Tanh and Sigmoid activation functions, the ReLU function accelerates the convergence of stochastic gradient descent. Additionally, it is more computationally efficient. However, this activation function is associated with the appearance of permanent inactivated units due to irreversible weights update. Note that this problem can be addressed with a low learning rate. [26]

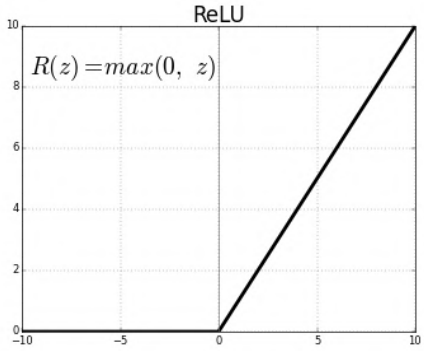


Figure B.1: Rectified Linear Unit Function (ReLU) activation function.

Appendix C

Clinical Study

C.1 Patient Data

Table C.1 presents the information of the 29 subjects enrolled in the study.

Table C.1: Patient information, including the patient ID, gender (M = Male, F = Female), gestational age and weight. The first number in the patient ID corresponds to the study (1 = Neonaten, 9 = Navpani)

Patient ID	**Gender**	**Gestational age** [weeks]	**Age at measurement time** [days]	**Weight at birth** [g]	**Weight at measurement time** [g]
S001_S002	F	35	17	2060	-
S001_S003	F	32	24	1290	-
S001_S004	M	32	24	1490	-
S001_S005	M	24	241	710	-
S001_S006	M	27	42	680	-
S001_S007	F	30	28	645	1145
S001_S008	M	30	8	1000	995
S001_S009	F	27	31	1125	1515
S001_S010	M	30	29	1100	1630
S009_S001	M	38	19	3300	3220
S009_S002	F	39	5	2790	2670
S009_S003	F	37	10	2460	2410
S009_S004	F	38	2	2620	2490
S009_S005	F	38	7	2500	2260
S009_S006	F	37	7	2908	3130
S009_S007	M	37	16	2790	2630
S009_S008	M	38	7	3374	3250
S009_S009	F	28	77	-	-
S009_S010	F	28	77	1020	2270
S009_S011	F	36	3	2300	2200
S009_S012	F	37	7	2646	2520
S009_S013	M	37	2	2020	1835
S009_S014	M	39	3	3004	2960
S009_S015	F	39	4	2680	2060
S009_S016	M	26	63	765	1720
S009_S017	F	30	29	1150	1320
S009_S018	F	39	4	2200	2230
S009_S019	M	31	1	2005	2000
S009_S020	F	29	1	2990	2800
Mean		33.4	27.2	1986.5	2228.7
SD		4.7	45.6	871.9	630.0

Appendix D

Region of Interest Selection

D.1 Neonaten-Navpani Dataset Distribution in Folds

Table D.1 reports the distribution of the Neonaten-Navpani dataset RGB and IR frames for each fold.

Table D.1: Summary of the dataset frames distribution in the five folds. M = Male, F = Female, W = White, B = Black, WB = Mixed White and Black, Su = Supine, P = Prone, Si = Side.

Fold	**nº of RGB dataset frames per skin colour**			**nº of IR dataset frames per skin colour**			**nº of RGB dataset frames per lying position**			**nº of IR dataset frames per lying position**		
	B	W	BW	B	W	BW	Su	P	Si	Su	P	Si
1	32	20	75	3	2	13	115	12	0	16	2	0
2	31	23	64	2	2	9	107	3	8	11	1	1
3	21	44	38	1	7	7	89	14	0	11	4	0
4	31	49	32	8	8	3	112	0	0	19	0	0
5	17	24	62	4	4	8	103	0	0	16	0	0
Total	132	160	271	18	23	40	526	29	8	73	7	1

D.2 Evaluation

D.2.1 Computational Complexity

Table D.2 reports the number of FMAs required for a single forward pass for an image with a resolution of 320×320.

The substantial increase of FMAs in the Oliveira et al. encoder-network with respect to the FMAs of the regular VGG-16 (31.51 GFMAs) derives from the padding increase, from one to 100, on the first convolutional layer of the network.

Table D.2: Total amount of FMAs and number of FMAs in the encoder and decoder network for each considered model. The reported values correspond to the number of FMAs required for a single forward pass for an image with a resolution of 320×320.

