# Machine Vision for Patient Monitoring in the Neonatal Intensive Care unit

by

Yasmina Souley Dosso

A doctoral research thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment for the degree of

Doctor of Philosophy, Biomedical Engineering

Carleton University
Ottawa, Ontario

# Abstract

Continuous patient monitoring of newborns in the neonatal intensive care unit (NICU) is often performed with wired sensors which can be cumbersome, can interfere with parental bonding, and can irritate the patient's fragile skin. Non-contact video-based patient monitoring systems are therefore a preferrable solution. While a multitude of high-performing machine vision technologies have been successfully implemented on an adult population, such methods often fail in neonatal population. In this thesis, we assess state-of-the-art adult-based methods to bridge the gap to an understudied neonatal population in the NICU environment. To this end, several important machine vision concepts are investigated, including scene understanding, image classification, face detection, semantic segmentation, motion detection, face tracking, and heart rate estimation. In each of these areas, we assess the state-of-the-art and identify its applicability to a neonatal population. In cases where serious limitations are observed, this thesis pushes the state-of-the-art and implements new techniques more suitable for newborns. Finally, a non-contact neonatal heart rate monitoring pipeline is created using multiple research contributions in this thesis. Doing so, we obtain a vital sign decision support tool for clinical use by estimating the uncertainty in each research contributions and demonstrating how errors can propagate from one to another.

Thirty-three newborns admitted to the NICU at the Children's Hospital of Eastern Ontario were recorded using a depth-sensing camera, which simultaneously captures color, depth and near-infrared videos, thereby acquiring pertinent data in all lighting conditions. Gold standard event annotations and physiologic data were recorded simultaneously as ground truth data for the development of machine vision models. Our proposed approach includes a combination of machine vision, deep learning, image processing, and signal processing techniques to overcome environmental factors such as lighting variations, occlusion, and motion artifacts. This thesis implements an efficient, robust, and reliable prototype neonatal monitoring system for potential future deployment in hospital settings. To this end, this research aimed to exploit transfer learning from state-of-the-art models to address problems such as complex NICU scenes, variations in newborn's visual features, data scarcity, and class imbalance which are often observed in clinical research and neonatal monitoring applications.

# Acknowledgments

Years of persistent and lengthy research culminated in the completion of this dissertation, which was only possible with the unwavering support, patience, and guidance of my supervisor Dr. James Green. Throughout my entire graduate studies, Jim has helped me grow as a researcher by allowing me to explore different research avenues, by giving me various opportunities to present my thesis research, and by being an exceptional mentor while teaching me to be a mentor to others. I am immensely grateful and proud to be part of Jim's lab, the Carleton University Biomedical Informatics Collaboratory (CU-BIC) lab, where I was able to conduct research that I've been truly passionate about for all those years.

I would like to thank members of my research committee including Eran Ukwatta and Pierre Payeur who have provided valuable feedback in my entire Ph.D. examination process. In addition, I would like to thank Zahra Moussavi and Doron Nussbaum for thoroughly evaluating my thesis for the defense.

I would like to thank Kim Greenwood from the Department of Clinical Engineering and JoAnn Harrold from the Department of Neonatology at the Children's Hospital of Eastern Ontario (CHEO), who have helped in the entire data collection process and continuously provided useful feedback pertaining to the clinical application of my research. I would also thank Randy Giffen, our collaborator from IBM, who has provided interesting ideas and feedback for this research.

Lastly, and most importantly, I would like to thank my parents, Ousmane Souley Dosso and Hélène Dossa, for their endless support, motivation, patience, and kindness. Especially in the last couple of years of my thesis, my parents have been my rock day and night, helping me get to the finish line, and I wouldn't have been able to complete this degree without their love, encouragement, and their continuous sacrifice to provide me with the life that I (and my siblings) deserve.

I would also like to thank my siblings Yazid and Hosnia for always being here for me, to the rest of my family encouraging me from France, Benin, and the USA, to our family friend Marcel Bellehumeur and his wife Lise, to Merisse Atencio and C.C. Tang who have been my best friends, persistent study partners, and emotional support. Finally, I'd like to thank my labmates and colleagues with whom I shared this amazing graduate studies experience, and to some of which I have built life-long friendships and have truly become family. All of these important people have positively impacted my life during my studies and helped me persevere to where I am today.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

AP - Average Precision

AUC - Area Under the Curve

BPM - Beats per minute

BFID - Bottle-Feeding Intervention Detection

CEA - Clinical Event Annotator

CE-CLM - Convolutional Experts Constrained Local Models

CHEO - Children's Hospital of Eastern Ontario

CNN - Convolutional Neural Network

CPAP - Continuous Positive Airway Pressure

CRF - Case Report Form

ECG - Electrocardiograph

EVM - Eulerian Video Magnification

GBRT - Gradient Boosted Regression Tree

FCN - Fully Convolutional Network

FFT - Fast Fourier Transform

SGDM - Stochastic Gradient Descent with Momentum

HR - Heart Rate

ICA - Independent Component Analysis

IOU - Intersection Over Union

LSTM - Long Short Term Memory

mAP - mean Average Precision

MCC - Matthews Correlation Coefficient

MEI - Motion Energy Image

MHI - Motion History Image

MS COCO - Microsoft Common Objects in Context

MTCNN - Multi-Task Cascaded Convolutional Neural Network

MVF - Motion Vector Field

NAS - Network Attached Storage

NELA - Nose-to-Eyeline Angle

NICU - Neonatal Intensive Care Unit

NIPPV - Nasal Intermittent Positive-Pressure Ventilator

NIR - Near Infrared

OHW - Overhead Warmer

PASCAL VOC - Pattern Analysis, Statistical Modelling and Computational Learning Visual Object Classes

PCA - Principal Component Analysis

PMDI - Patient Monitor Data Import

PSM - Pressure Sensitive Mat

R-CNN - Region-based Convolutional Neural Network

RGB-D - Red Green Blue - Depth

ROC - Receiver Operating Characteristic

ROI - Region-of-interest

RPN - Region Proposal Network

RR - Respiration Rate

SNR - Signal to Noise Ratio

SQI - Signal Quality Index

YOLO - You Only Look Once

# 1 Introduction

This chapter presents the motivation and problem statement for this thesis. This chapter also summarizes major contributions from this thesis before detailing the structure of this document.

## 1.1 Motivation

Patient monitoring systems in the neonatal intensive care unit (NICU) aim to continuously assess the patient's condition. Such systems typically require multiple wired sensors that can be cumbersome, uncomfortable, and irritating, particularly for neonates with fragile skin. Additionally, several false alarms may occur due to motion artifacts at the skin-sensor interface.

This has led to interest in developing non-contact patient monitoring techniques. Camera-based technologies are attractive due to their affordable price, ease of mounting on any bed type, and visual context they can provide from the hospital bed environment. While a single colour (RGB) camera is useful, several studies have used Red, Green, Blue, Depth (RGB-D) cameras [1]–[3], an array of multiple RGB cameras [4], infrared cameras [5], or thermal cameras [6] to capture additional information from the patient scene and to overcome the challenges specific to the NICU environment (e.g., complex scenes, low lighting environment, etc.).

Recent advances in camera technology and in deep neural networks create an opportunity to develop novel computer-vision-based solutions, especially for a neonatal population. This thesis evaluates the state of the art in machine vision technologies, originally developed for an adult population, to bridge the gap to an understudied neonatal population. The findings presented in this thesis demonstrate that simply using these technologies "out-of-the-box" on neonatal data often fails due to differences in visual features, overall shape, and subject poses (lying vs. standing). Many of these NICU-specific challenges have often been ignored or addressed manually in previous neonatal research (e.g., manual face detection). This thesis therefore bridges the gap between state-of-the-art technologies and robust neonatal research to create reliable patient monitoring pipelines. This goal is achieved by collecting unique and high-quality annotated data encompassing diverse complex NICU scenes; by assessing various algorithms; and by advancing the state of the art in key areas. The research presented here results in a number of solutions for machine vision tasks fine-tuned for the NICU environment to create multiple functional blocks that culminate in a comprehensive non-contact monitoring system. These blocks can contribute to diverse monitoring pipelines to address multiple applications in the NICU. As a demonstration of how these functional blocks may be composed into a full pipeline, this thesis presents

a non-contact, non-invasive, and unobtrusive heart rate monitoring system to improve continuous patient care.

Our overarching research initiative is exploring the use of an RGB-D camera positioned above the patient to record the bed environment, and a pressure sensitive mat (PSM), placed beneath the patient, to capture the time-varying contact pressure between the newborn and the mattress. With both modalities, our research aims to analyze the overall scene, detect patient's regions-of-interests (ROI), detect patient motion, and estimate patient vital signs, with the ultimate goal of improving neonatal patient care, providing useful context for clinical documentation, and potentially reducing false alarms. This thesis focuses on video-based analysis from the RGB-D camera, while other research students are investigating the use of PSM data and the multimodal analysis of fused video and PSM data. This research was completed in collaboration with neonatology and clinical engineering at the Children's Hospital of Eastern Ontario (CHEO) to obtain reliable data.

## 1.2 Problem Statement

State-of-the-art machine vision and deep learning algorithms have been successfully applied to several patient monitoring applications, especially in person and face detection [7]–[10]. These methods are typically developed for healthy upright standing adults and often perform poorly on neonatal population when used "out-of-the-box" [11]–[13]. In this thesis, we assess and advance the state-of-the-art in computer vision and video-based non-contact patient monitoring algorithms specifically for the NICU environment.

To that end, we first collected a dataset of RGB-D data from 33 actual patients in the NICU at CHEO, simultaneously with physiologic data and bedside-annotation of events of clinical significance. During the data collection process, we needed to develop methods to record video in all NICU beds, including open bed types (crib and overhead warmer) and closed incubators. In particular, closed incubators are expected to present significant challenges due to recording through reflective plexiglass surfaces. This thesis investigates how a single consumer-grade RGB-D camera and a custom silicone apparatus for closed incubators can be used to capture reliable data from all patients admitted in the NICU, in a practical and safe manner.

The resulting high-quality dataset contains data variations that are frequently observed in the NICU environment. Such NICU-specific data variations can, however, introduce challenges such as: lighting variations (bright, low, or phototherapy lighting), self-induced motion (when moving freely in the bed), temporary patient occlusion during clinical interventions, and other patient occlusions from beddings, self-induced due to free-moving limbs, or from hospital equipment (e.g., ventilation support). These NICU challenges are depicted in Figure 1.1, and among all of these complex scenes, the patient position


Figure 1.1: Common NICU-specific visual challenges. Top-left: The lighting intensity changes significantly and frequently in the NICU and may include phototherapy. Top-right: Self-induced motion of the limbs and head. Bottom-left: Ongoing interventions can occlude parts of the patient or modify the patient poses/viewpoint. Bottom-right: Other facial or body occlusions can occur from beddings, hospital equipment, or from the patient's limbs.

can also vary (supine, prone, or on one side). While assessing the state-of-the-art, these NICU challenges are identified and investigated to finetune models to a neonatal population captured in complex clinical scenes. This section details how each of these visual challenges occur during continuous patient care, and the associated goals of this thesis to address them when reviewing the state-of-the-art.

The lighting environment changes often in the NICU, depending on the time of day, level of artificial lighting, or on the patient status (incubator patients are often subjected to complete darkness to reduce sensory input). A continuous neonatal monitoring system should, however, be robust to these common variations. This thesis addresses this issue by identifying lighting conditions and ongoing phototherapy treatment to create a monitoring system efficient under various illumination environments, and informing this system during a change of lighting condition.

Patient movement can be challenging to an object-based monitoring system when the object frequently changes location. This thesis investigates whether identified motion can be exploited to track the patient overtime, thereby continuously obtaining a reliable ROI. Beyond the impact on a vision system, patient movement can also impact clinical care by causing false alarms on one or more physiologic signals due to motion artifacts created by electrode movement at the skin-sensor interface or due to cable motion [14]. The increase of false alarms can in part desensitize clinical staff, leading to alarm fatigue [15], [16]. A study conducted in the ICU showed that over 63% of generated alarms were false [16]. Other studies have directly identified motion as one of the leading factors in signal quality

reduction in the NICU, thereby increasing false alarms [15], [17], [18]. Contextual knowledge from a video-based system can be used to gate false alarms generated by the patient monitor due to motion-induced artifacts on physiological data streams. This thesis seeks to primarily detect and classify patient motion as a step towards gating motion-related false alarms.

Non-contact video-based monitoring is highly sensitive to what the system can "see", and self-induced patient motion or occlusion can negatively impact that vision. More specifically, partial or complete body occlusions can occur during an intervention, from beddings, from clothing items, from ventilation support devices, or from nasogastric feeding tubes and adhesive tape. These circumstances occur frequently in the neonatal environment and can hinder the detection of the patient's body or face using a vision-based system. A monitoring system should be robust to such cases. This thesis develops methods to detect ongoing clinical interventions and occlusions due to coverage by beddings to include awareness of these events in neonatal non-contact monitoring systems.

Types of routine care events include diaper change, feeding, checking temperature, checking vital signs, weighing, changing sensors, to name a few. All interventions must be carefully documented in the patient's chart indicating the date, time, personnel in charge, and any relevant details [19]. Such documentation is important since it is the primary source of communication between clinical staff across shifts. Ideally, all events would be documented contemporaneously. Realistically, the elevated nurse workload, combined with patient acuity requiring more time for patient care, often results in retrospective charting. Retrospective charting is problematic since it is often incomplete or inaccurate. A clinical assistive tool to automatically identify and chart interventions can help address this issue. This thesis aims to detect context variables from the scene and explores how a comprehensive NICU scene understanding can be used as a step towards semi-automated nurse charting to assist clinical staff in the documentation process and patient care.

This thesis assesses and advances the state of the art across multiple computer vision tasks related to patient monitoring, including scene understanding, image classification, face detection, semantic segmentation, motion detection, face tracking, and heart rate estimation. A video-based heart rate estimation pipeline is developed to demonstrate how these research contributions can be used in combination to form continuous patient monitoring systems. To this end, we demonstrate how a non-contact neonatal monitoring system can be implemented by integrating various research contributions presented in this thesis, while remaining robust to the complex NICU-specific environmental challenges such as lighting conditions, phototherapy treatments, patient coverage by beddings, occlusion from ventilation support, ongoing clinical intervention, multiple NICU bed types, varying patient poses, and natural diversity between patients of different sex, age, weight, and ethnicity.

## 1.3 Summary of Contributions

This thesis seeks to evaluate and extend the state of the art in computer-vision-based patient monitoring for the neonatal population. This first required the collection of a unique multi-modal and carefully annotated dataset from an actual NICU to better understand the challenges present in the environment. Data collection spanned two years and I was involved in all aspects of data collection. Based on these data, the thesis proposes several machine vision neonatal patient monitoring applications culminating with demonstrating a comprehensive non-contact RGB-D video-based neonatal patient monitoring system. This thesis develops and assesses all research contribution blocks required to build such a system by evaluating the state of the art on complex NICU scenes, and extends it when serious limitations are identified in existing computer vision and deep learning methods. A vital sign monitoring pipeline integrating multiple blocks demonstrates a specific instantiation of the entire patient monitoring system.

### 1.3.1 Non-contact neonatal monitoring pipeline

All research blocks were developed as reusable elements that could be combined into various video-based monitoring pipelines. In this section, contributions related to each block are first discussed (Figure 1.2), before combining them into a non-contact neonatal heart rate estimation system (Figure 1.3). The numbers in Figure 1.2 correspond to the chapter numbers in this thesis and letters represent contribution blocks implemented here (often divided in multiple sub-chapters for readability purposes).

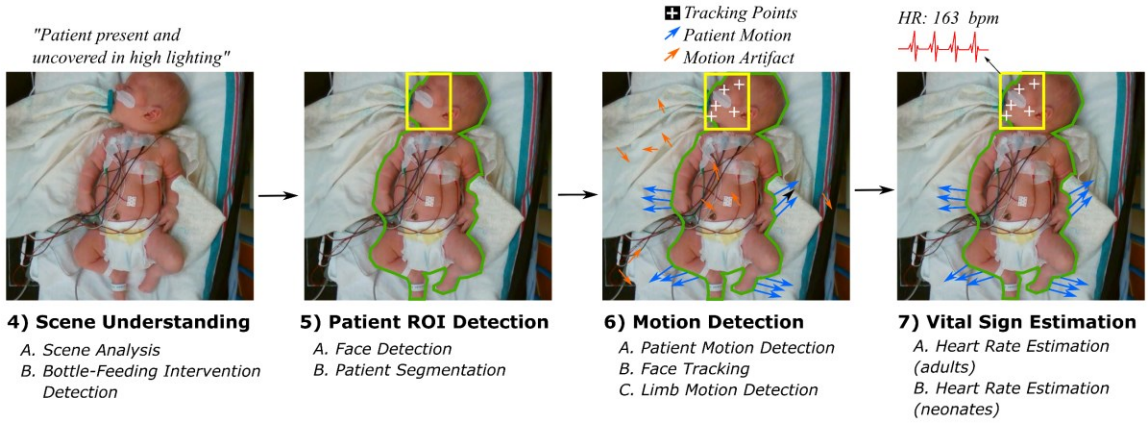



Figure 1.2: Thesis contributions in one picture. The numbers represent chapters in this thesis and letters indicate implemented contribution block.

#### *1.3.1.1 Scene Understanding*

As a first step, the overall NICU scene is analyzed to guide and inform the monitoring pipeline. The scene is examined within an image in terms of lighting conditions, ongoing phototherapy treatment, bed occupancy, patient coverage, ongoing intervention, and bottle-feeding intervention. A sensor-fusion approach is used to achieve robustness to lighting variations, patient coverage, and temporary occlusions

during clinical interventions. Bed occupancy specifically informs an algorithm to only run when the patient is present in the bed for computational efficiency. Bottle-feeding interventions are detected as part of routine care events for oral feeding assessment tools. The image is captioned using a generated sentence summarizing the scene, useful to assist nurses in a semi-automated documentation process. This work is illustrated in the first patient image in Figure 1.2 and is described in Chapter 4 of this thesis.

#### *1.3.1.2 Patient ROI Detection*

As a second step, ROI from the patient are detected including the newborn's face and body for subsequent analysis. This step restricts an algorithm to focus on the area of the patient and ignore the typically cluttered NICU environment including beddings, plush toys, cables, and hospital equipment. A robust face detector in complex NICU scenes is implemented to detect the newborn's face despite lighting variations, near complete occlusions (e.g., ventilation support) or challenging poses (e.g., near back view when in prone position). Patient segmentation techniques are explored to extract the outline of the patient in a cluttered background, with varying poses, and different levels of body coverage (e.g., unclothed, clothed, covered by beddings). This work is illustrated in the second patient image in Figure 1.2 and described in Chapter 5 of this thesis.

#### *1.3.1.3 Motion Detection*

As a third step, patient movements from the selected ROI (face or body) are detected.

From the segmented body of the patient, motion detection and classification techniques are used to identify patient movements and categorize moving limbs. Patterns of body movements are useful for detecting motion-related conditions, such as clonic seizures. The limb motion detection algorithm may be used to identify false alarms generated due to motion, for example, where a pulse oximeter is attached to a limb undergoing significant motion.

From the detected face of the patient, a tracking algorithm was designed to continuously obtain this ROI in video analysis, despite varying levels of patient movements and occlusions. Continuous detection of the facial area is important for continuous monitoring of vital signs. Analysis of neonatal facial expression can also leverage this tracked area for continuous pain assessment, sleep-wake cycle detection, or jaundice monitoring.

Given the selected ROI (face or body), the detected motion is restricted only to the patient, and ignores other movements in the scene (e.g., clinical movements visible in the scene). This work is illustrated in the third patient image in Figure 1.2 and described in Chapter 6 of this thesis.

#### *1.3.1.4 Vital Sign Estimation*

As a fourth step, the detected and tracked face ROI is analyzed for heart rate estimation. Leveraging the Eulerian video magnification (EVM) technique that amplifies subtle color variations in the scene

[20], cyclical changes in blood perfusion in the capillaries of the face are detected. The passband inherent within the EVM technique is determined automatically using a selective EVM approach to obtain a robust heart rate estimate. This work is illustrated in the fourth patient image in Figure 1.2 and is described in Chapter 7 of this thesis.

#### *1.3.1.5 Non-Contact Neonatal Estimation Pipeline*

Finally, steps 1-4 are combined sequentially to obtain comprehensive non-contact video-based neonatal monitoring pipelines.

This thesis demonstrates one instantiation of a video-based neonatal HR monitoring pipeline. From a detected patient in the bed (step 1), the face can be detected (step 2) then tracked over time (step 3) to achieve continuous vital sign estimation (step 4). We also evaluate how errors arising from each measurement step propagates to the final heart rate (HR) estimation, thereby informing the estimate with a measure of uncertainty. To do so, a measure of uncertainty is extracted at each step before combining it into a meaningful signal quality index (SQI). This final SQI is a useful decision support tool to clinical staff in a future deployment of our system in non-contact continuous HR monitoring of patients in the NICU. This work is illustrated in parts of Figure 1.3 and proposed in Chapter 8.

As illustrated in Figure 1.3, all research work accomplished in this thesis consists of research contribution blocks (in black) under each chapter (blue boxes), which can be integrated into several non-

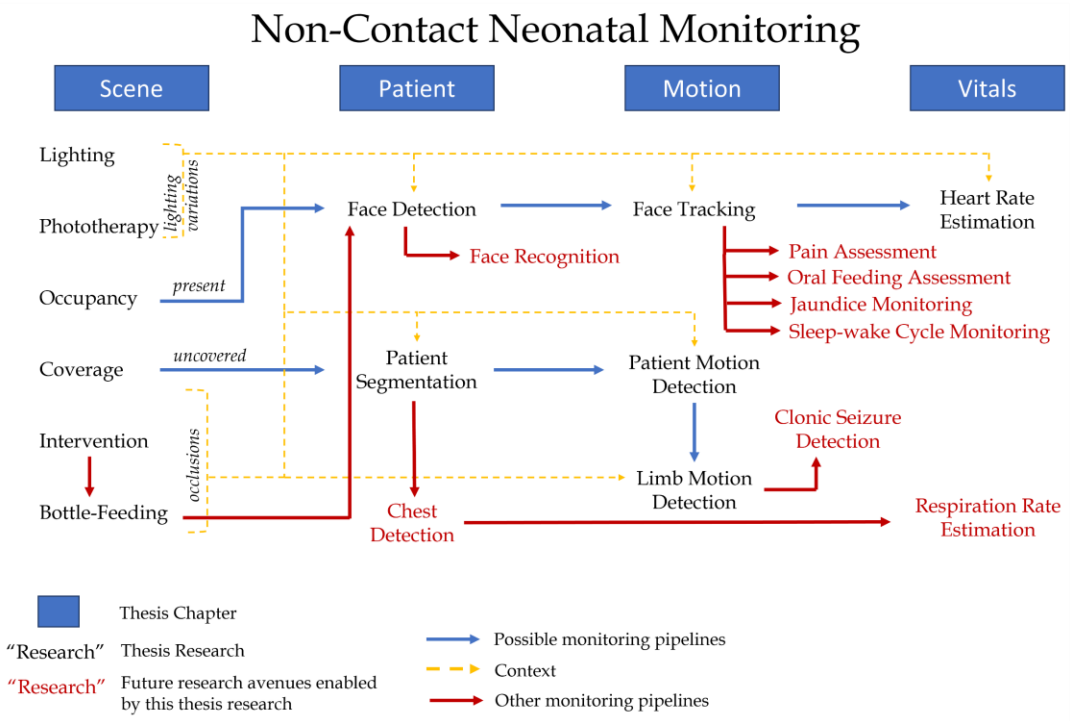



Figure 1.3: Non-contact monitoring pipelines that can be created from research contributions blocks developed in this thesis.

contact neonatal monitoring pipelines (blue arrows) that can be affected by context in the NICU environment (yellow arrows). This thesis addresses all contribution blocks in Chapters 4-7, before presenting how a robust "Heart Rate Estimation" pipeline can be created in Chapter 8.

Other neonatal monitoring research outside the scope of this thesis (in red) could also leverage our work for their own applications (red arrows).

This thesis overcomes several challenges commonly observed in the NICU environment and in neonatal population, and thereby advances the fields of machine learning and computer vision in neonatal monitoring applications. This represents the global impact of our work. Leveraging an affordable and consumer-grade RGB-D camera will enable the rapid and economical deployment of such systems in hospitals.

### 1.3.2 Thesis Contributions

Each building block of our non-contact monitoring system represents the culmination of multiple contributions. Specifically, this thesis makes the following contributions:

1) Introduced a multi-faceted and multimodal computer vision approach to scene understanding in the NICU. This comprises simultaneous classification of five different contexts in a single image. To the best of our knowledge, no such complex scene recognition model exists in the NICU. This work was performed by expanding existing image processing and deep learning state-of-the-art approaches. More specifically, color space transformations were implemented using new transformation functions for detecting patients in natural lighting or undergoing phototherapy treatments. Different RGB-D fusion schemes were explored for complex context variables (bed occupancy, patient coverage, and ongoing intervention), including a new image fusion technique to impute the depth data on all three RGB streams, thereby obtaining a 3-channel RGB-D image. This thesis also demonstrated a proof-of-concept system for generating textual summaries of a clinical scene in the NICU, as a step towards semi-automated patient charting. This work shows how image captioning can be applied in a clinical setting for continuous neonatal care to assist nurses. Finally, we developed a bottle-feeding intervention detection algorithm using a data expansion technique to address the class imbalance and data scarcity issue. Our approach presented limitations of naively utilizing transfer learning with a state-of-the-art CNN model, and proposed new data-driven approach to bridge the gap between the source and target domain when transfer learning using an imbalanced and scarce dataset. This work represents a novel contribution to the field of deep learning for clinical setting applications and to neonatal monitoring research, given that no previous studies have classified a clinical intervention in the NICU. This contribution is detailed in Chapter 4.

2) Detected the face of newborn patients under complex NICU scenes by assessing state-of-the-art face detection models, identifying NICU-specific challenges, and implementing new face detectors suitable for newborns in a clinical environment. In comparison to state-of-the-art face detectors, our proposed NICUface models address NICU-specific challenges such as varying levels of facial occlusions (from soother, ongoing clinical intervention, beddings, self-induced from moving limbs, phototherapy eye mask, and ventilation support), varying viewpoints (from a far distance, in profile view, in near top view when held in the bed, and in near back view when in prone position), and varying lighting (in bright, low lighting, and phototherapy light). These analyses culminate in the creation of robust NICUface detectors with improvements on our most challenging neonatal dataset relative to state-of-the-art CE-CLM, MTCNN, RetinaFace, and YOLO5Face models. NICUface also demonstrates exceptional performance in face detection among the two most complex scenes (prone position and ventilation support), and a solution for addressing the phototherapy lighting is proposed leveraging the phototherapy classification approach presented in Contribution 1. Comprehensively, NICUface is a novel robust neonatal face detector capable of reliably detecting newborn's faces in complex NICU scenes. This contribution is detailed in Chapter 5.
3) Developed a neonatal face tracking algorithm robust to patient motion and occlusion. We evaluate two tracking algorithms; one that would track the ROI by continuous detection of this area, and one that would track the ROI by identifying landmarks to track their displacement overtime. This thesis presents a new algorithm as a fusion of both methods. An ROI is detected, landmark points are identified and continuously tracked overtime until tracking deteriorates and the algorithm resets. This work expands on previous computer vision techniques by identifying when tracking becomes unreliable due to temporary occlusion or patient motion events. This contribution is detailed in Chapter 6.
4) Developed a sensor fusion approach for estimating heart rate using the EVM method. This thesis introduces a Selective EVM technique to automatically identify narrow passbands within an array of plausible HR values, instead of using a single wide passband. The selected narrow passband pertaining to the patient's HR then increases the signal-to-noise ratio (SNR). This method was implemented on adult subjects where HR is estimated among varying lighting conditions (4 brightness levels) and subject poses (lying, sitting, standing). A multi-modal system leveraging the RGB, depth and near infrared (NIR) streams was implemented and showed robustness to all lighting conditions and subject poses. The Selective EVM method was implemented and validated on adult subjects before being also validated on neonatal patients under similar conditions (patient at rest, in low to bright lighting environment, and from lying poses only), and

by modifying the range of passbands to encompass newborns' higher heart rate compared to adults. In addition, the Selective EVM is validated on resting patients undergoing phototherapy treatment and we demonstrated that best results are obtained under this condition for neonates, given the strong pulsatile signal observed in the blue color channel. This contribution is detailed in Chapter 7.

5) Created a non-contact neonatal heart rate monitoring pipeline by integrating multiple research contributions into one system. This pipeline consists of identifying bed occupancy, detecting the face of the patient, tracking the face of the patient overtime, and finally estimating their HR from the continuously detected and tracked ROI. This work demonstrates improved performance in HR estimation when the patient is at rest. During more challenging scenes (e.g., motion, occlusions, or varying lighting), the estimated HR can be impacted, and this thesis presents a method for quantifying the uncertainty at each stage in the pipeline, before obtained a fused SQI informing how the errors can propagate through the pipeline. This thesis demonstrated how the final fused SQI is more informative than only using the SQI from the final "HR Estimation" stage in the monitoring pipeline. To this end, we showed how the fused SQI can be used as a decision support tool for clinical use, by informing the clinicians about the uncertainty of the HR estimate. This contribution is detailed in Chapter 8.
6) In order to collect the data, this thesis developed a detailed design for an image data collection system for the NICU covering multiple bed types and ensuring patient safety. Careful experiments were performed to ensure the safety of the patient while acquiring high quality data, and recommendations for camera placement in the NICU across all bed types are suggested in this thesis. We investigated potential interference artifacts from active IR illumination from the structured light depth-sensor on the pulse oximeter typically used in neonatal care. This contribution is detailed in Chapter 3.

It should be noted that all these research contributions were enabled by the collection of a unique, multi-modal dataset from 33 actual patients in the NICU at CHEO. I personally collected data from the majority of patients in this study.

Portions of this thesis have appeared from the following publications, on which I was the lead author:

1. **Y. S. Dosso**, A. Bekele, and J. R. Green, "Eulerian Magnification of Multi-Modal RGB-D Video for Heart Rate Estimation," in Proc. of IEEE Int. Symp. Med. Meas. Appl. (MeMeA), Rome, Italy, 2018. ***Awarded the IEEE MeMeA "Women in Engineering Best Paper"*** [21]

   *This work presents a multimodal Selective EVM method used in RGB-D-NIR videos for heart rate (HR) estimation on adult subjects by applying different passbands among the*

*array of plausible HR ranges and selectively extracting the passband enclosing the person's true HR.*

2. **Y. S. Dosso**, A. Bekele, S. Nizami, C. Aubertin, K. Greenwood, J. Harrold, and J. R. Green, "Segmentation of patient images in the neonatal intensive care unit," in 2018 IEEE Life Sciences Conference, IEEE-LSC 2018, 2018, pp. 45–48. [22]

   *This work demonstrates a semantic segmentation method using the SegNet model for extracting the clothed patient from a cluttered NICU bed environment.*

3. **Y. S. Dosso**, S. Aziz, S. Nizami, K. Greenwood, J. Harrold, and J. R. Green, "Neonatal Face Tracking for Non-Contact Continuous Patient Monitoring," in 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2020, pp. 1–6. [23]

   *This work evaluates different tracking algorithms before proposing a novel face tracking algorithm tailored for newborns and robust to temporary occlusions and self-induced motion.*

4. **Y. S. Dosso**, S. Aziz, S. Nizami, K. Greenwood, J. Harrold, and J. R. Green, "Video-Based Neonatal Motion Detection," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2020, vol. 2020-July, pp. 6135–6138. [24]

   *This work detects patient motion perceived in the scene by evaluating state of the art techniques based on computer vision algorithms (Optical flow) and deep learning algorithms (LSTM).*

5. **Y. S. Dosso**, K. Greenwood, J. Harrold, and J. R. Green, "Bottle-Feeding Intervention Detection in the NICU." 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021. [25]

   *This work assesses the state of the art in image classification using the VGG-16 model and presents a data-driven approach to address the class imbalance and data scarcity problems observed in transfer learning a pretrained classifier, to bride the gaps between the source and target domain.*

6. **Y. S. Dosso**, R. Selzler, K. Greenwood, J. Harrold, and J. R. Green, "RGB-D Sensor Application for Non-Contact Neonatal Monitoring." 2021 IEEE Sensors Applications Symposium (SAS). IEEE, 2021. [26]

   *This work presents various experiments used in the study design stage to systematically evaluate the best camera placement above all NICU beds for neonatal monitoring*

*applications. Upon request, Roger Selzler created the silicone skirt suitable to mount the RGB-D camera atop closed incubators.*

7. **Y. S. Dosso**, K. Greenwood, J. Harrold, and J. R. Green, "RGB-D Scene Analysis in the NICU," in Elsevier – Computers in Biology and Medicine, 2021. [27]

   *This work demonstrates a multimodal and multi-faceted computer vision approach to identify contexts from the NICU scene using image processing techniques for the classification of lighting conditions and phototherapy treatment, and transfer learning with a deep learning model for the classification of ongoing intervention, bed occupancy, and patient coverage using RGB-D data.*

8. **Y. S. Dosso**, D. Kyrollos, K. Greenwood, J. Harrold, and J. R. Green, "Robust Neonatal Face Detection in Complex NICU Scenes". IEEE Access, 2022 ***(In review)***

   *This work evaluates the state of the art in face detection within complex NICU scenes and leverages this knowledge to implement robust neonatal face detectors by finetuning the YOLO5Face and RetinaFace models. Daniel Kyrollos provided annotations for the NBHR dataset and implementations of the RetinaFace model, while all other data annotation, method development, and performance analysis were completed by me.*

## 1.4 Thesis Structure

This research dissertation has nine chapters. Chapter 2 outlines the literature review for video-based neonatal monitoring, deep learning models, computer vision methods, and multi-modal algorithms leveraged to evaluate the state of the art. Chapter 3 presents the data acquisition set up in the NICU at the CHEO, and the important preliminary experiments conducted to ensure the safety of the data collection protocol at the hospital. Proposed research methodologies and results on scene understanding, patient ROI detection, motion detection, vital sign estimation, are demonstrated in Chapter 4-7 respectively, as depicted in Figure 1.2. More specifically, Chapter 4 includes an overall NICU scene understanding, with a scene analysis approach (4.1 – 4.3) identifying lighting conditions, ongoing phototherapy treatment, bed occupancy, patient coverage, and ongoing intervention. Within identified interventions, a bottle-feeding intervention detection technique (4.4 – 4.6) is developed. Chapter 5 describes neonatal face detection models implemented to address the complex NICU scenes (5.1) and a patient segmentation approach (5.2) to detect the body of the patient from the background. Chapter 6 presents multiple motion detection techniques including neonatal motion detection (6.1 – 6.2) where patient movements are identified. A neonatal face tracking method (6.3 – 6.4) demonstrates how the face of the patient can be tracked over time, while moving limbs are classified from uncovered patients (6.5 – 6.6). Chapter 7 proposes a non-contact HR estimation approach by detecting and evaluating perceived

blood flow on a person's face. The method is first developed and validated on an adult dataset (7.1) before being also validated on our neonatal dataset (7.2). Finally, one integrated pipeline combining multiple thesis research contributions is presented in Chapter 8. More specifically, the scene is analyzed to identify patient presence; when the patient is present in the scene, the patient's face is detected; the face is continuously tracked overtime; and HR is estimated within the tracked ROI (8.1). This pipeline implements a non-contact neonatal heart rate monitoring system, where SQIs are extracted at each stage to evaluate the propagation of uncertainty in final HR estimates (8.2). This pipeline presents a vital sign decision support tool for clinical use to inform the clinicians on the uncertainty of the HR estimate. Future work could investigate other monitoring pipelines using proposed work from this thesis as building blocks for other neonatal monitoring applications. These other research topics are outside the scope of this thesis.

For this research, the data collection was a lengthy process due to Health Canada regulations, ethics protocols, limited access to the NICU environment, and very specific inclusion criteria (mass, bed type, health status) and understandably hesitant parents during research participant recruitment. Data collection spanned two full years; hence, method development and validation occurred simultaneously with data collection. Our resulting dataset is unique with data acquired simultaneously from multiple modalities and high-quality real-time bedside annotations. Consequently, certain methods within our non-contact neonatal monitoring system were developed and validated with limited data or using supplementary data. Some early work has since been extended with our final substantially larger neonatal dataset. Chapter 9 summarized all results and contributions from this thesis, before presenting future work.

# 2 Background

This section outlines pertinent background and literature review for non-contact neonatal patient monitoring, scene understanding, ROI detection, motion detection, and vital sign estimation. A review of multimodal analysis and clinical charting is also presented.

## 2.1 Video-Based Neonatal Patient Monitoring

Continuous monitoring of NICU patients has been increasingly studied in recent years, particularly in the context of non-contact patient monitoring. Previous video-based studies in the NICU have primarily focused on vital sign monitoring [4], [5], [28], [29], behavioral and pain assessment from facial expression [30]–[32], detection of patient motion [1], [2], [33], and often in clonic seizure detection [2], [34], [35]. More recent non-contact studies have sought to *understand* the NICU environment by detecting patient presence in the bed [28], [36] or by detecting clinical interventions [28].

Fernando *et al.* investigated NICU conditions for vital sign estimation using two cameras mounted above the patient [4]. One camera, pointed at the face, perceived color changes for heart rate estimation, while the other observed the chest area for respiration estimation. Interestingly, they controlled patient coverage (covered vs uncovered), and lighting variations (ambient daylight, incandescent light, darkness, and phototherapy) to evaluate pulse rate extraction under different NICU-specific conditions. No automatic recognition of the patient coverage or lighting in the scene was implemented. Optimal vital sign estimation results were achieved for images acquired during ambient daylight, due to reduced color imbalance and overexposure; however, they ultimately did not discuss how patient coverage affected vital sign estimation. Klaessens *et al.* also used an RGB camera for heart rate estimation, but complemented their algorithm using an infrared camera to monitor the patient's body temperature and to detect variations perceived around the nose area for respiration estimation [5]. The depth information from an RGB-D camera has also been used to complement the color images. To monitor chest movements pertaining to breathing, Cenci *et al.* used the RGB images to manually locate the patient's chest area before extracting the corresponding depth images [1]. Rehouma *et al.* aimed to further reduce image noise by leveraging two depth-sensing cameras to obtain an aligned point cloud of the patient's chest area [3]. Villarroel *et al.* recently investigated similar tasks to our research, where they detected the patient from an image and interventions from a sequence of images (i.e., a video) [28]; however, they only used incubator-type beds and it is unclear if fully covered patients were included. Their experiments required drilling a hole in the top of the incubator for the camera to pass through. This thesis aims to gather all information from the scene within a single image, while still detecting patient presence, patient motion, and estimating vital signs despite their coverage, lighting conditions, or ongoing interventions, and without modifying the NICU bed types when recording with a single RGB-D camera.

Multiple studies aimed to select a pertinent region-of-interest from the patient to obtain and analyze data. In particular, for newborns in the neonatal intensive care unit (NICU), previous studies often focused on the patient's face or chest as ROI to estimate vital signs such as heart rate and respiration [1], [4], [37], of for detection of motion-related conditions [33]. Many studies have commonly detected this ROI either manually from controlled short-time monitoring periods [1], [4], [33], [38] or using skin detection algorithms [28], [33], [37]. Such methods are unreliable and unrealistic since a manual step would be time consuming to clinical staff and detecting the skin could be difficult for clothed patients or those covered by beddings.

The patient's facial area has also often been used to extract a pulsatile signal from which to estimate the patient's heart rate [4], [5], [39]. Fernando *et al.* [4] manually extracted the patient's face to track the changes in skin pixels using adaptive bandpass filtering and principal component analysis. Klaessens *et al.* [5] extracted regions of the face for subsequent HR estimation using the EVM technique [40]; although it is not discussed in detail, it appears that facial ROI were manually selected. Kyrollos *et al.* [41] also used EVM on facial video data, but for RR estimation. They leveraged the RetinaNet model [42] for automatic face detection. However, the automatic face detection was limited to within-patient testing and therefore the generalizability of their model is untested. Villarroel *et al.* [39] automatically detected video segments where patient skin is visible and manually identified a ROI (face, head, or neck) for subsequent HR estimation using independent component analysis. They later trained a multi-task convolutional neural network (CNN) for patient detection and skin segmentation to automatically detect the patient and all visible skin area for HR estimation [28]. Since they only performed skin detection, it is unclear how face detection would perform for their application given the varying amounts of visible skin during occlusions from beddings or hospital equipment. More recently, Huang *et al.* [13] manually detected the patient's face in video recordings for HR estimation. They discussed how challenging and inaccurate an automated face detector would be, considering the variations in patient posture and camera perspectives. Hence, a manual approach was used on the first video frame and a tracking algorithm would perform detection on subsequent frames.

Neonatal face detection has been applied for detecting jaundice, a condition that can be caused by hyperbilirubinemia (having high levels of bilirubin), thereby yellowing the skin [12], [43], [44]. These studies use image processing approaches based on the YCbCr color space for skin segmentation [43], followed by a manual ROI detection step to identify a specific facial region to quantify the yellowing of the skin [12]. Neonatal face recognition applications have also evaluated newborns' facial features to properly identify patients in hospital as a preventative measure to baby swapping or abduction. Bharadwaj *et al.* [45] performed manual face detection given that existing detectors failed to identify newborn faces. To overcome this issue, Awais *et al.* [46] leveraged the color palettes from the Fluke

TiX580 camera for automatic face detection and reported an accuracy of 98.5%. Their dataset used controlled head movements (-45° to 45° in yaw head tilt), close camera distance (0.25-0.36 m), and excluded occlusions from limb movements to obtain best quality data for face recognition.

Several studies have also observed the patient's face to assess their discomfort as seen from their facial expression [30]–[32]. These works all share a common preprocessing step where the patient's face must be extracted as an ROI, and they created a face detection method suitable for their controlled and short-time monitoring (from 10 second to a few minutes to obtain patient reaction from predefined stimulus). It is unclear how such ROI detection would perform in low-lighting or phototherapy environments, given that these studies were conducted in high lighting conditions using an RGB camera. An automated video-based neonatal face detection model robust to lighting conditions or partial occlusions is warranted for continuous patient monitoring in the NICU, and this thesis addresses this need.

In neonatal monitoring applications, skin segmentation is often used to segment the body of the patient from the background given that they typically use methods based on skin detection [28]. Antink *et al.* [47], however, leveraged a deep learning model and RGB + NIR data for body part segmentation (head, torso, left arm, right arm, left leg, and right leg). They proposed their own encoder-decoder architecture based on ResNet-50 [48] as the encoder and a modified version of fully convolutional network (FCN) [49] as the decoder. They only used patients where visible skin was observed, without occlusions from clothing, bandages, electrodes, or other equipment. They reported best results from RGB data with mean Intersection Over Union (IOU) of 82% for segmentation of the head, but the segmentation of the body revealed poor results (51% mean IOU) due to the difficulty in obtaining a good delineation of the patient against the background. This thesis explores limitations in obtaining proper patient delineation in segmenting the patient in various cluttered backgrounds, with varying patient coverage, and subject poses.

Asano *et al.* [50] also segmented body parts (head, body, arms, legs, and "other") from unclothed patients using a thermal camera as a step towards monitoring neonatal body temperature. They implemented a U-Net Generative Adversarial Network (U-Net GAN) and added a self-attention module (SA) that learns the relationship between pixels extracted in the map of body heat in the image. Best results were reported using the SA module in conjunction with the U-Net GAN to improve the segmentation of each body part, with a mean IOU of 70.4% for all body parts. While this study achieved promising results, segmentation of fully clothed patients remains to be investigated and that is addressed in this thesis.