Method	Total nº of FMAs (GFMAs)	Nº of encoder FMAs (GFMAs)	Nº of decoder FMAs (GFMAs)
Encoder-decoder-bilinear	8.44	8.40	0.04
Encoder-decoder-unconnected	8.42	8.40	0.02
Encoder-decoder-dropout	8.44	8.40	0.04
Encoder-decoder-batchnorm	8.44	8.40	0.04
Encoder-decoder-Oliveira	129.17	129.11	0.06

D.2.2 PASCAL Human Parts Dataset and Freiburg Sitting RGB Dataset

Quantitative results of the proposed FCNN model and Oliveira el al. model

Table D.3 compares the models' class prediction performance on the PASCAL human parts validation dataset.

Table D.3: Quantitative results on the validation RGB PASCAL human parts dataset. For each method, the IoU for each class and the overall mean IoU and accuracy is reported.

	IoU (%)					
Method	head	torso	arms	legs	**Mean IoU (%)**	**Mean Accuracy (%)**
Encoder-decoder-batchnorm	65	55	18	36	44	53
Encoder-decoder-Oliveira	83	79	74	77	78	86

Learning efficiency

Learning efficiency is a metric proposed by [12] that reflects the capacity of a architecture to better utilise its parametric space. The learning efficiency reported in Figure D.1, corresponds to the ratio between the top accuracy and the number of parameters of the considered architecture. For all the considered models, the verified top accuracy corresponds to the accuracy for the head class.

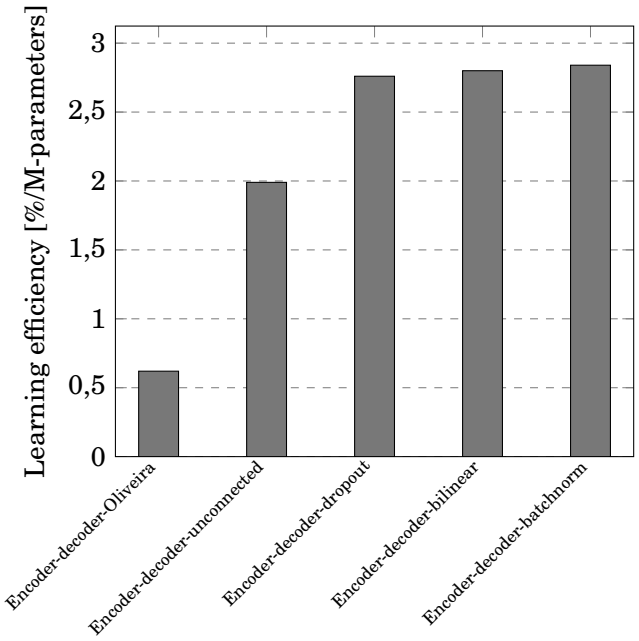


Figure D.1: Learning efficiency *vs*. model. The encoder-decoder-Oliveira detains the lower learning efficiency. On the other hand, the encoder-decoder-batchnorm has the higher learning efficiency, meaning that the model takes fully advantage of its small number of parameters.

D.2.3 Neonaten-Navpani-RGB Dataset

Table D.4 reports the class prediction performance of the proposed FCNN on the Neonaten-Navpani-RGB dataset for the five folds.

Table D.4: Quantitative results of the proposed CNN model (Encoder-decoder-batchnorm) on the Neonaten-Navpani-RGB dataset for the 5 folds. The training and validation images are downsampled by a factor of two in both dimensions (i.e. 576×960).

	IoU (%)						Accuracy (%)						Precision (%)								
Fold	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	**Mean IoU (%)**	**Mean Accuracy (%)**	**Mean Precision (%)**
1	84	39	52	35	68	47	87	45	68	57	75	60	96	50	66	49	89	67	54	65	70
2	85	36	55	35	59	37	92	42	69	52	68	44	92	64	65	40	78	61	51	61	67
3	69	24	62	26	62	38	72	40	72	38	78	45	80	35	75	44	72	47	47	58	59
4	77	46	29	41	41	22	82	54	40	66	46	28	89	66	43	49	63	32	43	53	57
5	82	59	47	40	48	55	91	65	61	60	54	66	89	73	57	47	60	73	55	66	67
Mean	80	41	49	35	56	40	85	49	62	55	64	49	89	58	61	46	72	56	50	61	64

D.2.4 PASCAL Human Parts Dataset and Freiburg Sitting Grey Scale Images Dataset

Table D.5 compares the class prediction performance of the proposed FCNN and the encoder-decoder-dropout model on the PASCAL human parts dataset and Freiburg sitting grey scale images dataset.

Table D.5: Quantitative results on the PASCAL human parts and Freiburg sitting validation grey images dataset. For each method, the IoU and accuracy for each class and the mean IoU and accuracy are reported. The Encoder-decoder-dropout, Oliveira et al. model, outperforms the proposed FCNN on the thinner body part classes (arms and legs).

	IoU (%)						Accuracy (%)							
Method	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	**Mean IoU (%)**	**Mean Accuracy (%)**
Encoder-decoder-dropout	61	50	31	30	43	44	67	63	37	36	46	50	43	60
Encoder-decoder-batchnorm	63	52	17	15	29	32	71	65	11	14	34	32	35	45

D.2.5 Neonaten-Navpani-IR Dataset

Neonaten-Navpani-IR datasets

Table D.6 compares the class prediction performance of the proposed FCNN model and the encoder-decoder-dropout model on both the Neonaten-Navpani-IR datasets. Similarly to the results reported in Section 6.6.2.2, the reported results are the combination of the models' validation results on each fold.

Table D.6: Quantitative results of the proposed FCNN model and the encoder-decoder-dropout model on both the Neonaten-Navpani-IR datasets for images of size 576×960. For the considered methods, the mean IoU, accuracy and precision across the five folds are reported for each class. Additional, the overall mean IoU, accuracy and precision are reported. The levels of image processing of the Neonaten-Navpani-IR datasets are represented by he second and third column.

				IoU (%)						Accuracy (%)						Precision (%)							
	Red Channel		Histogram manipulation																			**Mean IoU (%)**	**Mean Accuracy (%)**	**Mean Precision (%)**
Method	raw			head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg			
Encoder-decoder-dropout	✓			57	12	30	15	30	6	60	15	43	23	32	6	71	29	46	21	58	19	25	30	41
Encoder-decoder-dropout		✓		62	12	24	12	26	15	65	16	35	22	31	18	81	26	41	20	58	30	25	31	43
Encoder-decoder-batchnorm	✓			47	5	16	10	13	9	47	6	19	14	14	9	67	20	35	16	43	20	17	18	33
Encoder-decoder-batchnorm		✓		62	10	19	10	29	20	66	12	23	16	33	24	79	31	44	19	64	40	25	30	46

Model results on the Neonaten-Navpani-IR-manipulated dataset

Table D.7 reports the class prediction performance of the proposed FCNN model on the Neonaten-Navpani-IR-manipulated dataset for each fold.

Table D.7: Quantitative results of the proposed FCNN model on the Neonaten-Navpani-IR-manipulated datasets for images of size 576×960. The mean IoU, accuracy and precision are reported for each class. Additionally, the overall mean IoU, accuracy and precision are reported.

	IoU (%)						Accuracy (%)						Precision (%)								
Fold	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	**Mean IoU (%)**	**Mean Accuracy (%)**	**Mean Precision (%)**
1	75	19	19	4	43	19	76	22	23	12	45	24	92	36	42	6	72	42	30	34	48
2	63	13	11	24	32	19	71	17	17	34	34	21	81	22	32	34	73	30	27	32	45
3	50	0	27	7	28	9	52	0	34	13	38	9	56	0	52	18	61	33	20	24	36
4	55	3	9	10	18	6	59	3	10	12	21	11	76	41	54	18	45	22	17	19	43
5	68	17	25	6	23	46	70	18	34	8	25	53	90	54	41	20	69	74	31	35	58
Mean	62	10	19	10	29	20	66	12	23	16	33	24	79	31	44	19	64	40	25	29	46

D.2.6 Impact of Transfer Learning

Table D.8: Quantitative results on the Neonaten-Navpani-RGB dataset. The reported IoU and accuracy for each class corresponds to the mean IoU and mean accuracy of the five folds for each class.

		IoU (%)						Accuracy (%)						Precision (%)								
Method	pre-training	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	**Mean IoU (%)**	**Mean Accuracy (%)**	**Mean Precision (%)**
Encoder-decoder-batchnorm		68	13	8	1	15	8	84	32	11	1	20	14	75	19	32	19	39	23	19	27	71
Encoder-decoder-batchnorm	✓	80	41	49	36	56	40	85	49	62	55	64	49	89	58	61	46	72	56	50	61	64

D.2.7 Effect of Data Augmentation

Table D.9: Quantitative results on the Neonaten-Navpani-RGB dataset. The reported IoU and accuracy for each class corresponds to the mean IoU and mean accuracy of the five folds for each class.