Analysis of human motion has been explored in different fields, including camera surveillance, athletic performance, human-computer interaction, virtual reality and medical diagnosis [2], [51]. Specifically in medical applications, body movements have been analyzed to detect motion-related

diseases by analyzing the pattern of body and limb movements [1], [2], [33]. Recently, Cattany *et al.* have developed video-based motion detection algorithms to support diagnoses of motion-related diseases [2]. They have used multiple cameras in the NICU to detect rhythmic motions associated with apnea (from chest movements) and clonic seizures (from limb movements). Orlandi *et al.* identified atypical body movements for early detection of cerebral palsy [33]. An RGB camera was placed in a controlled environment to exclusively restrict the capture of patient motion in optimal conditions (no lighting variation, no intervention, no parent interaction, visible patient limb and skin, patient in supine position, no background objects nor equipment visible). Pertaining to the detection of moving limbs, Cenci *et al.* detected overall patient motion with a heat map corresponding to moving body parts; no specific identification/classification of moving limb was performed [1].

Beyond motion-related diseases, patient motion can also be analyzed to estimate respiratory rate from detected movements of the chest or abdomen area [1], [28], [29], [38]. Cenci *et al.* also detected specific patient movements from the manually selected chest area to estimate the patient's respiration [1]. Villarroel *et al.* monitored respiratory movements by tracking the change in segmented skin from the chest area during inspiration and exhalation periods [28]. Recently, studies have leveraged Eulerian video magnification to amplify movements from contraction and relaxation of the chest area [29], [38].

Patients in the neonatal intensive care unit (NICU) are typically monitored using several sensors to acquire physiologic data such as the electrocardiograph (ECG) and arterial oxygen saturation ($SpO_2$). Patient movements often lead to erroneous readings due to movement at the electrode-skin interface due to movement of the wires [52]. In the NICU, many false alarms are due to these motion artifacts, resulting in increased clinical interventions, alarm fatigue and parent concern [53]. Knowing when a patient is moving can inform physiologic signal quality estimates. Beyond movement detection, it is also necessary to classify the movement as being whole body or isolated to specific limbs. When combined with the known placement of sensors, improved signal quality estimates could be delivered. This thesis addresses this need by creating video-based patient limb movement detection and classification algorithms.

Other studies have aimed to avoid patient hospitalization by considering wearable home monitoring devices [54], [55]. Some commercially available home monitoring devices track sleep patterns from a camera paired with a wearable device for vital sign monitoring [56], [57]. These systems aim to provide continuous patient condition assessment; however, this approach is not always ideal for newborns or preterm infants. Their fragile skin and often precarious health condition make wearable sensors a suboptimal solution.

Recent camera-based monitoring studies have begun integrating depth-sensing cameras to obtain further information from the patient, especially in low lighting conditions [1], [2], [58], [59]. Among

RGB-D cameras, many past studies have utilized the Microsoft Kinect camera [2], [58], [59] due to its early availability, development SDK, and low price. These studies mainly evaluated perceived motion from the patient to monitor their respiration, detect apnea, or clonic seizures. Gleichauf *et al.* used a baby manikin with simulated breathing patterns to monitor respiratory rates using a Kinect camera [59]. The manikin was placed on a table and the camera was positioned 40 cm from the table to simulate the distance to the top of a closed incubator. Rehouma *et al.* conducted a similar respiratory monitoring study, however using two Kinect cameras placed 100 cm from a baby manikin in an open crib [58]. Although newborn respiration can be simulated with a manikin, other motion-related events cannot. Cattani *et al.* recruited real patients to detect body movements suggesting clonic seizures or apnea events using three Kinect cameras placed above an open crib [2].

Beyond detecting respiration or motion, the Kinect camera has also been used in conjunction with infrared thermal cameras to detect specific neonatal conditions. Shi *et al.* detected necrotizing enterocolitis (NEC), a condition seen in preterm newborns causing intestinal inflammation, using RGB-D data from the Kinect and thermal imaging to detect the abdominal region and identify thermal changes suggesting an early onset of the condition [6]. They used the Kinect camera for RGB-D imaging as a guiding tool for the segmentation of the abdominal area as an ROI. While the RGB stream is utilized for skin detection, the depth stream removes the background and extracts skeletal information. Rice *et al.* also investigated NEC detection by solely using thermal imaging for incubator patients [60]. They made a custom perforation in the top of the incubator, while covering the hole with plastic wrap for heat preservation during the 2-3 minutes of recording. In more recent studies, researchers have employed infrared thermal imaging for diverse neonatal monitoring applications such as respiration estimation by evaluating the change in body temperature in different ROIs from the patient [61], in the detection of various cardiovascular, abdominal, or pulmonary neonatal conditions observable from body temperature variation [62], in the identification of sleep propensity by analyzing distal skin temperatures before sleep [63], or in hypothermia prevention in preterm newborns post birth by monitoring the overall body temperature comparing peripheral and central temperatures from the foot and abdomen regions, respectively [64].

All device technologies presented here show their own usefulness for the intended neonatal application. Studies utilizing multiple cameras aim to capture the patient from different viewing angles or complement each modality for robust patient pose and ROI detection. The need to mount multiple sensors in the NICU room is, however, impractical for deployment in complex NICU room environments, especially when paired with the Kinect camera given its size, weight, cost per additional device, and the cumbersomeness of installing multiple cameras in the room. This thesis addresses the

effectiveness of using depth-sensing technologies by demonstrating how a single consumer-grade RGB-D camera can be suitable for non-contact neonatal patient monitoring.

Video-based neonatal patient monitoring is a complex task due to the challenging environment surrounding the patient. Such challenges include frequent, abrupt changes in lighting conditions, patient self-induced motion, externally-induced patient movement due to interventions, or the hospital bed may be encumbered in beddings and equipment; each of these can severely impact the outcome of a video-based patient monitoring system. The following section presents background on machine vision/computer vision methods, relevant to this thesis to address such challenges from the NICU environment.

## 2.2 Machine/Computer Vision

When analyzing images, multiple computer vision techniques can be used, depending on the complexity of the data and problem. For example, for detecting a stop sign, traditional image processing techniques may be used to detect red-colored and octagonal-shaped objects [65], as depicted in Figure 2.1. Previous studies were able to detect stop signs by extracting the red component from an RGB image and measuring the aspect ratio of that object. This approach can however be flawed in different lighting conditions, for example the red component would be hard to detect at night. Additionally, depending on the camera position, the stop sign's perceived shape could significantly differ. To address the latter, feature-based methods [66] used features from the letters 'STOP' that are complementary to the shape of the sign. More recently, more sophisticated computer vision approaches [67] have used a combination of these variations to address lighting variations and viewing angles. Additionally, deep learning methods [68] are generalizing the detection of stop signs in various countries by further expanding the feature set to be detected despite partial occlusions, weather condition, or written language. All above-mentioned techniques are independently successful; however, it depends on the application to decide which approach is most appropriate.

This research considers a comparable range of computer vision techniques, where traditional image processing methods are implemented in scene understanding using color space transformations for classifying brightness levels and ongoing phototherapy (Chapter 4.1-4.2), as a pre-processing step in face detection of phototherapy patients (Chapter 5.1), as a refining step using image filtering in patient segmentation (Chapter 5.2), in comparing applicable patient motion detection methods (Chapter 6.1), and as a face detection application using a skin-pixel face mask in neonatal face tracking (Chapter 6.3). For more complex tasks requiring the detection of specific features and larger variations in the dataset, a deep learning approach is implemented in most remaining work in this thesis.

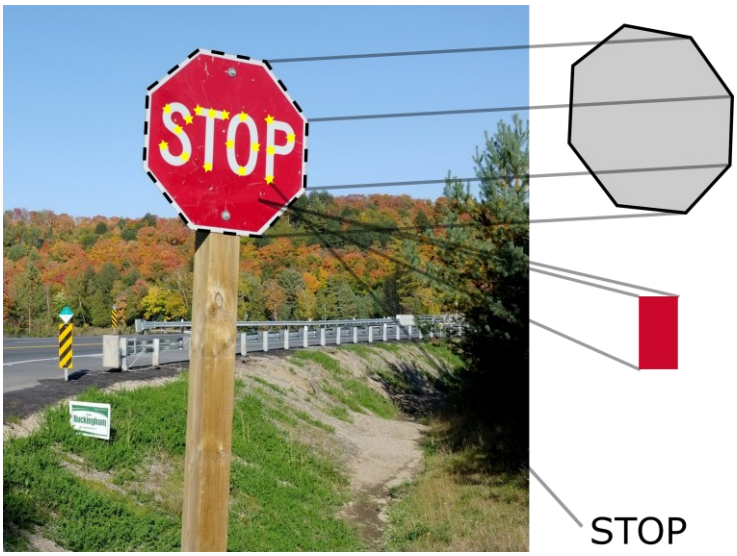



Figure 2.1: Computer Vision for Stop Sign Detection. Application of traditional image processing or machine learning for detection of a stop sign based on its red color, octagonal shape, or text-based features.

Recent studies have leveraged CNN models to perform various image analysis tasks including image classification [69], object detection [70], and scene recognition [71]. These models have mainly benefitted from large repositories of labelled data and from computational improvements in software (e.g., algorithms) and hardware (e.g., GPU). Numerous studies have used the ImageNet dataset consisting of over 14 million annotated images, where a subset of ~1 million images across 1000 different classes is often used to train networks from multiple classes such as *cat*, *guitar*, *plane*, and *cup* [72]. This benchmark dataset has underpinned the development of practical solutions for a range of application areas, such as face detection algorithms [73] or pedestrian and vehicle detection for traffic surveillance [74]. Despite evidence of broad applicability of models trained on these data, the NICU environment significantly varies from these applications (different appearance of adult vs. newborns, specific hospital equipment, etc.). In medical imaging, transfer learning has primarily been used on PET, CT, X-ray, or MRI images for the detection or diagnosis of diseases [75]–[77]. More specifically, the VGG-16 architecture has been widely used in melanoma screening [78], skin lesion classification [79], and cell nuclei classification from histopathology images [80].

A model specifically tailored to the physical neonatal environment is therefore warranted, and this thesis leverages transfer learning on state-of-the-art models to do so. In its most basic definition, *transfer learning* describes a process in which some pertinent information is passed from a certain source domain to a target domain [81]. In cases where both domains have labeled data, this procedure called *inductive transfer learning* is often performed by transferring pre-learnt weights from a model to another domain [82]. Both domains are assumed to be similar, so that key features can be seamlessly reused from a source task to perform a new target task, instead of learning all model weights from scratch. When

implementing a new model for a specific task, a transfer learning approach is beneficial, given that key features have already been learned from application areas that are data-rich (abundance of data easily labelled) which can be transferred to applications with insufficient and costly data. Our study falls into the latter category, where data are extremely difficult to acquire and label due to lengthy data collection procedures, equipment cost, data storage and management, ethics protocols, low subject recruitment rates, and data security and privacy, to name a few. This research aims to bridge the gap between data scarcity and NICU scene understanding through knowledge transfer proficiency.

This thesis demonstrates how NICU complex scenes are analyzed by leveraging state of the art deep learning models and image processing algorithms, and using transfer learning to finetune models thereby further investigating and overcoming observed failure cases. In some work, an image processing technique is complementary to the complex finetuned deep networks to achieve refined results. We also demonstrate how a machine vision system can be built for a non-contact neonatal monitoring application. In the following sections, background of each building block of this non-contact monitoring system is presented, and we introduce how various machine vision applications are leveraged to accomplish each task.

## 2.3 Scene Understanding

Automated scene understanding can be achieved through classifying the various contexts contained within it. For instance, the scene interpretation “a pine forest in winter at night” can be obtained by identifying four different classes: tree type, tree density, season, and time of day. Each context provides valuable information regarding different aspects of the scene. This thesis aims to provide analogous information from the neonatal intensive care unit (NICU) environment to assist clinical staff in patient care and monitoring, as illustrated in Figure 2.2.


Figure 2.2: Scene Understanding Examples. The scene can be interpreted from a landscape (left) or clinical setting (right) based on context identified in the image.

In the NICU, critically ill newborns often require a high level of care, which is carefully documented by the patient's nurse. Routine care can be administered multiple times during a single shift, including feeding, diaper change, temperature check, and recording of vital signs. Maintaining clear and complete patient care documentation enables communication between health care providers in an NICU team. During a shift change, the incumbent nurse must relay the patient's information to their successor, thereby enabling continuity of care. Clinical staff must carefully and frequently chart interventions, routine care events, patient status updates, and any physical assessment performed on the newborn while indicating the date and time and any significant observation [19]. Unfortunately, this protocol can cause substantial workload for the nurse to document patient condition, routine care events, and interventions, which can lead to overload when nurses have multiple patients under their care or when some patients require high levels of care due to their precarious condition. Time spent charting is time not spent caring for the patient. Furthermore, when clinicians are overloaded, the charting process becomes retrospective and events are often inaccurately recorded [19], [83]. This is a common problem in hospitals and previous studies on incomplete nursing documentation have highlighted patient safety concerns [19], [84]. A recent study highlighted hypoglycemia, hyperbilirubinemia, and sepsis among high-acuity patients commonly affected by inadequate charting [19]. Recent studies digitizing portions of medical health records have shown considerable decrease in nursing workload. Hence, development of a system to help with charting of events of clinical significance is warranted [85], [86].

We herein propose a camera-based scene recognition model for analyzing context variables perceived in the NICU during patient care. A simple generator for scene summary text is created as a proof-of-concept. This represents a step towards automatically assisting clinical staff in the charting process.

An image processing approach is used to differentiate between different lighting conditions, by extracting different color space components in the image. Color space transformations are typically useful when standard RGB components are not sufficient to obtain required information from the image [87], [88]. A deep learning approach is used for the other contexts by leveraging transfer learning [81] of pre-learnt network weights from a VGG-16 [89] image classification network, pre-trained for a different data-rich classification task. This research investigates if pre-learnt weights are transferable to multiple context classification tasks relevant to patient care in the NICU. A sensor fusion approach is investigated to make optimal use of color and depth images collected from the Intel RealSense SR300 RGB-D camera. We also explore transfer learning in this context, examining whether network weights from a color-based task can be usefully transferred to an RGB-D model. We examine two fusion approaches: image fusion or network fusion. Previous multi-modal studies have merged color and depth data directly within the image [90], [91] or at some point within the network [90]–[92]. Given the differences between the source and target tasks in our application, we herein examine all depth fusion

schemes to evaluate transfer efficiency for each target classification task.


Figure 2.3: Context Classification Example Data. Contexts are bolded and classes are italicized.

Five binary classification tasks are performed to determine the lighting conditions (high or low), phototherapy treatment (therapy or natural light), ongoing intervention (nurse present or absent), bed occupancy (patient absent or present), and coverage (patient covered by beddings or not). Sample data from context variable are illustrated in Figure 2.3. Knowledge of these five contextual variables can be leveraged in subsequent computer vision tasks (e.g., applying non-contact vital sign monitoring only when the patient is present, or adapting a vision-based algorithm based on lighting conditions). Each classification task is discussed below:

The **"lighting"** and **"phototherapy"** variables indicate periods where a non-visible-light imaging modality may be preferred, such as the depth or near-infrared modalities available on the SR300 camera. For many patients, especially those who are premature, attempts are made to limit sensory stimulation and promote rest; thus, lighting conditions are kept low whenever possible. Some patients are also affected by jaundice and this condition is treated using phototherapy treatments where the patient is subjected to visible light in the 460-490 nm range [44]. Changes in lighting condition or interruptions to phototherapy can be detected and directly recorded into the patient's chart, thereby increasing temporal precision of annotations while reducing workload on the clinical staff.

In order to reduce the number of sleep interruptions, routine care events typically occur consecutively within a block of about 5-30 minutes, depending on the patient's status. Realistically, clinical staff often record all events retrospectively at the end of the intervention block, increasing the chance of a charting error or omission. Automating the detection of each routine care or **"intervention"** event would be helpful in this regard. As a proof-of-concept, this thesis aims to detect the presence of clinical staff within the field of view of the camera as an indication of ongoing intervention.

It is recognized that many intervention events occur outside of the bed and field of view of the camera, such as weighing the patient or breastfeeding. Using the **"bed occupancy"** classification task, the rough

timing and duration of these events can be automatically captured for documenting out-of-bed care.

In conjunction with the "**intervention**" classification, when the patient is present, the **"coverage"** class is useful for identifying interventions, given that they typically start by removing blankets and undressing the patient, and end by dressing and replacing blankets. Specifically detecting when the patient is covered can also inform on motion detection models from free and uncovered limbs. Determining "coverage" is less useful for patients in temperature-controlled incubators since they normally remain uncovered for optimal temperature control and observation of patient tone and activity.

While these five context variables are useful for subsequent computer vision analysis, they are also directly useful for producing textual summaries of the patient environment. As a proof-of-concept, this thesis combines the five context variables to produce simple textual summaries, such as "Between 16:05-16:20 the *uncovered* patient in *low light* received an *intervention*".

### 2.3.1 Clinical Intervention Detection

During continuous monitoring, clinical interventions can sometimes pose a problem in the development of non-contact monitoring systems since clinical staff can naturally occlude portions of the patient. In many cases, intervention periods are actually excluded from analysis [24], [28], [29], [33]. One recent studies has, however, detected interventions from color video data [28]; this thesis shows how this task can be achieved from a single image and how RGB-D data are robust to NICU lighting variations for continuous and reliable intervention detection. To detect clinical interventions, this thesis investigates how the arm/hand of the nurse is best detected/observed at a close distance to the camera from RGB-D data.

Clinical interventions may represent pivotal moments in newborns' continuous care in the NICU; further investigation of these events is therefore warranted. We previously explained routine care events and the significance of charting contemporaneously: this section focuses on bottle-feeding events.

Previous studies have established the importance of monitoring the oral feeding process (breast-feeding and bottle-feeding) in newborns less than 6-7 months old [93]–[95]. In fact, newborns' acquired feeding skills are crucial in these first few months as they directly impact their nutrition and brain development, especially for those born prematurely or those in critical health condition. Assessment tools have then been implemented to help guide nurses or parents in evaluating the newborn's feeding skills [96]–[98]. This includes evaluating the newborn's state when bottle-fed, such as muscle tone, readiness, sucking, swallowing, breathing, fatigue, and oral-motor patterns. These assessment tools can help in deciding when the patient can safely be discharged from the hospital, and for transitioning to solid food in the NICU or at home [95].

This research focuses on video-based detection of bottle-feeding interventions as a step towards automated clinical documentation in the NICU, while supporting neonatal oral feeding assessments. To this end, we aim to identify in each image whether a bottle-feeding event is occurring or not. Recent neonatal monitoring studies have reported detection of clinical interventions; however, no specific detection of a particular intervention types, such as feeding, has been investigated [28].

We herein leverage transfer learning on a pretrained neural network, VGG-16 [89], to classify "bottle-feeding" vs "no-feeding" events in an intervention image.

## 2.4 Patient ROI Detection

### 2.4.1 Object Detection

Over the last few years, an increased interest in human detection and face detection models have been perceived motivated by the need to identify, quantify and track individuals observed in various scenes [99]–[101]. In the last few years, object detection models have improved in both speed and accuracy. Several families of object detection models are described below.

The R-CNN family of two-stage CNN object detectors, which include the R-CNN [102], Fast R-CNN [103], and Faster R-CNN [104] models, leverage region proposal methods that suggest areas of the image where an object of interest is suspected to reside, followed by an object localization stage that uses bounding box regression.

Instead of relying only on selected proposed regions of the image, the You Only Look Once (YOLO) family of object detectors look at the entire image and simultaneously generates class probabilities within each predicted bounding box [105]–[109]. The object of interest corresponds to the highest probability region, thus only requiring to "look once" at the image before making a prediction. Such one-stage object detectors have gained popularity due to their fast computation, especially in real-time applications. Redmon *et al.* created three versions of this YOLO architecture from 2015 to 2018, by incrementally improving the model's speed and accuracy [105]–[107]. In the past couple of years, other researchers have extended Redmon's work to achieve even better and faster real-time performance with YOLOv4 by leveraging additional methods used during training and testing. Among them, significant detection improvement was noticed using a new mosaic data augmentation which creates a tile of four training images thereby helping the model detect small objects while reducing the required mini-batch size during training. Compared to the mean square error (MSE) used in YOLOv3 for bounding box regression, YOLOv4 [108] uses a complete IOU (CIOU) loss which compares the predicted and ground truth bounding boxes area by considering the distance between each center points and aspect ratio, in addition to evaluating their overlap from traditional IOU. Compared to the four previous versions, Glenn Jocher introduced YOLOv5 [109] implemented on PyTorch instead of the Darknet framework, thereby

allowing the implementation of models of various sizes including small and lightweight ones for easy deployment to mobile devices. YOLOv5 also introduces a Focus Layer made up of YOLOv3's first three layers only to reduce layers, parameters, and CUDA memory, while improving speed during forward propagation and backpropagation. Overall, YOLOv5 is faster, more lightweight, and more accurate among the entire YOLO family.

Other prominent recent object detectors include the single-stage object detector RetinaNet which introduces a new Focal Loss optimization that focuses on extreme foreground-background class imbalance during training [42], the EfficientDet [110] model that uses the EfficientNet [111] classifier as a backbone for model scaling, and the DEtection TRansformer (DETR) network that leverages a CNN and transformer encoder-decoder architecture to perform end-to-end object detection with bipartite matching for generating direct predictions [112].

To train these above-mentioned detectors, research groups have often relied on the Pattern Analysis, Statistical Modelling and Computational Learning Visual Object Classes (PASCAL VOC) dataset [113] and/or the Microsoft Common Objects in Context (MS COCO) dataset [114]. These two object detection benchmark datasets were created for various object recognition challenges including classes such as *person*, *cat*, *bicycle*, etc.

More recently, the need to detect the individuals' face has surged due to interest in camera surveillance and facial analytics for detection, identification, and verification systems [115], [116]. These studies have leveraged state-of-the-art object detectors for face detection algorithms, mostly trained and validated on healthy upright standing adults, and reused by various other researchers for different tasks. In our case, blindly reusing such models with our neonatal dataset comes at a cost given the essential difference in the overall appearance of adult and newborns' face and body. A re-adaptation of state-of-the-art face detection models is implemented here, to account for the challenging data from neonates in the NICU environment.

### 2.4.2 Face Detection

Neonatal face detection in a clinical setting can be a difficult task due to complex NICU including varying patient poses, facial occlusions, and lighting environment, as depicted in Figure 2.4.

Generally, detecting the facial area is often performed in three different ways: the detection of the entire face enclosed within a bounding box (face detection), the detection of the geometric structure of the face outlined by specific landmarks (face alignment), or the detection of every pixel pertaining to the person's face (face segmentation). All these applications are depicted in Fig 2.5.


Figure 2.4: Face Detection Difficulty in Complex NICU Scenes.

Facial alignment is typically applied using 5-point landmarks including the center of left eye, center of right eye, tip of nose, left corner of mouth, and right corner of mouth [7], [117]. In other cases, finer facial structure is extracted with 68-point landmarks including eyebrow line, eye contour, length and width of nose, upper and lower lip contour, and jawline [118]. In face segmentation, the whole face is either segmented as a whole [119] or is segregated into different facial regions (*e.g.*, eyes, nose, mouth, skin, hair) [120]. Face alignment and segmentation are particularly useful in further facial analysis applications such as face recognition or facial expression detection; however, they are more difficult tasks to achieve compared to detecting bounding boxes. Only face detection results are investigated quantitatively in this thesis, while facial alignment methods are evaluated qualitatively.

To train and evaluate face detection models, several benchmark face image datasets are available. *Face detection benchmark datasets* include:

- WIDER FACE [121]: Images include faces with variations in scale, pose, occlusion, expression, makeup, and illumination (393,703 annotated faces from 32,203 images). Different subsets are included as Easy, Medium, and Hard data, based on the increasing level of difficulty to detect the face due to varying scale, occlusion, and pose.
- FDDB [122] (Face Detection Dataset and Benchmark): Images including faces with variations in occlusions, poses, resolution, and out-of-focus faces (5,171 annotated faces from 2,845 images).

*Face alignment benchmark datasets* include:

- AFLW [123] (Annotated Facial Landmarks in the Wild): Real-world images including faces with variations in pose, lighting, expression, ethnicity, age, and gender (25,993 annotated faces from 21,997 images).

- 300-W [124] (300 Faces-In-The-Wild): In-the-wild images from indoor and outdoor scenes including variations in identity, expression, illumination, pose, occlusion, and face size (600 annotated faces from 399 images).

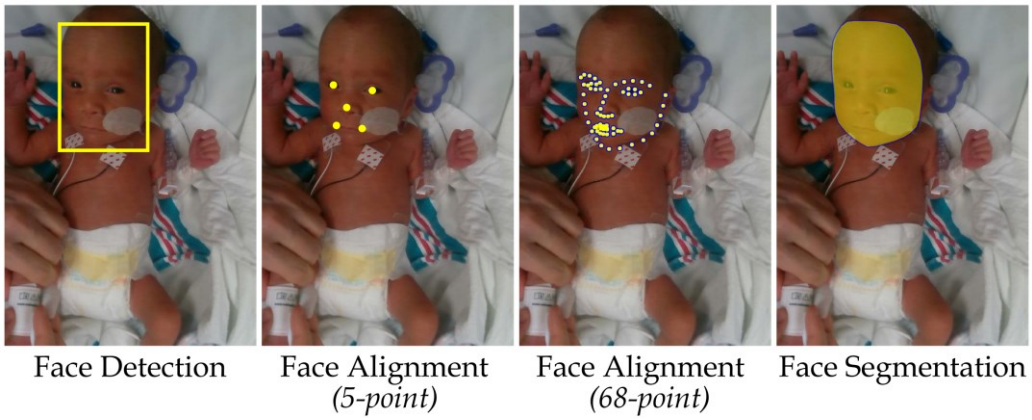



Figure 2.5: Visual representation of different face detection techniques.

Among state-of-the-art face detection and alignment models, the Multi-Task Cascaded Convolutional Network (MTCNN) [117] has a cascaded CNN architecture of three different networks: (1) A Proposal Network (P-Net) where several facial regions in the image are proposed as candidates; (2) A Refinement Network (R-Net) where all candidate regions are rejected or retained for further analysis by the following network; (3) An Output Network (O-Net) where remaining candidates are further refined to obtain a final selected region corresponding to the face region with landmarks. At each stage, bounding box regression vectors and non-maximum suppression are computed to obtain corresponding outputs. This model was trained on three different datasets (WIDER FACE, FDDB, and AFLW) and performs joint face detection and alignment with resulting 5-point landmarks.

As opposed to MTCNN's regression-based approach, the Convolutional Experts Constrained Local Model (CE-CLM) uses a model-based approach where the appearance of facial landmarks are computed to obtain an output [118]. Traditional CLMs use local detectors to model each facial landmark and shape them from constrained optimization techniques. Although this approach can be robust to occlusions or subject pose (especially faces in profile), it is severely impeded by complex variation in facial appearance such as facial hair, makeup, or accessories. The Convolutional Experts Network (CEN) can model such adult facial variations using a mixture of experts. While most of these complex variations should not occur in NICU-based data, a different set of NICU-specific variations exist, thus warranting model fine-tuning for a neonatal population.

The CE-CLM framework can be considered a three-fold process; first, a face detector is applied to obtain landmark positions (CLM); second, each landmark is accurately localized (CEN); and third, all landmarks are properly aligned using point distribution models to create a 68-point facial landmarks.

CE-CLM can use different model architectures for its backbone including cascade detectors, tree-structured models, and more recently the MTCNN model. The CE-CLM model was trained on four different datasets (300-W, 300-VW, IJB-FL, and Menpo Challenge), that were selected due to the presence of challenging environment such as varying lighting, occlusions, different image quality, varying poses, profile faces, and video data [118].

Aiming to obtain dense face localization, the RetinaFace model [8] is a single-stage detector that uses a multi-task network for face classification, face box regression, 5-point facial landmark regression and 1k 3D vertices regression. The RetinaFace model uses the ResNet-50 model as a backbone for generating a feature pyramid, applies a context module to each pyramid level to increase the receptive field to help detect smaller faces, and uses different anchor sizes at each level to detect faces of varying sizes. A multi-task loss is computed as a linear combination of the loss of each corresponding task. Deng *et al.* demonstrated that each of these tasks can contribute to one another. The RetinaFace model is trained and tested on the WIDER FACE dataset for face detection evaluation, with an emphasis on the Hard subset.

Most recently, the YOLO5Face [7] has redesigned the YOLOv5 [109] object detection model into a face detector. Important modifications were implemented such as adding a 5-point landmark regression to obtain facial alignment, reducing the kernel sizes in the spatial pyramid pooling (SPP) block to enable detection of smaller faces, replacing YOLOv5's Focus layer with a Stem block to improve generalization and reduce computational complexity, and tailoring the data augmentation techniques to face detection. Qi *et al.* have provided different YOLO5Face models based on various YOLOv5 backbones for computer or mobile device applications. The overall loss function of YOLO5Face extends from YOLOv5 as a compound loss of bounding box location regression loss, confidence loss, classification loss, plus a Wing loss for the added landmark regression. YOLO5Face used WIDER FACE to train and test the face detection task.

### 2.4.3 Semantic Segmentation

Semantic segmentation of a scene consists of partitioning an image into distinct regions by assigning each pixel to a meaningful classification label, as depicted in Figure 2.6. Multi-segmentation applications extend this concept to cases where more than two segments are obtained in a given frame. Such approaches have been applied to autonomous vehicle systems by labeling a scene [125], segmentation of common objects in a given scene [126], and to biomedical imaging, such as the recognition of neural membranes in electron microscopy images [127]. With increasing availability of computational power and training data, deep learning methods have been successfully used to accurately recognize and segment objects within heterogeneous scenes.

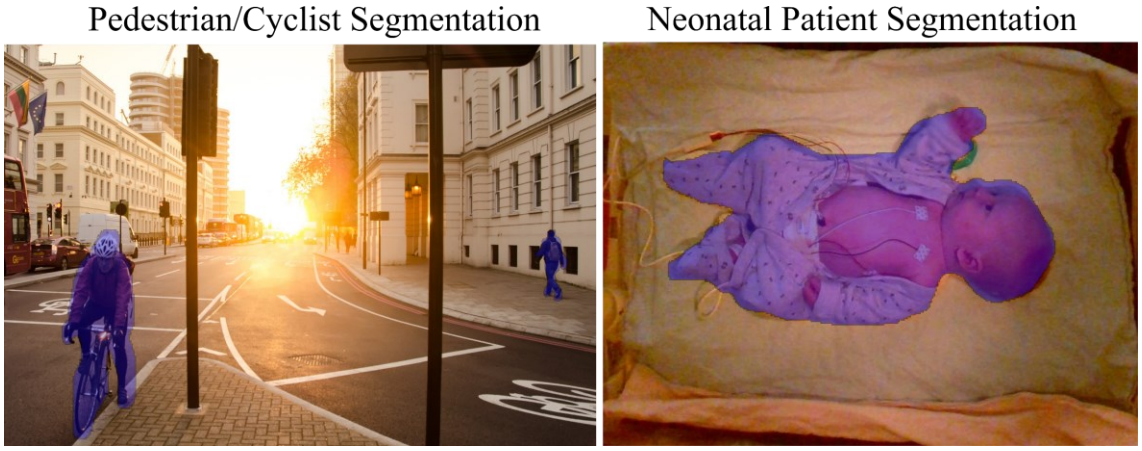



Figure 2.6: Semantic Segmentation Examples. Purple-colored pixels represent the segmented person. Left) Person segmentation application typically used for detecting pedestrian or cyclist on the street for autonomous vehicle applications. Right) Segmentation of lying newborn inside a neonatal bed for patient monitoring applications.

This thesis leverages deep learning to segment patients from the background in actual overhead video images from a NICU environment. Using a semantic segmentation technique, we label pixels in the frame as either belonging to the neonate or the background class, where the background comprises bedding, sensors, wires, and equipment. Correctly segmenting patients from background is a critical prerequisite for subsequent machine vision tasks such as detecting self-induced motion, categorizing clinical events performed on the patient, and detecting conditions such as apnea or clonic seizures [2]. Additionally, obtaining a segmentation of the patient could provide accurate ROI information for other applications, such as the estimation of physiological parameters as part of an unobtrusive patient monitoring system [128]. In our scenario, we aim to segment a neonatal patient laying on bedding from overhead video, whereas recent person detection methods have mostly been implemented on walking or standing adult subjects, generally for video games or autonomous vehicle applications.

Several image processing techniques have been applied to segmentation of persons addressing challenges such as cluttered backgrounds, partial occlusions, and lighting variations. A successful technique must be robust to such situations, especially in the complex NICU environment.

Li *et al.* aimed to tackle the cluttered background issue by exploring a data-driven graph-cut method [129]. This consisted of using top-down inferences for body pose estimation and bottom-up cues to fine-tune the segmentation using visual cues. By initially detecting the face of the person, it consequently guides the detection of the torso, followed by the detection of the limbs using a human kinematic tree. Although strong performance is reported by the authors, this algorithm can only be employed in controlled environments and hinges on an initial face detection step that often only functions on frontal poses. Another study generated blob models by forming a network of vertices mimicking the human body skeleton, by considering each body part's plausible range of motion [130]. This method may be

unsuccessful in presence of occlusions or absence of such intrinsic body parts. From *a priori* knowledge of the shape of the object of interest, Borenstein and Malik developed a method that pre-emptively generates a model of the target segmentation [131]. Theoretically, this concept appears to be beneficial for rapid and accurate execution of a model made specifically for the given application. Lin and Davis [132] adopted this technique by creating shape models for the detection and segmentation of human bodies leveraging a database of human poses and template-matching accordingly. This method is not typically applied to segmentation of complex object shapes and positions, such as humans, and is more often adapted for limited or constrained poses, body shape, and appearance such as those seen in animals including horses or cows [129], [131].

More recently, researchers have explored machine learning or deep learning approaches using combinations of human detection, pose estimation and/or face detection to obtain a segmentation of the people in a given scene [133]. From a large and complex database, CNNs can examine and learn this varied data to provide appropriate results. Deep learning semantic segmentation can be perceived as two simple tasks: classification and localization. These two problems can be solved when considered independently, and many CNNs have been developed to tackle these individual tasks for popular competitions such as the ImageNet Large Scale Visual Recognition Competition (ILSVRC) [134] and the PASCAL VOC Challenge [135]. Seeking to investigate the influence of depth of a network on large-scale image recognition tasks during the 2014 ILVSRC, the Visual Geometry Group (VGG) at Oxford University developed very deep CNNs containing 16 and 19 layers, resulting in the now widely used VGG-16 and VGG-19 models, respectively [136]. Results from the VGG-16 model showed significantly improved performance in image classification compared to previously published networks. Consequently, one could leverage this model to achieve pixel-level classification as part of a more complex segmentation network. Badrinarayanan *et al.* [125] explored this further, where they altered the last fully connected layers of the VGG-16 (*i.e.* the layers responsible for delivering the detected class), by transforming them into a decoder network responsible for assigning positions of each classified pixels based on their corresponding indices within the image. The resulting model, labelled SegNet, can therefore fully exploit the pixel-categorizing abilities of the VGG-16 as an encoder section of the network, before performing semantic segmentation from the decoder section of the network [125]. A final softmax layer ultimately generates the output image with two distinct classes: foreground and background.

Other models have also exploited transferred capabilities from object detection to semantic segmentation, notably with the Mask R-CNN model [10] leveraging the Faster R-CNN object detector [104]. The Mask R-CNN model copies the Faster R-CNN model while adding a FCN branch, parallel to classification and bounding box regression to simultaneously perform pixel-wise semantic segmentation

and object detection. With this FCN branch, each ROI predicted from the RPN network is upsampled to create masks. The Mask R-CNN can then perform instance segmentation where multiple instances of the same object can be detected in the image. This model is thus useful for semantic segmentation of multiple objects vs the background.

Due to best performance reported on deep learning networks, state-of-the-art segmentation models (SegNet and Mask R-CNN), are evaluated in our neonatal population. Considering that patients in the NICU are often encumbered with diverse wired sensors and are surrounded by bedding and equipment, a custom solution is warranted.

## 2.5 Motion Detection

### 2.5.1 Human Motion Detection

Different computer vision methods have been utilized to perceive motion. Notably, human motion is often detected using background subtraction techniques [137]. Such methods select an initial frame as the background and update this frame only if the detected motion exceeds a predefined threshold. Zhang and Liang [137] investigated several approaches to selecting an initial frame. Furthermore, they updated the background using a dynamic threshold since their application involved real time human motion estimation under varying lighting conditions. Bobick and Davis describe a method to identify motion based on temporal intensity differentiation [138]. Their method creates a binary motion-energy image (MEI) indicating the location of detected movement within the frame. Secondly, they presented a greyscale motion-history image (MHI) based on the recency of motion. A third approach to movement detection is the optical flow which exploits the observed intensity patterns in an image due to relative motion created by the object of interest [139]. Consequently, a motion vector field (MVF) quantifies the magnitude and direction of the detected motion. Bauer and Pathirana [140] leveraged optical flow, in conjunction with object state parameters, to estimate movement with a stationary or moving camera.

We herein propose to address common monitoring issues by leveraging a motion detection and classification algorithm for non-contact, non-invasive and unobtrusive monitoring of patients in the NICU. By comparing the three above-mentioned detection techniques, we formulate an integrated solution comprising pertinent components from each.

Human motion is often detected in consumer-grade camera systems using computer vision approaches. For example, background subtraction techniques use an initial frame to define the background and movement represents a change from that frame. The background frame can be optionally updated if the detected motion exceeds a predefined threshold [137]. Additionally, temporal intensity differentiation [138] and optical flow [139] techniques are popular approaches to locate the detected movement within the frame, given that they provide the direction and magnitude of that movement. This thesis leverages the optical flow technique as one computer vision approach in detecting motion.

More recently, studies have leveraged deep recurrent neural networks for motion detection and video captioning [141], [142]. More specifically, long short-term memory (LSTM) networks have been used, given their ability to retain long-term dependencies in sequence data [143]. In video classification applications, these networks can differentiate between multiple complex scenes by first using a pretrained CNN to extract key features from each image in a video sequence, before classifying that sequence using a recurrent neural network. For the latter component, many studies involving human motion activity detection and recognition have used LSTMs [144], [145]. This thesis employs an LSTM network for detecting motion events, specific to neonatal patient monitoring.

### 2.5.2 Face Tracking

Using a camera placed above the patient's bed, video data can be captured and analyzed in real-time to estimate physiologic signals such as heart and respiratory rate. The accuracy of video-based physiologic monitoring solutions is often predicated on having a well-defined ROI, usually corresponding to the patient's face [21], [40], [146]. Often, the ROI is selected manually at the beginning of data collection by drawing a bounding box over the patient's face and keeping this position constant throughout the monitoring process [5], [39], [147]. However, in realistic scenarios, this approach is not suitable due to patient movements, occlusions by limbs and/or beddings, and temporary clinical intervention events. This thesis proposes to address this issue by detecting the face of the patient and tracking it over time, thereby ensuring a continuous correct definition of ROI.

While the effectiveness of face tracking algorithms has been established for adult users [148]–[150], it is unclear whether such approaches will be effective on neonatal patients in the NICU. Common face tracking algorithms have been trained on a plethora of adult faces, often using the appearance of the eyes, nose, and mouth as a template to guide the feature extraction and tracking process. Newborns' facial appearance are substantially different, thereby making this task more complex to achieve. A modification tailored to newborn faces is therefore warranted. A re-adaptation of state-of-the-art face detection and tracking models is implemented here, using data from neonates in the NICU.

An ROI tracking system can be considered three-fold; the detection of an ROI, the identification of useful features as landmarks within this ROI, and tracking landmarks as they change in position over time to arrive at a dynamic ROI [148], [151], [152]. A successful tracking algorithm must accomplish all three steps. A robust tracking algorithm would additionally indicate when the algorithm might fail and be robust to temporary interruptions to correct tracking. Common challenges leading the tracking process to fail include changes in subject pose due to movements, change in lighting conditions, and temporary occlusions. Several recent tracking algorithms overcame these challenges by adopting object detection for *tracking by detection* [153], [154], where the object is continuously detected in successive

frames, as opposed to *tracking by displacement* of identified landmarks. This thesis develops a new tracking algorithm by combining both techniques for neonatal face tracking. To this end, we implemented a robust ROI tracking algorithm for patients in the NICU during periods of patient rest, head motion, and occlusion from limbs and/or beddings. This thesis enables subsequent video-based patient monitoring applications that require the face area as ROI.

## 2.6 Vital Sign Estimation

Non-contact estimation of heart rate (HR) has been investigated using diverse sensor modalities and under different conditions [155]–[160]. Most studies aimed to create non-contact, unobtrusive, and non-invasive systems for continuous real-time patient monitoring [161], [162]. Successive algorithms have sought to achieve higher accuracy and improved implementation efficiency. Among them, EVM has shown to be highly effective leading to promising results in various applications, notably for heart rate estimation [156], [163], [164]. EVM can enhance the visual effect of time-varying blood flow in a person's face, or other ROI, by amplifying color channels in the video and thereby estimating the patient's HR.

The EVM approach was developed at the Computer Science and Artificial Intelligence Lab of the Massachusetts Institute of Technology (MIT CSAIL) [20]. The purpose of EVM is to amplify subtle or low-amplitude variations in video that are often invisible to the naked eye. Input video is first decomposed into spatial frequency bands using a Laplacian pyramid decomposition. A temporal filter is then applied to each spatial frequency band to further magnify faint color variations, while suppressing low-frequency noise such as motion induced by respiration. The magnitude of each pixel now provides a time series corresponding to variations in color or intensity over time. One or more frequency bands of interest are then selected using band-pass filters. Selection of the passbands is application-specific and typically requires prior knowledge of the expected frequency range of interest. The resulting band-passed signal is then multiplied by a magnification factor before re-adding this new magnified signal to the original one. The video is then reconstructed by collapsing the spatial pyramid, producing an EVM-enhanced output video.

EVM has been demonstrated to amplify changes in color and/or motion. It has largely been applied to color (RGB) video, although it has recently been applied to thermography video for HR estimation [155]. A recent study examined the application of EVM to the multi-modal Intel RealSense SR300 camera, which measures color, depth, and NIR information. In that study, the authors applied EVM to the NIR and RGB streams for HR estimation, but only used the depth stream for locating regions of interest on the subject's face and did not apply EVM nor estimate HR from the depth stream directly. Another study evaluated the estimation of heart rate under low lighting condition using an infrared night vision camera

and the EVM framework [160]. The authors emphasized the advantages of infrared cameras, in that they tackle the issue of using optical cameras in darkness and avoid the high cost of thermal cameras.

In this thesis, we examine the application of EVM to all three streams of the multi-modal Intel SR300 RealSense camera; RGB, depth and near-infrared. Data fusion is used to arrive at a consensus estimate of heart rate, leveraging all three sensor modalities. Furthermore, we address the issue of selecting the optimal passband within EVM. The original proposed use of EVM for HR estimation required the passband to closely match an *a priori* known subject heart rate [40]. Subsequent studies have suggested the use of a broader passband, possibly followed by repeated application of EVM with a narrower passband focused on the suspected HR [155], [157], [164]. Furthermore, little guidance is given on how to set the passband in the absence of *a priori* knowledge of the expected HR, and this thesis addresses this issue.