		IoU (%)						Accuracy (%)						Precision (%)								
Method	data augmentation	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	**Mean IoU (%)**	**Mean Accuracy (%)**	**Mean Precision (%)**
Encoder-decoder-batchnorm		82	38	58	38	62	49	88	44	68	45	72	57	91	60	72	58	73	66	55	62	70
Encoder-decoder-batchnorm	✓	80	41	49	36	56	40	85	49	62	55	64	49	89	58	61	46	72	56	50	61	64

D.2.8 Receptive Field Analysis

In table D.10, the results of the proposed FCNN on the Neonaten-Navpani-RGB dataset using different input image resolutions are displayed.

Table D.10: Quantitative results on the Neonaten-Navpani-RGB dataset when changing the input image resolution. The reported results for each class corresponds to the mean IoU, mean accuracy and mean precision of the five folds for each class.

		IoU (%)						Accuracy (%)						Precision (%)								
Method	input resolution	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	head	torso	right arm	left arm	right leg	left leg	**Mean IoU (%)**	**Mean Accuracy (%)**	**Mean Precision (%)**
Encoder-decoder-batchnorm	288 × 480	73	35	46	29	52	38	78	45	53	50	60	45	86	49	64	36	65	54	46	55	59
Encoder-decoder-batchnorm	576 × 960	80	41	49	35	56	40	85	49	62	55	64	49	89	58	61	46	72	56	50	61	64
Encoder-decoder-batchnorm	1184 × 1920	71	24	31	31	39	31	77	33	39	42	46	39	85	49	56	45	62	49	38	46	58

D.2.9 Refinement Algorithm

The box plot of Figure D.2 compares the precision obtained before and after the application of the refinement algorithm. The statistical analysis has demonstrated that the quartiles for the head, torso, right and left leg classes do not change substantially before and after the refinement algorithm application. However, for the right and left arm, the statistical analysis revealed a generalized increase of the quartiles' values for the results obtained after the refinement algorithm when compared with those obtained before.

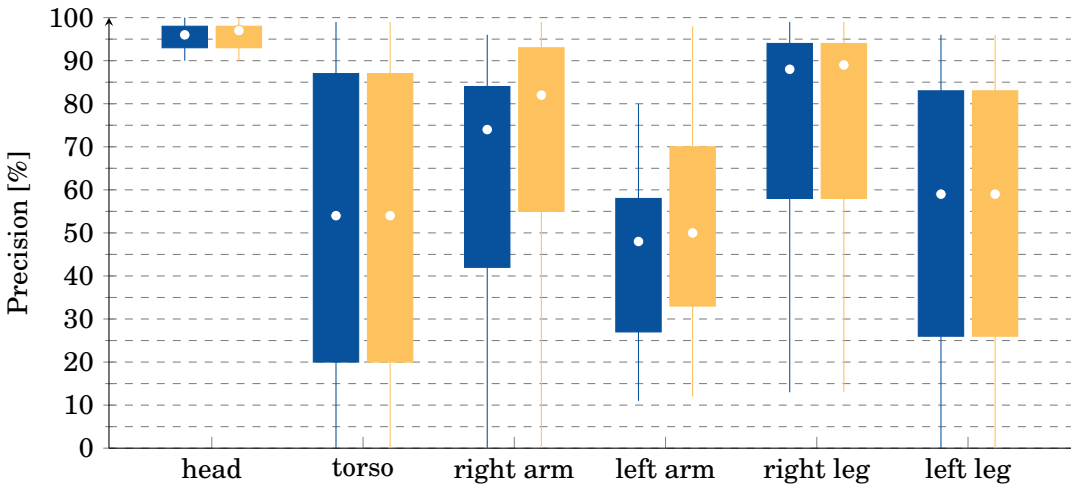


Figure D.2: Box plot comparing the class prediction precision before (blue) and after (yellow) the refinement algorithm application for each class.

Appendix E

PPGI and Heart Rate Extraction

E.1 PPGI extraction

E.1.1 Notation

Table E.1: PPGI notation.

Body part	i
Head	1
Torso	2
Right arm	3
Left arm	4
Right leg	5
Left leg	6
head lower section	7
head middle section	8