## 2.7 Multimodal Analysis

Multiple studies have used vision-based devices, other than traditional RGB cameras [90], [91], [100], [165]. More specifically, depth-sensing cameras can capture important details from the scene due to its 3D representation. An RGB image with corresponding depth representation is illustrated in Figure 2.7.

RGB-D data has often been used for semantic segmentation of indoor or outdoor scenes [90]–[92] to accurately detect and recognize multiple objects from a crowded scene. From outdoor scenes, studies have detected pedestrians and vehicles for obstacle detection with autonomous vehicles [101]. Other studies have paired RGB with thermal data to further detect pedestrians, especially at night [100]. Since depth sensors can provide near-infrared (NIR) data, some studies performed in low lighting environments have paired RGB with NIR data for various applications [165], [166].

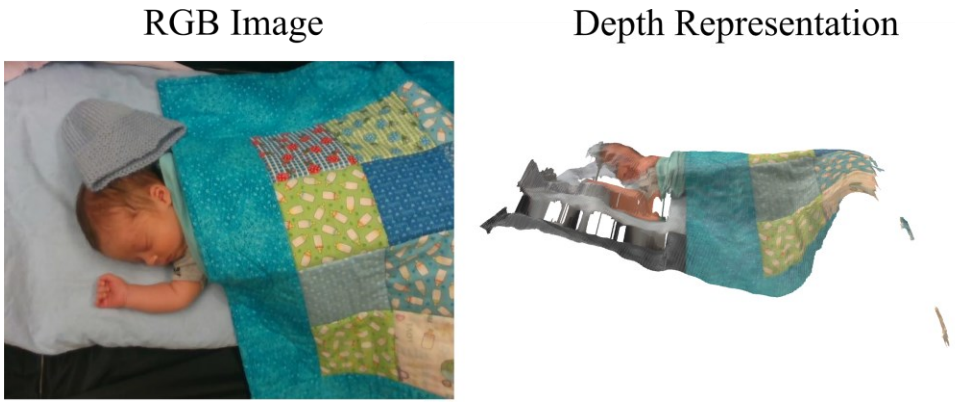



Figure 2.7: RGB Image with Corresponding Depth Representation. The 3D image captures one side of the entire estimated depth, as viewed from the patient's anterior coronal plane.

In comparison with the ImageNet dataset [72], no such large-scale RGB-D dataset currently exists. In 2012, the NYUv2 [167] dataset was created using the Kinect v1 and includes 1,449 labelled RGB-D

images among 464 different indoor scenes. In 2015, the SUN RGB-D [168] dataset included the NYUv2 and additional images captured with the Kinect v2, Intel RealSense and Asus Xtion camera, for a total of 10,335 RGB-D images among at least 464 different scenes (actual number unreported). Since then, synthetic datasets were created to drastically increase the dataset size due to limitations in capturing real-world data. These include ScanNet [169] with 2.5 million images among 1,513 scenes, and SceneNet RGB-D [170] with 5 million images among 16,895 scenes. Some studies have leveraged one or more of these datasets for their application [91], others have collected their own data [165]. In comparison, our neonatal dataset comprises of 14,892 RGB-D images among 27 scenes (detailed later in Chapter 4.1). Since our dataset size is comparable with previous real-world datasets, we proceed with using our own for our application.

In multimodal studies, it is important to identify the usefulness of each modality and the effectiveness of fusing them. When using a single stream of data, some studies showed color-based models outperforming the other modalities [90], [171]; other studies showed the opposite results [165], [171]. However, following data fusion, multimodal analyses proved to be more robust than unimodal ones [90], [165], [171].

Some studies have explored the effectiveness of multi-modal fusion directly in the image (e.g., concatenate image channels from different modalities) or at some point within the network. The work of [90], [91] benefitted from a fusion at a network level while [165] found better results from image fusion.

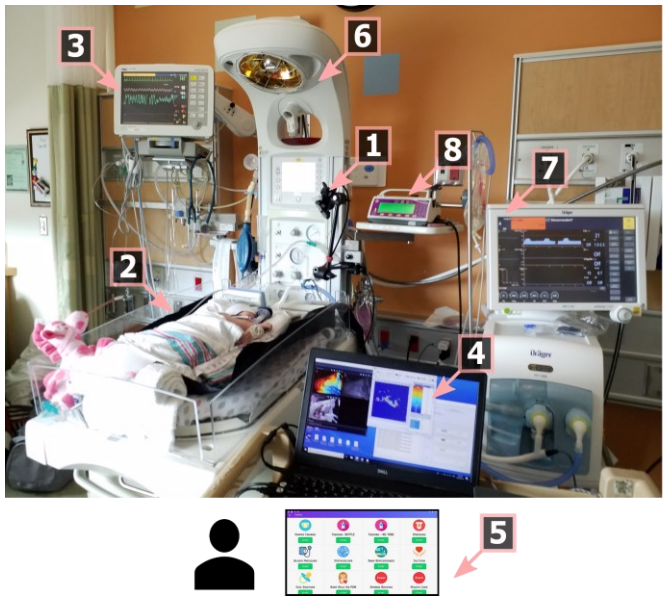
Within the network, fusion can occur at different levels. There is no consensus on the best point at which to implement the fusion: [90] found better results from an early fusion stage, [100] from a middle stage, and [92] from a later stage. In this thesis, early fusion represents a fusion after the first convolutional layer, late fusion occurs after the last convolutional layer, and middle fusion denotes any fusion between the second and penultimate layer. The work of [91] compared all plausible middle layers and found negligible differences between them.

Within the image, previous work concatenated the color channels with the channel(s) of the other modality [90], [91], [165]. This typically demands a network modification for traditional RGB-based models with a 3-channel input layer to process a fusion-based model with an x-channel input, where $x > 3$. To the best of our knowledge, no work has been done to fuse depth data within an image as a 3-channel input, rather than expanding the input to greater than three channels. Such a 3-channel fusion would prevent requiring modification of the original 3-channel CNN architecture and is explored in this thesis.

# 3 Data Acquisition

Before any machine vision methods can be implemented, validated, and improved for the neonatal patient population, a suitable dataset was required. For our study, we needed simultaneous data from an RGB-D camera, gold standard of physiologic data, and careful bedside annotation of all events of clinical interest, while encompassing variation in data (e.g., bed types, weight, ongoing interventions, etc). Since no such dataset existed, we undertook to collect these data in close collaboration with CHEO. This section presents all steps in the study design leading to the dataset used for machine vision applications. This includes an overview of each modality used to record patient data (3.1), a description of the collected data (3.2 – 3.3), various experiments on the application of the RGB-D camera in the NICU before data collection could officially begin (3.4), a description of apparatus designed to safely mount the camera on all bed types (3.5), and data pre-processing steps for machine vision application of neonatal monitoring.

## 3.1 The NICU Environment




1. Depth-sensing Camera
2. Pressure Sensitive Mat
3. Patient Monitor
4. Data Acquisition Laptop
5. Bedside Annotation Application
6. Neonatal Bed
7. Ventilator
8. Syringe Pump

Figure 3.1: NICU Environment & Experimental Set Up at CHEO.

As part of a larger overarching study, a video dataset was obtained from the NICU at CHEO. This study was approved by the Research Ethics Boards of both the hospital and Carleton University. Unfortunately, we are not currently able to publicly release the dataset due to restrictions from the hospital's Research Ethics Board. Figure 3.1 illustrates all modalities used to obtain data for the entire study, and they are detailed below.

1. **Depth-sensing Camera**: The Intel RealSense SR300 camera [172] is safely mounted above the patient to capture overhead videos from all three streams. The camera simultaneously records color, depth, and near-infrared data using a customized Intel RealSense Viewer software. Video recordings are used for non-contact video-based monitoring applications as described in this thesis. Color videos are also used for post-processing annotations for data quality assurance.
2. **Pressure Sensitive Mat** (PSM): A total of 10,000 sensors in a 50.8 x 50.8 $mm^2$ area placed underneath the patient and above the mattress. Pressure-based data were recorded using the XSensor X3 Pro V8 software for monitoring applications (not presented in this thesis).
3. **Patient Monitor**: Draeger Infinity Delta patient monitor. Gold standard vital sign data were imported from the patient's bedside monitor, including the heart rate, respiration rate, and oxygen saturation levels. A custom Patient Monitor Data Import (PMDI) software was created to capture and decode vital signs data from the RS232 serial port on the monitor [173].
4. **Data Acquisition Laptop**: All data from the camera, PSM, and patient monitor were saved and visualised on a data acquisition laptop during data collection. Upon completion, all data were safely backed up in a network attached storage (NAS) device at Carleton University.
5. **Bedside Annotation Application**: A research assistant at the patient's bedside annotated any other events in a tablet in real time, such as clinical interventions, patient motion, routine care, and alarms. A custom clinical event annotator (CEA) Android application was created to facilitate this task [174]. Color video from the camera were used for post-processing quality assessment of real-time annotations. Missed or new annotated events were inserted with refined start/end times.
6. **Neonatal Beds**: Patients were admitted to the NICU of CHEO in one of the three bed types (Open crib, Giraffe overhead warmer, and closed incubator).
7. **Ventilator**: For some patients, a ventilation support was used in addition to the patient monitor. Ventilation modes could be nasal prongs, continuous positive airway pressure (CPAP), nasal intermittent positive-pressure ventilator (NIPPV), or a conventional ventilator where this device helps regulate the patient's respiration. No data were extracted from the ventilator, but clinical interventions associated with this device were annotated (e.g., adjusting cables, respiration check from nurse, etc.).
8. **Syringe Pump**: Some patients were fed through nasogastric tubes using a syringe pump periodically monitored by nurses. During such feeding event, the patient would remain present in the bed. No data were extracted from the syringe pump, but clinical interventions associated with this device were annotated (e.g., adjusting cables, changing tubes, changing tapes, documenting feeding intake and schedule, etc.).

## 3.2 Data Collection

A total of 33 patients were recruited and up to six hours of data were recorded per patient. When a patient was recruited and ready to be part of the study, a research assistant would obtain the data acquisition laptop and the tablet from Carleton before heading directly to CHEO. Upon arrival to the hospital, all remaining equipment (RGB-D camera with mounts, PSM, cables for connecting to the patient monitor) were obtained from a cart stored in a secure room in the Department of Clinical Engineering. The entire cart and equipment were wiped down with Virox before leaving the room and heading to the NICU. At the entrance of the NICU, a study binder in the physician room included important information from the recruited patient. To be enrolled, the patient would need to fall within one of the weight/bed type strata, as depicted in Table 3.1. Three different weight types were collected to analyze pressure changes on the PSM from lighter to heavier patients. Three different bed types were explored to generalize our non-contact monitoring system to all NICU beds. Note that we originally planned to enroll five patients in each weight/bed type strata; however, due to the COVID-19 pandemic and the reduced admission of patients weighing <1500 g in overhead warmer, only three patients were recruited in this category.

Table 3.1: Enrolled patients per bed type and weight

| Mattress | Patient weight in grams | | |
|---|---|---|---|
| | <1500 | 1500-2500 | >2500 |
| Incubator | 5 | 5 | - |
| Overhead Warmer | 3 | 5 | 5 |
| Crib | - | 5 | 5 |

Before going to the patient's location, the researcher must double-check the patient ID, name, pod number, bed type, weight, and consent forms filled by the parents. Doing so confirmed the eligibility criteria and identity of the enrolled patient before verifying the patient's location at the front desk. Once the front desk assistant granted the researcher access to the patient, the researcher would meet with the patient's nurse to establish a safe time to setup all recording equipment. During the study, the nurse provided some patient information which were documented in a case report form (CRF) by the researcher, and other medical information were transcribed by a neonatologist (e.g., ventilation support settings, medication, clinical diagnoses). Throughout data collection, the researcher would periodically verify that all data were correctly being captured on the laptop. All events were annotated in the bedside annotation application, while sometimes inquiring the nurse about clinical interventions when unsure about the event or for identifying false alarms. The annotation process was very time-consuming and required constant focus from the researcher to capture events every second. After about 2-3 hours, a second research assistant would take over to complete another 2-3 hours of annotations, thereby

preventing annotator exhaustion which could affect the quality of annotations overtime. Upon completion of recording, all equipment were removed from the NICU bed, cleaned with Virox, and stored again in the secure room. Before leaving the NICU, some paperwork was filed for the neonatologist and the recruitment team for constant communication of the study progress. The laptop and tablet were then returned to Carleton where ~ 800 GB of data was transferred to the NAS. All data from the laptop were subsequently erased to reduce the risk of data theft/loss, and the laptop was ready for recording the next patient.

In total, data collection took approximately eight hours per patient and all patients' data were collected over the span of about two years. I personally participated in the data collection for 20 of the 33 patients in addition to extensive data cleaning, verification, management, and quality control thereafter.

Table 3.2: CHEO Neonatal Dataset Breakdown

| Pt ID | Weight (g) | Age (weeks, days) | Sex | Bed Type | Vent | Feed | Min Dist. (cm) | Head | Body Position | Challenges | | Video Length |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Low Light | Occl. | |
| 1 | 2465 | 37w5d | M | OHW | CV | Unsure | 77 | E | Supine | X | Min | 5:15:28 |
| 2 | 3620 | 38w6d | M | Crib | RA | Breast | 50 | W | Supine | | | 5:12:22 |
| 5 | 3480 | 38w0d | M | Crib | RA | Bottle | 55 | E | Supine/ Prone | X | | 4:01:03 |
| 6 | 3500 | 42w1d | M | OHW | RA | Breast | 60 | E | Supine | X | | 4:25:14 |
| 8 | 2340 | 38w0d | F | OHW | RA | Bottle | 52 | E/S | Supine | | Maj | 5:58:23 |
| 9 | 2520 | 39w0d | F | Crib | RA | Unsure | 50 | N | Supine | | Maj | 4:29:04 |
| 10 | 2530 | 32w2d | M | Crib | NIPPV | NG | 50 | E | Right side | X | | 4:23:55 |
| 11 | 2415 | 36w2d | F | Crib | RA | Breast | 45 | W/S | Supine | | Maj | 4:32:39 |
| 13 | 4435 | 41w4d | M | Crib | RA | NG | 53 | S/E | Supine | X | Maj | 4:51:19 |
| 14 | 2800 | 38w5d | F | OHW | RA | NG/ Bottle | 64 | E | Supine | | Min | 5:51:03 |
| 15 | 3250 | 39w5d | F | OHW | RA | Breast | 58 | E | Supine | | | 5:24:41 |
| 16 | 3460 | 41w3d | F | OHW | NIPPV | NG | 40 | E | Left/ Right side | | | 5:07:03 |
| 17 | 2650 | 36w5d | M | OHW | RA | Bottle | 40 | E | Supine | X | Min | 4:32:59 |
| 18 | 1290 | 35w1d | F | Inc. | RA | Breast | 25 | E | Prone/ Supine | X | | 5:16:49 |
| 19 | 1604 | 32w2d | M | Inc. | RA | Unsure | 25 | E | Supine | Photo | | 3:31:32 |
| 21 | 1100 | 33w0d | M | Inc. | RA | NG | 27 | E | Supine | X | Maj | 4:11:49 |
| 22 | 1391 | 31w6d | F | OHW | CV | NG | 71 | E | Supine | | Maj | 4:12:57 |
| 23 | 1690 | 33w3d | M | OHW | CV | NG | 68 | E | Left/ Right side | X | Maj | 4:15:54 |
| 24 | 1280 | 31w2d | M | Inc. | RA | NG | 27 | W | Supine | X | Min | 5:24:10 |
| 25 | 889 | 29w1d | M | Inc. | CPAP | NG | 30 | W | Right side | X | Min | 5:31:08 |
| 26 | 2045 | 33w5d | M | OHW | NP | NG/ Bottle | 55 | W | Right side | X | Min | 3:46:05 |
| 27 | 2300 | 39w0d | M | OHW | RA | Breast | 42 | W | Right side | | Min | 4:10:08 |
| 28 | 2050 | 35w2d | F | Crib | RA | NG/ Breast | 44 | E | Supine | X | | 5:42:43 |
| 29 | 1930 | 37w1d | F | OHW | RA | Bottle | 40 | W | Supine | X | | 4:30:12 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 30 | 2310 | 36w5d | M | Crib | RA | NG/ Breast | 45 | S/E | Supine/ Right side | | Min | 3:55:52 |
| 31 | 1260 | 31w3d | M | Inc. | CV | Unsure | 30 | E | Right side | X | Maj | 4:10:41 |
| 32 | 1787 | 35w3d | M | Crib | RA | NG | 37 | E | Right side | | Min | 4:11:04 |
| 33 | 1825 | 37w4d | F | Inc. | RA | Unsure | 26 | E | Supine | X | | 5:08:23 |
| 34 | 1727 | 31w6d | F | Inc. | RA | NG | 27 | W | Supine | X | | 4:11:53 |
| 35 | 1220 | 31w6d | M | OHW | NP | Unsure | 73 | S/W | Supine | | | 3:54:30 |
| 36 | 1605 | 33w4d | M | Inc. | RA | NG | 27 | E | Supine | | | 4:07:22 |
| 37 | 1720 | 33w2d | M | Inc. | RA | NG | 26 | E | Supine | | | 4:42:22 |
| 38 | 1430 | 33w2d | F | OHW | CPAP | NG | 74 | E | Right side | | | 4:07:20 |

**Pt ID** = Patient ID. **Weight** = Weight on the day of the study. **Age** = Postmenstrual age. **Bed Type** (OHW: Overhead warmer, Inc.: Incubator, Crib: Crib). **Vent** = Ventilation Support (RA: Room air, CV: Conventional ventilator, NS: Nasal prongs, CPAP: Continuous positive airway pressure, NIPPV: Nasal intermittent positive-pressure ventilator). **Feed** = Feeding Method (NG = Nasogastric feeding, Bottle = Bottle feeding, Breast = Breastfeeding, Unsure = Unsure of the feeding method). **Min Dist.** = Minimum depth distance between the camera and recorded patient. **Head** = Head orientation (N = North, S = South, W = West, E = East). **Low Light** = Patient under low lighting environment (X = low lighting during data collection, photo = phototherapy lighting). **Occl.** = Patient occlusion during data collection (Maj = Major occlusion, Min = Minor occlusion).

Table 3.2 summarizes a detailed description of our collected dataset, with important information pertaining to the video recordings. Note that patient numbers absent from the table belonged to those who were recruited but ultimately not enrolled. These patients either did not meet the recruitment criteria, were discharged from the NICU before data collection could be completed, or could not be recorded due to equipment or technical difficulties.

Table 3.3: Summarized Patient Demographic

| Category | Sub-category | Number |
|---|---|---|
| Bed type | Crib | 10 |
| | Incubator | 10 |
| | Overhead Warmer | 13 |
| Sex | Female | 13 |
| | Male | 20 |
| Weight | <1500 g | 7 |
| | 1500 – 2500 g | 16 |
| | >2500 g | 10 |
| Age | <37 weeks | 19 |
| | 37-40 weeks | 11 |
| | >40 weeks | 3 |
| Recorded Video | Average | 4:38:26 |
| | Sum | 153:08:07 |

The final patient demographic data of newborns enrolled in our study are shown in Table 3.3 where we have a proportional distribution of patient bed type, sex, weight, and age categories. The reported weight corresponds to the patient's weight on the day of data collection. The reported age corresponds to the post-menstrual age (gestational + chronological age), where the data collection day per patient ranged between 4 to 64 days after birth. On average, patients were recorded for 4.5 hours, with a total of about 153 hours of video for our entire dataset. Patient ventilation was either *room air* (*i.e.,* no ventilation assistance) or supported by one of the four different ventilation devices detailed in Table 3.2.

Many patients were also fed by nasogastric tubing or bottle-fed while in the bed, or breastfed outside of the bed. During a breastfeeding event, the bedside annotator would give the parents privacy and all annotations were paused for that time. Note that the newborn could be fed by breast or bottle under the parents' care. For simplicity, all out-of-bed feeding events were categorized as breastfeeding with the patient outside the bed, while bottle-feeding events were annotated when visible inside the bed.

Figure 3.2: Neonatal beds. Left: Crib, Middle: Overhead warmer, Right: Incubator.

Patients were recorded using the Intel RealSense SR300 depth-sensing camera given that it is relatively small, lightweight, and affordable (further explanation in Section 3.4.1). The camera was easily and safely mounted on all three different bed types (crib, overhead warmer, and incubator), as demonstrated in Figure 3.2. The design of equipment used to mount the RGB-D camera is detailed in Chapter 3.5. Sample video data are illustrated in Figure 3.3. For the depth images, colors are used to illustrate distances from the object to the camera where blue pixels represent closer objects and red pixels further ones. Video recordings were captured at a resolution of 640x480 and a rate of 30 frames per second from all three streams. This depth-sensing camera uses structured light with a rolling shutter to determine depth at each pixel, as illustrated in the top part of Figure 3.4. By projecting a pattern of black and white stripes over the scene, feature points are obtained along each stripe thereby gaining useful information pertaining to the distance of the pixels from the camera. As for the infrared videos, a similar

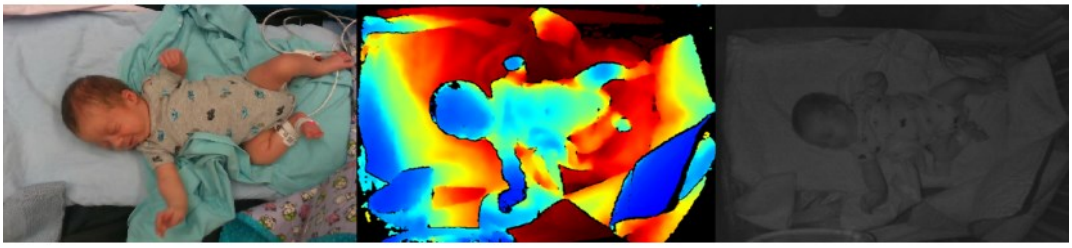


Figure 3.3: Sample video frame from the patient dataset. Left: Color, Middle: Depth, Right: NIR images.

process is used where a white frame is projected on the scene, as illustrated in the bottom part of Figure 3.4.

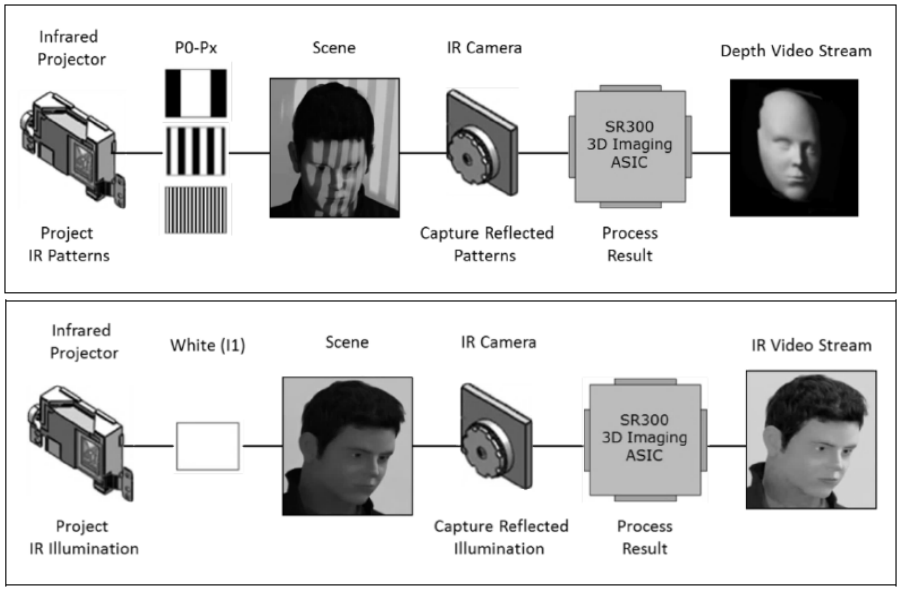



Figure 3.4: Depth-sensing mechanisms of the Intel RealSense SR300 camera. Top) Depth video data flow. Bottom) Infrared video data flow. Reproduced from [170].

When setting up all equipment in the NICU, the nurse removed the patient from the bed and the PSM was placed atop the mattress on bed sheets, underneath beddings. Figure 3.5 demonstrates each step when placing the patient on the neonatal bed. Coverage from blankets would vary among patients (e.g., some incubator patients would be completely uncovered), but the other positioning steps would remain similar across patients.


Figure 3.5: Patient in the NICU bed (With PSM Placement)

Multiple events were annotated during data collection using the clinical event annotator (CEA) application to capture any changes to the video, PSM, or physiological data. These include routine care and clinical interventions, patient motion with specific moving body part, physiological events from the patient, changes in the NICU environment, alarms from the patient monitor, and external changes in pressure from the PSM. A list of corresponding events is tabulated in Table 3.4. For video data specifically, some of these events were extracted and analyzed as a post data collection step to prepare the data for a vision-based application. For instance, periods where the equipment was moved (e.g., repositioning the camera) were excluded from our machine vision dataset, while other important events (e.g., change in lighting condition) were further analyzed. Additional data collection steps are detailed in the following section.

Table 3.4: All event names and categories

| **Routine Care** | **Clinical Intervention** | **Physiological Events** |
|---|---|---|
| Diaper Change | NG Tube – Added | Apnea |
| Feeding – Bottle | NG Tube – Removed | Cry |
| Feeding – NG Tube | NG Tube – Fixed | Cough |
| Dressing | Endo Tube | Hiccup |
| Blood Pressure | X-Ray | Seizure |
| Stethoscope | Temperature | Fever |
| Baby Repositioned | Intravenous | Sleep |
| Suction | Blood Draw | Sneeze |
| Give Soother | EEG | Yawn |
| Sponge Bathing | Echocardiogram | Burp |
| Mouth Care | Heel Prick | |
| | | **Pressure Sensitive Mat (PSM)** |
| **Motion** | **Environment** | Patient Removed from PSM |
| Full Body | Change in Lighting Condition | Baby held on PSM |
| Head | Bed Moved | Pressure Applied – Head |
| Right Arm | Equipment Bumped | Pressure Applied – Chest |
| Right Leg | Photo Taken | Pressure Applied – Left Arm |
| Left Arm | | Pressure Applied – Right Arm |
| Left Leg | **Alarm** | Pressure Applied – Left Leg |
| | Heart Rate | Pressure Applied – Right Leg |
| **SpO2 Sensor** | Respiratory Rate | Pressure Applied – Back |
| SpO2: Left Arm | SpO2 Sensor | Registration |
| SpO2: Right Arm | | |
| SpO2: Left Leg | **Interruptions** | |
| SpO2: Right Leg | Pause Session | |

## 3.3 Machine Vision Dataset

Once our patient video was recorded, post-processing techniques on the raw collected data were used to prepare a dataset for our machine vision application. This includes video and image data, color and

fused RGB-D data, and any modification to obtain a usable, relevant, and reliable machine vision dataset for non-contact neonatal monitoring.

Upon observation of the data, we detected a decline in the frame rate from 30 to 15 frames per second as recording progressed. This was due to an auto-exposure adjustment, causing the frame rate to vary over the recording period. To standardize our data, we re-encoded all videos to 15 frames per second by interpolating frames within each 30-second period.

From video recordings, multiple images were extracted to perform image analyses. From each patient, one image was extracted per 30 second to encapsulate enough diversity between images while excluding repeating or visually similar images. For a six-hour recording session, about 720 images are obtained per stream, resulting in a total of 2,160 color, depth, and near-infrared images.

Given that the field of view from RGB is narrower than depth and NIR, all streams were aligned to the RGB field of view. This alignment step allowed for proper correspondence of data across all streams. For various machine vision applications, images often needed to be annotated, and the RGB images were easier to work with for this annotation procedure. Identical annotations could then be performed on the corresponding depth and NIR data for multimodal analysis. Additionally, the head orientation of the patient was noted, such that the recorded images could later be standardized, as needed (e.g., all oriented North). Typically, one head orientation would be captured per patient throughout the entire recording session, except after severe patient motion, or when the camera or patient would be repositioned.

Figure 3.6: NICU Lighting Conditions.

Given that the NICU environment presents visual challenges to a machine vision model due to complex and realistic scenes, further information from the patient was noted to identify such limitations. The lighting environment can change drastically from high to low lighting (and vice versa), or from a patient undergoing phototherapy treatment, as depicted in Figure 3.6. Also, depending on the patient's health status, they can be completely engulfed in beddings and clothing items making it impossible to be perceived in the bed. Additionally, minor to major facial and body occlusions can occur from hospital equipment including a ventilation support, nasogastric feeding tapes, phototherapy eye mask, soother, or various beddings. Other temporary occlusions can also occur from the patient's upper limbs covering the face, or from clinical staff during an intervention. Examples of such occlusions are depicted in Figure 3.7. Most annotations collected from the CEA application were used to extract these events in video recordings; further annotations were performed thereafter to identify such specific occlusion periods.


Figure 3.7: Patient Occlusions observed in the NICU. Upper) All ventilation systems in increasing occlusion severity. Conventional ventilators, NIPPV, and CPAP systems are equally and most severe. Lower) Other occlusions frequently seen in neonatal care.

The machine vision and deep learning applications in this thesis were implemented on MATLAB and Python using multiple computing resources:

- NVIDIA GeForce GTX 1070 GPU
- NVIDIA Tesla V100 16GB GPU
- Compute Canada - Cedar v100l cluster including 192 nodes, 32 CPU cores per node, 4 GPUs per node (V100-SXM2), and 32GB GPU memory.
- Tesla P100-PCIE-16GB

## 3.4 RGB-D Camera Application on Neonates

Before starting data collection at CHEO, some experiments were conducted in the NICU and at Carleton University to select a camera suitable for acquiring all necessary data for video analysis, while ensuring that the selected device is safe for patients. This section describes such preliminary research and experiments.

### 3.4.1 RGB-D Camera Selection

Table 3.5: RGB-D Camera Comparison

| | **Microsoft Kinect V2** | **Intel RealSense SR300** |
|---|---|---|
| Price | $140 | $109 |
| Weight | 1400 g | 9.4 g |
| Dimensions (LxWxH) | 24.9cm x 6.6cm x 6.7cm | 11.0 cm x 1.26 cm x 0.38–0.41 cm |
| Resolution | Up to 1280 x 960 (< 30 fps)<br>640 x 480 pixels (30 fps) | RGB: up to 1920 x 1080 (30 fps)<br>Depth: 640 x 480 |
| Field of View (Horizontal x Vertical) | RGB: 70°x 60°<br>Depth: 70°x 60° | RGB: 68° x 41.5°<br>Depth: 71.5° x 55° |
| Depth capture | Time-of-flight (Distance estimated from time for emitted light to travel from the camera to the object and back) | Structured lighting (Distance estimated from distortions of projected pattern) |
| Power | Laser output power of 60mW | Laser output power of 180mW.<br>Class 1 laser compliant coded light infrared projector system |
| Preferred distance | 0.5m – 4.5m considering one person (could work up to 8 m) | 0.2 – 1.5m |
| Storage Temperature & Humidity | When moved to a location with a temperature difference of 20 degrees or more from the previous location, allow the console to come to room temperature before turning it on. | 0°C to 40°C (sustained, controlled)<br>-30°C to 65°C (short exposure)<br>90% relative humidity at 30°C (non-condensing) |
| Operating Temperature & Humidity | 5°C to 35°C | 0°C to 35°C |
| Additional features | • Multi-array microphone: four microphones to capture sound, record audio, and find location of the sound source and direction of audio wave.<br>• Infrared capabilities<br>• Body tracking | • Face analytics and tracking<br>• Scanning and mapping<br>• Scene segmentation<br>• Hand and finger tracking<br>• Augmented reality |
| File size | ~ 1 min of still video: 3 GB<br>~ 1 min of moving video: 6 GB | ~ 1 min of still video: ~ 3 GB<br>~ 1 min of moving video: ~ 3 GB |

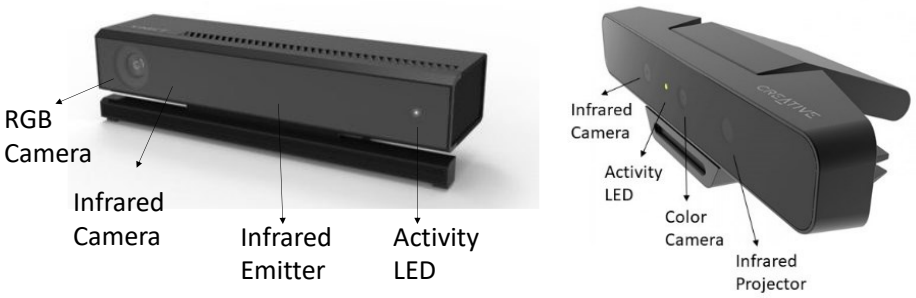



Figure 3.8: RGB-D Cameras. Left) Microsoft Kinect V2. Right) Intel RealSense SR300. (not to scale)

Multiple factors were considered when selecting a camera for neonatal machine vision application. Our study would collect data from all NICU bed types at different times of the day depending on when parents would consent for our research team to perform our experiment. The lighting variation factor was considered since incubator patients are often subjected to complete darkness to promote their growth, and since some patients could be recorded at night where lights are dimmed during their sleep. A depth-sensing camera can capture useful data in all lighting conditions while providing valuable information from objects in the scene for machine vision analysis. We were faced with two depth-sensing camera options, the Microsoft Kinet V2 and the Intel RealSense SR300, as depicted in Figure 3.8. These two were considered due to their affordable price and ease of mounting on all NICU bed types; RGB-D camera comparisons are presented in Table 3.5.

Depth estimation measures the distance from the camera to an object in the scene; two techniques are typically used in RGB-D cameras. With a structured light camera, an infrared projector illuminates the scene with a structured pattern of (typically non-visible) light and the camera captures the reflected light in an image. Deformations in the observed light pattern indicate changes in depth, and such depth estimation technology is used in the Intel RealSense SR300 camera. The Microsoft Kinect v2 camera, however, uses time-of-flight technology which measures the distance from the time it takes for emitted light to travel from the camera's emitter to the object and back. Most previous RGB-D neonatal monitoring studies used the Kinect camera as their depth-sensing technology [2], [6], [58], [59]; the applicability of the Intel camera for neonatal monitoring remained to be investigated.

While both devices present compelling specifications, the Kinect's built-in skeleton extraction has been advantageously used in some neonatal monitoring applications [6]. At the study design stage, a preliminary experiment was conducted at CHEO using a baby manikin and simulating complex NICU scenes on all bed types (e.g., applying various degrees of occlusions from beddings). For the incubator bed, limitations were observed with the depth estimation and skeleton extraction while the camera was positioned atop the incubator at ~25-30 cm from the patient. In contrast, the Intel camera provided much

better depth quality given that this device can capture depth at a much closer range (min 20 cm compared to min 50 cm for the Kinect). This range is important since it permits the camera to be mounted close to the patient for certain bed types (discussed in Chapter 3.4.2). While the skeleton extraction could be useful under RGB imaging, capturing NIR and depth-based images is most useful in low lighting environments. For these reasons, we ultimately selected the Intel RealSense SR300 camera in this thesis. Even at a close range, the neonate's entire body is still captured by the Intel camera given the wide enough field of view from the color stream (68° x 41.5°) and similarly for the depth stream (71.5° x 55°). Additionally, the SR300 is significantly smaller and lighter, hence would be easier and safer to be mounted on all bed types (e.g., in the overhead warmer where there is no protective plexiglass between the camera and the patient in case of an equipment failure).

Since the Intel has differences in field of view from the color and depth stream, both streams are aligned before data analysis to obtain adequate data correspondence. Also, the size of recorded files is extremely large, reaching ~ 3 GB per minute for the Intel camera. Data processing, management, and storage is therefore an important part of this study, as detailed in Chapter 3.3.

### 3.4.2 Impact of Camera Distance on RGB-D Data

Most previous neonatal monitoring studies used an open bed style (crib or overhead warmer); those studies that were completed with a patient in a closed incubator required perforating the top incubator surface for the camera to record through an unimpeded hole [28], [60]. This thesis proposes a custom silicone apparatus to secure the camera to the Plexiglass surface without requiring modification of the incubator. When employing a depth sensor, it is important to consider reflection artifacts that can arise when imaging the patient through the Plexiglass surface encasing the incubator bed. This thesis explores reflection artifacts and/or depth distortions from RGB-D data when recording through the Plexiglass at various distances with respect to the incubator surface (on or away from the surface). To this end, we evaluate the impact of the incubator plexiglass material on the RGB-D image for single wall and double wall designs. Also, we assess the impact of increasing camera distance from the plexiglass surface on the RGB-D image. To the best of our knowledge, no work has examined these important limitations in neonatal non-contact monitoring of patients in closed incubators.

Figure 3.9 illustrates a closed incubator and example images of patients captured in varying lighting conditions and camera positions. While the RGB images are affected by lighting variations, depth images are robust to this challenging environment. Validating the depth data quality while ensuring patient's safety is therefore warranted. This thesis suggests a minimum distance from the camera to the patient for all bed types to prevent from depth distortions in the image (when the camera is too close to the

object) or reflection artifacts in the image (when recording through the incubator plexiglass surface) thereby ensuring RGB-D data quality.

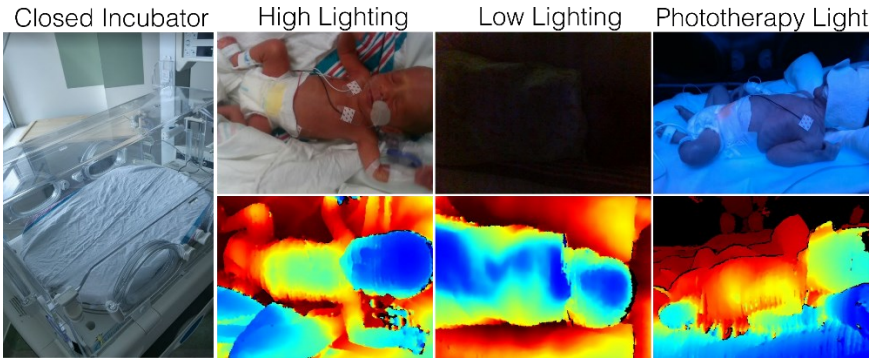



Figure 3.9: Closed incubator with example images of patients captured during varying lighting conditions using the RGB stream (upper) and depth stream (lower). The depth data is robust to lighting conditions. Note that the phototherapy light is placed on top of the incubator so the RGB-D camera is mounted on one of the upper slanted sides. A pulse oximeter is seen around the patient's right foot in the "High Lighting" image.

### 3.4.3 Impact of Camera Distance on Oximetry

In the NICU, a pulse oximeter is typically used to monitor the patient's oxygen level in the blood, or more specifically in hemoglobin (proteins in red blood cells carrying oxygen). Pulse oximetry technologies include two light-emitting diodes (LEDs) at different wavelengths to illuminate the skin and evaluate the absorption of light [175]. One LED emits around 660 nm, which is highly absorbed by deoxygenated hemoglobin, while the other emits around 940 nm, which is more absorbed by oxygenated hemoglobin. To measure blood oxygen (SpO2), the ratio of light that is transmitted (i.e., not absorbed) is calculated. Monitoring SpO2 of patients in the NICU is important to inform clinicians of potential hypoxia or hyperoxia (low or high oxygenated blood) [175], [176]. While a finger clip is typically used for adult patients, for neonates, the oximeter is typically wrapped around one of the newborns' hands or feet. The "High Lighting" image in Figure 3.9 illustrates the pulse oximeter placed around the patient's right foot. Given that both the pulse oximeter and the RGB-D camera emit and/or project an infrared light, a non-contact monitoring research must ensure that the infrared light from the camera does not interfere with the SpO2 readings.

In neonatal monitoring, clinical staff typically assign alarm thresholds for normal SpO2 values such that alarms are triggered when thresholds are exceeded [177]. There are, however, other factors affecting the reliability of SpO2 readings such as low perfusion (reduced blood circulation), motion artifacts, or other noise artifacts [175], [176]. The SpO2 value alone is insufficient to identify these factors; however, the pulse waveform does include useful information, as depicted in Figure 3.10. Clinical staff can visually examine the waveform to differentiate artifact from true photoplethysmogram (PPG) signal [175], [176]. This thesis examines whether infrared light emitted by an RBG-D camera can cause noise

or interference artifact in the PPG waveform from all bed types. We explore how SpO2 readings and PPG waveforms can differ from varying camera distance from the pulse oximeter in open beds, from the plexiglass surface in closed incubators, and from the outer wall surface in single and double wall incubator designs.

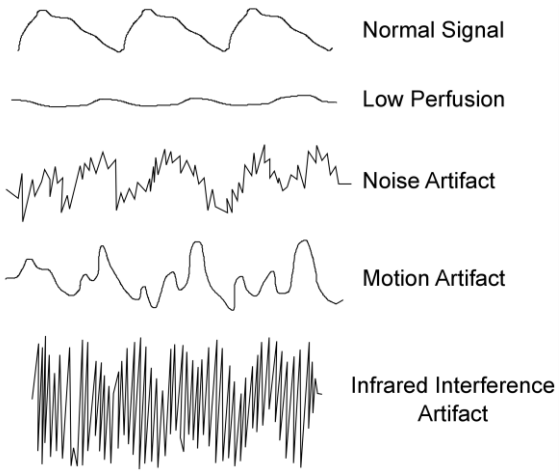



Figure 3.10: Waveforms from different signals. Normal signal to motion artifact (adapted from [175]) and interference artifact observed when the RGB-D active infrared projector is placed near the pulse oximeter.

We herein explored the use of the RealSense RGB-D camera in non-contact neonatal monitoring by evaluating a safe distance from the camera to the patient and addressing the suitability in all NICU bed types, especially when recording through a Plexiglass surface for closed incubators.

### 3.4.4 Methods & Experimental Setup

To investigate the suitability and safe use of the Intel RealSense camera, four different experiments were performed. Section 3.4.4.1 presents an experiment for open beds (crib and overhead warmer), while Sections 3.4.4.2 to 3.4.4.4 demonstrate experiments for closed incubators with plexiglass walls. The first experiment (3.4.4.1) evaluated a safe distance from the camera to a patient wearing a pulse oximeter. The second experiment (3.4.4.2) considered patients in a closed incubator, where the camera recorded through a plexiglass surface. Section 3.4.4.3 examines the distance from the plexiglass surface and Section 3.4.4.4 investigates the impact of single- vs. double-wall incubators on the RGB-D camera. Figure 3.11 depicts all experimental setups that simulated the NICU environment during neonatal monitoring. For all experiments, potential infrared interference artifacts on SpO2 measurements were reviewed and RGB-D data quality was assessed.

#### *3.4.4.1 Patient Distance in Open Bed*

This experiment seeks to determine a safe distance from the RGB-D camera to the pulse oximeter for patients placed in open beds, as illustrated in Figure 3.11-A. To identify this oximeter-camera

distance, the camera is placed at increasing distance from an adult volunteer wearing a neonatal pulse oximeter around one finger. A neonatal pulse oximeter and Draeger patient monitor were used to capture SpO2 values for all experiments. The RGB-D camera records the subject's hand using the depth stream for about 1 minute. Distance values range from 5, 10, …, 50 cm while the SpO2 readings and PPG waveforms are extracted at each distance. These distances were selected to simulate a range of possible camera positions. Even greater distances are plausible for open beds; however, keeping the camera as close as safely possible to the patient is desired to capture high resolution data for subsequent image analysis.