E.2 HR Estimation

E.2.1 Cross-correlation Plots

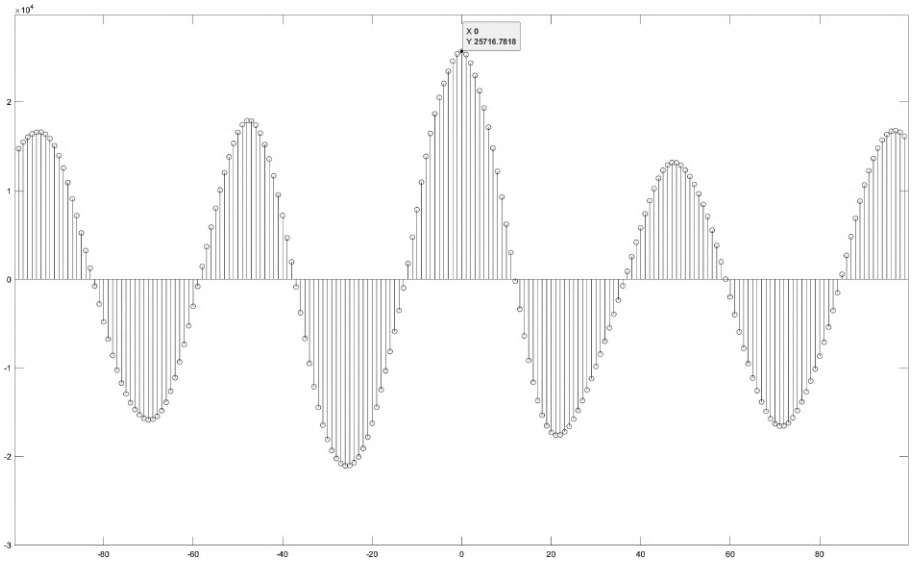


Figure E.1: Cross-correlation plot between the PPGI extracted from the head and the PPGI extracted from the torso.

E.2.2 Implementation Details

Algorithm 2: Quality index algorithm

procedure qualityindex($f_{\max}$, DFT, $\hat{\text{PPGI}}_i^1(win)$);
if $(1.3 < f_{\max} < 4)$ *and* $(\max(\hat{\text{PPGI}}_i^1(win)) - \min(\hat{\text{PPGI}}_i^1(win)) < 100)$ **then**
  $QI_i \leftarrow$ Equation 7.2
end if
else
  $QI_i \leftarrow 0$
end if
return QI;

Algorithm 3: HR estimation algorithm

procedure hrestimator($\mathrm{PPGI}_i(t)$) ← Input: set of resampled raw PPGI signals;
$\hat{\mathrm{PPGI}}_i^1(t)$ ← Set of filtered PPGI signals;
$\hat{\mathrm{PPGI}}_i^2(t)$ ← Set of filtered PPGI signals;
W ← Lenght of the time window;
J ← Window jump;
$L \leftarrow lenght(\mathrm{PPGI}_i(t))$;
for $k : J : L - W + 1$ **do**
  $win \leftarrow k : k + W - 1$;
  for $i = 1 : 8$ **do**
    $\mathrm{DFT} \leftarrow \mathrm{FFT}(\hat{\mathrm{PPGI}}_i^1(win))$;
    $f_{\max} \leftarrow \max(\mathrm{DFT})$;
    $QI_i \leftarrow \mathrm{qualityindex}(f_{\max}, DFT, \hat{\mathrm{PPGI}}_i^1(win))$;
    $f\mathrm{PPGI}(win) \leftarrow f\mathrm{PPGI}(win) + \hat{\mathrm{PPGI}}_i^1(win) \times QI_i$;
  end for
  $\mathrm{DFT} \leftarrow \mathrm{FFT}(f\mathrm{PPGI}(win))$;
  $f_{\max} \leftarrow \max(\mathrm{DFT})$;
  $QI_9 \leftarrow \mathrm{qualityindex}(f_{\max}, \mathrm{DFT}, \hat{\mathrm{PPGI}}_i^1(win))$;
  $g\mathrm{PPGI}(t) \leftarrow \mathrm{PPGI}(t)$ with the higher QI;
  SST ← wavelet synchrosqueezed transform of $g\mathrm{PPGI}(t)$;
  HR ← maximum energy time-frequency ridge per sample from SST;
end for
return HR;

E.3 Results

E.3.1 Multiple Regions of Interest

Table E.2: Performance results of the proposed method for HR estimation relying exclusively on the PPGI extracted from the head region. HMI stands for high motion intensity. The first columns of the RMSE, MAE and prediction accuracy columns are the results when the complete recording is considered, the second columns refer to the results when HMI periods are excluded.

Patient ID	**Measurement**	**Recording time** [s]	**HMI time** [s]	**RMSE** [bpm]		**MAE** [bpm]		**Prediction accuracy** % of time	
S009_S009	1	587	105	26	17	16	9	54	62
S009_S010	1	587	55	20	20	15	15	43	43
S009_S012	1	587	21	3	3	2	2	94	95
S009_S014	1	587	522	15	14	9	7	57	71
S009_S016	2	587	79	13	11	9	6	63	69
Mean				16	13	10	8	62	68
SD				9	7	6	5	19	19