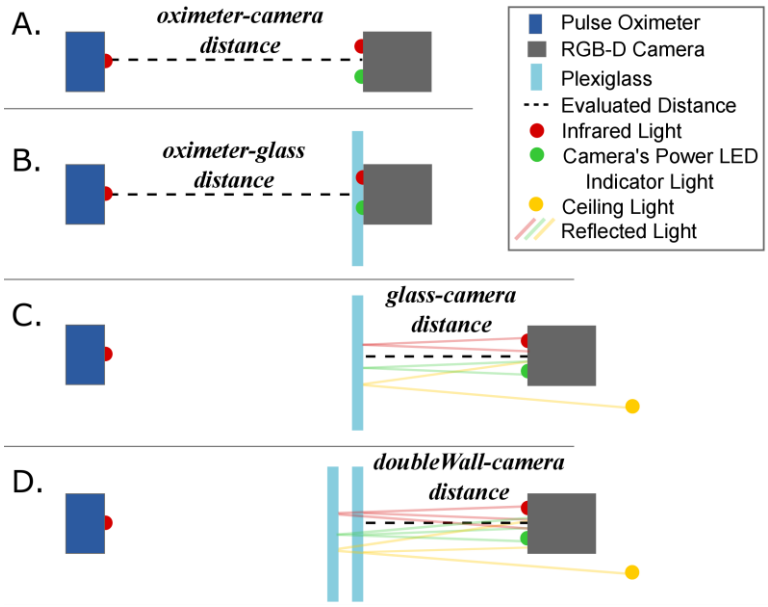



Figure 3.11: Sensor application experiments. Different distances are evaluated including A) the distance between the oximeter and the camera, B) the distance between the oximeter and the plexiglass with the camera placed on the glass' surface, C) the distance between the glass and the camera with the plexiglass at a fixed position, and D) distances from the glass at a fixed position, with a single or double wall surface.

### *3.4.4.2 Patient Distance in Closed Incubator*

While the first experiment in the previous section simulates recording a patient in an open bed, this section investigates a similar approach for patients placed in a closed incubator, as illustrated in Figure 3.11-B. Since the surface of the incubator is made of plexiglass, we simulate this setting by placing a piece of plexiglass directly in front of the camera. This experiment will specifically help identify if interference between the infrared projector and pulse oximeter would differ with the addition of a plexiglass material in front of the projector. Assuming that the camera is mounted directly on the surface of the incubator, the distance between the camera and the oximeter is dependent on the interior height of the incubator. The top of the incubator is trapezoidal-shaped where the camera can be mounted directly above the patient, 35 cm from the mattress (Figure 3.9 "High Lighting" shows the camera view), or on one of the slanted sides at ~30 cm from the mattress if a phototherapy lamp is already mounted on

top (Figure 3.9 “Phototherapy Light” shows the camera view). Depending on additional blankets placed underneath the patient, and the pulse oximeter location on the patient, the oximeter-glass distance can be decreased. This experiment evaluates minimum to maximum glass distances at 20, 25, 30, and 35 cm from the oximeter.

#### *3.4.4.3 Camera Distance from Incubator Surface*

For incubator patients, the previous experiment assumed that the camera was mounted directly on the surface of the incubator. This section explores the case where the camera is positioned at some distance from the incubator surface, as illustrated in Figure 3.11-C. This experiment will specifically help identify if reflection artifacts can be seen at varying distances from the plexiglass material, thereby informing the design of a camera mount suitable for this bed type. We evaluate several camera distances at 0, 5, 10, or 15 cm from the glass surface to illustrate the impact of increasing distances. The plexiglass is placed at a fixed distance of 35 cm from the oximeter to ensure that only the glass-camera distance factor is evaluated here.

#### *3.4.4.4 Single- vs. Double-wall Incubators*

Finally, the impact of single- vs. double-walled plexiglass incubator designs on an RGB-D camera is explored. Some NICU rooms utilize double wall incubators to reduce heat or air loss [178]. To simulate this bed type, a second piece of plexiglass is placed 3 cm in front of the original glass at 35 cm from the oximeter, as illustrated in Figure 3.11-D. This experiment evaluates the camera distance from the outer glass surface. We selected the off-surface distance as 10 cm away without loss of generality among distances > 0 cm. These findings can then be compared with corresponding distances from single-wall incubators presented in the previous section.

Note that the pieces of plexiglass used here were 2.5 mm thick. Depending on the manufacturer, the thickness of one incubator wall can vary between 1-5 mm [179]–[183]. Newer incubator designs are being researched to be portable [180], energy efficient [183], [184], and affordable for accessibility in developing countries [183]. To that end, the number of walls, the thickness of one wall, and the gap between the two walls in a double design can differ and substantially increase to address these concepts. This thesis uses 2.5 mm thick walls as a proof-of-concept.

#### *3.4.4.5 Evaluation*

All experiments are evaluated from the SpO2 values, the PPG waveforms, the RGB image, and the depth image.

From all experiments, the SpO2 values are extracted from the pulse oximeter to observe the effect of the infrared projector required to measure depth. A baseline SpO2 is obtained when recording with the RGB stream (i.e., when the infrared projector is inactive). Readings from the baseline vs. experiment

are compared overtime while PPG waveforms are compared by their visual appearance using the guidelines from the signals depicted in Figure 3.10.

Additionally, the RGB and depth data are inspected to identify if depth distortions would occur when recording at a close distance from the person's hand, and if reflection artifacts would occur when recording through a plexiglass surface for incubators. Distortions in the depth image can be characterized by rough edges around the captured object, vertical lines spanning across the image, or granular black spots. These represent areas in the scene where depth cannot be estimated. The Intel RealSense SR300 camera suggests a depth range of 0.2 – 1.5 m; this thesis investigates if a camera-patient distance at the minimum depth range (20 cm) is suitable for our application in all bed types. When recording through a plexiglass surface, reflection artifacts in a depth image show similar spots or speckles from the infrared projector being reflected on the glass, and the RGB image would mirror this observed light. Other overhead lights could also appear, including the camera's power LED indicator or ceiling lights. This thesis evaluates if such reflection artifacts can be observed at any distance from the plexiglass.

## 3.4.5 Camera Placement Results

### *3.4.5.1 Patient Distance in Open Bed*

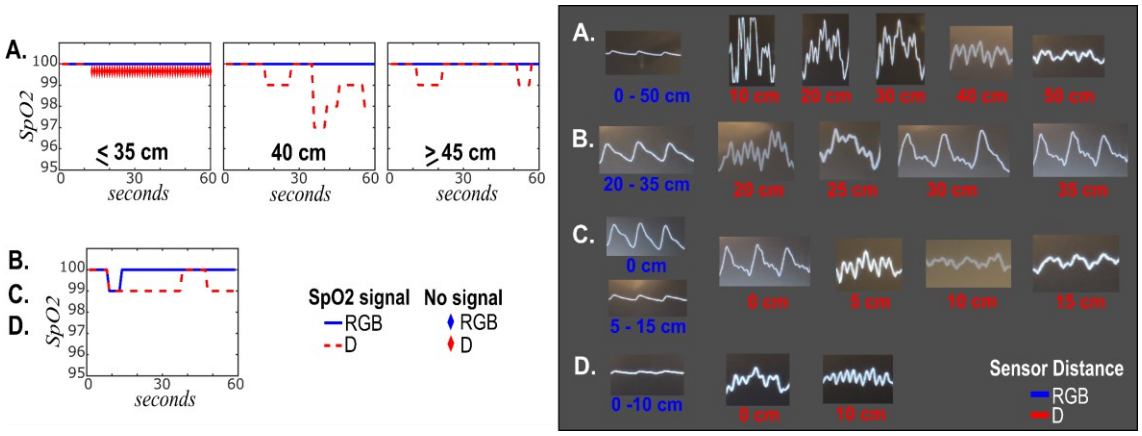



Figure 3.12: SpO2 values and PPG waveforms at various camera-oximeter distances during RGB imaging (infrared projector off) and depth imaging (projector on). See Figure 3.11 for descriptions of the experiments reported in parts A through D. Left) The SpO2 signal during depth imaging is lost when the camera is placed ≤ 35 cm from the pulse oximeter in experiment A. SpO2 signals from experiment B-D are similar and shown in a single graph for visualization purposes. Right) PPG waveforms are constant during RGB imaging at varying distances while some infrared interference artifacts can be observed during depth imaging at closer oximeter-camera distances and further glass-camera distances.

For open beds, the SpO2 readings during depth imaging fluctuate for all distances while the SpO2 readings during RGB imaging are constant. For RGB-D camera distances ≤ 35 cm, interference artifacts are quite severe as demonstrated by the corresponding waveforms in Figure 3.12. Consequently, the SpO2 signal during depth imaging is lost after a few seconds. At distances greater than 40 cm, the

estimated SpO2 continues to fluctuate, and PPG waveforms are also impacted, but correctly exhibit a periodic, rather than sporadic, signal. As for the RGB-D images, the RGB data remains unaffected at all distances, while depth distortions are observed in the depth image when the camera is ≤ 20 cm from the person's hand. Resulting images are depicted in Figure 3.13 where only distances between 10 and 30 cm are shown to visualize the increasing impact of artifacts. Shorter distances would reveal more artifacts while longer distances would show less.

#### *3.4.5.2 Patient Distance in Closed Incubator*

For this experiment, SpO2 values show minor fluctuations at all distances, but the SpO2 signal is never lost. Compared to the open bed readings, the SpO2 signal ≤ 35 cm is not lost thereby suggesting that the plexiglass might have absorbed some infrared projections [185]. As for PPG waveforms, those obtained during depth imaging only have a similar periodicity to waveforms during RGB imaging after 30 cm, and minor interference artifacts are seen at the end of each cycle. Similar to the open bed, depth image distortions are observed at 20 cm from the hand while the RGB image data remains unaffected.

#### *3.4.5.3 Camera Distance from Incubator Surface*

In this experiment, the camera placement resulted in no significant change in SpO2 readings. However, the only reliable PPG waveform signal from the depth stream corresponds to the camera mounted directly on the surface, while the other distances demonstrate some noise artifacts. Results from this experiment reveal that placing the camera directly on the glass surface also creates better RGB-D data quality, as depicted in Figure 3.13-C. No RGB-D data artifacts are seen at 0 cm from the surface. At 10 cm, reflection artifacts appear in the RGB image (green dot from the camera's power LED

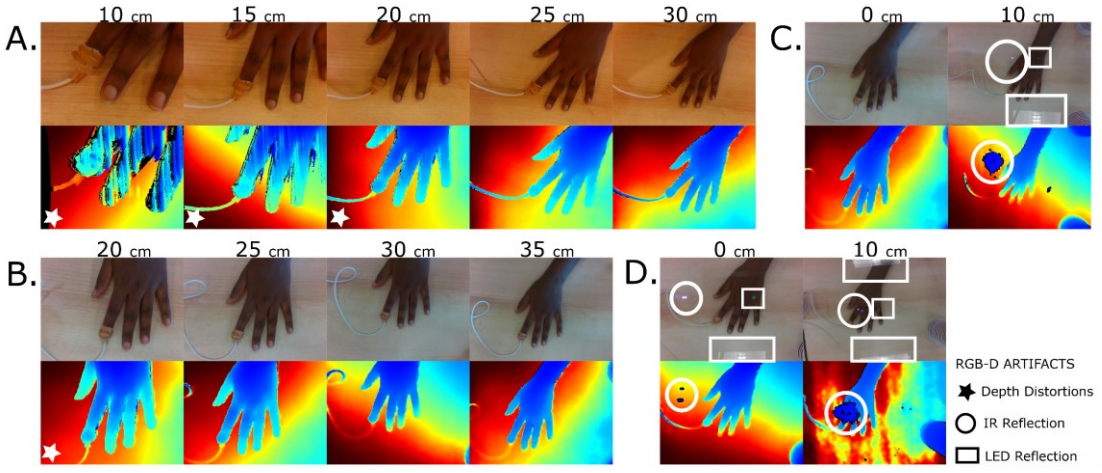



Figure 3.13: RGB and depth data with corresponding artifact from all experiments. Depth distortions are observed when recording patients in open bed (A) or incubator (B) at ≤ 20 cm. C) When the camera is placed several distances from the incubator surface, LED reflections are seen in the RGB image and IR reflections on both the RGB and depth image. D) Double wall incubators exhibit similar artifacts as single wall ones regardless of camera position.

indicator light, red dot from the infrared projector, and ceiling lights) and in the depth image (a speckle-like noise from corresponding reflected infrared light). Figure 3.13-C only shows the 0 and 10 cm results for comparison of on-surface and off-surface results, respectively. Increasing distances would increase the severity of the reflection artifacts.

#### *3.4.5.4 Single vs Double Wall Incubators*

When comparing SpO2 readings from single and double walls, values are similar and fluctuate slightly. However, PPG waveforms differ significantly where noise artifacts can be seen for double walls at any camera position (on or off the surface). Similarly, the RGB-D data for double walls suffer from image data artifacts, at any camera position. In comparison with a single wall design, the camera placed at 0 cm from the outer surface suffers from RGB-D artifacts due to reflections from the inner wall. At 10 cm, these artifacts worsen and are duplicated due to reflections on two surfaces (two red dots from the infrared projection are seen in the RGB image and the depth image has a larger speckle with two dots in the middle). All SpO2 and RGB-D artifacts are substantially more severe with the camera away from the surface as depicted in Figure 3.12 and Figure 3.13-D. Keeping the device on the surface would significantly reduce artifacts for double wall incubators.

### 3.4.6 Camera Placement Recommendations

Table 3.6 summarizes the best experimental results to avoid SpO2 infrared interference signal or RGB-D data artifacts. Comprehensively, results demonstrate that the RGB-D camera should be placed at a minimum distance of 40 cm from the patient in an open bed and at 30 cm from the patient within a closed incubator. Since the top of the incubator is 35 cm high, this gives sufficient room for the addition of mattress and blankets while keeping the pulse oximeter at a safe distance. The camera should also be mounted directly on the plexiglass incubator surface, especially for double-wall incubators to eliminate or minimize reflection artifacts. Note that depth distortions were observed with the camera at 20 cm from the person's hand (at the RealSense SR300's suggested minimum depth distance). We therefore recommend mounting the camera 5 cm beyond the vendor's suggested minimum distance (at minimum 25 cm) to ensure best depth data quality should a neonatal monitoring application not require pulse oximetry readings.

For non-contact monitoring applications in the NICU, two important factors are required; a camera suitable to capture all necessary information for computer vision research, and a mounting apparatus to affix the camera to closed incubators and open bed types in the NICU. These two factors should also consider patient safety and prevent obstruction of patient care. This section presented how the Intel RealSense SR300 can fulfill the camera requirement and provided suggestions for camera placement. Note that studies employing other depth-sensing devices could use our recommendations as a starting

point; however, the minimum distance of each camera may differ according to the specific artifacts observed, depending on the depth-sensing technology used (e.g., one may observe different artifacts with time-of-flight depth estimation). This thesis provides valuable and easily replicable experiments in such cases, while providing valuable insights for different camera placements in the NICU.

Table 3.6: Best Experimental Results for Camera Placement Suggestions

| Experiment | | Evaluation | | | |
|---|---|---|---|---|---|
| | | ***SpO2 values*** | ***PPG waveforms*** | ***RGB Data*** | ***Depth Data*** |
| **A** | Open bed | ≥ 40cm | ≥ 40cm | All | ≥ 25cm |
| **B** | Closed incubator | All | ≥ 30cm | All | ≥ 25cm |
| **C** | Distance from surface | All | On surface | On surface | On surface |
| **D** | Single vs double wall | All | Single wall | Single wall | Single wall |

## 3.5 Device Apparatus Design

Since our research aimed to collect patient data from all three NICU bed types, specific mounts needed to be created for each of them. This section presents designs of equipment for mounting the camera on all bed types. Figure 3.14 depicts all apparatuses mounted on each bed.


Figure 3.14: Device mounts on all NICU bed types.

### 3.5.1 Device Apparatus for Open Beds

To facilitate positioning the camera on open beds, an apparatus was designed using a 60-cm articulated arm with multiple degrees of freedom paired with a super clamp for cribs, as illustrated in Figure 3.15. The clamp allows a firm grip on vertical bars on the sides of the crib, regardless of the varying width between each bar. The articulated arm permits to reach around and above the bed to position the camera directly overhead the patient.

For overhead warmer (OHW) beds, the same articulated arm is used, paired with a custom side rail attachment, as illustrated in the right side of Figure 3.15. The side rail attachment was obtained from the

NICU at CHEO (right side of Figure 3.15), from a Giraffe bed and was customized to easily interlock with the articulated arm, thereby allowing to be used on all OHW beds. A heater positioned atop the OHW is sometimes turned on for patients requiring increased body temperature. To prevent the camera from overheating, it must be mounted outside of that heat field. The length and degrees of freedom from the articulated arm ensures that the camera can remain above the OHW bed at a safe distance for the camera, and especially for the patient's safety.

Figure 3.15: Device Apparatus for Open Beds. Left) Articulated arm with a clamp for cribs (upper) or overhead warmer side rail attachment (lower). Right) Closer view of the overhead warmer side rail attachment.

### 3.5.2 Silicone Skirt for Closed Incubators

For closed incubators, we developed a silicone-based skirt firmly adhering the camera to the plexiglass surface, as illustrated in Figure 3.16. For use in a clinical environment, the skirt must be easily cleaned, flexible, and durable to sustain multiple handlings. We opted for a dyed silicone-based skirt that provided ease in adhering to the plexiglass surface while blocking passage of overhead lights. Remaining ambient light may be blocked by blankets draped on the incubator. This silicone skirt provides a practical, reliable, and affordable solution for closed incubators. Previous studies either did not include incubators in their research [1], [2], [58], [59], or required drilling a hole in the top of the incubator [28], [60].

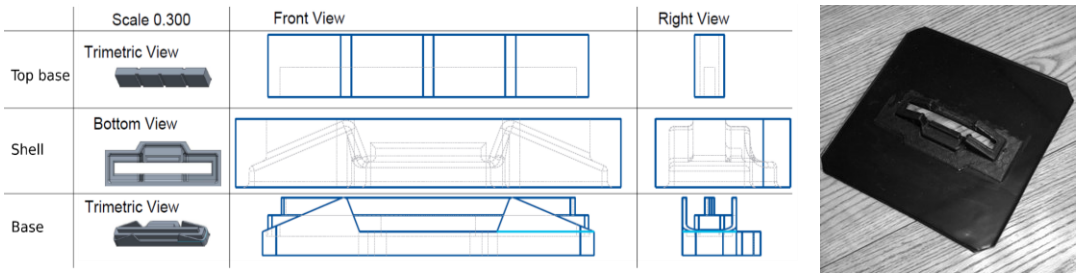



Figure 3.16: 3D model of silicone skirt design with resulting device apparatus. The silicone skirt both secures the camera and blocks room light.

Figure 3.16 shows the 3D models for the *Top base*, *Shell*, and *Base* components of the mold. The models were 3D printed using a Ultimaker 2+ 3D printer. First, a rectangular shape was glued to a glass pane to form the outer edge of the mold. The *Base* was fixed at the center of the rectangle using double sided tape. Easy Release® 200 was applied to all components for easy demolding. Ecoflex™ 00-50 silicone rubber, coloured using black silicone pigment, was poured to a depth of 3 mm. When this layer of silicone hardened, the *Shell* was positioned on top of the base. Prior to filling the *Shell*, the *Top base* was connected to the *Base*, through the *Shell*, to create an accessible space at the back of the camera for cable access and cooling.

## 3.6 Data Acquisition - Conclusions

This chapter described how data were collected from all 33 patients from multiple devices (RGB-D camera, PSM, patient monitor, and annotation tablet). To clarify my role in the data collection, I participated in the creation of a detailed data acquisition protocol for all recording devices, provided guidance for the creation and refinement of both the bedside annotation application (CEA) and PMDI software packages. I also participated directly in the collection of 20 patients, trained other research annotators, and oversaw the detailed data quality assessment undertaken by a contract researcher post data collection.

This chapter presented required preliminary experiments by selecting an RGB-D camera and identifying recommended camera distances to ensure that patients are recorded in a safe and secure manner, while guaranteeing best quality of RGB-D data. To facilitate recording on all bed types, custom camera mounts were designed. Note that additional experiments were performed with the Intel RealSense SR300 to ensure that the emitted power by the device would not harm the patient, as described in Appendix A and B.

All these contributions culminated in the acquisition of a unique, multi-modal, and reliable high-quality neonatal dataset. This dataset enabled subsequent contributions in this thesis, where robust evaluation of state-of-the-art machine vision technologies are completed for a neonatal population, thereby assessing the gaps in the field for neonatal monitoring applications. Unless otherwise stated, research contributions implemented in this thesis used our neonatal dataset collected at CHEO. In some instances, supplemental data was acquired from other sources to address a specific problem (as noted in the appropriate sections).

# 4 Scene Understanding

This chapter includes methods and results from Scene Analysis (4.1 – 4.3) where various contexts are recognized in the scene: lighting conditions, phototherapy treatment, ongoing intervention, bed occupancy, and patient coverage. A sentence generation method for semi-automated clinical charting is presented by combining these contexts. Within the ongoing intervention context, a method for bottle-feeding intervention detection is presented in Section 4.4 – 4.6. Conclusions of our scene understanding methodologies for non-contact neonatal monitoring are presented in 4.7.

## 4.1 Scene Analysis - Methods

As previously mentioned, multiple computer vision techniques can be used when analyzing images, depending on the task. An image processing technique is developed for classifying "**lighting**" and "**phototherapy**" contexts, as detailed in Section 4.1.1. A deep learning approach is leveraged for classifying the "**intervention**", "**occupancy**", and "**coverage**" contexts, while exploring RGB-D applications, as demonstrated in Section 4.1.2. Corresponding baseline methods for deep learning methods are described in Section 4.1.3. A comprehensive combination of all classified contexts is then summarized using a rule-based sentence generator explained in Section 4.1.4. The remainder of this section highlights the dataset used here and evaluation metrics for all models. The multimodal computer vision model demonstrating the comprehensive RGB-D scene analysis network is depicted in Figure 4.1.

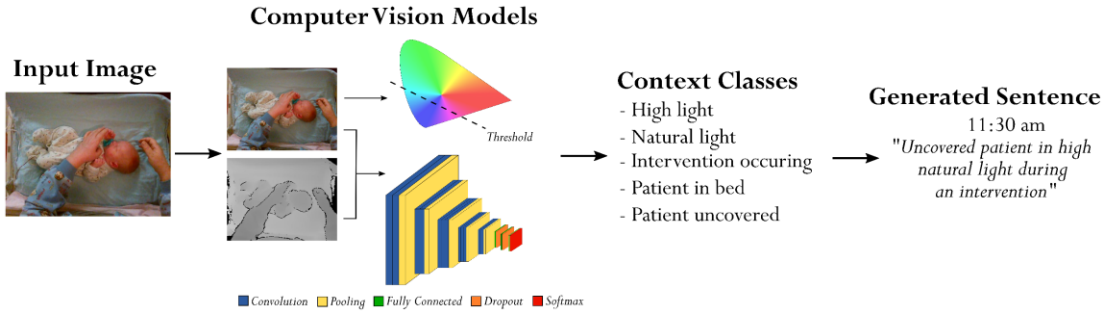



Figure 4.1: RGB-D Scene Analysis network. Image processing using color space transformation by thresholding within the space to analyze RGB images. Deep learning using VGG-16 by transfer learning on new target tasks to analyze RGB-D images. Combination of all context classes summarized as a generated sentence with the image timestamp for semi-automated nursing documentation.

### 4.1.1 Image Processing Models

Classifying lighting levels of a scene is a straight-forward task to achieve using traditional image processing techniques. In fact, from color space transformation of the image, we can transform the image from the Red, Blue, Green (RGB) color space to the Hue, Saturation, Value (HSV) color space, as depicted in Figure 4.2-A-B. This RGB-to-HSV transformation is essential to obtain the Value channel

and leverage it here for identifying the lighting level in the image. The Value component (also known as luminance), averaged across the image, can then be used to differentiate between high value (*high* lighting) and low value (*low* lighting). To this end, low lighting images can be classified as:

$$Lighting_{low} = \frac{1}{255\,x\,w \times h} \times \left(\sum_{i=1}^{w}\sum_{j=1}^{h} V_{ij}\right) \leq V^{*}_{index}\,, (4.1)$$

where $V_{ij}$ is the value component for pixels at the $i^{th}$ position among image width ($w$), and $j^{th}$ position among the image height ($h$). The $V^{*}_{index}$ represents a Value index differentiating *low* vs *high* lighting. Otsu's thresholding is employed to separate the bimodal distribution by finding a threshold that best minimizes the within-class variance in each distribution [186]. The model is validated using 5-fold cross-validation, repeated five separate times using random assignments of patients per fold. Data from the same patient never appeared in both the training and testing fold.

Similar to the "lighting" context, patients undergoing phototherapy treatments can be identified from color space transformation. Since the light emitted by the phototherapy lamp is in the range of 460-490 nm, it illuminates the scene with a blue color, as depicted by the top left image in Figure 4.2-D. This wavelength range typically mainly depicts the blue and cyan colors from the light spectrum where a few shades of green can be found near the 490 nm upper bound. Pictorially, this color distribution can be seen across the RGB channels, as illustrated in Figure 4.2-D. In phototherapy patients, since the image has a high number of blue-colored pixels, the Blue channel has high pixel intensities (appears nearly white) compared to a small number of green-colored pixels (Green channel appears gray) and negligible number of red-colored pixels (Red channel appears nearly black). In comparison, patients under a natural light have a near uniform distribution of red-, green-, and blue-colored pixels across the image, and thus respective channels appear gray uniformly. This knowledge can be exploited by measuring the Blue-Red difference to discriminate between patients undergoing the *phototherapy* light (high difference) vs. *natural* light (low difference). Within the RGB color space, the Blue and Red channels are extracted, and pixel intensities are averaged respectively, each with a scalar number ranging from [0, 1]. We created

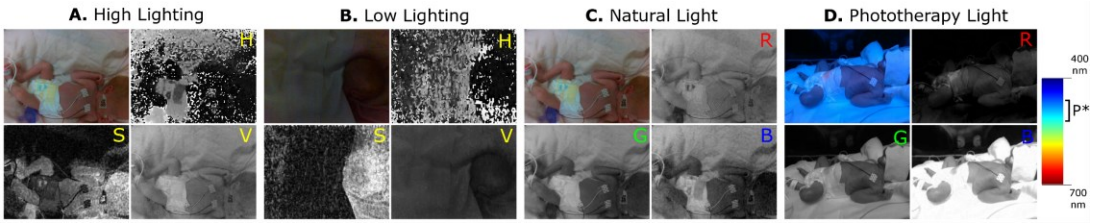



Figure 4.2: Image processing for lighting variation analysis. A-B) The "lighting" classification converts RGB (Red, Green, Blue) images to HSV (Hue, Saturation, Value) color space and uses the Value channel to differentiate between high and low perceived brightness. C-D) The "phototherapy" model measures the difference between the Blue and Red channel from the original RGB image to differentiate between uniform RGB distributions (natural light) and high Blue pixels intensities (phototherapy light, P* from the light spectrum).

a *phototherapy index* metric as the difference between the averaged Blue and averaged Red channel. To this end, phototherapy images can be classified as:

$$Phototherapy = \frac{1}{255\ x\ w \times h} \times \left(\sum_{i=1}^{w}\sum_{j=1}^{h} B_{ij} - \sum_{i=1}^{w}\sum_{j=1}^{h} R_{ij}\right) \leq P^{*}_{index}\ , \qquad (4.2)$$

where $B_{ij}$ and $R_{ij}$ are the Blue and Red channels, respectively, for pixels at the $i^{th}$ position among image width ($w$), and $j^{th}$ position among the image height ($h$). The $P^{*}_{index}$ represents a Phototherapy index differentiating *natural* vs *phototherapy* lighting. This scalar number ranges from [-1, 1] where an increasing number demonstrates greater perceived phototherapy light in the image, a value around 0 would suggest natural lighting, while an overhead warmer light (~760-1400 nm, red/infrared) would provide a negative phototherapy index due to greater red pixel intensities. All phototherapy index values are differentiated using Otsu's thresholding method.

To correctly develop a model for each context, we require representative data from both classes during training and testing. Since our dataset only included one phototherapy patient, that model was supplemented by 9 publicly available images of babies undergoing phototherapy treatment [187]. Using the search term "phototherapy" on Flickr, we included an image if it contained a baby undergoing phototherapy treatment in an NICU environment. This small, curated dataset was then added to ensure that *phototherapy* data would be present in the training and test set. Given the significantly large class imbalance between the *therapy* and *natural* class (1:32), the *natural* class was undersampled by creating five distinct subsamples of 5-6 patients. Within each subsample, a 2-fold cross-validation was performed such that our phototherapy data collected from CHEO would reside in one fold, and the phototherapy Flickr data in the other. Patients in natural lighting were also split in these two folds such that patients used in training would not appear in testing. All five subsamples are trained separately thereby ensuring that all data is analyzed. Final model evaluations would consist of the averaged metrics over all subsamples.

### 4.1.2 Deep Learning RGB-D Models

In its basic architecture, an image classification CNN contains an input layer, followed by various combinations of convolutional layers where features are extracted, activation function layers (often using rectified linear unit -- RELU), and max-pooling layers to downsample the outputs of previous layers. The CNN typically culminates with a fully connected layer that assembles all information before classification probabilities are calculated by a softmax output layer, thereby indicating the class associated with the input image. Numerous CNN model configurations have been explored, including adding dropout layers to avoid overfitting [69].

This thesis explores the use of the VGG-16 deep model [89], originally developed for the ILSVRC-

2014 object recognition competition [72], to the novel application of scene understanding in the context of video-based patient monitoring in the NICU. A pictorial description of our comprehensive RGB-D scene analysis model is depicted in Figure 4.1 where a model is trained on each context before concatenating outputs. The VGG-16 model comprises 13 convolutional layers with filters of size 3x3, and 3 fully connected layers. Further details about its architecture can be found in [89]. The VGG-16 network is selected due to its relatively simple architecture (relative to other deep networks), while achieving very high object classification accuracies in a number of application areas, especially in medical imaging [78]–[80]. This work leverages the "fine-tuning" approach which initializes weights using transfer learning and subsequently modifies weights in all model layers during training (i.e., not only in the final dense layers, but also in the convolutional layers).

Our dataset exhibits significant class imbalance for all three classification tasks, where the positive class, explained in section 4.1.6, tends to be scarce. A weighted classification layer is thus implemented to emphasize the minority class. During training, this layer calculates a weighted cross entropy loss, which emphasizes incorrectly labelled positive classes by penalizing this error more. Each variable exhibited different class imbalance; thus cross-entropy weights were optimized separately for each task. Network training parameters were optimized during preliminary examination using a subset of four patients for each context variable. A mini-batch size of 32 and initial learning rate of 1e-5 with a maximum of 20 epochs were used for all classification tasks. As a loss function, we used the stochastic gradient descent with momentum (SGDM) of 0.9, defined as

$$\theta_{l+1} = \theta_l - \alpha \nabla E(\theta_l) + \gamma(\theta_l - \theta_{l-1}) , \qquad (4.3)$$

with parameter vector θ updated at every iteration $l$, with coefficient $\gamma$ assigning the contribution of previous gradient descent from previous iteration (*l-1)* to the current one ($l$), and with non-negative learning rate $\alpha$ and loss function $E(\theta_l)$. All deep learning models were evaluated using 5-fold cross-validation with distinct patients in every fold, resulting in an average of ~ 3000 samples per cross-validation fold.

As reviewed by previous studies [188], [189], data augmentation can substantially increase the dataset by creating synthetic images based on transformations applied to the original images. These transformations include translation, rotation, reflection, scaling, or shearing. Training images were augmented using reflections along the X and Y axis, and rotation from 0-360 degrees. Given the nature of our dataset (patient positioned in different orientations and varying viewpoints), our selected forms of image augmentation are sufficient and appropriate. Since objects of interest in our dataset were sometimes small or found near the edges of the image, we did not perform translation, scaling or shearing transformations to prevent from losing valuable information from the image.

This thesis also investigates if pretrained weights are transferable from a color-based model to a depth-based model or one using fused RGB-D data. Different fusion methods are explored to evaluate various combinations for using all of these channel data, as depicted in Figure 4.3.

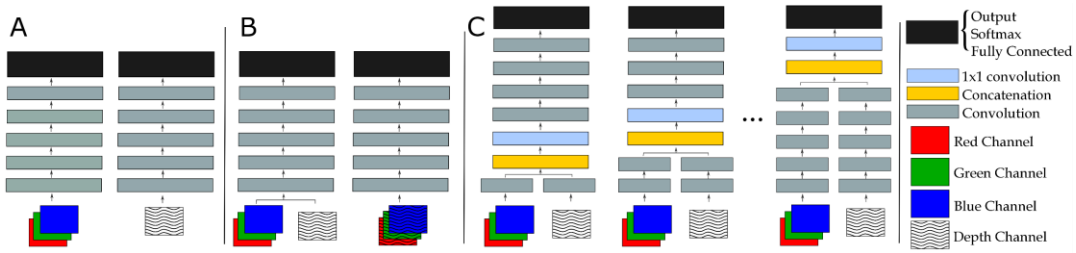



Figure 4.3: Deep Learning Models. A) Independent models using original color images (RGB model) and depth images (D model). B) Image fusion models using concatenation of depth channel after RGB channels resulting in a 4-channel fused image (RGBD4 model), or composite of depth channel on RGB channels resulting in a 3-channel fused image (RGBD3 model). C) Network fusion models using parallel branches to independently process the color and depth data before fusing after the 1st convolutional layer (early fusion: $RGBD_{conv1}$ ), the last 13th convolutional layer (late fusion: $RGBD_{conv13}$), or any layer in between (middle fusion: $RGBD_{convi}$, $i$ = 2, …,12).

First, we independently evaluate the RGB and depth (D) channels within two distinct unimodal networks, as illustrated in Figure 4.3-A. For the RGB model, the VGG-16 architecture is kept constant to transfer learn on our original color images. For the D model, given that the VGG-16 model expects three input channels, the first convolutional layer is modified to only keep one weight channel and discard of the other two. Without loss of generality, we kept the first weight channel to transfer learn on our depth images.

Second, we consider fusing the three RGB channels with the depth channel to create a new fused image before it is fed to the model, *i.e.,* ***image fusion***, as depicted in Figure 4.3-B. Two different image fusion approaches are implemented. Given that the VGG-16 model expects three input channels, we explore one approach to impute fused data to the network by imposing the depth across all three RGB channels, resulting in a 3-channel fused image. The fused RGB-D image is created from a composite of color and depth images by overlaying the depth channel on each of the color channels and scaling their intensity using the depth information. In this way, all data are fused into a 3-channel image simultaneously containing color and depth information. These fused images are then directly applicable to the original 3-channel input layer from the VGG-16 architecture to analyze the RGB-D data. We label this network RGBD3 model given that it processes RGB-D data within a 3-channel image.

In contrast, RGBD4 refers to a 4-channel image comprising red, green, blue, and depth channels. The fourth channel is accommodated by adding a copy of one weight channel from the input layer of VGG-16 to the first convolutional layer; again, without loss of generality, we selected the first (red) weight channel.

Third, we integrate the depth data directly within the network, *i.e.,* ***network fusion***, as demonstrated in Figure 4.3-C. Doing so, both the RGB image and the depth image are independently fed in the network through parallel branches before being merged. This fusion can occur early in the network after the first convolutional layer, in a process called *early fusion*. Fusion can also occur much later in the network after the last (13th) convolutional layer, *i.e., late fusion*. A fusion in any other layers is then considered *middle fusion*. We label a RGBD$conv_i$ model as a fusion of RGB and depth data after the $i^{th}$ convolutional layer, where $i$ = 1, …, 13. The weights of the fusion layer were modified similarly to the RGBD4 model.

All fusion models are implemented and compared for each context variable (i.e., occupancy, coverage, intervention). RGB-D fused images were augmented in the same way as RGB images and other model parameters were kept constant (learning rate, mini-batch size, max epochs, loss function, weighted classification layer).

### 4.1.3 Baseline Methods for Deep Learning Models

For each context, a baseline method strictly using RGB data is compared to each deep learning context model by applying image processing techniques or the pretrained VGG-16 model as-is. The baseline models serve to assess the usefulness of transfer learning, deep CNN networks, and multimodal RGB-D analysis.

#### *4.1.3.1 Intervention – Baseline Method*

When the nurse is present, a hand or an arm is frequently seen reaching into the scene to perform an intervention on the patient. A baseline approach could simply detect the presence of a person's hand/arm using a skin color model. Such an image processing method identifies a variety of human skin tones and detects pixels-of-interest. The baseline method examines the total number of skin-colored pixels in the image to differentiate interventions (high number of skin pixels) from none (low number of pixels). Our baseline model uses a recently developed range of skin pixel colors derived from skin color data obtained from 582 human individuals, using a spectrophotometer for HSV estimation [190]. To detect skin pixels, the RGB image is converted to HSV color space. Predefined threshold values from [190] are applied to each channel to identify pixels exhibiting skin tones. Detecting skin pixels in an image is not limited to nurse's hands; some pixels from the patient's skin may be detected as well. As a baseline model, we hypothesized that a higher number of skin pixels would be obtained from the nurse during an intervention with a patient present in the bed, a lower number when only the patient would be present with no intervention, and a number near 0 when the patient would be absent. The Otsu's thresholding method is applied to differentiate among skin pixel distributions. A 5-fold cross-validation using patients assigned similarly as the deep learning model is used.

#### 4.1.3.2 *Occupancy – Baseline Method*

As a baseline for the "occupancy" classification, we can leverage the pretrained VGG-16 model to classify objects in the scene associated with classes where a baby would be found in the scene or not. The classes used to train VGG-16 such as *diaper*, *pajama*, *bib,* and *bonnet* are associated with the presence of a baby. Although no newborns in clinical setting would wear a bib or a bonnet, these classes included images of a baby wearing that item, and are therefore considered here. The class *oxygen mask* was also included since some images showed a baby on a ventilation support. This list of baby-related subset of classes is used for a baseline of occupancy classification. Since multiple objects can be found in the image, the top-3 ranking predictions are evaluated. An image is classified as patient *present* if one of the baby-related classes is found in the top predictions, and classified as *absent* patient otherwise.

#### 4.1.3.3 *Coverage – Baseline Method*

Considering that this task is a subset of "occupancy", we can use a similar baseline model to identify classes related to an uncovered or covered patient among the baby-related classes. Thus, a baseline method could identify an *uncovered* patient as an image containing *diaper* or *pajama* objects since these would only be seen when the patient is not covered by beddings. A diaper is seen in unclothed patients and a pajama for fully clothed patients. These clothing items could therefore differentiate between patient coverage. An image is classified as a patient *uncovered* if a diaper or pajama are found in the top-3 predictions, and classified as a *covered* patient otherwise.

### 4.1.4 Sentence Generation for Scene Recognition

Once binary classifications are obtained for the five context variables, these predictions can form the basis for generating an informative sentence summarizing the scene from a single image. While complex natural language text generation models are available to create sentences in human-readable and legible forms [191], [192], our scene analysis application has a rather small number of variables in comparison.

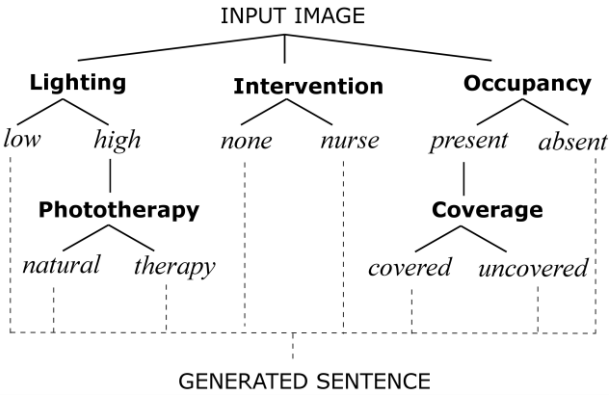



Figure 4.4: Context Variable Hierarchical Tree Representation. From an input image, first and second level context variables are obtained before being assembled into a sentence captioning the scene.

Therefore, sentence structures can easily be controlled and standardized across all possible classifications. We opt for a simple rule-based approach to generate a sentence caption for all images. Given a hierarchy among variables, sentences are generated for images residing at the leaf nodes of the tree diagram illustrated in Figure 4.4. The pseudocode in Algorithm 1 is implemented for sentence generation. Note that the impact of the first level variables (occupancy, lighting, and intervention) is

```
if (Occupancy == present) && (Coverage == covered)
    patient = "Covered patient"
elseif (Occupancy == present) && (Coverage == uncovered)
    patient = "Uncovered patient"
else
    patient = "Empty bed"
end

if (Lighting == high) && (Phototherapy == natural)
    light = "high natural light"
elseif (Lighting == high) && (Phototherapy == therapy)
    light = "phototherapy light"
else
    light = "low light"
end

if (Intervention == nurse)
    care = "during an intervention"
else
    care = ""
end

sentence = "string(patient) in string(light) string(care)"
```

Algorithm 1: Pseudocode for Sentence Generation.

greater than the second level variables. This is carefully executed to address ambiguity cases between hierarchical predictions. An absent patient cannot be covered or uncovered in the scene, and low lighting environment means that no or little light is perceived (not natural nor phototherapy). A comprehensive data description of each context variable is detailed in Section 4.1.6.

### 4.1.5 Model Evaluation

To evaluate the performance of all five classification models, the following performance metrics are calculated;

$$Sensitivity = TP/(TP + FN) \quad (4.4)$$

$$Specificity = TN/(TN + FP) \quad (4.5)$$

$$Precision = TP/(TP + FP) \quad (4.6)$$

$$Total = TP + FP + FN + TN \quad (4.7)$$

$$Accuracy = (TP + TN)/Total \quad (4.8)$$

$$F1\ score = 2\left(\frac{precison \times sensitivity}{precision+sensitivity}\right) \quad (4.9)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4.10)$$

The sensitivity is calculated to evaluate how well our system is able to detect positive classes – *i.e.,* those revealing clinical importance. On the other hand, precision captures the proportion of positive label predictions that are truly correct. Given the imbalance in our dataset, the F1 score is also calculated. Since we both want to correctly predict positive samples (precision) and increase the number of captured positives (sensitivity), the F1 score, often referred to as the harmonic mean evaluates the model's ability to perform both tasks. Matthews Correlation Coefficient (MCC) measures the overall agreement between predicted and ground truth labels. This measure, often used in imbalanced binary classification, calculates the agreement between the predicted and ground truth labels. It ranges between 1 (complete agreement) and -1 (complete disagreement), with a 0 value considered as no agreement between ground truth and predicted labels. All metrics are measured on all context models (baseline, RGB-based, and depth fusion-based).

Evaluating our sentence generation model can be done quite differently than common natural language processing practices [193]. Since the image textual summary is organized in a controlled rule-based manner, sentence structure, grammar, or punctuation are not factors of concern. Simply, the correct binary classifications across all five contexts need to be evaluated (i.e., the accuracy). We evaluate: 1) first-level classification by calculating the accuracy among "occupancy", "lighting", and "intervention" context classes, 2) independent second-level classification by calculating the accuracy among "phototherapy" and "coverage", 3) conditional second-level class measuring the accuracy of correct second-level contexts given the correct first-level prediction, and 4) independent predictions to evaluate all or most correct contexts in a subsequent generated sentence.


Figure 4.5: Dataset illustrating the five context variables (bolded) with corresponding binary classes (italicized).

### 4.1.6 Scene Analysis Dataset

Data from 29 patients from our CHEO dataset were used to implement our scene analysis methods. Patient demographics are found in Table 4.1. Only color and depth images are considered in the scene analysis dataset. A single image was extracted per 30 seconds of data per patient, resulting in a total of 15,954 images per modality. For our deep learning models, two patients were excluded from the training and test set: the phototherapy patient (Patient 19) and one patient exhibiting inconsistencies in depth capture during data collection (Patient 18). Sample images of each context with corresponding classes are depicted in Figure 2.3, reproduced here as Figure 4.5.

Table 4.1: Patient Demographic in Scene Analysis Dataset

| Category | Sub-category | Number |
|---|---|---|
| Bed type | Crib | 10 |
| | Incubator | 8 |
| | Overhead Warmer | 11 |
| Sex | Female | 12 |
| | Male | 17 |
| Weight | < 1500 g | 5 |
| | 1500 – 2500 g | 14 |
| | > 2500 g | 10 |
| Age | < 37 weeks | 15 |
| | 37-40 weeks | 10 |
| | > 40 weeks | 3 |

Table 4.2: Scene Analysis Dataset Distribution

| Context Variable | Class | # Patients | # Images | Total # Images | Description |
|---|---|---|---|---|---|
| Lighting | + *low* | 17 | 1,781 | 15,954 | Dark environment |
| | - *high* | 29 | 14,173 | | Bright environment |
| Phototherapy | + *therapy* | 1 | 424 | 14,182 | Patient under phototherapy treatment (1 patient from CHEO dataset supplemented by 9 subjects obtained from Flickr) |
| | | 9 | 9 | | |
| | - *natural* | 28 | 13,749 | | Patient under natural lighting (e.g., Sunlight, LED, fluorescent lamp, bed light) |
| Intervention | + *nurse* | 27 | 1,260 | 14,892 | Presence of nurse (or adult) hand/arm reaching into the scene |
| | - *none* | 27 | 13,632 | | Only the patient is visible in the scene |
| Occupancy | + *absent* | 17 | 982 | 14,892 | Patient absent from the bed |
| | - *present* | 27 | 13,910 | | Patient present in the bed |
| Coverage | + *uncovered* | 27 | 1,952 | 13,910 | Visible head/torso and at least three free limbs |
| | - *covered* | 26 | 11,958 | | All limbs and body covered by beddings |

To create the gold standard annotations, each context variable and corresponding class were categorized according to the guidelines in Table 4.2. The positive and negative classes are identified with a '+' or '-' mark, respectively. It is important to note that a hierarchy exists between some of these contexts, as demonstrated by Figure 4.4. Specifically, all positive instances of "**phototherapy**" also

represent *high* "**lighting**", and the "**coverage**" dataset was obtained from the *present* class in "**occupancy**".

## 4.2 Scene Analysis - Results

### 4.2.1 Image Processing Models

#### *4.2.1.1 Lighting*

Results obtained from the "lighting" model are presented in Table 4.3, where a high sensitivity (99.83%) and specificity (96.29%) demonstrate the model's excellent ability in differentiating between high and low lighting conditions. As also presented in Figure 4.6-A, the distributions of high and low luminance values can be easily discriminated by a threshold among $V^*_{index} = [0.24, 0.27]$, with a middle-point suggestion of 0.255.

$$Lighting_{low} = \frac{1}{255\,x\,w \times h} \times \left(\sum_i^w \sum_j^h V_{ij}\right) \leq 0.255 \quad (4.11)$$

It is important to note that when the patient was absent from the bed and the image would only show an empty bed, the model was still able to measure the lighting environment, regardless of bedding color (lighter or darker). The use of luminance (value) as opposed to the hue ensures that such cases are correctly classified.

Table 4.3: Scene Analysis Results from Image Processing Models

| **Model** | **Evaluation Metrics (%)** | | | | | |
|---|---|---|---|---|---|---|
| | *Sens* | *Spec* | *Prec* | *Acc* | *F1* | *MCC* |
| Lighting | 99.83 ±0.05 | 96.29 ±0.44 | 76.85 ±2.68 | 96.68 ±0.39 | 86.82 ±1.70 | 85.93 ±1.67 |
| Phototherapy | 98.66 ±0.10 | 100 ±0.00 | 100 ±0.00 | 99.81 ±0.03 | 99.33 ±0.05 | 99.22 ±0.06 |

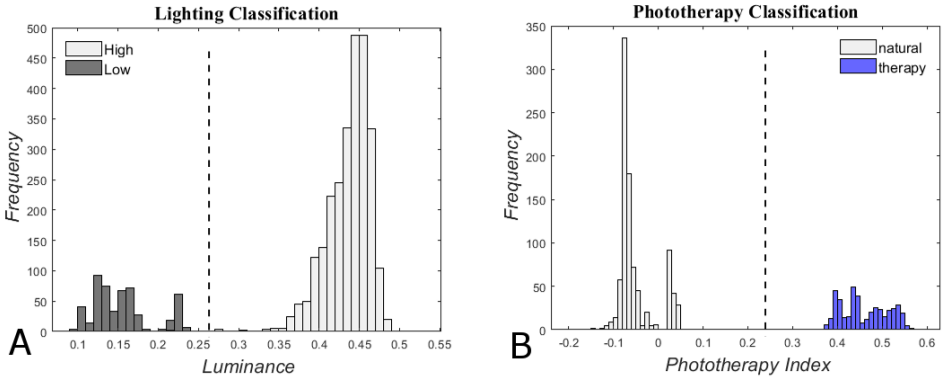



Figure 4.6: Image Processing Models Results. A) Thresholding between higher (high lighting) vs lower (low lighting) luminance distributions. B) Thresholding between higher (phototherapy light) vs lower (natural light) phototherapy index distributions.

#### 4.2.1.2 Phototherapy

The phototherapy model performed very well with all metrics above ~98%. Here, the phototherapy index distributions are even better separated than were luminance distributions (see Figure 4.6-B). Phototherapy images can be discriminated by a threshold among $P^*_{index} = [0.10, 0.35]$, with a middle-point suggestion of 0.23.

$$Phototherapy = \frac{1}{255\ x\ w \times h} \times \left(\sum_i^w \sum_j^h B_{ij} - \sum_i^w \sum_j^h R_{ij}\right) \leq 0.23\ , \qquad (4.12)$$

Using $P^*_{index}$ can also classify images with natural light (< ~0.10) and overhead warmer light (-0.50 to -0.20). It is important to note that supplementing the dataset using the 9 publicly available images helped train and test the model properly. Given that online repositories are greatly limited in specific NICU data, actual patient data collection is highly valuable, albeit costly.

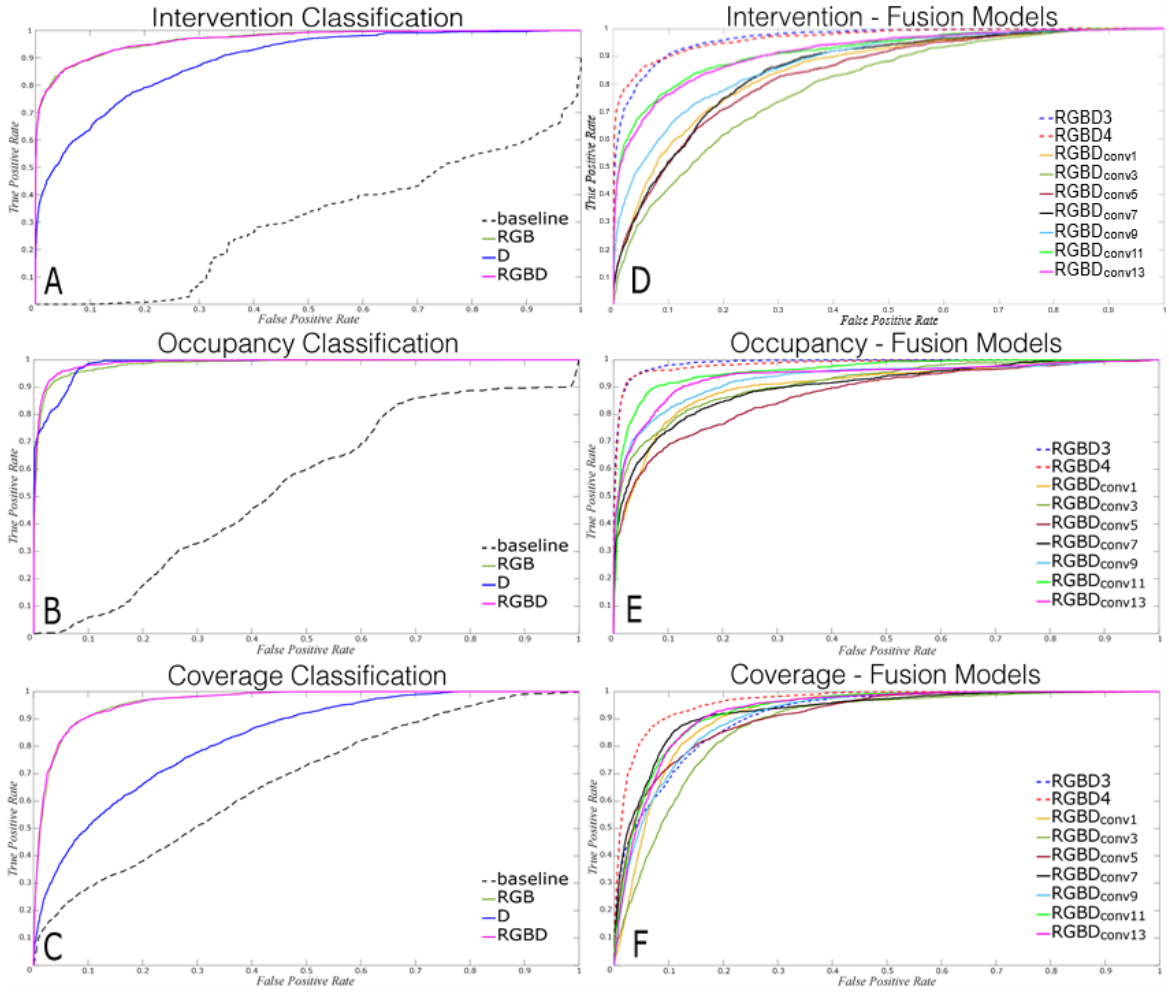



Figure 4.7: Deep Learning Models Results with ROC Curves per Context Classification. A-C) Context Classification from RGB, D, and best RGB-D model. D-F) Image Fusion (RGBD3, RGBD4) and Network Fusion (RGBD$conv_i$, $i$ = 1, …,13) models per context.

### 4.2.2 Deep Learning RGB-D Models

In the following section, deep learning models are evaluated across all data fusion schemes. A baseline method is compared to each context using image processing techniques or the pretrained VGG-16 model as-is to assess the usefulness of transfer learning, deep CNN networks, and multimodal analysis. Results from the baseline, RGB, D, and best RGB-D fusion models are presented here and depicted in the Receiver Operating Characteristic (ROC) curves in Figure 4.7-A-C. All fusion models (image vs network fusion) are then discussed in detail and illustrated in the ROC curves in Figure 4.7-D-E.

#### *4.2.2.1 Intervention*

Interestingly, the RGB-D model performed similarly to RGB, as demonstrated in Table 4.4 and Figure 4.7-A. We expected that the depth channel could provide valuable information since the hand/arm of the nurse is elevated in the scene; however, complex scenes appear to impede this depth perception. Such cases occurred more often in overhead warmer patients when the camera is placed higher, thereby capturing the entire bed frame and hospital equipment. Such misleading depth interpretation would need to be addressed to differentiate these objects captured at a close camera distance for improved RGB-D model performance, and is left as future work outside the scope of this thesis.

Among lighting environments, we expected that the RGB model would perform better in high lighting where hand features are more visible, as opposed to lower illumination environments. Since all lighting variations were included in the "intervention" dataset, we evaluated separate subsets of interventions in high vs. low lighting. Since nurses typically turn on the lights before performing their routine care, most interventions occurred in high lighting (~98% of interventions). The remaining 2% would often consist of quick routine events (e.g., replacing blankets, giving soother, checking respiration). Further studying the behaviour of the model in low lighting is therefore warranted.

Table 4.4: Scene Analysis Results from Intervention Models (Baseline and proposed methods are presented in Section 4.1.3 and 4.1.2, respectively)

| **Model** | **Evaluation Metrics (%)** | | | | | |
|---|---|---|---|---|---|---|
| | *Sens* | *Spec* | *Prec* | *Acc* | *F1* | *MCC* |
| Baseline | 18.20 ±20.89 | 93.13 ±13.97 | 59.01 ±43.36 | 86.60 ±10.98 | 16.63 ±9.22 | 20.52 ±14.28 |
| RGB | **83.44** ±0.07 | **95.76** ±0.15 | **64.55** ±0.80 | **94.72** ±0.14 | **72.79** ±0.53 | **70.63** ±0.54 |
| D | 66.11 ±0.16 | 89.25 ±0.07 | 36.24 ±0.16 | 87.29 ±0.06 | 46.82 ±0.16 | 42.64 ±0.18 |
| RGBD | **84.25** ±1.03 | **95.70** ±0.51 | **64.54** ±2.51 | **94.73** ±0.45 | **73.06** ±1.60 | **70.98** ±1.62 |

A comparison between interventions in high lighting (*high*-intervention) low lighting (*low*-intervention) are depicted in Figure 4.8. The RGB and RGB-D models in *high*-intervention performed similarly, while the RGB-D model performed best in *low*-intervention. These findings suggest that the RGB-D model is more robust to lighting conditions than the RGB model.

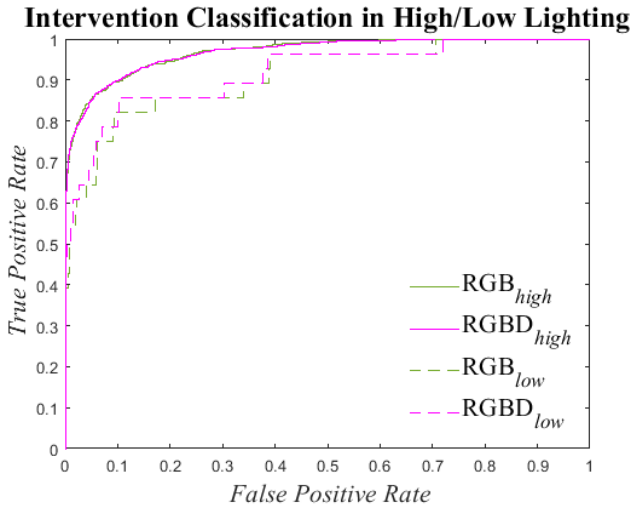



Figure 4.8: Intervention Results in Different Lighting Conditions. RGB and RGB-D perform similarly in high lighting, while RGB-D outperforms RGB in low lighting due to the depth channel.

Results from the skin-detection baseline model are quite poor with sensitivity of about 18% compared to deep learning models (> 66%). The weak performance of the baseline model may be due to multiple factors, such as overlayed skin pixels between the patient and the nurse during an intervention, or the skin color model being sensitive to noise from pixel intensities included in the human skin color gamut. The latter can be perceived in Figure 4.9, where parts of the patient's clothing are falsely detected as

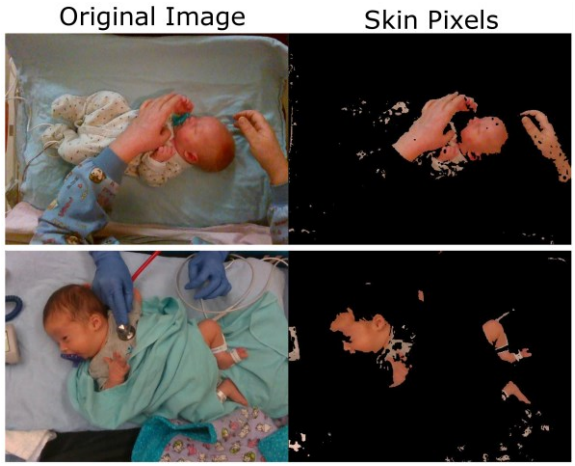



Figure 4.9: Skin pixel detection from "intervention" baseline model. Extracted skin-colored pixels from transformation in the HSV color space. Upper) Skin from patient and nurse are detected. Lower) Only patient skin is detected when the nurse is wearing gloves.

skin pixels. Most importantly, failure to detect skin pixels when the nurse is wearing gloves would increase false negatives while a patient wearing only a diaper, thereby generating many skin pixels, would contribute to false positives. A model adding a time factor where skin pixel variations are tracked between frames, thereby detecting when additional pixels from a "nurse" object is entering the frame, would be more suitable for this model. Similarly, a hand detector model could identify when the nurse's hand is visible in the scene, however, differentiating between the nurse's and patient's hand may be an issue.

In comparison, the deep learning approach seems to overcome the challenges above, given the substantial increase in performance. More specifically, the RGB and RGBD models have learnt features that differentiate an adult hand or arm from the patient and background. Given that a few images of nurses wearing gloves were included in the dataset, the model learned to identify the hand from its shape and the context of the scene rather than simply the skin color. Note that skin detection is still a reliable baseline method for comparison with our deep learning methods since only 7.4% of the intervention images included a nurse wearing gloves (among 15/27 patients). Most of the time, they would wear gloves while changing a diaper and performing other care. Other times, they would not wear gloves, or only one person would wear gloves while other nurses would simultaneously perform different care on the patient (*i.e.*, skin detection would still be valid).

#### *4.2.2.2 Occupancy*

The deep learning models performed well on this task, with the RGB-D model performing best, followed by the RGB, and D models, as demonstrated by Table 4.5 and Figure 4.7-B. When detecting if a patient is present in the bed, a color-based model would need to find features from the patient's head, face, limbs, or entire body. A depth-based model would need to identify the 3D shape of the patient vs. a uniform bedding surface when the patient is absent. Such "ideal" examples are illustrated in the left side of Figure 4.10 (patient clearly present or absent). Less ideal examples are shown on the right side of Figure 4.10, including low lighting conditions and patient coverage that hinder the RGB model. In these cases, the depth-based model would still correctly detect the patient. Conversely, the depth perception may be hindered from surrounding objects in a crowded environment, especially when the camera is positioned further from the patient. In comparison with the depth or RGB models, the RGB-D fusion model is more robust to lighting conditions and patient coverage despite depth noise from crowded environments.

It is clear from the results in Table 4.5 and Figure 4.7-B that the deep learning models have again significantly outperformed the baseline method for this context variable. The baseline's low sensitivity and high specificity suggests a bias towards predicting the negative (present) class, although both metrics are quite poor in comparison with the deep learning models' performance.

Table 4.5: Scene Analysis Results from Occupancy Models (Baseline and proposed methods are presented in Section 4.1.3 and 4.1.2, respectively)

| Model | Evaluation Metrics (%) | | | | | |
|---|---|---|---|---|---|---|
| | *Sens* | *Spec* | *Prec* | *Acc* | *F1* | *MCC* |
| Baseline | 6.42 | 62.01 | 1.18 | 58.34 | 1.99 | -16.34 |
| RGB | 84.46 ±1.50 | **98.17** ±0.09 | 76.56 ±1.04 | **97.27** ±0.15 | 80.31 ±1.10 | 78.96 ±1.19 |
| D | 77.11 ±0.42 | 98.08 ±0.36 | 74.10 ±3.76 | 96.70 ±0.35 | 75.54 ±2.05 | 73.81 ±2.21 |
| RGBD | **85.64** ±0.56 | **98.49** ±0.24 | **80.10** ±2.49 | **97.64** ±0.22 | **82.76** ±1.31 | **81.56** ±1.38 |

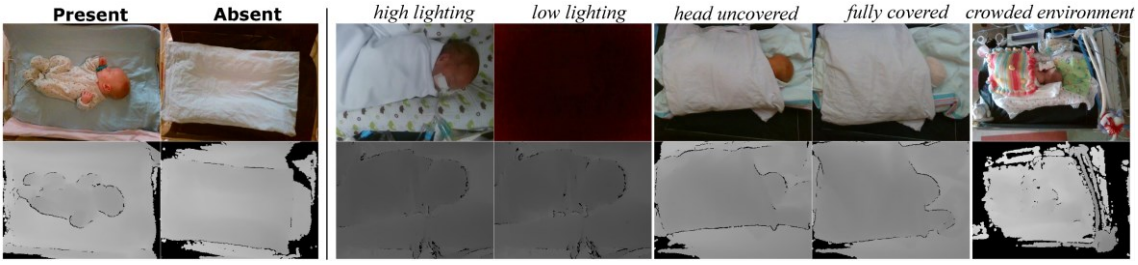



Figure 4.10: Occupancy Examples with Corresponding RBG (upper) and Depth Data (lower). Left: Ideal cases where the patient is clearly and visibly present and absent from the bed in RGB and D images. Right: Challenging cases: RGB is limited by lighting conditions and patient coverage, while crowded environments mainly affect depth perceptions (stronger depth data from ventilation hoses).

#### *4.2.2.3 Coverage*

For this context, the RGB and RGB-D models performed best and similarly, as shown in Table 4.6 and Figure 4.7-C. The "coverage" context aims to further categorize the presence of the patient. Since an uncovered patient would have visible head/torso and most free limbs (at least three for minor occlusions), the RGB model would identify features from these visible parts. Low lighting conditions would impede this model, similarly as it did for the occupancy context. A depth-based model would further classify the obtained 3D shape as a 3D object in the shape of a patient (*uncovered*) vs. a trapezoidal prism-based shape outlining the patient and hanging blankets on each side (*covered*). In ideal cases, lighting conditions are an issue in detecting these shapes. However, depth noise from surrounding or occluding objects can severely impact the depth perception of these shapes regardless of the lighting conditions. For this context, both RGB and RGB-D models performed similarly since they each suffered their own limitations. While the RGB model's low lighting problem cannot be avoided from a strictly color-based model, a fusion model can tackle depth noise from occluding objects. More specifically, occlusions often consisted of the nurse's arm/hand during an intervention. While intervention events are implicitly included in the coverage dataset, an explicit analysis of "coverage level during an intervention"

were not investigated, but are suggested as future work outside the scope of this thesis. Such analysis could inform on the type of intervention, especially in video analysis. For example, a patient who remains fully covered during an intervention could suggest an equipment-based intervention (e.g., adjusting nasogastric or intravenous tubing).

For the task of detecting patient coverage, the higher sensitivity compared to specificity would suggest over-prediction of the minority positive (uncovered) class. More specifically, the *diaper* class from VGG-16 was triggered about 43% of the time, where the patient *present* class only comprised of 14% of the coverage dataset. The baseline model then largely fails to capture patient coverage, as demonstrated by all metrics being poor, aside from sensitivity. Transfer learning however overcame this minority over-prediction issue as shown by deep learning models' improved performance for all evaluation metrics.

Table 4.6: Scene Analysis Results from Coverage Models (Baseline and proposed methods are presented in Section 4.1.3 and 4.1.2, respectively)

| **Model** | **Evaluation Metrics (%)** | | | | | |
|---|---|---|---|---|---|---|
| | *Sens* | *Spec* | *Prec* | *Acc* | *F1* | *MCC* |
| Baseline | 85.19 | 15.33 | 14.11 | 25.13 | 24.21 | 0.51 |
| RGB | **88.22** ±0.82 | **92.30** ±0.23 | **65.16** ±0.50 | **91.73** ±0.11 | **74.95** ±0.23 | **71.29** ±0.28 |
| D | 62.40 ±0.71 | 83.02 ±0.55 | 37.51 ±0.58 | 80.13 ±0.40 | 46.85 ±0.38 | 37.30 ±0.45 |
| RGBD | **88.15** ±0.48 | **91.84** ±0.40 | 63.83 ±1.02 | **91.32** ±0.29 | 74.04 ±0.59 | 70.29 ±0.61 |

#### *4.2.2.4 Comparing RGB-D Fusion Models*

Among unimodal models, the RGB data alone was more useful than depth alone. Each context demonstrated strengths and caveats from each stream of data, while fusion networks would generally overcome respective limitations.

For the "intervention" and "coverage" contexts, the best performing fusion model was the RGBD4, while "occupancy" had slightly better RGBD3 results, as presented in Figure 4.7-D-F. Since previous studies have not performed depth-controlled alpha-blending as a composite of depth on RGB channels, it is important to properly compare the RGBD3 vs. RGBD4 models. Given that only a small margin difference was observed in the "occupancy" context for RGBD3 vs. RGBD4, but much larger with "coverage", we can conclude that the choice of fusion approach is case-dependent.

In terms of testing runtime, the RGBD3 model is much faster given that only weights from three channels are processed instead of four. For the "occupancy" context, images were tested at an average rate of ~0.008 seconds/image for RGBD3, compared to ~0.037 seconds/image for RGBD4. In cases where both fusion approaches lead to comparable performance, one may be inclined to use the RGBD3 network for faster processes, especially in real-time applications.

Overall, image fusion techniques performed better than network fusion. This shows that training fused

data from the start works better than processing RGB and depth separately. Interestingly, when streams are processed separately, results indicate that fusing at a later stage was more beneficial. The "intervention" context revealed optimal network fusion at the 11$^{th}$ then 13$^{th}$ layer, "occupancy" at the 11$^{th}$ then 13$^{th}$ layer, and "coverage" at the 7$^{th}$ then 11$^{th}$ and 13$^{th}$ layer. Comprehensively, fusion at a later stage (around the 11$^{th}$ layer) seems to work well for all contexts, while middle fusion at the center point of the network (7$^{th}$ layer) is only ideal for "coverage". Similar results were obtained for early fusion across all contexts revealing poor performance overall.

### 4.2.3 Sentence Generation

To generate a sentence comprising of all predicted classes, the top-performing models among each context were selected. Table 4.7 presents the computed accuracy among all contexts including the hierarchy between them, as depicted in Figure 4.4. We label first-level predictions as correct lighting (L), intervention (I), and occupancy (O) predictions. In second-level predictions, a measure of these contexts independent of hierarchy is reported for phototherapy (P) and coverage (C). Since the hierarchical structure could affect lower levels as the error propagates downward, we also report conditional accuracy, including correct phototherapy given lighting (P|L) and correct coverage given occupancy (C|O). The final generated sentence is only considered to be correct when all five contexts are correctly classified.

Table 4.7: Scene Analysis Results from Sentence Generation

| Level | Correct Context | Accuracy (%) |
|---|---|---|
| 1 | L | 96.83 |
| 2 | P | 99.83 |
| | P\|L | 96.55 |
| 1 | I | 94.74 |
| | O | 97.67 |
| 2 | C | 91.46 |
| | C\|O | 90.04 |
| All (pred = 5) | L,P,I,C,O | 73.05 |
| All (pred >=4) | L,P,I,C,O | 97.97 |

Results from the generated sentence are obtained from 5-fold cross-validation of deep learning models to include the entire dataset and they demonstrate great promise in providing meaningful contextual summaries to images given the rather high accuracies at each level. The second-level accuracy is more affected than the first-level one since the prediction errors carried forward from a higher to lower hierarchical point; however, the performance difference is not significant. As for independent predictions, an image had five correct predictions ~73% of the time or at most four correct predictions

~98% of the time. This near 25% difference is due to the "coverage" context's increased difficulty, relative to the four other contexts.

## 4.3 Scene Analysis - Discussion

Traditional image processing methods were shown to be highly successful for lighting-specific contexts, but inadequate for context classification of complex patient scenes within the NICU. The baseline vs. deep learning models of the "intervention" context demonstrate this claim. The complexity of images requires extraction of meaningful features from each context. The deep learning approaches using transfer learning proved to be more suitable for such tasks. More specifically, multimodal analysis leveraging the depth channel was successful for all contexts, especially in low-light environments. Using the depth data alone was not effective, since analyses based on the perception of objects from the camera can be hindered from surrounding objects. In best-case scenarios, the camera would capture the patient closely from an uncrowded environment. This is, however, not feasible in realistic scenarios since bedding, soothers, plush toys, and hospital equipment are often found in NICU bed environments and can be captured when using a wide field of view. The camera had to be positioned at a higher distance from the patient in some cases when the overhead warmer heater was turned on to prevent the camera from overheating and posing a risk to the patient. To tackle this depth noise issue for such patient data, an attention mechanism could focus on certain objects of interest rather than the whole scene, thereby pivoting from a classification to object detection task. Future work is recommended to investigate object detection applications in the NICU.

Once classes from context variables are correctly categorized, a time stamp can be extracted at given points in time for charting purposes. Descriptive text was generated in the present thesis using a simplified model based directly on the five context variables predicted by the multimodal network. This was created as proof-of-concept for the semi-automated nursing documentation.

Other patient monitoring studies could also benefit from the findings of this research. By capturing contexts, subsequent computer vision algorithms can be selected in a context-dependent way. For example, knowing the overall lighting condition could inform the choice of imaging modality (near-infrared vs. natural light), jaundice monitoring applications could start executing upon detection of phototherapy lighting, applications using patient segmentation could leverage “uncovered” patients as input, or knowledge of an ongoing intervention may silence an otherwise-false alarm.

In comparison with previous neonatal research, no other study has accomplished such a multi-faceted multi-modal framework for analyzing the NICU bed environment. Villarroel *et al.* however performed research on bed occupancy and intervention detection. In their work, they achieved 94.50% accuracy for intervention detection from classification of color video recordings. Our work outperformed their

research with 94.73% accuracy from classification of a single RGB-D image. Additionally, their bed occupancy method had 98.80% accuracy with 98.20% area under the curve (AUC), excluding interventions from their dataset, only with patient uncovered, and by including a single bed type (incubator). Our bed occupancy classification included multiple variations in the data (during ongoing intervention, varying levels of patient coverage, camera at different distances from the patient, and across all three NICU bed types). For bed occupancy, we obtained 97.64% accuracy and 98.98% AUC, thereby revealing a great promise for our model to generalize among a larger neonatal population, while presenting high model performance.

In the scene analysis models, we detected an intervention in an image; future work remains to further classify between different types of interventions (e.g., diaper change, feeding, temperature check). The next section addresses this need and focuses on one type of intervention: bottle-feeding.

## 4.4 Bottle Feeding Intervention Detection - Methods

Once an intervention is detected using the multimodal scene analysis model, we could further categorize this intervention among multiple routine care events occurring in the NICU. Among them, bottle-feeding intervention is the focus of the present section. Compared with the scene analysis methods, the following section presents a different data-driven method to train the VGG-16 model with supplemented focused data to close the gap between the source and target domain. This approach presents a solution when class imbalance and data scarcity are important factors in one's dataset.

From the ImageNet dataset used to train the VGG-16 model, some classes included the cooccurrence of a baby and an adult, such as *cradle*, *diaper*, or *bib*. These object classes are present in both our positive and negative classes, however, the additional occurrence of a nursing bottle would classify our image as "bottle-feeding" if the nursing bottle is present, and "no-feeding" otherwise. Can a network, pretrained in a source domain devoid of bottle-feeding events, efficiently classify bottle-feeding events? Transfer learning can partially address the domain gap [82], but we have very limited data available in the target domain. We herein address this question by adding a third domain comprising images extracted from publicly available sources, similar in key features to a bottle-feeding intervention image. This supplemented data domain will help bridge the gap between the source and target domains for knowledge transfer proficiency. To this end, we investigate how the knowledge acquired from millions of images in a source domain, complemented by an expanded data domain, can be transferred to a significantly smaller dataset in the target domain. We evaluate the impact of data expansion in transfer learning to address data scarcity in the target domain. We additionally visualize and quantify that gap to demonstrate

the influence of the data expansion domain. The following sections explains this method with detailed results, and our model is depicted in Figure 4.11[1].

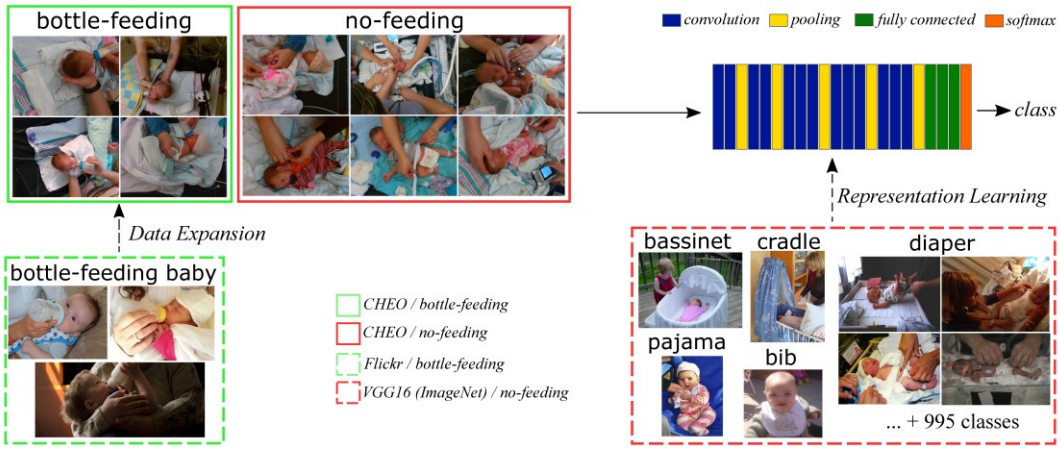



Figure 4.11: Bottle-feeding transfer learning model trained on VGG-16 using a subset of ImageNet [72], [134] with data expansion from publicly available images [194].

### 4.4.1 Transfer Learning & Data Expansion

As previously mentioned, to perform classification of "bottle-feeding" vs "no-feeding" events, we selected the pretrained VGG-16 model due to its strong performance on previous image classification tasks [69]. The model comprises 13 convolutional layers, often referred to as the "feature extraction" layers, followed by three densely connected layers responsible for arriving at a final classification. Model parameters were tuned using preliminary experiments. Training of images involved a mini-batch size of 32 with a learning rate of 1e-5 over 20 epochs. Due to a class imbalance among images, a weighted classification layer was used to emphasize the minority class. The model is evaluated using 5-fold cross-validation where different patients were selected per fold. Additionally, given that "bottle-feeding" event were observed in only 6 out of 27 patients, these patients' data were distributed separately among the five distinct folds (one fold included two of these patients' data).

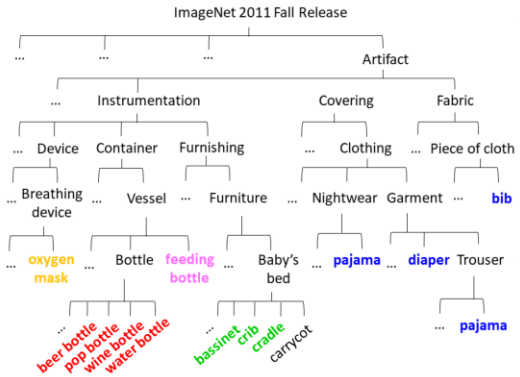
During model training, data augmentation is used to improve model performance and generalization [188], [189]. Training images were augmented using reflections along the X and Y axis, and rotations from 0-360 degrees. Due to the nature of our dataset, where objects of interest are often small or can be found near the edges of the image, we refrained from performing translation, scaling or shearing transformations.

While traditional data augmentation produces synthetic copies of original images, we also explore a data expansion approach to extract similar-feature images from external sources to further supplement

[1] Due to licensing issues, the "bottle-feeding baby" images portrayed in Figure 4.11 were obtained from pixabay.com and only shown here for visualization purposes.

the training dataset. As previously mentioned, collecting and labelling clinical data is a laborious task due to a multitude of reasons detailed in Chapter 2.2. This often results in unique and rich data, albeit in limited quantity. We address this issue by gathering publicly available images from Flickr [194] by carefully searching and curating images showing similar objects and contexts. To this end, we obtain supplemented data from similar-feature images sufficient for performing binary classification in the target domain when the source domain is significantly deficient in data from one of these classes.

The BFID model (bottle-feeding intervention detection) was trained and tested on our original dataset. Then, the BFID*exp* (BFID with data expansion) model was trained on our expanded dataset, including Flickr images, and tested on our original dataset. Model performance is compared before and after data expansion. As a baseline method, we tested our dataset on the pretrained VGG-16 model using baby- and bottle-related classes drawn from the model's 1000 classes. These classes were extracted from a more complete list of words in ImageNet, which was structured according to the WordNet hierarchy [195]. WordNet is a lexical database of English words grouped into synsets or synonym sets if they share similar concept and semantic relations. Pictorially, these relations can be demonstrated in a tree map, and a subset of this tree highlighting a few classes of interest are depicted in Figure 4.12. Here, we have added a *feeding bottle* class to show word similarity. A conceptual relation can be drawn from the tree map, and we additionally visually inspected a subset of images used to train VGG-16 from related classes to inform on feature-based relations. These classes were selected due to the environment, a piece of clothing, or an object typically seen with babies. To this end, we can curate a list of classes used to train VGG-16 related to our data, as demonstrated in Figure 4.12. From both representations, we can clearly see the close conceptual-semantic relationship between *crib*, *cradle*, and *bassinet* class, while *diaper*, *pajama*, and *bonnet* share a feature relationship to baby-related images. Most of these corresponding images used to train VGG-16 contained a baby and sometimes an adult present but no nursing bottle, thereby similar to our "no-feeding" class. As for the "bottle-feeding" class, bottle-related classes such




| Related VGG-16 Classes | | | Visual Description |
|---|---|---|---|
| Baby-Related | *Environment* | bassinet | Empty object or baby in the bassinet. Person (with visible face) standing by the bassinet. |
| | | crib | Empty object or baby in the crib. |
| | | cradle | Empty object or baby in the cradle. Person (with visible face) standing by the cradle. |
| | *Clothing* | diaper | Baby wearing a diaper or image of a diaper object. Person (face visible or not) changing a baby's diaper. |
| | | pajama | People of all ages wearing pajamas or picture of pajamas. |
| | | bib | Baby wearing a bib often held by an adult or empty object. |
| | *Object* | oxygen mask | Empty object or adult wearing oxygen mask. Some babies on ventilation support. |
| Bottle-Related | *Object* | pop bottle | Empty object or collection of objects on tables or in store. |
| | | water bottle | Empty object or collection of objects. Some adults drinking from a water bottle. |
| | | wine bottle | Empty object or collection of objects in wine cellar or rack. |
| | | beer bottle | Empty object or collection of objects on tables/shelves. One picture of baby and toddler playing with beer bottles. |

Figure 4.12: Concept relations from tree map of some classes used to train VGG-16. Feature relations from description of 7 baby- and 4 bottle-related classes.

as *water bottle*, *pop bottle*, *beer bottle*, and *wine bottle* shared semantic relationships with each other and close relations to a feeding bottle object. However, since very few images used to train VGG-16 contained our baby + reaching hand + bottle condition, this pretrained model contained negligible association with our bottle-feeding images. As discussed below, the baseline prediction model labels an image as "bottle-feeding" if any of the bottle-related object classes are detected in the image.

All models are evaluated using the performance metrics described in Chapter 4.1.5 (sensitivity, specificity, precision, accuracy, F1 score, and Matthews Correlation Coefficient, from Equation 4.1-4.7), where the positive class corresponds to "bottle-feeding" events.

### 4.4.2 Domain Distance Mapping

As previously mentioned, transfer learning can be useful in cases where the source domain contains a large amount of labeled data, to extract and transfer that knowledge to a smaller amount of labeled data. To this end, we repurpose the source task (classification of 1000 classes from VGG-16 model) to a new target task (classification of bottle-feeding interventions). Both domains are typically assumed to be similar; however, it is not always the case. This thesis presents such a scenario where the dataset used to train the VGG-16 model shares more similarity with the "no-feeding" class than the "bottle-feeding" class, as qualitatively demonstrated by Figure 4.11. Although both classes share similar features from the baby, nurse, and overall bed environment, the principal distinction remains in the presence or absence of a nursing bottle. Supplemental training data using similar-feature images including a nursing bottle can then bring both distributions closer together. The distance between domains can in fact be estimated for the BFID and BFID*exp* models. Measuring distances between domains is commonly performed in domain adaption, which is an unsupervised approach to transfer learning used when the source domain has labeled data, but the target domain does not [82], [196]. Domain adaptation seeks to minimize the gap between domain distributions during training by learning shared key features. More recently, this technique has also been used in multi-source domain adaptation where labeled data originate from multiple sources [197]. We here leverage this concept to measure the distances from the source domain to the domains of BFID or BFID*exp* models. Doing so demonstrates how the similar-feature data

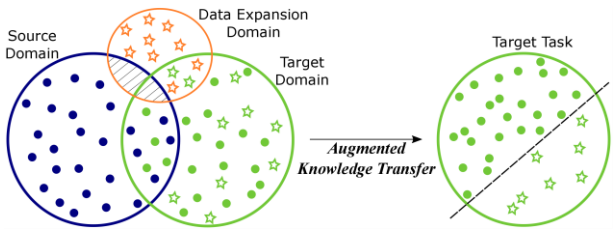



Figure 4.13: Similar-Feature Domain Expansion. Dots represent the negative class "no-feeding" and stars represent the positive class "bottle-feeding".

expansion can help narrow the distances between the source and target domain. This concept is illustrated in Figure 4.13, highlighting how the data expansion domain augments the knowledge transfer between the source and target domain to accomplish a specific target task.

To quantify and visualize the domain distances, feature maps from the BFID and BFID*exp* models are individually extracted. These correspond to the activation obtained across all training samples and the activations map these samples to the feature space. In order words, the stronger an activation in an area of the image, the greater the number of detected features in that area. Given that activation maps are provided as a greyscale image, the Otsu's thresholding technique is used to differentiate between high and low intensities. The resulting blobs represent areas of heightened activation. Since the principal difference between the "bottle-feeding" and "no-feeding" events is the presence of a nursing bottle, we can estimate the domain distance within the "bottle-feeding" class as the distance between the centroid of the nearest detected blob in the feature map and the actual nursing bottle object. The closer the blob to the bottle, the closer the domains.

To measure domain distances, all nursing bottles in our image dataset were first manually segmented and represented by their centroid. The domain distance is measured by the Euclidean distance (4.13), (4.14) between the bottle and feature map centroids, since previous domain adaptation studies showed negligible difference in domain distance mapping when using different distance metrics [196], [197]. When more than one activation is detected (*i.e.,* could detect two different objects) the closest activation to the object is selected.

Domain distance mapping can also be visualized per image by overlaying feature maps, gold standard bottle centroid, and distances on the original image. By visualizing these data, we can examine the impact of expanding our model using similar-feature data by evaluating changes in activations from feature maps. Domain distance mapping is evaluated as the percentage of closer bottle-activation distances in BFID*exp* compared to BFID among bottle-containing images. Additionally, the *pixelDistance* calculates how much closer the activation is to the bottle object in number of pixels using the following metrics:

$$dist_{BFID} = \sqrt{(centroid_{BFID} - centroid_{bottle})^2} \quad (4.13)$$

$$dist_{BFIDexp} = \sqrt{(centroid_{BFIDexp} - centroid_{bottle})^2} \quad (4.14)$$

$$pixelDistance = dist_{BFID} - dist_{BFIDexp} \quad (4.15)$$

### 4.4.3 Bottle Feeding Intervention Detection Dataset

Data from 27 patients from our CHEO dataset were used in the bottle-feeding dataset. The start and stop times for all intervention events were annotated, including bottle-feeding interventions. To ensure variations among images, one image was extracted every 30 seconds and only the color data from the

camera were analyzed. Given that our dataset shares similar features (e.g., patient present, a hand from the nurse or parent reaching into the frame, NICU bed environment), an image is classified as “bottle-feeding” if a nursing bottle is present at or near the patient’s mouth. If the nursing bottle is absent, the image is classified as “no-feeding”. Bottle-feeding events were only seen in six of out of the 27 patients. Other patients were fed by nasogastric tube or breast-fed. The complete dataset is summarized in Table 4.8. To supplement these hospital-based images, we extracted 60 Flickr images showing similar objects and context in the scene using the search word “bottle feeding baby”, as depicted in Figure 4.11. To simulate our CHEO data, a bottle-feeding image was included if it contained a baby, a hand reaching into the frame, and a nursing bottle at or near the baby’s mouth. The image environment could differ, where the baby would be placed on a pillow, blankets, a cradle, a feeding table, a baby bouncer, or in someone’s arms. Images showing an adult person’s face were excluded to closely simulate the NICU bed environment.

Table 4.8: Bottle-Feeding Dataset Breakdown

| **Class** | **Data source** | | | |
|---|---|---|---|---|
| | *CHEO* | | *Flickr* | |
| | *# images* | *# patients* | *# images* | *# subjects* |
| Bottle-feeding | 73 | 6 | 60 | 60 |
| No-feeding | 1187 | 27 | 0 | 0 |
| **Total** | **1260** | **27** | **60** | **60** |

## 4.5 Bottle Feeding Intervention Detection - Results

In this section, transfer learning results for all models are reported. In particular, the impact of data expansion on the knowledge transfer using the pretrained VGG-16 network is demonstrated. The domain distance mapping concept is finally presented to support and further explain our findings.

### 4.5.1 Transfer Learning & Data Expansion

As a baseline, the VGG-16 model was directly applied to our dataset and the top five predicted object classes were extracted since multiple objects can be found in the scene. This 1000-classification model outputs seven baby-related object classes and four bottle-related object classes. If the predicted object classes contained baby-related AND bottle-related classes, they were classified as "bottle-feeding". Otherwise, they were classified as "no-feeding".

In comparison with the two transfer learning models, the baseline method performs quite poorly, as depicted in Figure 4.14. Unsurprisingly, the baseline model has high specificity and accuracy values, strongly suggesting that the model is classifying images as the "no-feeding" class and cannot detect bottle-feeding events. Although no bib was used in clinical settings, that concept was useful due to association with babies, but it only appeared 1.9% of the time among the top-5 predictions. Similarly,

the *oxygen mask* class including a person wearing a breathing device and some images of babies on ventilator support leading to this class being predicted for 40.6% of images. The most frequently detected baby-related class was *diaper* (77.4%), while *water bottle* was the most frequent bottle-related class (1.6%).

In comparing with BFID and BFID*exp* models, results demonstrated a significant increase in performance after transfer learning, and even further improvement after similar-feature data expansion is applied. As displayed in Table 4.9, transfer learning results overperform the baseline and overall results are better for the BFID*exp* compared to BFID, especially in sensitivity (18.63% increase) and F1-score (3.4% increase). These two metrics are most pertinent in evaluating our methods, given the high class imbalance and the greatest concern in detecting bottle-feeding events.

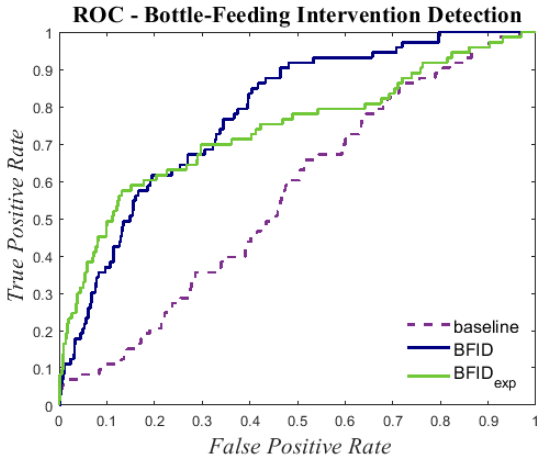



Figure 4.14: Bottle-Feeding Intervention Detection Results

Table 4.9: Results from Bottle-Feeding Intervention Detection

| **Model** | **Evaluation Metrics (%)** | | | | | |
|---|---|---|---|---|---|---|
| | *Sens* | *Spec* | *Prec* | *Acc* | *F1* | *MCC* |
| Baseline | 09.59 | **98.90** | 35.00 | **93.73** | 15.05 | 15.88 |
| BFID | 32.88 ±1.94 | 91.41 ±0.72 | **19.10** ±1.28 | 88.02 ±0.64 | **24.14** ±1.28 | 18.93 ±1.44 |
| BFID *exp* | **51.51** ±4.06 | 85.96 ±3.33 | **18.94** ±3.68 | 83.96 ±3.13 | **27.54** ±4.10 | **24.11** ±4.44 |

These findings thus corroborate how the shortage of "bottle-feeding" images used in training the VGG-16 model impacts the knowledge transfer ability to our classification task. Given that the "bottle-feeding" and "no-feeding" classes share many similar features, distinguishing between the absence or presence of the nursing bottle object is a difficult task to achieve. We have however demonstrated that our similar-feature data expansion technique can solve this issue.

### 4.5.2 Domain Distance Mapping

To evaluate the distance between the source and target domain, we opted for a heuristic approach where we measure the distance between the object of interest (nursing bottle) and the models' strongest activation from the feature map. When identifying key features of an object, we can naively consider it as a whole or focus on the more salient parts. For example, when seeing a torch, we typically focus our attention on the flame, not the handle. Similarly, we explore if the attention in a nursing bottle is focused on the whole bottle or the most salient part, i.e., the bottle cap. Both objects are manually annotated for evaluation.

Results reveal that the BFID and BFID*exp* models focused more on the bottle cap than the entire bottle. This suggests that the bottle cap shows greater saliency information attributed to the bottle object, as hypothesized by our torch object analogy. In fact, the BFID*exp* model detected the bottle cap in ~60% of the images, compared to ~54% for the BFID model. Interestingly, both models sometimes detected the soother object which shares very similar features to a nursing bottle cap (~7% for BFID*exp* and ~11% of images for BFID). In many cases, the BFID model still detected the soother, while BFID*exp* model learned to detect the bottle cap instead. Some of these examples are illustrated in Figure 4.15. This shows how data expansion can further teach our classifier to detect the correct object among two very similar ones. Spatially within the image, on average the BFID*exp* model detected an object at 111 pixels in distance to the nursing bottle while the BFID model detected the bottle at 126 pixels. This averaged 15-pixel difference may seem small, but it was observed with a maximum of 254 pixels when the BFID model is closer, and a maximum of 511 pixels for the BFID*exp* model (over twice as close). Our data expansion technique thus positively influences our model, given the closer distance to the bottle object. Other detections would include the patient's or nurse's arm or hand, the patient's head, toy, blankets, or cables, with comparable results from both models, and in absence of the bottle.

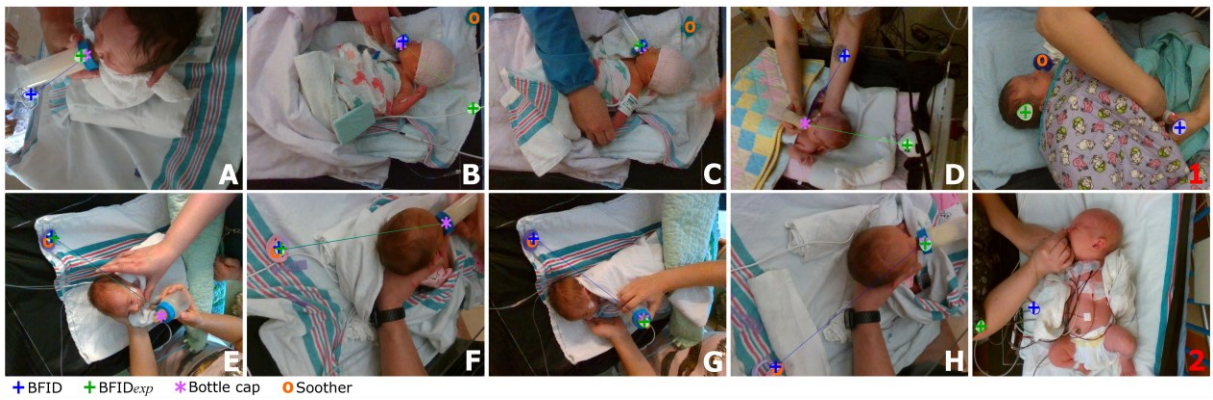



Figure 4.15: Domain Distance Mapping Results. Better performing model is A) BFID*exp*, B) BFID, C) both, and D) neither. The soother is detected in E-F) by both models, G-H) in BFID while BFID*exp* detects the bottle cap. 1-2) In absence of a bottle, both models detect other objects (ex: patient, nurse, cables).

## 4.6 Bottle-Feeding Intervention Detection - Discussion

Overall, we obtained promising findings from our similar-feature data expansion method. When applying this technique, it is important to exclude the supplemented data in the evaluation since it could systematically learn to distinguish between one's own dataset and the outside sources.

Our data expansion method can provide content but not always context. For example, an image of an adult holding a baby in one hand and a beer bottle in the other could satisfy our inclusion criteria (bottle & baby). Likewise, a photo of a child playing with empty beer bottles. However, these images have different meaning than a nursing baby. The original dataset obtained from data collection remains important to gain context for the classification task, while our similar-feature data expansion technique adds sufficient relevant content to address data scarcity and class imbalance. Not only is it time and cost effective compared to collecting new data, but it can substantially improve results for a difficult context-rich classification task.

## 4.7 Scene Understanding - Conclusions

This thesis addressed binary classification tasks pertaining to five context variables observed in the NICU. Image processing techniques were used for classification of lighting conditions. RGB-D deep learning approaches were investigated, revealing effectiveness of data fusion at different stages, depending on the context. Subsequently, we demonstrated how text generation from different contexts identified in NICU scene understanding can constitute a step towards semi-automated clinical charting. Within the "intervention" class, a categorization of one intervention (bottle-feeding) was achieved using a data expansion method with RGB images only. This data expansion technique was useful when addressing the data scarcity and class imbalance issue.

Note that further details of the dataset description and obtained results from this chapter are presented in Appendix E.

Overall, among scene analysis and bottle-feeding contexts, notable classification performance was achieved for all six context variables. However, additional experimentation would be required to increase precision and recall to the point where clinical deployment becomes feasible. In summary, this work shows great promise in scene recognition in hospital environments while providing a step towards automating the nurse documentation process, ultimately improving patient care.

# 5 Patient Region-of-interest Detection

After analyzing the scene, certain ROI can be detected from the patient, this thesis mainly identifies the face and body of the patient. This chapter addresses two ROI detection problems in the NICU: face detection and whole-body segmentation. First, several state-of-the-art face detection algorithms are explored and assessed for application to the NICU environment in Section 5.1. Key NICU-specific challenges are identified and addressed by advancing the state of the art in neonatal face detection through fine-tuning over NICU-specific data. Second, patient segmentation approaches are developed (5.2) using semantic segmentation of the patient vs the background for uncovered patients. Conclusions arising from this exploration of neonatal patient ROI detection are summarized in Section 5.3 and we assert that both research works are essential for robust non-contact continuous monitoring.

## 5.1 Neonatal Face Detection

In this section, state-of-the-art face detectors are assessed to detect the face of newborns in complex NICU scenes. Challenges arising from this environment that impede the performance of face detectors are identified and are later addressed through fine-tuning, resulting in two neonatal face detectors robust to such scenes. To evaluate and advance the state of the art in neonatal face detection, two public datasets are extracted from other studies to supplement our highly curated data from our CHEO dataset. Taken together, these comprehensive data represent various levels of scene complexity. This section describes the image data preparation for evaluating and improving face detection from video data collected at CHEO and the two public databases (5.1.1). Face detection methods identified from the state-of-the-art face detectors, along with the proposed fine-tuned models, are presented in Section 5.1.2, followed by model evaluation (5.1.3), and results and discussion (5.1.4).

### 5.1.1 Neonatal Datasets

This section describes the three neonatal datasets used for evaluating and advancing the state of the art in neonatal face detection.

Table 5.1: Face Detection Neonatal Datasets

| **Dataset** | *Tot Imgs* | *Unique Imgs* | *Patients* | *Age* | *Resolution* | *Avg. BBox Area* | *Viewpoint* |
|---|---|---|---|---|---|---|---|
| COPE | 288 | 183 | 27 | 18h – 3d | 3008 x 2000 | 23% | Close up face |
| NBHR | 889 | 565 | 257 | 0 – 6d | 640 x 480 | 15% | Close up face |
| CHEOopt | 2,048 | 111 | 16* | 4 – 64d | 640 x 480 | 8% | Full body |
| CHEOch | 11,517 | 1,855 | 33 | 4 – 64d | 640 x 480 | 3% | Full body |
| **Total** | 14,742 | 2,714 | 317 | 18h – 64 d | - | 3 – 23% | Multiple views |

*Subset of patients from the entire CHEO dataset, including only images representing optimal conditions (see text).

The four datasets are listed in Table 5.1 where *Tot Imgs* represents the total number of images extracted from the dataset, before finalizing the data to obtain unique images (described below). For the

distinct number of patients in each dataset, the *Age* represents their age on the day of the study, image resolution is also reported with *Avg. BBox Area* as the average area of the bounding box encapsulating the face of the patient in the image, and captured at various viewpoints.

#### *5.1.1.1 CHEO*

**Data extraction:** All ~153 hours of RGB video data were used from our 33 collected patients. One image was extracted per 30 seconds of video data. This provided substantial variation during events (*e.g.*, clinical intervention, patient motion) but insufficient variety when the patient is at rest. Therefore, images were further filtered to eliminate highly similar images.

**Image hashing:** To remove visually similar images, an average hash method was used. Each image was resized to 8x8, grayscaled, and the average of this new image is computed. Each pixel is then compared to the calculated average to compute a bit value (*e.g.*, set to 1 if above the average, and 0 otherwise) and all bits are extracted sequentially to form a 64-bit integer as the image hash. Images were then hierarchically clustered using hamming distance to compare hash values and only one image from each cluster was retained such that no two images had a hamming distance $\leq 5$. As shown in Table 5.1, image hashing reduced the original number of images (*Tot Imgs*) to visually distinct images only (*Unique*

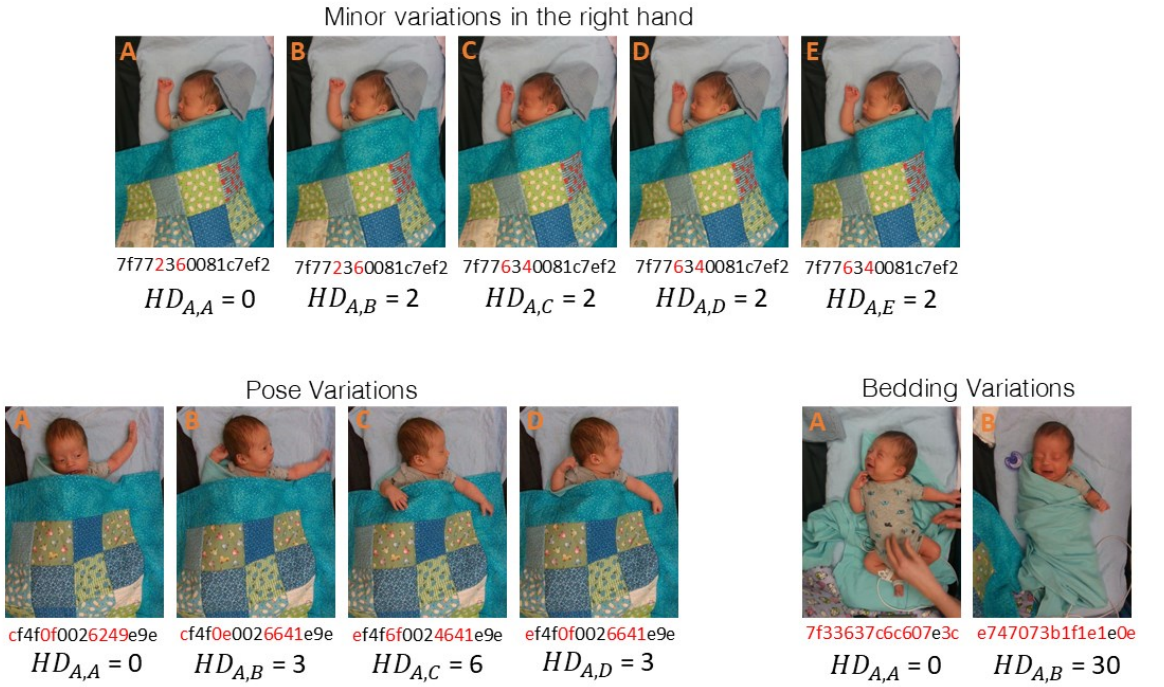



Figure 5.1: Image hashing and hamming distance evaluation using sample images from the CHEO dataset. The average hashing value is computed (below the image) and the hamming distance (HD) is calculated at different video frames with respect to the first frame (A) in the three distinct scenarios. A larger HD would suggest larger visual variations in the image.

*Imgs*). Sample images from this dataset are illustrated in Figure 5.1 to demonstrate the visual comparison among images with different hashing values, and evaluation of hamming distance.

**Data curation:** The CHEO image set was subdivided into "optimal", "challenging", and "negative" data subsets. The "optimal" subset ($CHEO_{opt}$) includes images where the patient's face is clearly visible, with high lighting, no facial occlusion, clear frontal view, close distance from the camera (max 60 cm), no ongoing phototherapy treatment, no ongoing clinical intervention, and no blur due to patient motion. The "challenging" subset ($CHEO_{ch}$) included the opposite cases from the "optimal" subset. The "negative" set was excluded from further analysis and contained those images where the face of the patient is not visible, such as complete facial occlusion, face out of frame, patient absent from bed, or complete darkness making it impossible to visualize the patient's face for a human observer.

**Standardized face orientation:** The camera is typically at a fixed position and orientation for the entire recording session, but the orientation varied between patients. As a preprocessing step, all images were rotated such that the head is at the top of the image (referred to as the "North" orientation). In Appendix D, a method to automatically determine head orientation is developed and assessed.

**Face annotation:** Faces within each image were manually annotated. Bounding boxes captured the area from forehead to chin and ear to ear, and only visible parts were selected in cases of partial occlusions.

#### *5.1.1.2 COPE*

The Infant Classification of Pain Expressions (COPE) dataset was obtained from Brahnam *et al.* [198], [199] where their research focused on neonatal pain assessment. The database contains 288 images of 27 newborn faces in the NICU, collected during a painful procedure (*e.g.*, heel lancing), vs non-painful ones (*e.g.*, light puff of air on the nose, friction from rubbing alcohol). To finalize data preparation, **standardized face orientation**, **image hashing**, and **face annotation** were performed similarly as for the CHEO dataset described in Section 5.1.1.1.

#### *5.1.1.3 NBHR*

The newborn baby heart rate estimation database (NBHR) was obtained from Huang *et al.* [13] where they collected synchronized video recordings and physiologic signal for non-contact neonatal heart rate estimation. The database includes 9.6 h of facial videos collected among 257 patients, with photoplethysmograph (PPG) signals, heart rate values, and oxygen saturation levels. The dataset consisted of 1130 videos, where for each video, the first frame was extracted as an image. To finalize data preparation, **standardized face orientation**, **image hashing**, and **face annotation** were performed similarly as described in Section 5.1.1.1.

### 5.1.2 Face Detection Methods

This section describes the pretrained models used in this thesis (5.1.2.1), in addition to supplemental analysis of complex scenes (5.1.2.2). Finetuned models are then created using the best pretrained networks to create our NICUface models (5.1.2.3).

#### *5.1.2.1 Pretrained Models*

Four pretrained models are used here: MTCNN, CE-CLM, RetinaFace, and YOLO5Face.

**MTCNN:** The pretrained MTCNN model is tested without modification (see Section 2.4.2) on our four neonatal datasets.

**CE-CLM:** The CE-CLM can use different face detectors as a backbone of the CEN network to obtain landmark positions. This thesis leverages the MTCNN model for the CEN backbone. Predictions differ from the MTCNN model in that they are further refined and also include the 68-point face alignment.

**RetinaFace:** The pretrained RetinaFace was used with a ResNet-50 backbone model, as described in Section 2.4.2.

**YOLO5Face:** The YOLO5Face model was used with the "large" Stem block since this was shown to be one of the most accurate by Qi *et al.* [7]. The YOLOv5l6 is not used here since they reported that, while the P6 block addition improved performance on the WIDER FACE's Easy and Medium subsets, it can decrease the performance on the Hard subset (which more closely resembles our data).

#### *5.1.2.2 Complex NICU Scenes*

Complex scenes are further analyzed by evaluating face detection performance under various clinical challenges. Using the $CHEO_{ch}$ dataset, we extract challenging cases based on varying levels of occlusions, viewpoints, and lighting. In terms of occlusions, they can occur when the patient is sucking on a soother, from the nurse's hand or arm during a clinical intervention, when the patient is wearing a phototherapy eye mask during treatment, from a ventilation support device, or from free-moving limbs or beddings. Viewpoints are considered to be challenging when the camera is positioned at a far distance from the patient (> 1m); when the patient is being held in the bed; or when the face is only visible in profile view, from a near-top view, or from near-back view when the patient is in prone position. Lighting conditions are challenging during dimly lit periods (*e.g.*, patient sleeping or reduced sensory input environments) or during phototherapy treatment. To evaluate these complex scenes, the best performing pretrained models (RetinaFace and YOLO5Face) are tested on each challenging case before being finetuned. Previous neonatal monitoring applications have discussed how challenges from clinical scenes can pose a problem to the face detection performance. This thesis quantifies the impact of these individual complex scenes.

#### *5.1.2.3 Finetuned Models*

We create the NICUface detectors from finetuning pretrained RetinaFace and YOLO5Face models. For both models, models were trained and evaluated on different patient subsets to quantify model generalization, as described in Table 5.2. Note that the same 16 patients from $CHEO_{opt}$ were present in $CHEO_{ch}$, only with different challenging scenes. For example, Split 1 is trained only on COPE + NBHR and is tested on $CHEO_{opt}$ to ensure that testing data arise from entirely different patients from those seen during training. For the $CHEO_{ch}$ data, given its variety of challenging conditions, the 17 unique patients in this dataset (not present in $CHEO_{opt}$) were divided into three folds. Each fold contained a proportional number of complex scenes, especially considering very challenging scenes assessed from the pretrained models (e.g., low lighting and patients on ventilation support). The final reported performance on the $CHEO_{ch}$ dataset is reported as the average of these three folds for the pretrained and finetuned models to provide a fair comparison, as demonstrated by "Avg(4:6)" in the last row of Table 5.2

Table 5.2: Face Detection Train & Test Sets

| **Split** | **Train** | **Test** |
|---|---|---|
| 1 | COPE + NBHR | CHEOopt |
| 2 | COPE + CHEOopt + CHEOch | NBHR |
| 3 | NBHR + CHEOopt + CHEOch | COPE |
| 4 | COPE + NBHR + CHEOopt + CHEOch_notF1 | CHEOch_F1 |
| 5 | COPE + NBHR + CHEOopt + CHEOch_notF2 | CHEOch_F2 |
| 6 | COPE + NBHR + CHEOopt + CHEOch_notF3 | CHEOch_F3 |
| Avg(4:6) | --- | CHEOch |

##### 5.1.2.3.1 NICUface-RF

The RetinaFace model was finetuned using a ResNet50 backbone and the RetinaFace weights. Finetuning occurred over 10 epochs with a batch size of 8 with an initial learning rate of 0.001 with a warmup to 0.1 at epoch 1 and then 0.1 decay for epochs 2 and 5. Anchors were matched to an object when the IOU was larger than 0.45 and to the background when the IOU was less than 0.3. Training data were augmented with random horizontal flip and photo-metric colour distortion. The loss function was not changed from the original RetinaFace model; however, during training landmark regression error was ignored by setting all landmark inputs in the training data to -1. Only bounding box error was used.

##### 5.1.2.3.2 NICUface-Y5F

The YOLO5Face model was finetuned using the YOLOv5l weights, trained over 10 epochs with batch size of 16, initial learning rate of 0.0032 and final learning of 0.12, optimized using stochastic gradient descent with 0.5 momentum in the first 2 epochs and momentum of 0.843 after, and an IOU threshold

of 0.2 during training. The loss function of NICUface-Y5F is similar to the loss function of YOLO5Face as

$$loss = loss_{box} + loss_{conf} + loss_{cls} + \lambda_{land} \cdot loss_{land} \quad , \qquad (5.1)$$

where $loss_{box}$ is the bounding box regression loss, $loss_{conf}$ is the confidence loss, $loss_{cls}$ is the classification loss, and $loss_{land}$ is the landmark regression loss with weighting factor $\lambda_{land}$. This $\lambda_{land}$ was set to only 0.005 to pay less attention to the landmarks given the unsupervised landmark localization. Similarly to YOLO5Face, the $loss_{conf}$ and $loss_{cls}$ were optimized using the cross-entropy loss function. In terms of data augmentation, YOLO5Face reported that Mosaic augmentation and removal of up-down flipping improved their performance, but only on the Hard WIDER FACE subset without ignoring small faces or random cropping. Given the difficulty of our neonatal dataset we also applied Mosaic and removed up-down flipping.

The training set was divided into two sets of data used during training and validation stages where different patients were used for training and validation. As can be seen in Table 5.2, each face detector was tested on a completely different dataset from that used to train the models.

### 5.1.3 Face Detection Evaluation

For face detection performance of pretrained and finetuned models, all models are evaluated using the average precision metrics with varying IOU requirements. The AP is calculated with IOU $\geq$ 0.5 as a standard evaluation metric (AP50), while the mAP captures the mean over AP with IOU=0.5:0.05:0.95 to reward models producing more specific bounding boxes. The facial landmarks are not evaluated quantitatively here since no gold standard landmark annotations were performed on our neonatal datasets; however, landmarks are reviewed qualitatively to generally assess the performance of the face alignment task and to identify challenging cases where the alignment would fail.

In evaluating all models, we opted to only output the prediction with the highest confidence score. This approach is feasible since only one face is assumed to be present in each image. Given the difficult task of finding neonatal faces in complex scenes, this approach allows low confidence predictions of the patient's face to still be considered while ignoring other irrelevant false predictions in the scene.

We also look at cases where we decrease the IOU threshold to 0.30 (AP30) to include slightly overestimated or underestimated bounding boxes around the face. Although most object detectors report AP with IOU of at least 0.5, recent applications leveraging these detectors have opted for lower IOU threshold in cases where the objects are small and hence the AP/mAP metric would be drastically impacted by marginal errors [200], [201].

### 5.1.4 Results & Discussion

This section assesses the state of the art in face detection for neonatal patients in NICU environments. Experiments cover neonatal face detection challenges from multiple experiments with different pretrained models, datasets, and complex NICU scenes.

Table 5.3: Face Detection Results from All Models and Datasets

| Model | COPE | | | NBHR | | | CHEOopt | | | CHEOch | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AP30 | AP50 | mAP | AP30 | AP50 | mAP | AP30 | AP50 | mAP | AP30 | AP50 | mAP |
| MTCNN | 74.31 | 74.31 | 51.79 | 60.07 | 59.74 | 38.18 | 49.95 | 48.83 | 26.41 | 7.32 | 4.93 | 1.62 |
| CE-CLM | 92.08 | 91.19 | 31.94 | 79.34 | 77.29 | 25.81 | 64.10 | 57.77 | 16.35 | 16.19 | 8.95 | 1.75 |
| RetinaFace | **100** | **100** | 76.73 | **100** | **100** | 78.60 | **99.95** | **99.95** | **79.12** | 52.47 | 52.12 | 29.56 |
| YOLO5Face | **100** | **100** | 83.80 | 99.56 | 99.67 | 77.95 | 95.73 | 95.73 | 76.39 | 50.78 | 48.58 | 28.65 |
| NICUface-RF | **100** | **100** | 86.55 | **100** | **100** | 80.53 | 99.10 | 99.10 | 77.92 | 86.12 | 79.73 | 43.67 |
| NICUface-Y5F | **100** | **100** | **88.30** | 99.95 | 99.95 | **82.39** | 93.16 | 93.16 | 76.73 | **88.61** | **87.98** | **61.75** |


Figure 5.2: Visualization of Face Detection Results from All Models and Datasets. The level of complexity in the NICU scene increases from left to right, and the model performance increases from top to bottom. Predictions are labelled as correct (IOU ≥ 50, Green), partial (IOU ≥ 30, Yellow), or incorrect (IOU<30, Red).

#### *5.1.4.1 Pretrained Models & Neonatal Datasets*

Among all pretrained models presented in Table 5.3 (blue-shaded) and depicted in Figure 5.2, MTCNN performs worst, and interestingly, the CE-CLM model using MTCNN as a backbone detector performs better in comparison. The fact that landmark positions are refined in the CE-CLM model before applying the denser 68-point distribution model strongly suggests the advantage of the CEN layers in the localization task. Using 68-point landmarks could, however, be application-dependent since there is

no real need to get a fine-detailed facial structure for HR estimation or jaundice detection (which primarily look at the skin), but it would be highly relevant for pain assessment or sleep-wake detection (which primarily look at the facial expression). This could open a door to retraining an $x$-point landmark distribution model suitable for neonatal population with $x$ salient facial features observed in newborns, where x is 5, 68, or, another suitable number of landmarks specifically for newborns in a clinical setting.

Figure 5.2 depicts results from all models with increasing level of scene complexity from left to right, and increasing performance of each model from top to bottom. Bounding box predictions are labelled as correct (IOU ≥ 50, Green), partial (IOU ≥ 30, Yellow), or incorrect (IOU<30, Red). As illustrated in Figure 5.2 some false negatives with MTCNN have become true positives with CE-CLM for the COPE dataset (with correct detection and decent facial alignment despite the partial occlusion by blanket), for the NBHR dataset (with partial detection and misaligned facial landmarks due to profile view), and for the $CHEO_{opt}$ dataset (with partial detection and proper facial alignment). For the $CHEO_{ch}$ dataset, no detection is obtained with MTCNN and CE-CLM for most scenes, except for a few with very minor occlusions and viewpoints where all facial landmarks are visible (e.g., patient imaged from a far distance).

By a rather large margin, the pretrained RetinaFace and YOLO5Face models outperform MTCNN and CE-CLM in a complementary manner. While RetinaFace performs best on NBHR, $CHEO_{opt}$, and $CHEO_{ch}$, YOLO5Face performs best on COPE.

Across all models, a consistent pattern exists in dataset performance with COPE > NBHR > $CHEO_{opt}$ > $CHEO_{ch}$. This pattern agrees with a qualitative assessment of the level of difficulty among our datasets in analogous fashion to the WIDER FACE dataset's easy, medium, and hard subsets [121]. Our COPE data represent our "easy" subset with close up facial views, NBHR has "medium" difficulty with close up faces and more challenging poses and occlusions, $CHEO_{opt}$ is "medium-hard" where the image includes the full body and bed environment. Finally, $CHEO_{ch}$ is a "hard" dataset as it includes the entire bed environment and complex scenes, such as low lighting, ventilation support, pose variation, etc. Considering that the performance of all models is significantly reduced on the $CHEO_{ch}$ dataset, the complex scenes therein are further analyzed in Section 5.1.4.2. The datasets are then leveraged for the implementation of NICUface using the best competing pretrained models (RetinaFace and YOLO5Face) in Section 5.1.4.3.

#### *5.1.4.2 Complex NICU Scenes*

Among all datasets, $CHEO_{ch}$ led to the lowest face detection peformance for all models due to the complexity of scenes included therein. The increasing level of difficulty among these complex scenes is

presented in Table 5.4 and Figure 2.4 (reproduced here as Figure 5.3) with varying levels of occlusions, viewpoints, and lighting.


Figure 5.3: Face Detection Difficulty in Complex NICU Scenes

Both RetinaFace and YOLO5Face demonstrated a similar pattern of performance across the complex scenes. Mouth occlusions from a soother are not as challenging as partial occlusions from the nurse's arm/hand or from beddings. Near complete facial occlusions from the ventilation support remain the most challenging occlusion-based scenes. Among viewpoints, far distance and profile view performed best, but near-top view or prone position are most challenging given that only a small portion of the face is visible. From lighting environment, a low lighting environment doesn't affect the model as severly as the phototherapy light. Note that the phototherapy eye mask is also an occlusion-based challenge, however, we dimmed the blue-colored lighting more important to this unique scenario. All NICU-specific complex scenes remain to be learnt. Having established the limits of the state of the art in face detection for complex NICU scenes, we turn our attention to addressing the remaining NICU-specific challenges through finetuning, leading to the NICUface models.

Table 5.4: AP30 Face Detection from Complex NICU Scenes on CHEOch using RetinaFace & YOLO5Face

| **Challenges per categories** | | **RetinaFace** | **YOLO5Face** | **Num Images** |
|---|---|---|---|---|
| Occlusions | Soother | 99.50 | 96.58 | 24 |
| | Intervention | 60.96 | 64.48 | 88 |
| | Bedding/self | 61.29 | 53.95 | 286 |
| | Ventilator | 20.33 | 3.04 | 195 |
| Viewpoint | Far distance | 67.90 | 84.17 | 134 |
| | Profile | 68.83 | 69.96 | 47 |
| | Near top view | 40.45 | 42.24 | 24 |
| | Prone position | 7.34 | 1.11 | 18 |
| Lighting | Low lighting | 75.93 | 41.39 | 36 |
| | Phototherapy | 33.93 | 33.33 | 19 |

#### *5.1.4.3 NICUface*

In this section, we report on the performance of the NICUface models, where we have finetuned the top-performing RetinaFace and YOLO5Face models for NICU-specific challenges (see Table 5.2 for datasets used for finetuning and evaluation). From the results of the pretrained models, it was established that the RetinaFace and YOLO5Face already performed very well across COPE, NBHR, and $CHEO_{opt}$ datasets. Given the near-perfect performance of these models across these datasets, it is unsurprising that comparable results were obtained with the NICUface models, with near 100% AP30 and AP50 values among these three datasets. The advantage of fine-tuning becomes apparent on the $CHEO_{ch}$ dataset, where the NICUface models exhibit large improvements on this challenging data, as demonstrated in Table 5.4. NICUface-RF showed an increase of +33.65, +30.67, and +17.74 in AP30, AP50, and mAP respectively compared to RetinaFace. NICUface-Y5F showed an increase of +37.83, +39.40, and +33.10 in AP30, AP50, and mAP respectively compared to YOLO5Face. Between both NICUface models, NICUface-Y5F slightly outperformed NICU-RF on $CHEO_{ch}$ with a difference of +2.49, +8.25, and +18.08 in AP30, AP50, and mAP, respectively.

As illustrated in Figure 5.2 and in Table 5.5, the NICUface models showed robustness to the presence of ventilation support and patients in near-back view when in prone position, while pretrained models were impaired by these scenes. Given that these two complex scenes are the two most challenging ones, NICUface-RF demonstrates impressive performance with an improvement in AP30 of +68.74 and +35.47 for the prone position and ventilation support, respectively. NICUface-Y5F also improves drastically with AP30 of +78.31 and +62.83 for the prone position and ventilation support, respectively.

Table 5.5: AP30 Face Detection Results from Complex NICU Scenes on CHEOch using NICUface-RF and NICUface-Y5F)

| **Challenge** | **NICUface-RF** | **Challenge** | **NICUface-Y5F** |
|---|---|---|---|
| Profile | **100** | Profile | **100** |
| Soother | **100** | Soother | 99.23 |
| Near top view | **100** | Near top view | 98.24 |
| Bedding/self | **98.79** | Bedding/self | 97.57 |
| Low lighting | **89.61** | Far distance | **97.02** |
| Far distance | 89.01 | Intervention | **95.21** |
| Intervention | 87.92 | Low lighting | 87.42 |
| Prone position | 76.08 | Prone position | **79.42** |
| Ventilator | 55.80 | Ventilator | **65.87** |
| Phototherapy | **0** | Phototherapy | **0** |

Moreover, both models are highly complementary to one another. NICUface-RF presents strengths in detecting patients in low lighting conditions (with +13.68 improvement in AP30), while NICUface-Y5F is better at detecting smaller faces (with +12.85 improvement in AP30). These conditions are illustrated in Figure 5.2, where NICUface-RF was able to correct RetinaFace's false positive by correctly detecting

the face of the patient under vary low lighting. In the same scenario, YOLO5Face also made an incorrect prediction but NICUface-Y5F was not able to rectify this error. On the other hand, during a clinical intervention, the face of the patient captured from a far distance was detected with NICUface-Y5F, while NICUface-RF avoided a previous false positive but overestimated the bounding box area. This improvement is still remarkable, given that it was now able to make a detection in the general location of the face, however it fails to reach the precision of NICUface-Y5F.

Among all evaluation metrics, the AP30 is the most reliable measure of model performance for neonatal monitoring applications. In our case, the frequent presence of small faces in the CHEO dataset warrants evaluating with a smaller threshold than the standard AP50. As seen in Table 5.1, our most challenging dataset has an average bounding box area that only makes up 3% of the image. In such cases, NICUface would tend to slightly overestimate the bounding box area which would severely affect the AP metric, despite the relevant prediction. Slight overestimation is not an issue for monitoring applications requiring the entire face for facial expression analysis in pain assessment, sleep-wake cycle detection, or face recognition. Slight underestimation is also not an issue when small facial ROI can be sufficient in some applications such as HR estimation or jaundice detection relying on visible skin patches. Due to high level of facial occlusions in the NICU, some non-contact neonatal monitoring applications have opted for different techniques to only obtain visible facial area (*e.g.*, skin segmentation). The AP30 metric is therefore a most reliable measure since lowering the IOU threshold permits considering slightly overestimated or underestimated predictions which can still be useful in a wide array of neonatal applications.

Among all neonatal monitoring applications presented in Chapter 2, Awais *et al.* [46] was the only one performing automatic face detection and reporting its performance (to the best of our knowledge). They achieved 98.5% accuracy using the Fluke TiX580 camera for intensity-based face detection on patients with 0-degree head tilt (i.e., frontal view). In comparison, NICUface-RF and NICUface-Y5F achieve 100% accuracy on our COPE dataset which most closely compared to their dataset. For more challenging scenes, NICUface-RF still performs remarkably well with 100%, 99.10%, and 73.87% accuracy for NBHR, $CHEO_{opt}$, and $CHEO_{ch}$, respectively. NICUface-Y5F also performs exceptionally well with our most challenging data with 100%, 88.29%, and 83.77% accuracy for NBHR, $CHEO_{opt}$, and $CHEO_{ch}$, respectively.

Among the pretrained face detectors with facial alignment, YOLO5Face performed best on our most challenging dataset and was therefore used to create a face orientation estimation method (described in Appendix D). This method exploits the confidence scores from predicted bounding boxes in conjunction with the landmark position from facial alignment to rotate the image such that face of the patient is orientated North. Such standardization approach is required in most non-contact monitoring

applications, and often performed manually or programmatically as a trial-and-error approach. This thesis presents a simple but reliable face orientation approach as a pre-processing step.

For both NICUface models, the blue-colored light during phototherapy treatment (in addition to the facial occlusion from the eye mask) posed a problem in face detection performance. Interestingly, their pretrained counterparts were able to detect a few images when the nose and face were visible, resulting in an AP30 of ~ 33 for both. NICUface-Y5F did show promise with very small detections from the visible skin in a few images, however with an IOU < 0.3. The following section presents a solution to improve face detection of patients undergoing phototherapy treatment, without having to retrain the entire model.

### 5.1.5 Face Detection of Phototherapy Patients

Given that the blue-colored light in phototherapy images can be challenging for face detection, we present a solution to modify the image during inference time before face detection is attempted using NICUface detectors. Doing so, we leverage the phototherapy classification presented in Chapter 4.1 to differentiate from images captured in natural lighting. This technique can be performed in three simple steps:

1. Detect phototherapy images (presented in Chapter 4.1-4.3)
2. Pre-processing blue filtering for phototherapy images (presented here)
3. Face detection using NICUface (presented in Chapter 5.1.1-5.1.4)

#### *5.1.5.1 Phototherapy Detection*

Recall that Chapter 4.2 presented results in classifying phototherapy images with Equation 4.4 describing a lower and upper bound of 0.10 and 0.35 for the phototherapy index, with 0.23 suggested as the final threshold value. These index values are tested across different datasets including images with natural lighting (COPE, NBHR, $CHEO_{opt,}$), phototherapy lighting (only phototherapy images from $CHEO_{ch}$, flickr images used in Chapter 4.1-4.3), and a mix of both lighting environment ($CHEO_{ch}$). Using data with varying lighting to finetune the phototherapy index ensures the generalization of this phototherapy detection method. Results presented in Table 5.6 and Figure 5.4 demonstrate the accuracy of phototherapy detection using different phototherapy index thresholds and datasets. Bolded values represent suggested phototherapy index threshold. The CHEOch_photo dataset represents the phototherapy patients obtained in the $CHEO_{ch}$ dataset (resulting in 19 images), while flickr_photo represents 11 images obtained from online sources used here for comparison.

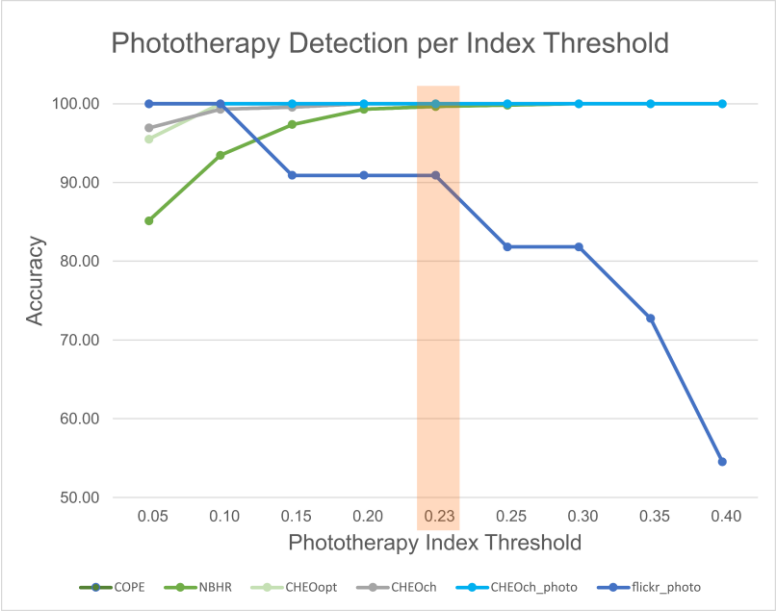



Figure 5.4: Phototherapy Detection Rate vs. Index Threshold across Six Datasets

Table 5.6: Phototherapy Detection Accuracy using Different Phototherapy Index across Varying Datasets

| Lighting | Dataset | Accuracy per Phototherapy Index Threshold | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *lower bound* | | | | *suggested* | | | *upper bound* | |
| | | **0.05** | **0.10** | **0.15** | **0.20** | **0.23** | **0.25** | **0.30** | **0.35** | **0.40** |
| Natural | COPE | 100 | 100 | 100 | 100 | **100** | 100 | 100 | 100 | 100 |
| | NBHR | 85.13 | 93.45 | 97.35 | 99.29 | **99.65** | 99.82 | 100 | 100 | 100 |
| | CHEOopt | 95.50 | 100 | 100 | 100 | **100** | 100 | 100 | 100 | 100 |
| Mix | CHEOch | 96.93 | 99.30 | 99.57 | 100 | **100** | 100 | 100 | 100 | 100 |
| Phototherapy | CHEOch_photo | 100 | 100 | 100 | 100 | **100** | 100 | 100 | 100 | 100 |
| | flickr_photo | 100 | 100 | 90.91 | 90.91 | **90.91** | 81.82 | 81.82 | 72.73 | 54.55 |

In accordance with the suggested middle-point, we select 0.23 as the final phototherapy index threshold to favour the detection of phototherapy images (100% and 90.91% testing accuracy), while limiting the false positives in natural images given that face detection in natural lighting already performed well using NICUface models.

### *5.1.5.2 Pre-processing blue filtering.*

In Chap 4.1-4.3, we discussed how the Red, Green, and Blue channels in the natural images are almost uniformly distributed. In comparison, phototherapy images are heavily weighted with blue-colored pixels, relative to red-colored pixels. This important knowledge is exploited here to perform a color space transformation on the phototherapy images to equalize the colour channels. Our "Blue

Filtering" method scales pixel intensities of the red and blue channels to the pixel intensities of the green channel to equalize the image as

$$Scale_R = \frac{1}{w \times h}\left[\sum_i^w \sum_j^h G_{ij} - \sum_i^w \sum_j^h R_{ij}\right] \quad (5.2)$$

$$Scale_B = \frac{1}{w \times h}\left[\sum_i^w \sum_j^h B_{ij} - \sum_i^w \sum_j^h G_{ij}\right] \quad (5.3)$$

$$R_{ij}^* = \begin{cases} R_{ij} + Scale_R, & \text{if } R_{ij} + Scale_R \leq 255 \\ 255, & \text{otherwise.} \end{cases} \quad (5.4)$$

$$B_{ij}^* = \begin{cases} B_{ij} - Scale_B, & \text{if } B_{ij} - Scale_B \geq 0 \\ 0, & \text{otherwise.} \end{cases}, \quad (5.5)$$

where $Scale_R$ measures the scaling factor using $G_{ij}$ and $R_{ij}$, the Green and Red channels, respectively, for pixels at the $i^{th}$ position among image width ($w$), and $j^{th}$ position among the image height ($h$). $R_{ij}^*$ represents the updated Red channel in the "equalized-phototherapy" image, as illustrated

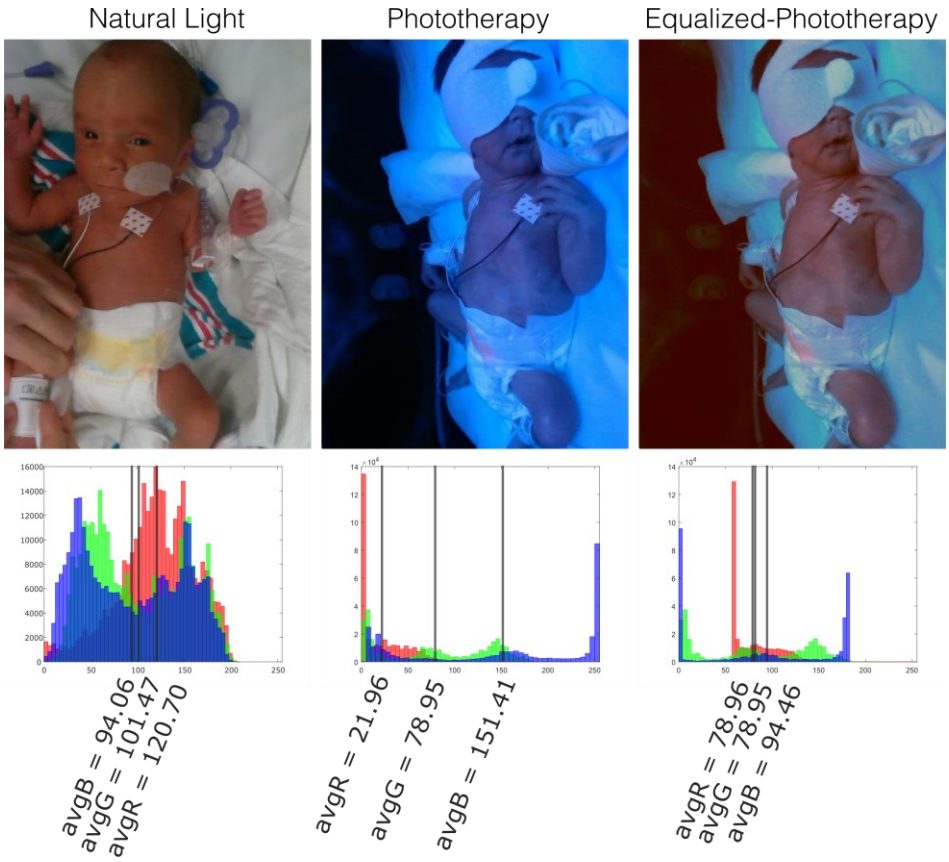



Figure 5.5: Blue Filtering of Phototherapy Images. The average of each corresponding RGB channel are sparse in the Phototerapy image, and thereby attempts to narrow the gaps across channels in the Equalized-Phototherapy image simulating the Natural Light condition.

in Figure 5.5. Given that the $Scale_R$ value is added to each pixel, the updated $R^*_{ij}$ caps all pixels exceeding 255 as an intensity of 255. Similar steps are performed to obtain an updated Blue channel $B^*_{ij}$ by scaling down the pixels by $Scale_B$, and capping all pixels inferior to 0 as an intensity of 0.

As we can see in Figure 5.5, the predominant blue hue is successfully filtered out, especially in areas of the visible skin. Other surfaces still have a slight blue tint; this is apparent in areas known to be truly white, such as the bedding or eye mask.

#### *5.1.5.3 Face Detection with NICUface and Blue Filtering*

During inference, the phototherapy detection is applied directly on the image to differentiate between lighting environments. Images deemed to represent ongoing phototherapy are processed using the Blue Filtering method and NICUface detects the face in the modified image. Images deemed to reflect natural lighting are unchanged.

To validate this face detection approach on phototherapy patients, the 19 images from the only patient in our CHEOch dataset are used and evaluated with AP30 metric. This method is validated using the best performing face detectors (RetinaFace, YOLO5Face, NICUface-RF, and NICUface-Y5F).

#### *5.1.5.4 Results & Discussion*

Face detection results are demonstrated in Table 5.7, upon using phototherapy detection and blue filtering or not. On the state-of-the-art models, an improvement in AP30 of +2.47 and +16.22 is observed on RetinaFace and YOLO5Face, respectively, with blue filtering. An even more impressive improvement is observed with the NICUface models with +50.00 and +41.49 AP30 for NICUface-RF and NICUface-Y5F, respectively. Qualitative results are shown in Figure 5.6, where both NICUface models can now correctly detect the face or close regions.

Table 5.7: Face Detection Results on Phototherapy Patients after applying Blue Filtering

| **Model** | **Without Blue Filtering** | **With Blue Filtering** |
|---|---|---|
| RetinaFace | 33.93 | 36.40 |
| NICUface-RF | 0 | **50.00** |
| YOLO5Face | 33.33 | 49.55 |
| NICUface-Y5F | 0 | 41.49 |

The results of the Blue Filtering pre-processing method shown in Figure 5.6 are remarkable given that the face of the patient is now detected despite being highly occluded by the phototherapy eye mask. Also, these results were obtained without having to retrain the entire NICUface model. These results show great promise in the application of pre-processing methodologies for color-based challenges in other machine vision technologies during inference time, without requiring retraining an entire model.

This is particularly useful in cases when obtaining new data can be a difficult or expensive task to achieve.

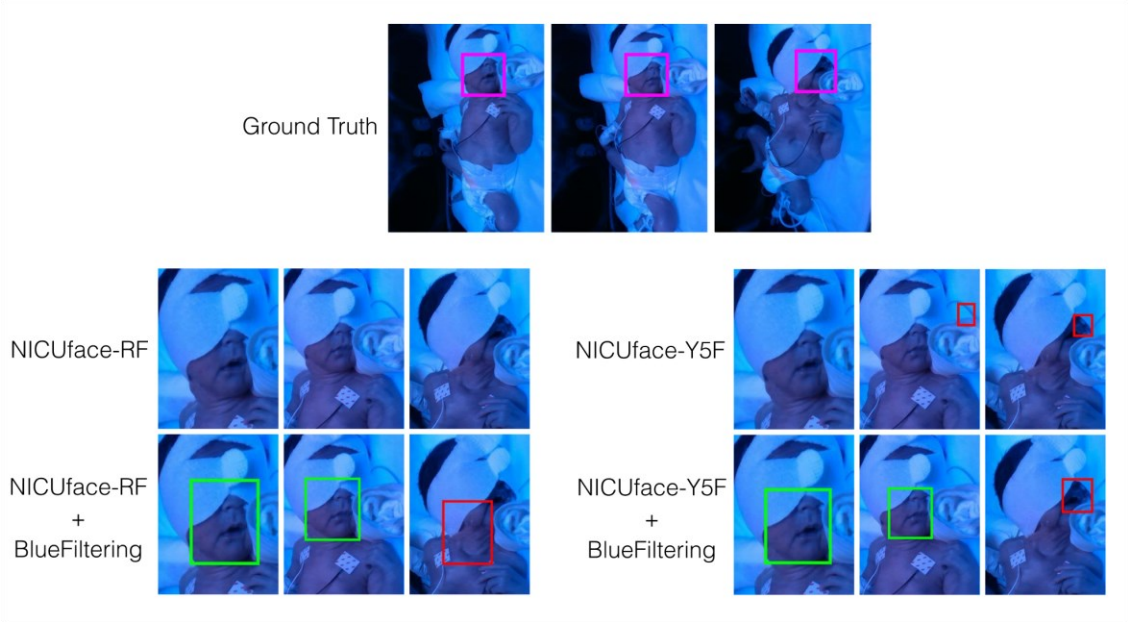



Figure 5.6: Examples of Face Detection on a Phototherapy Patient using NICUface models with and without Blue Filtering.

### 5.1.6 Neonatal Face Detection - Conclusions

Neonatal face detection is a difficult task to achieve, and this thesis demonstrated several scenarios when it might fail due to complex NICU scenes. By evaluating the state of the art with four different models and three different datasets, we found that patients on ventilation support and in prone position were most difficult scenes, and NICUface detectors were finetuned to address this issue. Furthermore, phototherapy patients present a variety of challenges due to the blue-colored lighting and occlusion from eye mask. A blue filtering approach leveraging the phototherapy detection is therefore a suitable and reliable solution for such cases. This technique showed to be effective with the images from the one phototherapy patient collected at CHEO; however, having only one phototherapy patient in our study presented its limitations. Although images were extracted at different times throughout the 6-hour recording to provide further variations in the scene, the blue filtering approach would need to be validated on more patients with varying skin tones and level of skin coverage to validate the robustness of our blue filtering technique.

Obtaining a reliable ROI is an important step in neonatal monitoring, and this chapter presented solutions for detecting the face, while the following section will demonstrate techniques for segmenting the full body of the patient.

## 5.2 Patient Segmentation

Patient segmentation is an import ROI step in neonatal monitoring given that it focuses on the full body of the patient for subsequent monitoring applications such as respiration rate estimation, full body motion detection, or limb motion detection. Many studies have opted for skin segmentation given that it is often challenging to find a proper delineation of the patient against a crowded background in the NICU bed environment. Solely detecting the skin can however be problematic for clothed patients, or those covered/wrapped in beddings. This thesis therefore investigates semantic segmentation of patients under varying levels of coverage (unclothed, clothed, wrapped in beddings, and partially covered with blankets).

In this section, two state-of-the-art segmentation models, SegNet and Mask R-CNN, are evaluated for the task of patient segmentation. While SegNet has an older architecture and performs pixel-wise semantic segmentation, we evaluate its applicability in patient segmentation. With the newer Mask R-CNN performing instance segmentation, we evaluate its performance using an existing pretrained model and using a larger and modified dataset. All these techniques are applied to assess the state of the art in semantic segmentation of patients in a clinical setting. A post-processing image processing approach is also presented in conjunction with these CNNs to adjust the patient-background boundary.

### 5.2.1 SegNet

The SegNet model depicted in Figure 5.7 demonstrates the segmentation process with a sample image previously used for the implementation. For this patient segmentation work, we used translation, scaling, and reflection to augment the dataset.

To prevent the gradient from biasing, all training data are randomly shuffled before each epoch. Moreover, a small learning rate of 0.01 is used allowing adequate time for variations to be learnt, with SGDM with momentum of 0.9 is added to prevent from possible oscillations of the optimizer by

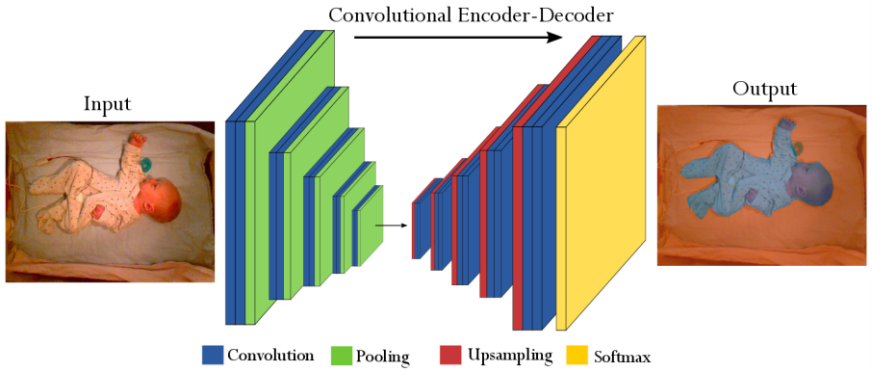



Figure 5.7: SegNet: Deep Convolutional Neural Network for semantic segmentation. An image is inputed to the network that encodes each pixel within the frame by assigning a label. Specific locations of each pixel are assessed by transferring indices to corresponding decoder levels where the image is upsampled, thereby recreating the scene as a semantic segmentation output. Adapted from [9].

providing a boost to overcome potential fluctuations. The maximum number of epochs is limited to 300 and we added a stopping criterion upon increase of validation loss. After training the network, a new set of images is used for evaluation.

### 5.2.2 Mask R-CNN

The pretrained Mask R-CNN is used here (as described in Chap 2.4.3), and this model was implemented on multiple objects for instance segmentation. For our application, we extract the top-scoring "person" class object given that we only expect to segment one person in the image. Though bounding boxes and masks are provided, only masks are used to evaluate the semantic segmentation performance.

### 5.2.3 Post-processing Image Filtering

Given the noisy results presented at the delineation between the patient and the background, a 17x17 *median filter* is applied on the segmented images. Since SegNet performs pixel-wise segmentation, it produces grainy outputs, thus, **morphological closing after semantic segmentation** is applied on the filtered image using a 16-pixel radius disk as the structuring element. For the Mask R-CNN model, segmentations are smoother given the mask-based outputs for instance segmentation, and random patches are often observed near the delineation. **Morphological opening after instance segmentation** is therefore applied on the filtered image using the 16-pixel radius disk.

### 5.2.4 Algorithm Evaluation

Several segmentation metrics exist, suited for different applications. We selected the Jaccard index and Dice coefficient to assess the performance of our model since they are appropriate for 2D segmentation problems. Note that Jaccard and IOU are used interchangeably in this thesis. In the following equations, the gold standard annotated image is labelled as $G$ while the resulting segmented image is labelled as $F$.

$$Jaccard\ Index = \frac{|F \cap G|}{|F \cup G|} \qquad (5.6)$$

$$Dice\ Coefficient = \frac{2|F \cap G|}{|F| + |G|} \qquad (5.7)$$

Both metrics are computed on images belonging to the test set, providing realistic estimate of the expected performance of the SegNet and Mask R-CNN models when applied to new future images.

### 5.2.5 Patient Segmentation Dataset

This research focuses on the segmentation of patients where head, chest and limbs are mostly visible within the frame. To train the SegNet model, a small dataset was used. We partitioned the dataset into a training set containing 86% of the data across three patients (total of 78 images) and a testing set

containing 14% of the dataset (one new patient with a total of 16 images). Variations in the dataset include patients in an overhead warmer or a crib.

While validating the Mask R-CNN, several modification and additions were made with the small dataset used for SegNet. The same 4-patient data is used (U4), these images are rotated for standardized head orientation (UR4), and a larger dataset is acquired containing 30 patients and standardized head orientation (UR30). From the 33 collected CHEO uncovered patients, images were curated to remove clinical interventions thereby only obtaining the person in the image, visually similar images were removed using the image hashing approach presented in Chapter 5.1, and images were rotated for North-facing head orientation using the method described in Appendix D. This resulted in 222 images among 30 patients presenting variations in pose, bed type (incubator, crib, and overhead warmer), lighting (including phototherapy), skin tone, and skin coverage (unclothed, clothed, wrapped in blanket).

### 5.2.6 Results & Discussion

The SegNet model demonstrated highly promising results. Figure 5.8 demonstrates a sample segmentation acquired from our method as evaluated on the test set. Having learned from the three training patients, the model is now aiming to obtain a segmentation based on learnt features and variants from other images. We can safely claim that this task was accomplished since the resulting segmentation enclosed all body parts of the patient, however the overall outcome appears heavily noisy, especially around the edges. This can be observed by the granular segmentation at the boundary between the patient and the background. Not unexpectedly, these boundary regions proved to be more difficult to differentiate, compared to the central area of the patient. Post-processing image filtering helped improve the overall segmentation resulting in smoother edges and a large improvement in performance metrics from 62.86% Jaccard and 77.20% Dice from model predictions alone to 82.05% Jaccard and 82.40% Dice with post-processing technique, as demonstrated in Table 5.8.


Figure 5.8: Example of SegNet semantic segmentation results. Left: Gold standard annotation; pixels belonging to background are shown in orange. Middle: Algorithm segmentation obtained using SegNet (purple = patient). Right: Model predictions with post-processing filtering.

Table 5.8: Patient Segmentation Results

| Model | Dataset | Test patients | Model Predictions | | With Post-processing | |
|---|---|---|---|---|---|---|
| | | | Jaccard (%) | Dice (%) | Jaccard (%) | Dice (%) |
| SegNet | U4 | 1 (pt6) | 62.86 | 77.20 | **82.05** | **82.40** |
| Mask R-CNN | U4 | 1 (pt6) | 48.53 | 60.76 | 48.36 | 60.60 |
| Mask R-CNN | UR4 | 1 (pt6) | 63.81 | 77.66 | 63.62 | 77.53 |
| Mask R-CNN | UR4 | 4 | 69.18 | 80.23 | 68.83 | 79.97 |
| Mask R-CNN | UR30 | 30 | **74.09** | **83.35** | **83.35** | **83.41** |

With the Mask R-CNN model, testing the pretrained model on the same patient tested from finetuned SegNet revealed worst performance with decrease of -14.33% Jaccard and -16.44% Dice. This can in


Figure 5.9: Example of best Mask R-CNN results.

part demonstrate the utility of finetuning a semantic segmentation model on our neonatal dataset despite the small dataset size, and especially given that both SegNet and Mask R-CNN models were implemented on adults, mostly in an upright standing position. Upon rotation of images for standardized head (and therefore body) orientation, comparable results are observed from the model predictions, again supporting the fact that up-right position are preferable for this model. With the larger dataset of 30 patients, interesting results were observed given the high performance and wide variation of complex poses. Some of the best examples are presented in Figure 5.9 where Mask R-CNN can accurately detect outline of the patient clothed or not, from a close or far distance, and with varying levels of occlusions from blankets. For more challenging scenarios depicted in Figure 5.10 (e.g., prone position, phototherapy lighting, partial occlusion from blanket), the model is still able to detect the majority of the patient, however delineations are not always clear, especially when blankets and clothes are confused as part of the patient or not. For this dataset, the post-processing method improved the pretrained Mask R-CNN results with 83.35% Jaccard and 83.41% Dice. Such results are comparable as the finetuned SegNet, thereby suggesting that training on few data in conjunction with an image processing approach can be sufficient when limited data is available.

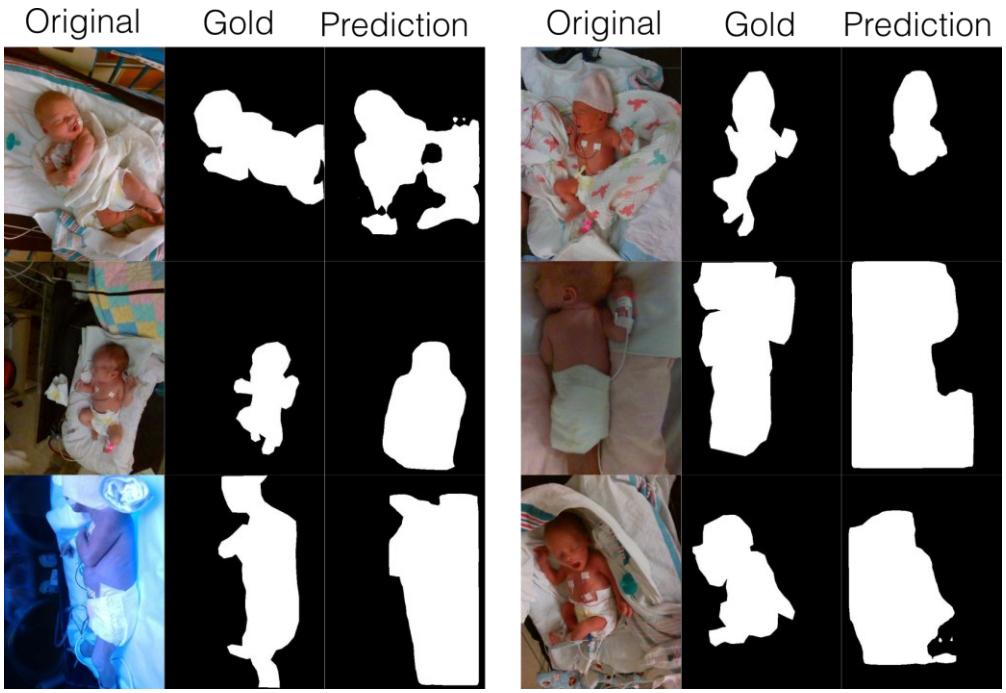



Figure 5.10: Example of worst Mask R-CNN results.

Other neonatal monitoring applications performed specific body part segmentation, as described in Chapter 2.1. While Antink *et al.* [47] achieved 51.0% mean IOU, Asano *et al.* [50] obtained 70.4% mean IOU for overall body segmentation. In comparison, this thesis achieved 83.4% mean IOU, which

surpasses both approached by a large margin. Previous studies have struggled to perform body part segmentation due to different boundaries from one part to another, in addition to the delineation from the background. This thesis demonstrated how similar challenges were observed during occlusions from beddings to properly segment the unoccluded body region, during complex positions where limbs and body region are confounded together, and with clothed patients where the boundary between the clothes and bedding is hard to define. Despite these challenges, the Mask R-CNN model in conjunction with image filtering + morphological opening presents very promising results for patient segmentation in complex NICU scenes.

## 5.3 Patient Region-of-Interest Detection – Conclusions

This thesis assessed the state of the art and developed methods for detecting the face and body of the patient despite the challenging scenes observed in the NICU. We have demonstrated how NICUface models are robust to the most complex scenes including patient on ventilation support or in prone position and proposed a post-processing image filtering approach to improve results of phototherapy patients.

We have also demonstrated the potential of using a finetuned deep learning model with limited data or a pretrained model with data variation for patient segmentation in the NICU. We also addressed the issue in patient-background delineation using an image processing approach in conjunction to the predictions obtained from a CNN model. Since the body shape of newborns in a clinical setting is quite complex with limb boundaries often hard to define, standardizing the body orientation is recommended.

In both ROI detection tasks, this thesis shows how combinations of deep learning and image processing techniques are greatly advantageous when faced with complex data. To this end, the identified patient ROI can serve as an initial step for developing a patient monitoring system for neonates in the NICU, and Chapter 8 demonstrate one use case for HR estimation.

# 6 Motion Detection

This chapter includes methods and results from three motion detection research works useful for non-contact monitoring. The first work demonstrates Neonatal Motion Detection (6.1 – 6.2) where the absence or presence of patient motion is identified. The second work presents a method for Neonatal Face Tracking (6.3 – 6.4) by continuously tracking the face of the patient after it has been detected by methods such as those from Section 5.1. Lastly, Limb Motion Detection (6.5 – 6.6) recognizes which limb is moving, as one approach to further categorize patient motion. Concluding remarks (6.7) describe how each motion detection work can impact the overall neonatal monitoring process. With each method, the state of the art is assessed for detecting motion from different ROI and under varying NICU scenes to identify gaps in these methodologies.

## 6.1 Neonatal Motion Detection - Methods

To detect patient motion in the scene, two different approaches are explored: a computer vision model using the optical flow algorithm and a deep learning network using an LSTM. For both methods, we address the complex NICU scenes, including video recorded from varying viewpoints, of neonatal patients often covered by clothing or bedding, and with varying levels of patient movements.

### 6.1.1 Optical Flow Technique

The optical flow method can determine the presence of motion between subsequent frames by estimating the change in perceived intensity at each pixel [139]. This change is demonstrated by a vector field with orientation and magnitude corresponding to the detected motion. We leverage this motion vector field (MVF) to differentiate between presence of motion (dense MVF) vs. absence of motion (sparse MVF), as

$$Flow_x = median\left(\sum_{i=1}^{n-1} M_{t+i}\right), \qquad (6.1)$$

with clip number $x$, time step $t$, magnitude $M$, calculated over $n$ number of frames, where a high Flow value would suggest presence of motion. Figure 6.1 illustrates this approach, where the MVF is calculated for all frames within a video clip. The sum of the magnitudes in the MVF is computed and represents the perceived degree of motion in the scene. To estimate motion in the entire video clip, the median of the distribution of all summed magnitudes is used. The median is selected here to avoid undue influence by outliers resulting from lighting variations and non-continuous motion observed in the video. All Flow values are obtained per video and the distribution of "motion" vs "no motion" is differentiated using Otsu's thresholding.

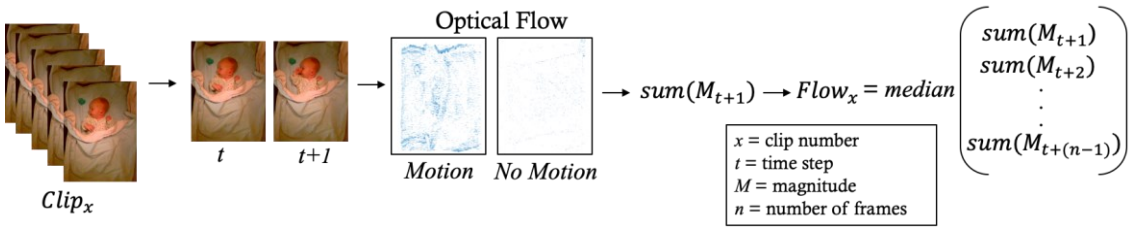



Figure 6.1: Motion Detection from Optical Flow-based Pipeline.

### 6.1.2 Long Short-Term Memory Network

To differentiate between motion classes, we developed a system that leverages a long short-term memory (LSTM) network for video classification, paired with a pre-trained classifier (GoogLeNet [202]) for feature extraction, as illustrated in Figure 6.2. For proper training and testing data representation, as a preprocessing step, all videos were standardized to have the head of the patient oriented North. From an input video, the sequence of frames is unfolded and the CNN extracts features from each frame. Given that the classifier was previously trained on a very large dataset [72], we can leverage the transfer learning approach where pre-trained weights are used. Once key features are extracted, the video is reassembled according to the original sequence structure. Each processed video is then passed into an LSTM recurrent neural network for video classification. The network comprises of 200 LSTM hidden units, followed by a dense feedforward network to classify each video as "motion present" vs "no motion".

The LSTM network is used in the present thesis given its demonstrated ability to monitor long-term dependencies within sequential input data. Doing so, information passed at the beginning of a video can

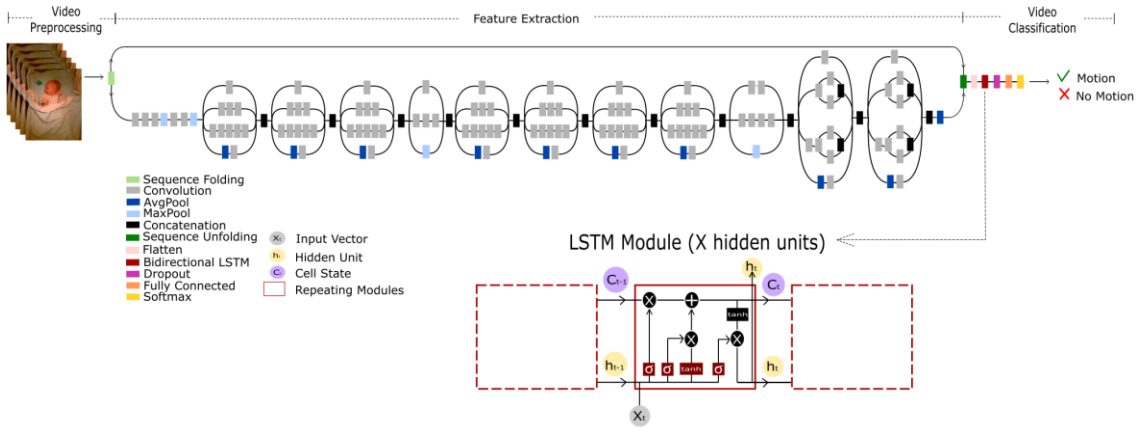



Figure 6.2: Motion Classification from LSTM-based Pipeline. Parameters for classification of "Motion" vs "No Motion" classes. Video preprocessing includes North-facing body standardization of images. Feature extraction is performed using the GoogleNet model before being fed to the LSTM module for video classification.

be retained, if deemed valuable for the understanding of the entire sequence. To this end, relevant context is extracted from the entire video before it is eventually classified.

### 6.1.3 Model Evaluation

For both above-mentioned techniques, the performance of the method is evaluated using the precision and recall metrics, where the positive class is defined as videos with motion present.

### 6.1.4 Neonatal Motion Detection Dataset

As a proof-of-concept for this thesis, three patients from a crib and three patients from an overhead warmer bed were included in the dataset. This work only analyzed color videos for motion detection. Ground truth data were obtained from real-time bedside annotations provided by the CEA app. All "no motion" events correspond to instances when the patient is resting and not moving at all. All "motion" events correspond to those where visible limbs are moving and/or the visible head is moving. This distinction is important, given that patients often have partial occlusions from bedding, and only the visible part of the body is detected for movements. Exclusion criteria include patients in ventilation support, low lighting environment, periods of time when the patient was removed from the bed, and scenes when the camera was moved. An event is defined as a period of time during motion, or between two motion events. From each event, one or more short videos were extracted, using a non-overlapping 16-second window. While each "motion" vs "no motion" video maximized the presence vs absence of motion, respectively, a mix of both events could occur in the selected video. These videos were used to develop and evaluate both motion detection approaches. Furthermore, the patients were divided into training and testing subsets, such that videos from any patient were never used to both train and evaluate a method. A detailed description of the dataset is provided in Table 6.1. Each patient had ~2-6 hours of usable data.

Table 6.1: Neonatal Motion Dataset Description

| **Data Split** | **Bed Type** | **Patient** | **Event** | | **Total Clips/ Pt** | **Total Clips** | **Total Hours** |
|---|---|---|---|---|---|---|---|
| | | | ***No motion*** | ***Motion*** | | | |
| Train (71%) | Crib | 5 | 545 | 185 | 730 | 3569 | 14:59:51 |
| | | 11 | 547 | 81 | 628 | | |
| | OHW | 14 | 933 | 209 | 1142 | | |
| | | 15 | 958 | 111 | 1069 | | |
| Test (29%) | Crib | 2 | 705 | 293 | 998 | 1411 | 5:51:26 |
| | OHW | 6 | 305 | 108 | 413 | | |

## 6.2 Neonatal Motion Detection – Results & Discussion

### 6.2.1 Optical Flow Technique

The greatest challenge with the optical flow technique was overcoming noise from outliers caused by changing lighting conditions and non-continuous motion. A video including a change in environmental lighting would impact the motion vector field generated between adjacent frames, resulting in a highly dense vector field. Similarly, halted or discontinuous patient movement during a "motion" event would cause a sparse motion vector field to be generated, potentially confusing the classifier. By computing the median of summed magnitudes, we limit the impact of these outliers; however it could not be completely avoided, as depicted by the partial overlapping distributions of median summed flow in the left panel of Figure 6.3.

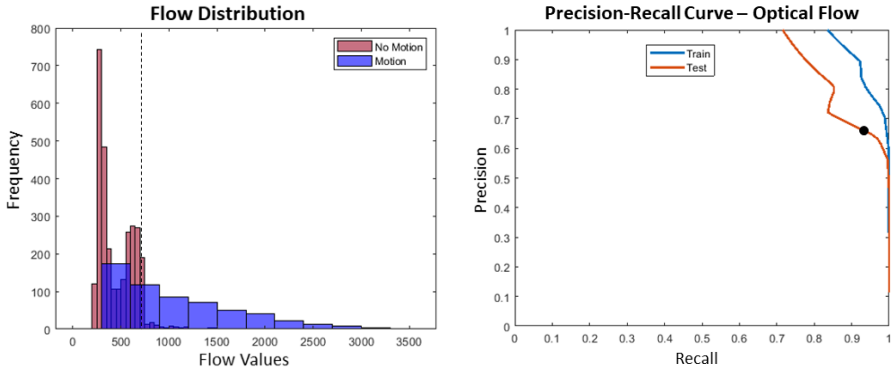



Figure 6.3: Motion Detection Results using Optical Flow. Distribution of flow values with selected threshold (left) and Corresponding precision-recall curve for train and test dataset for the optical flow technique (right).

The right panel of Figure 6.3 illustrates a precision-recall curve for the optical flow technique. Here, the decision threshold is varied from being highly conservative (high precision, low recall) to highly permissive (low precision, high recall). As shown in Figure 6.3 (right), performance is excellent for this technique. A high testing recall strongly suggests that the model is able to detect presence of motion accurately, as depicted in Table 6.2.

Table 6.2: Motion Detection Results from Optical Flow and LSTM methods.

| Method | Precision (%) | Recall (%) |
|---|---|---|
| Optical Flow | 68.02 | 90.51 |
| LSTM | 31.02 | 82.54 |

### 6.2.2 Long Short-Term Memory Network

Results obtained using the LSTM technique are shown in Table 6.2. When compared to the optical flow technique, a similar recall is obtained, but with substantially lower precision value. The significant

difference in performance can be seen in the lower panel of Figure 6.4, where a higher testing recall was selected (at cost of a lower precision) to increase the correct prediction of motion events. These performance outcomes could be explained by multiple factors which must be taken into consideration when training and evaluating a deep network. By directly testing the pretrained GoogLeNet model on a sample image in the left upper panel of Figure 6.4, the network predicted the class "bassinet" with a score of 40%. This "bassinet" class is selected among the 1000 classes used to train the GoogLeNet model given that this class includes images where a baby would be present in the image (i.e., in the feature extraction step to detect the patient before detecting motion with the LSTM layers). Areas in the image contributing to this prediction are depicted in the middle panel of Figure 6.4, where overlayed color pixels represent classification scores to the predicted class. This figure aims to visually represent the feature extraction process from the CNN, given that it depicts the gradient of the classification scores generated by the last convolutional layer of the GoogLeNet network. The output of this layer is then used to feed our video classification pipeline using an LSTM. With the sample image, we can see that the facial area has greatest influence on the classification.

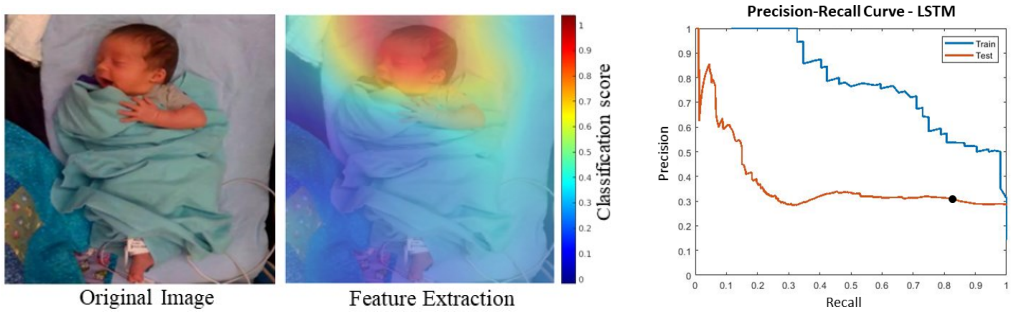



Figure 6.4: Motion Detection Results using LSTM. Sample image from dataset (left) with illustration of pixel influence on CNN classification (middle). Precision-recall curve for train and test dataset for the LSTM technique (right).

Other factors contributed to the network performance, including the 1:5 class imbalance between the "motion present" vs "no motion" class which was not addressed in this application. Considering that a classifier will be biased towards the dominant class in such situations, this represents a possible explanation for a lower recall value. Additionally, given the complexity of the LSTM network and the number of free parameters that must be learned from the data, the overall dataset size could be larger to provide more data for the model to learn.

### 6.2.3 Neonatal Motion Detection – Discussion

This thesis presented how optical flow or LSTM based methods can be used for detecting presence of patient motion in the scene. Although the optical flow technique performed well in detecting motion, one limitation of this approach is that one cannot distinguish between motion-related events. Such

application would require a more complex method that is not constrained to aggregating a score across the entire frame. A deep learning approach using an LSTM can, however, be employed to overcome this issue and is left for future work outside the scope of this thesis.

Now that we have established the feasibility of detecting overall patient motion in the scene, the next two sections will address how specified ROIs can be leveraged for tracking a moving object (face tracking in Section 6.3 – 6.4) or for classifying a moving object (limb motion detection in Section 6.5 – 6.6).

## 6.3 Neonatal Face Tracking – Methods

Detected movement of a patient's facial region is required for several patient monitoring applications including HR estimation from facial ROI or pain assessment from facial expression. This section presents such a face tracking method; each step is discussed before detailing the evaluation process. Two facial ROI tracking approaches are described here: 1) tracking by displacement only, 2) tracking by detection only. Finally, a proposed novel fusion approach of these two techniques is demonstrated to overcome the weaknesses of each. By tracking the face, we are able to continue to extract accurate facial ROI even when face detection may fail (e.g., during transient obstructions). Figure 6.5 depicts the proposed method that combines these two tracking techniques.

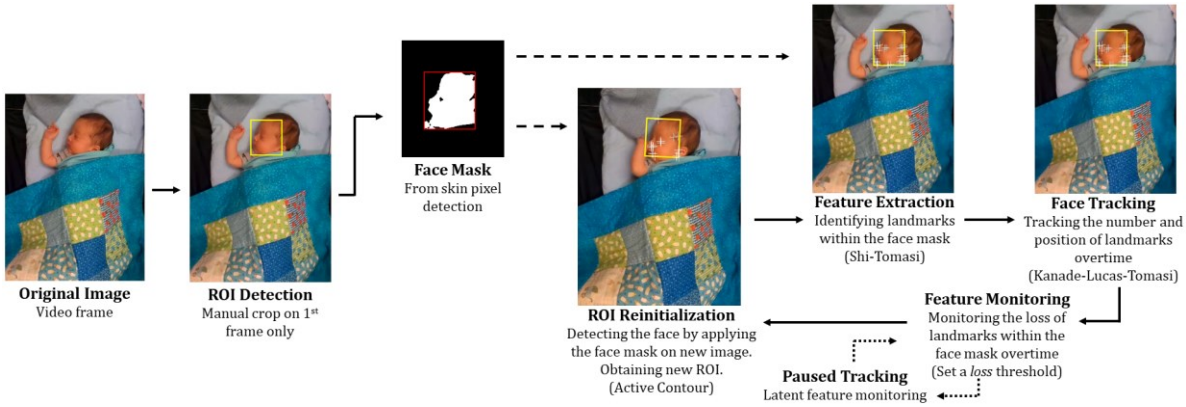



Figure 6.5: Proposed method for neonatal facial ROI tracking. Tracking by detection reinforced by displacement.

This thesis addresses two main challenges: 1) changes in patient pose and 2) temporary occlusions. In our dataset, patient poses mainly vary due to roll, with the head moving from side to side (frontal and profile). Occlusions are largely self-induced, as arms wave in front of the face or when the patient places their hands on their face. External occlusions during periods of clinical interventions are excluded; tracking during such events will be investigated in future work. During occlusion periods, pre-identified tracking landmarks can be lost, thereby hindering the tracking process. Upon significant loss of these landmark features, the tracked ROI can become incorrect or be lost entirely. This work addresses this

problem by identifying *when* tracking is degrading by monitoring landmarks, *why* a degeneration is observed by identifying events contributing to this loss, and *what* can be done to resume proper tracking by reinitializing a correct ROI.

While a tracking algorithm can account for translational changes due to strict object displacement or rotation, significant changes in object appearance due to changes in pose must also be addressed. In the neonatal patient monitoring context, significant head movements (e.g., from left profile to frontal to right profile view), can make facial ROI tracking challenging. We therefore propose to periodically detect the face of the patient using a 2D mask as a template. By defining a collection of templates corresponding to multiple expected poses, reliable ROI detection can be achieved for neonatal patients.

### 6.3.1 Tracking by Displacement only - KLT Algorithm

#### *6.3.1.1 Initial Region-of-Interest Detection*

As previously mentioned, a tracking process begins with identification of a facial ROI. Using the NICUface-Y5F model presented in Chapter 5.1, the face of the patient is automatically detected from the first frame.

#### *6.3.1.2 Feature Extraction*

Once an ROI is detected, features must be selected within this region to be tracked thereafter over time. Multiple algorithms are available to select features based on texture or morphology information (e.g. pixel intensity, corners, change in texture) [203], [204]. Although this approach successfully identifies key landmarks in an image, these features are not necessarily optimal for tracking. Shi and Tomasi have explored this issue at length, resulting in a method to identify strong features that can be reliably tracked [205]. Features are selected by comparing the eigenvalues of the gradient at a given pixel to a pre-defined threshold. This threshold is typically measured as the half-way point between an eigenvalue of uniformly-textured region and an eigenvalue of a highly-textured region.

#### *6.3.1.3 Face Tracking*

Once reliable features have been selected, they must be tracked between consecutive frames. Rather than tracking individual pixels between frames, Lucas and Kanade [151] proposed a method using spatial intensity gradient information to track features between successive frames. The combination of the Shi and Tomasi method of feature selection, with Lucas and Kanade's tracking algorithm, is referred to as the Kanade-Lucas-Tomasi (KLT) approach and is used here.

### 6.3.2 Tracking by Pose Detection only - Mask Application

As an alternative to tracking by displacement using the KLT approach, tracking a facial ROI can be accomplished from tracking by detection. In this approach, a facial mask (defined as all skin pixels within the original bounding box) is detected in each frame.

#### *6.3.2.1 2D Face Mask Generation*

In the tracking-by-detection approach, an initial detection of the object is obtained, followed by repeated detection in subsequent frames. Here, the correspondence problem is solved given that the object in frame $t$ is matched by finding its new location in frame $t+\alpha$. We use a skin detection method for identifying all pixels attributed to the patient's skin perceived in the image. The initial ROI provided in Section 6.3.3.1 is used here to define the region corresponding to the patient's face at any given moment.

#### *6.3.2.2 Defining Face Mask using Active Contour*

To detect the face of the patient using the 2D mask, skin pixel detection is first applied on the video frame resulting in one or more binary blobs. The active contour method is then employed to identify the blob most likely to correspond to the patient's face, beginning with the known face location in the first frame. This approach utilizes an initial snake-like shape and aims to fit boundaries of the desired object by minimizing the average difference in pixel intensities between the snake and the object, inside and outside of that snake boundary. Active contour techniques are then used to shift the initial contour to match the new face region, which may have both translated and morphed slightly. From a resulting segmented contour, a bounding box can be drawn enclosing the object, indicating the detected object.

### 6.3.3 Tracking by Detection Reinforced by Displacement

Our final proposed approach combines tracking by displacement (described in section 6.3.1 above) with periodic reinitialization when tracking fails by leveraging tracking by detection (described in section 6.3.2 above).

#### *6.3.3.1 Region-of-Interest Detection*

The initial ROI is obtained similarly to Section 6.3.1.1. Additionally, to provide sufficient masks to enable reinitialization of the tracking algorithm (see below), we propose to generate three masks using two manually defined frames when the patient monitoring is first started: one "frontal" mask captured during a frame where the patient is facing the camera and one "profile" mask from another frame. A horizontal flip of the profile frame provides the mask for the case where the patient has rolled their head fully from one side to the other. This approach does assume that a patient's face is largely symmetric.

#### 6.3.3.2 *Feature Monitoring*

Throughout the tracking procedure, tracked features can be lost due to a surge in affine motion during pronounced motion events or due to occlusions. During and following these periods, a degeneration in tracking can be observed. We propose to monitor the number of tracked features as an ROI tracking quality metric. A substantial loss of tracked features would suggest that a considerable variation has occurred in the scene (e.g., the patient has changed poses), thereby indicating that the algorithm should reinitialize the ROI and restart tracking from that point. Without triggering such a reinitialization, the ROI can be lost over time, as demonstrated by our results.

#### 6.3.3.3 *Region-of-Interest Reinitialization*

In the case where a loss of tracking is detected, we propose to resume tracking by automatically reinitializing the ROI, thereby reidentifying a sufficient number of features within this new region. The three facial skin masks created during initialization can be leveraged to detect the face in this new frame. By considering all three face masks, our approach allows for changes in patient pose (e.g., head rolling from side to side, or frontal to profile). In this way, the tracking pipeline is reinitialized using the most suitable ROI and new features can be detected and tracked.

#### 6.3.3.4 *Paused Tracking*

Despite all efforts to achieve robust tracking even with patient motion, occlusions, and changes in pose, there are times where face tracking remains infeasible (e.g., due to a combination of these challenges, prolonged occlusions, or during clinical intervention events). In these cases, tracking should be halted until suitable conditions are resumed. For example, identifying the patient's face for vital sign monitoring should be paused when the patient is absent from the bed, having nasogastric tubes changed, or when the face is completely occluded. This work therefore proposes a latent feature monitoring approach, where the number of features are periodically tracked while tracking is paused. Upon an increase in a number above a certain threshold, the ROI is deemed reliable and detection and tracking is resumed.

#### 6.3.3.5 *Evaluation*

To evaluate the performance of the tracking algorithm, gold standard data from manual ROI annotations were compared with tracking-based detections. These detections are evaluated by measuring the *Jaccard Index* between the gold standard and the obtained ROI.

From the plot of the Jaccard Index over time, we can additionally calculate the area under this curve (AUC) as a metric for evaluating the performance of the tracking algorithm during continuous monitoring. These measures are used to compare a baseline method that uses an unchanging initial ROI,

the KLT tracking-by-displacement approach, the tracking-by-detection approach, and our proposed hybrid approach.

To better understand the limitations of the tracking algorithm, a retrospective inspection of the events contributing to the degeneration of tracking is conducted. By determining the type, length, and number of events leading up to the degeneration point, we gain insight into why a loss in tracking occurred.

### 6.3.4 Face Tracking Dataset

The face tracking application used color videos from three neonates, where a continuous 20 minutes of data were analyzed per patient. Video data includes consecutive series of "rest" periods when the newborn's head and body are not moving significantly, "motion" periods when their head is moving, and "occlusion" periods when their face is occluded by moving limbs or beddings. A breakdown of the data used per patient is detailed in Table 6.3, describing total continuous monitoring and length of event periods (divided across many periods during the 20-min continuous monitoring).

Table 6.3: Face Tracking Dataset Breakdown

| **Patient** | **Event periods (MM:SS)** | | | **Total Continuous Monitoring (MM:SS)** |
|---|---|---|---|---|
| | ***Rest*** | ***Motion*** | ***Occlusion*** | |
| A | 12:35 | 04:25 | 03:00 | 20:00 |
| B | 17:31 | 01:05 | 01:24 | |
| C | 06:59 | 03:46 | 09:15 | |
| **Total** | **37:05** | **09:16** | **13:39** | **---** |

Leveraging the real-time bedside event annotations from the CEA application, "Rest" and "Motion" events were obtained from absence or presence of patient head motion respectively. Events specifically including face occlusions were annotated after data collection by carefully reviewing the recorded video. To determine a gold-standard ROI for evaluation purposes, video frames were extracted every 5 seconds, and a bounding box was manually drawn over the face area using the *Image Labeler* application in MATLAB. During periods of occlusion, the ground truth ROI was restricted to the visible portion of the face, as close as possible.

## 6.4 Neonatal Face Tracking - Results

In the following section, results are presented for four methods. 1) A baseline approach using a static ROI initialized from the first frame and repeated overtime, 2) a tracking-by-displacement-only method using the KLT algorithm, 3) a tracking-by-detection-only method using active contour with a face mask, and 4) a comprehensive method using fusion of the two tracking approaches.

### 6.4.1 Baseline Approach

A baseline approach using an initial ROI defined using the first frame and continuously applying this region over time would identify periods when tracking is necessary.

Given that rest periods are observed more frequently than motion or occlusion events in neonates, one could assume that newborns do not produce significant movements. However, this assumption does not always hold, especially for continuous monitoring. The displacement field in Figure 6.6 highlights an initial ROI (yellow box) and all areas covered by the actual bounding box corresponding to patient displacement over the 20-minute period (red box). The displacement was calculated from computing the Euclidean distance between the center of the initial ROI and the center of the current frame, at 5 second intervals in the video. This figure clearly demonstrates the effect of patient motion on ROI displacement. A tracking algorithm from displacement monitoring is therefore warranted.

Figure 6.6 depicts a number of roll poses, ranging from 0 – 180 degrees, for a typical neonatal patient. Given the wide variations in patient poses, identifying an initial ROI and reusing without tracking is not sufficient for continuous patient monitoring; dynamic ROI tracking is clearly required.

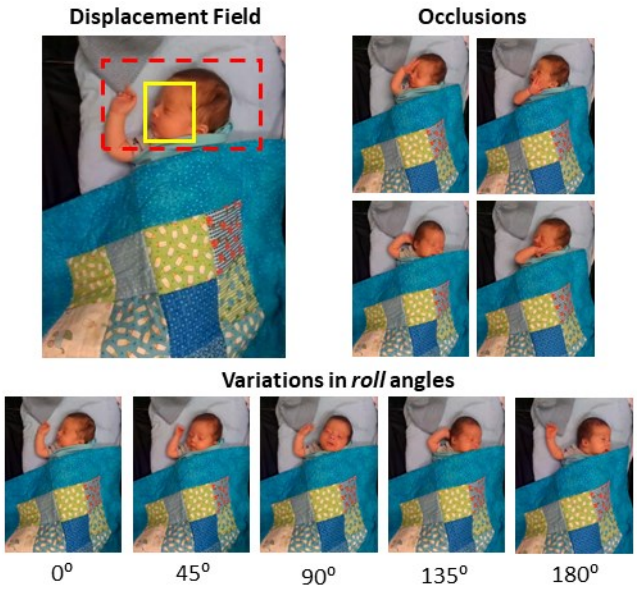



Figure 6.6: Pose variation and challenges from face tracking dataset. Displacement field (red) represents the total displacement due to occlusions and variations in roll angles.

### 6.4.2 Tracking by Displacement only - KLT Algorithm

As previously mentioned, many tracking algorithms account for displacement of features within the ROI over time to measure the trajectory taken by the moving object. We herein leverage the KLT algorithm, and furthermore propose the use of number of tracked "good features" over time as a measure of KLT tracking quality. This thesis investigates the longevity of the KLT tracking algorithm (how long can the algorithm perform before failing), and the impact of the number of features (how many features are required for constant high-quality monitoring). The former is evaluated by measuring the decline in Jaccard index between the ground truth and the resulting bounding box, as depicted in Figure 6.7 (left). The latter is measured by monitoring the decline in the number of features over time, as depicted in Figure 6.7 (right). There is good correlation between the number of tracked features (right) and the quality of tracking (left), indicating that the number of tracked features could serve as a surrogate tracking quality estimator (investigated later in Chapter 8.2.2.3).

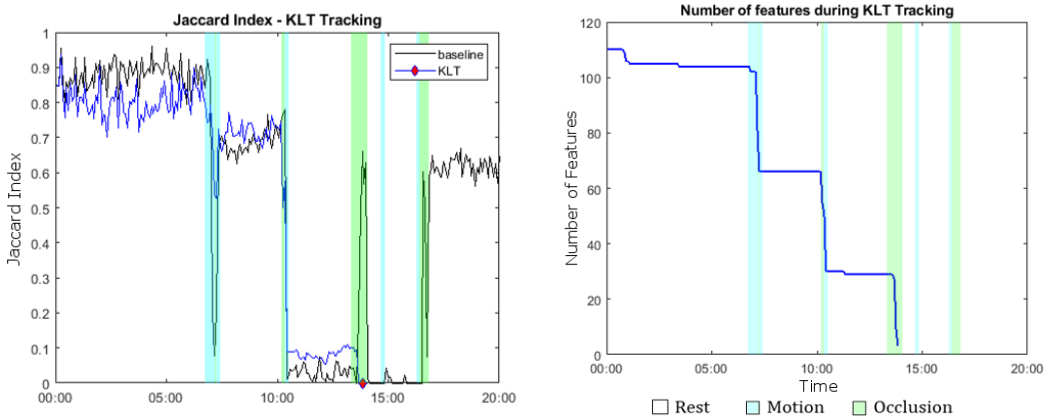



Figure 6.7: Patient B results from tracking by displacement only. Left: Jaccard index measured over time. Right: Number of features remaining during tracking.

Clearly, the KLT tracking algorithm is not sufficient for continuous monitoring given that it fails after a short period of time (between ~ 6 -14 min across all three patients) when an ideal algorithm should be capable of continuous tracking for several hours without requiring manual interventions.

Table 6.4: Tracking Degeneration

| Patient | Events leading to degeneration | | | |
|---|---|---|---|---|
| | ***Number of events*** | ***Event Type*** | ***Event Length*** | ***Failure Point*** |
| A | 4 | Motion | 51 sec | 6:31 during Motion |
| | 1 | Occlusion | 25 sec | |
| B | 3 | Motion | 45 sec | 13:54 during Occlusion |
| | 3 | Occlusion | 49 sec | |
| C | 9 | Motion | 215 sec | 11:32 during Occlusion |
| | 7 | Occlusion | 162 sec | |

Table 6.4 summarizes the number and type of event that leads to the failure of the KLT tracking algorithm for each of the three patients in this face tracking application. Across all patients, the failure

in tracking occurred during either a motion or occlusion period. Overcoming this failure by reinitializing the ROI is therefore warranted.

### 6.4.3 Tracking by Detection only – Mask Application

Aside from tracking an ROI using its change in displacement, we can track by detection only. A mask is applied on video frames and the active contour algorithm is used to optimize the resulting shape. Detecting an ROI using a mask is highly dependent on the appearance of that mask, and this thesis investigated three main patient poses: right profile, frontal view, and left profile.

Given the constant change of patient poses, selecting the best mask at any given point is required. This selection is obtained by applying all three masks on the given image and selecting the minimum error from the active contour approach. Typically, the active contour algorithm aims to minimize the average difference pixel intensities inside and outside of the object. Given that skin pixel detection can be noisy (i.e., a frame would contain detected face plus any other visible body parts), we limit the error to the average difference pixel intensity between the contour and the mask, inside the resulting contour only. The selected contour would correspond to the one providing minimum error.

Results of each method are outlined in Table 6.5, where results within each patient are consistent; one profile outperforms the other and the frontal view is a middle ground. Between each patient, however, results vary significantly where the selection of the optimal mask remains a difficult task to achieve, mainly due to occlusion periods seriously affecting the detection the facial area. Patients exhibiting greater occlusion events suffered in continuous ROI tracking. Pausing tracking during such periods is necessary to maintain a reliable ROI throughout continuous monitoring.

Table 6.5: Face Tracking Results

| **Method** | **Evaluation per Patient (AUC)** | | |
|---|---|---|---|
| | ***Patient A*** | ***Patient B*** | ***Patient C*** |
| Baseline | 0.7870 | 0.5366 | 0.6312 |
| Displacement_KLT | 0.2221 | 0.4137 | 0.2053 |
| Detection_Profile1 | **0.7895** | 0.5440 | 0.2766 |
| Detection_Frontal | 0.4083 | 0.3157 | 0.2983 |
| Detection_Profile2 | 0.3861 | 0.1092 | 0.3088 |
| Detection_OptimalMask | 0.5346 | 0.5855 | 0.2589 |
| Proposed Approach | 0.6312 | **0.6836** | **0.6315** |

### 6.4.4 Tracking by Detection Reinforced by Displacement

Leveraging both tracking techniques, we can combine the feature tracking approach to continuously know *when* the algorithm is performing well or degenerating, and identify *what* ROI to reinitialize by providing an optimal mask for detection. This approach is labelled *tracking by detection reinforced by*

*displacement* given that a fusion of these techniques is used to obtain a comprehensive tracking algorithm. This proposed method is compared to other techniques and illustrated in Figure 6.8.

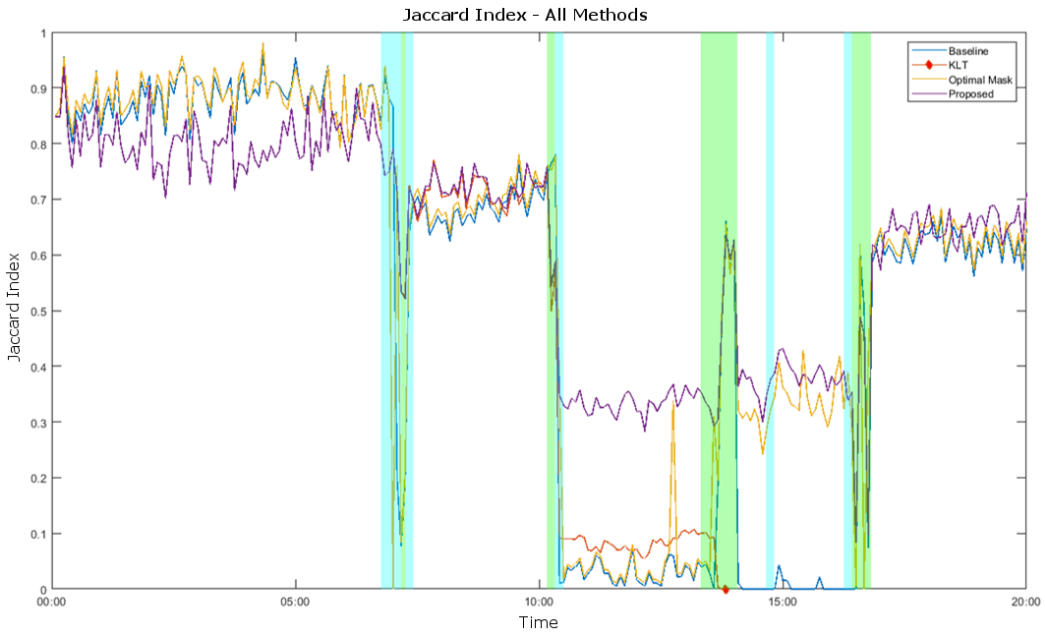



Figure 6.8: Patient B tracking results from all methods.

Periodical detections are used here only when required, i.e., when a loss in tracking is observed. During periods of occlusions, a significant loss in extracted features is observed, thereby suggesting to halt tracking until redeemed safe. We employed a retrospective systematic approach where a pause would take 15 seconds before resuming. Results presented in Table 6.5 demonstrate the usefulness in leveraging both displacement and detection for continuous ROI tracking. Patient B and C have greater results with our proposed method, while Patient A has comparable results with the other methods. Overall, all results across patients remains consistent. This suggests an overall improvement of the tracking methods despite motion and occlusion events. Facial tracking during such conditions can remain a difficult task to achieve, however, this work demonstrates its strong potential.

### 6.4.5 Face Tracking Discussion

This work addressed a multitude of challenges encountered in object tracking, including occlusions and motion events. A reinitialization step is introduced in cases where such challenges were severely impacting the tracking process, thereby ensuring continuous monitoring. To overcome the variations in patient poses, a face template was generated leveraging a skin pixel algorithm. By including two profile and one frontal view, this research addresses the change in roll angles while ensuring that the face tracking algorithm could always proceed.

## 6.5 Limb Motion Detection – Methods

Detecting patient movement can also be individualized as the detection of moving limbs. Similarly to the previous motion detection subchapters, this section reviews state-of-the-art computer vision methods for identifying motion, and proposes an application to the detection of patient limbs (left arm, right arm, left leg, right leg).

Multiple computer vision methods have been used to detect motion in the scene, as detailed in Chapter 2.5. Among them, the background subtraction, temporal intensity differentiation and optical flow techniques are evaluated here for detecting and classifying patient's moving limb.

Before selecting one of the three mentioned detection algorithms, we compared their abilities using a small sample of the dataset as proof-of-concept. Specifically, we visually evaluated their capabilities in detecting relevant primary motion in the presence of noisy secondary motion; denoted as movement of inanimate objects (e.g., sheets or sensor cables) due to patient motion. The selected detection method should therefore reduce secondary motion while accurately detecting the moving limb within the frame. The background is initially excluded by manual segmentation on the first frame only replicated on subsequent ones. This allows standardization of different viewing points across patients, and a focused view on the patient; sensor cables, beddings, bed frame and toys are excluded. Background lighting will periodically change when, for example, the nurse dims or turns off the lights. Dynamic thresholding is required to avoid false movement detection during lighting changes. The dataset utilized in this research did not include such severe lighting variations, thus a dynamic threshold term was not required.

Once motion estimation is obtained from the selected detection method, a MEI is created, as inspired from [138], to localize where the movement appeared within the frame. A limb classification template is subsequently applied to the binary MEI to classify individual limb movements. The classification template is created manually as follows: From the location of the shoulders and hips, we can estimate the region in which each limb is expected to occupy by estimating the range of movement and the limb length. A selection of each limb's ROI is performed on the initial frame and this same template is applied to all subsequent frames. Given neonatal patients do not significantly move their core position while in the supine position, a single definition of each limb's ROI can be used.

### 6.5.1 Limb Motion Detection Dataset

As a proof-of-concept, three patient data were used in this work, where we only examined segments where the patient was fully uncovered, thereby enabling visualization of each limb. Gold standard movement event data were obtained from the real-time bedside annotations. The patient dataset contains continuous movement of the limbs with negligible displacement of the patient's core while in the supine position. The relative frequency of each limb movement was roughly equal in the tuning, training, and

testing sets, as shown in Table 6.6. One patient's data were utilized for tuning (three videos) and training (14 videos) the model, while the testing set consisted of two new patient's data (two videos). This holdout validation was performed such that a variation of data would reside in the training set with multiple videos including individual moving limbs, before testing on new patient's data.

Table 6.6: Limb Motion Dataset Distribution

| **Stages** | **Moving limb (number of frames)** | | | | **Total number of frames** |
|---|---|---|---|---|---|
| | *Left Arm* | *Right Arm* | *Left Leg* | *Right Leg* | |
| Tuning | 31 | 35 | 34 | 39 | 87 |
| Training | 90 | 43 | 55 | 78 | 406 |
| Testing | 41 | 45 | 38 | 36 | 160 |

## 6.6 Limb Motion Detection – Results

### 6.6.1 Comparing Motion Detection Methods

Careful comparison of the two sequential frames depicted in Figure 6.9-a and 6.9-b indicate that both arms and the patient's left leg moved between frames. The results of the three motion detection algorithms (Figure 6.9-c-e) illustrate that only the right arm was correctly detected by all three methods. The patient's left side included a large amount of noise from secondary motion, resulting in the wide distribution of white pixels in that area for the background subtraction (Figure 6.9-c) and temporal differentiation (Figure 6.9-d) methods. Lowering the threshold with these techniques would permit detection of all true moving limbs, while also admitting a greater amount of noise. Conversely, the optical flow approach (Figure 6.9-e) resulted in correct detection of all three moving limbs. Given that it appeared more robust to noise, we adopted the optical flow as the foundation for the motion detection stage of our algorithm. The tuning set was used for this calibration step.

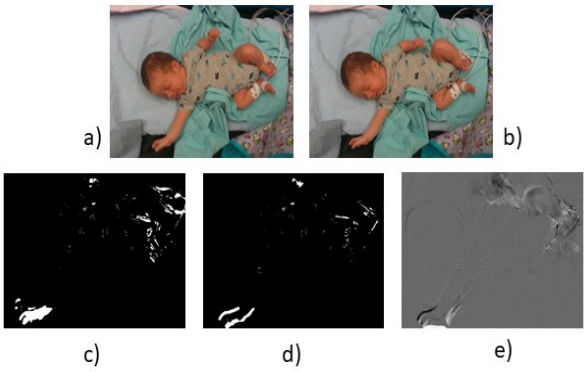



Figure 6.9: Comparison of motion detection techniques. This sample included both arms and the patient's left leg moving. a) Frame at time t. b) Frame at time t+1. c) Background subtraction. d) Temporal differentiation. e) Optical flow.

### 6.6.2 Motion Detection & Classification

As explained above, movements are primarily detected using the optical flow method. To demonstrate the results of our algorithm, Figure 6.10 uses the sequential frames from Figure 6.9a)-b) as an example. The optical flow in Figure 6.10-a is represented by a MVF comprising of arrows indicating the direction and magnitude of the detected motion. Given that secondary motion noise can result in a denser MVF, the image is subsampled into blocks using a 8x10 grid summarizing the average motion within each block. Provided that the magnitude of movement exceeds a predefined threshold, it is projected onto the MEI in Figure 6.10. A classification template, as explained in section 6.5, is finally utilized for limb-specific motion classification, as depicted in Figure 6.10-c.

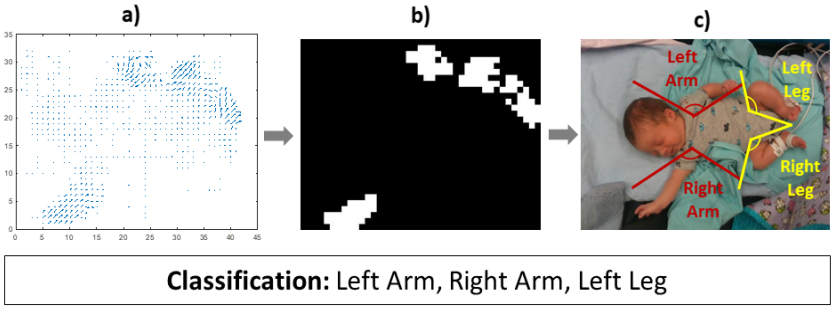



Figure 6.10: Limb Motion Detection Results. Results obtained from a sample image pair showing the optical flow (a) and corresponding motion-energy image (b). A classification template (c) is used to classify the moving limb.

The acccurate classification of this example was obtained from careful selection of threshold from the MVF to the MEI. A threshold calibration was obtained using the tuning set, which was excluded from the training and testing set. To determine the appropriate threshold for our application, at each level we calculated the sensitivity, specificity, accuracy, and precision metrics. Here, a positive motion is one that truly contains a moving limb while a negative motion is one where the limb is not moving.

To visualize the effect of varying threshold, an ROC curve is plotted, as depicted by Figure 6.11-b, where the left upper corner denotes a perfect sensitivity and specificity. The closest coordinate to this point is represented with a threshold of 6 having a sensitivity of 88.3%, specificity of 93.3%. An ideal threshold should not be too permissive (would increase noise) nor too conservative (would decrease true motion detection), as also illustrated in Figure 6.11-a; threshold of 6 provides an accuracy of 91.4% and precision of 90.5%, thereby corroborating our decision.

### 6.6.3 Algorithm Results & Evaluation

Given the 4-way classification of limbs, the algorithm is evaluated by converting qualitative data into quantitative ones (*i.e.,* for each frame, limbs determined to be moving are labelled as 1, while static limbs are labelled as 0). Each limb is evaluated independently and Table 6.7. depicts such outcomes.

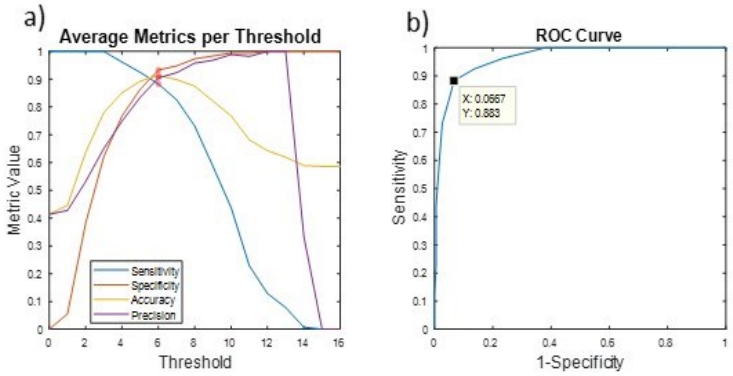



Figure 6.11: Evaluation metrics with varying threshold levels for limb motion detection.

This presented algorithm was effective at detecting presence of limb motion, however, it underperformed at correctly identifying absence of motion, as demonstrated by the high sensitivity of 100% and low specificity of 31.5%. Despite this difference, this method was efficient for overall detection and classification of limb motion, as shown by the high accuracy of 85.3% and precision of 82.0%. The selected holdout validation procedure was however not sufficient and will need to be revisited with more rigorous evaluation to properly measure the model's performance given its wide difference during the training and testing stages. For instance, the high average specificity of 97.1 % during training compared to low average specificity of 31.5 % during testing might suggest that the model was overfitting. Even though testing on two new patients who were not seen during training was a step towards generalizability, further evaluation (e.g., leave-out-patient-out cross validation) would improve the model's performance and generalization ability.

Table 6.7: Results From Each Moving Limb

| **Metric** | | **Moving limb** | | | | **Average** |
|---|---|---|---|---|---|---|
| | | *Left Arm* | *Right Arm* | *Left Leg* | *Right Leg* | |
| Training | Sensitivity | 79.6 | 80.4 | 77.6 | 90.8 | **82.1** |
| | Specificity | 97.7 | 99.5 | 92.0 | 99.3 | **97.1** |
| | Accuracy | 93.4 | 95.3 | 88.0 | 97.7 | **93.6** |
| | Precision | 93.3 | 92.5 | 51.4 | 97.2 | **83.6** |
| Testing | Sensitivity | 100 | 100 | 100 | 100 | **100** |
| | Specificity | 29.2 | 45.0 | 26.9 | 25.0 | **31.5** |
| | Accuracy | 85.7 | 96.4 | 80.4 | 78.6 | **85.3** |
| | Precision | 82.7 | 95.6 | 76.8 | 72.9 | **82.0** |

## 6.7 Motion Detection – Conclusions

This chapter first investigated the applicability of two techniques for detecting motion in neonates from video. As a proof-of-concept, we implemented a simple but precise approach using optical flow, and a more complex but extendable LSTM approach. This work is a step towards identifying motion

events contributing to false alarms due to motion artifacts, while also being pertinent to a multitude of motion-related applications.

Secondly, in this chapter we investigated three algorithms for face tracking; tracking by displacement only, tracking by detection only, and a fusion method. Results demonstrate that the proposed fusion method combining both techniques is more robust to motion and occlusion events during continuous ROI tracking. This approach is therefore useful for tracking the face of the patient in continuous monitoring; Chapter 8 shows how this work can be particularly useful in non-contact vital sign monitoring.

Finally, non-contact limb-specific motion estimation is valuable for deriving physiologic signal quality estimates and for evaluating neonatal motion-related conditions. This research developed a proof-of-concept motion detection and classification algorithm to detect moving limbs.

---

# 7 Vital Sign Estimation

Given that data collection at the NICU of CHEO started over a year after the beginning of my thesis, some preliminary work was conducted on adults to build a heart rate estimation approach using all streams from the depth-sensing camera. Accordingly, this chapter reports methods and results for a novel method of heart rate estimation on adults (7.1) and it is validated on a neonatal population with other method modifications (7.2).

This thesis develops an algorithm that examines multiple passbands covering the full range of plausible heart rates. Combined with sensor fusion, the algorithm effectively estimates the true HR from all passbands. We here present an automated, non-contact system capable of HR estimation using a multi-modal depth-sensing camera. Previous studies have suggested that EVM on visible light video is highly sensitive to lighting variations resulting in poor HR estimations [160], [162], [163]. We therefore performed experiments under controlled lighting conditions to systematically explore the effect of ambient lighting on HR estimation via EVM, and thereby quantitatively explored its impact on the accuracy of estimations. Furthermore, most studies tracking changes from the face of the subject to estimate HR would require them to sit or stand in front of the camera [155], [157], [163]; such scenarios are further explored. We therefore investigate the HR estimation accuracy by considering different subject poses, thereby demonstrating reproducibility in different environments.

## 7.1 Heart Rate Estimation on Adults

When using EVM for enhancing video, certain parameters must be optimized, such as a magnification factor (a multiplicative variable in temporal filtering), a frequency passband range (range within which to amplify variations), and a sampling rate (rate at which to apply the process). When employing EVM for enhancing changes in color, a narrow passband can be selected. On the other hand, a broad passband may be more suitable for motion magnification [40]. Typically, previous studies have used a wide passband including all plausible HR values. Narrower passbands are however preferable for HR estimation given that they increase the SNR, but little guidance has been provided to assign such passbands *de novo*. To address this issue, we developed a Selective EVM approach to automatically determine the narrow passband, without resorting to very wide (meaningless) passbands. Additionally, multi-modal video can be leveraged for this task.

For our experiment, we used an ideal band pass filter for magnification of changes in blood flow in the face. The naïve approach would be to use broad frequency ranges, such that all conceivable HR values will fall within the amplification range. However, this incurs a cost, since it can potentially amplify motion artifacts in frequencies near the HR. Thus, one must ensure that this range is not too wide, which is often done by trial and error. Clearly, once a true heart rate is known, using a narrower

frequency range would result in a more accurate estimation, though this is not feasible for *de novo* estimation of HR.

We developed a heuristic approach, where multiple different narrow passbands were applied to the same input resulting in multiple corresponding HR estimates. Data fusion is used to combine these estimates based on a confidence metric. This method, which we call *Selective EVM*, is investigated here as a means to estimate HR in each band and select the most plausible ones as our actual HR estimate.

### 7.1.1 Multimodal Selective EVM

The selective EVM method described above provides a natural way to achieve robust HR estimation from a multi-modal camera, such as the Intel SR300 camera. In order to estimate HR, the selective approach is performed on each of the three streams independently before forming a final estimate through multi-modal sensor fusion. We call this process the *Multimodal Selective EVM*, which is illustrated in Figure 7.1.

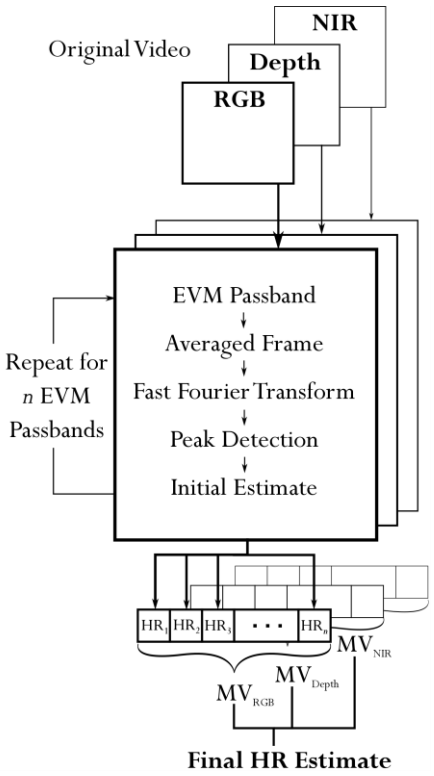



Figure 7.1: Multimodal Selective EVM. This framework processes a video by applying EVM with one given magnification passband at a time to obtain an initial HR estimate. This procedure is repeated for n number of passbands to arrive at a collection of heart rate estimates, to which a majority vote (MV) provides an HR estimate for that specific modality. After replicating this entire system on all modalities (color, depth, and near-infrared), a final HR estimate is obtained.

#### *7.1.1.1 HR estimation using EVM with multiple passbands*

Before EVM can be applied to enhance a video, the passband must first be selected. Without knowing the actual HR, it is impossible to pre-select an appropriate passband while avoiding overly broad passbands. The Selective EVM is used where EVM is applied to each video using each of 13 passbands listed in Table 7.1. These passbands were selected to cover all expected adult resting HR while ensuring that any actual HR would be included in three separate bands. For example, a true HR of 52 bpm would be enhanced by applying EVM with ranges 2, 3, and 4 from Table 7.1. For each range, the EVM-enhanced output video is converted to a time-varying scalar value by averaging pixel intensities over each frame. For RGB video, the red channel is used to emphasize changes in blood flow over that period. For depth and NIR videos, a simple average of the single channel is used. The power spectrum of these time-varying data is computed via fast Fourier transform (FFT). An additional band-pass filter is then applied on the Fourier spectrum to remove low-frequency noise and ignore irrelevant high frequencies. Peak detection is finally applied, with the highest peak representing the HR estimate for this EVM range number.

Table 7.1: Bandpass Filters

| Range Number | Heart Rate range (bpm) | Frequency Range (Hz) |
|---|---|---|
| 1 | 35 - 50 | 0.583 – 0.833 |
| 2 | 40 – 55 | 0.667 – 0.917 |
| 3 | 45 - 60 | 0.750 – 1.000 |
| 4 | 50 – 65 | 0.833 – 1.083 |
| 5 | 55 – 70 | 0.917 – 1.167 |
| 6 | 60 – 75 | 1.000 – 1.250 |
| 7 | 65 – 80 | 1.083 – 1.333 |
| 8 | 70 – 85 | 1.167 – 1.417 |
| 9 | 75 – 90 | 1.250 – 1.500 |
| 10 | 80 – 95 | 1.333 – 1.583 |
| 11 | 85 – 100 | 1.417 – 1.667 |
| 12 | 90 – 105 | 1.500 – 1.750 |
| 13 | 95 – 110 | 1.583 – 1.833 |

Figure 7.2 demonstrates the benefits of magnifying signals through EVM for a sample video with a true HR of 77 bpm. It is apparent that the signal is much noisier in the original video: Figure 7.2-a fails to display a clear pattern corresponding to HR and therefore does not produce well-defined peaks in the frequency domain (Figure 7.2-b). Figure 7.2-c illustrates an EVM-enhanced time series using a magnification passband of 1.25 - 1.50 Hz. Its corresponding Fourier spectrum is depicted in Figure 7.2-d with a clear distinction of peaks corresponding to the HR. The shaded area in Figure 7.2-d corresponds to 0.583 - 2 Hz or 35 - 120 bpm, thereby filtering noise. The dominant peak within this range is used as the HR estimate, as illustrated in Figure 7.2-e. In this particular experiment, the true heart rate is 77 bpm

and the frequency of greatest magnitude is 1.284 Hz, resulting in an estimate of 77 bpm (1.284 Hz x 60 s).

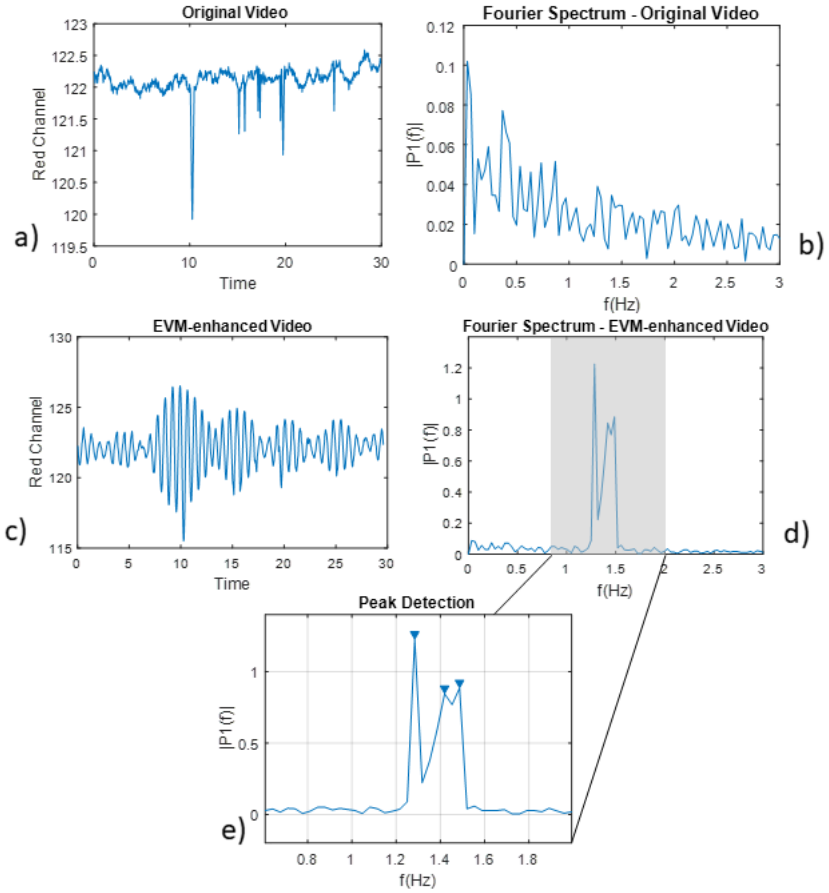



Figure 7.2: HR estimation from EVM-enhanced video. Time-average red channel of original video (a) and corresponding Fourier spectrum (b); averaged red channel of EVM-enhanced video (c), with its Fourier spectrum (d); and peak detection graph of EVM video (e).

### *7.1.1.2 HR estimation in each video modality*

Now that 13 different initial HR estimates are obtained for each video modality, majority voting is utilized to identify the most frequently occurring HR estimate. As illustrated in Table 7.2, it is possible to have multiple modes. This reiterates the importance of defining frequency ranges such that any true HR is included in multiple ranges. In this way, the true HR is expected to be magnified by multiple ranges, increasing the chance of being selected as the dominant estimate. An example of the majority vote process can be seen in Table 7.2, where the color and NIR streams produced two different initial HR estimates.

Table 7.2: Example of Majority Vote

| Range Number | Estimated HR (bpm) | | |
|---|---|---|---|
| | ***RGB*** | ***D*** | ***NIR*** |
| 1 | 47 | 43 | 47 |
| 2 | 53 | 53 | 53 |
| 3 | 53 | 53 | 53 |
| 4 | 53 | 53 | 53 |
| 5 | 55 | 67 | 67 |
| 6 | 61 | 67 | 67 |
| 7 | 75 | 67 | 67 |
| 8 | 82 | 75 | 81 |
| 9 | 82 | 75 | 81 |
| 10 | 82 | 85 | 81 |
| 11 | 86 | 85 | ~~75~~ |
| 12 | 92 | 101 | 101 |
| 13 | 108 | 101 | 101 |
| **Gold Standard** | **53** | | |

* Lying position in "Full" lighting condition (subject 5)

Prior to computing the mode of all HR estimates, a data quality check can be performed as follows. From preliminary experiments, we observed that HR estimates that fell outside the designed passband for that range number tended to be false. In the example illustrated in Table 7.2, the underestimated HR estimate of 75 bpm for Range 11 of the NIR stream should be discarded, since the corresponding passband of 1.417 - 1.667 Hz should only amplify true HR between 85 - 100 bpm.

Considering that the multi-modal selective EVM may result in multiple modes corresponding to multiple putative HR estimates, we first select a single mode for each imaging modality prior to arrive at the final HR estimate. To this end, all possible sets of modes are computed by drawing one mode from each video modality. For example, there are twelve possible sets corresponding to the data illustrated in Table 7.2 including: {53,53,53}, {53,53,67}, {53,67,81}, {82,53,53}, etc., for RGB, depth and NIR respectively. For each set, the standard deviation is computed, as formulated by Equation 7.2, where $s$ is standard deviation, $x$ is mode in question, $\bar{x}$ is the mean of the modes across modalities, and $n$ is the number of video modalities (three for the Intel SR300 RGB-D camera).

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \qquad (7.2)$$

The set of three modes with the minimum standard deviation is selected as the RGB, depth, and near-infrared HR estimates. This approach ensures that the chosen estimates consist of those with highest degree of agreement between all modalities. In our example from Table 7.2, this would correspond to 53 bpm for each separate modality since it represents lowest standard deviation when compared to one another (s = 0 here).

#### *7.1.1.3 Final HR estimate using data fusion*

To estimate the true heart rate, a combination of evaluations obtained from all three streams (RGB, depth and near-infrared) is considered. The final heart rate estimate is computed as the average of the set of stream estimates with smallest standard deviation.

Figure 7.3: Sample of video frames displaying RGB, depth and near-infrared images (left to right) for adult-based HR estimation.

### 7.1.2 Dataset & Experimental Setup

As a proof of concept, we collected a dataset from five different individuals of varying skin tones, sex, and age. Videos were captured using the Intel RealSense SR300 camera. A sample of the dataset is displayed in Figure 7.3 showing all three imaging modalities.

The face of each participant was captured during 30 seconds while they were instructed to remain as still as possible under several lighting conditions and in multiple poses, as detailed in Table 7.3. Doing so, we aimed to investigate the impact of environmental changes and unintentional subject movements on EVM-based HR estimation from RGB-D video. All data were recorded in a room with strict control over lighting levels which were consistent across three subject poses (lying, sitting, standing). Four lighting levels were used: low, medium, high, and full. The first three used incandescent lighting while the fourth used overhead fluorescent lighting. Table 7.3 reports the total illumination of the subject's face for each lighting level. It was expected that the RGB stream would perform best in brightest light while the depth and NIR streams would be largely invariant to lighting variations. In fact, we anticipated that the NIR stream would perform best in low lighting conditions.

Table 7.3: Lighting levels in the Adult-based HR Estimation Dataset

| **Positions** | **Lighting levels (lux)** | | | |
|---|---|---|---|---|
| | ***Low*** | ***Medium*** | ***High*** | ***Full*** |
| Lying | 21 | 44 | 63 | 255 |
| Sitting | | | | |
| Standing | | | | |

Although research participants were instructed to hold still, minute unintentional movements are still expected, relating to respiration, swaying, or ballistocardiac motion, etc. We hypothesize that depth-

based HR estimation will be most effective when the subject is free to move in the sagittal plane (towards the camera). We therefore recorded videos in three positions with different levels of expected movement: free-standing (most movement expected), sitting, and lying (least movement). The lying position was specifically used to control movement in the sagittal plane to examine whether HR could be estimated from the depth stream without movement.

Throughout the experiment, an adult SpO2 sensor from a Draeger monitor registered the person's true heart rate as gold standard. The gold standard is used to calculate the HR estimation accuracy for all videos, using equation (7.3), where $HR_{actual}$ is the gold standard data while $HR_{exp}$ is the final HR estimate derived from the proposed multi-modal selective EVM approach.

$$Accuracy\ \% = \left(1 - \frac{|HR_{actual} - HR_{exp}|}{HR_{actual}}\right) \times 100 \quad (7.3)$$

A per-subject accuracy measure is also calculated, by averaging the accuracy observed over all videos relating to each individual research participant.

### 7.1.3 Results & Discussion

Our multi-modal method was evaluated by comparing results obtained from the consensus HR estimate against results when restricting ourselves to a single sensor stream. Results are shown in Table 7.4, when averaged across all lighting conditions and subject poses. We can see that both the depth stream and our consensus approach led to improved performance compared to the color stream or the near-infrared stream alone.

Table 7.4: Performance of the Multimodal Selective EVM

| **Subject** | **Accuracy** | | | |
|---|---|---|---|---|
| | ***Color*** | ***Depth*** | ***Near-infrared*** | ***Consensus*** |
| 1 | 79.5 | 84.7 | 79.5 | 83.1 |
| 2 | 85.5 | 83.6 | 85.1 | 87.0 |
| 3 | 84.4 | 84.9 | 85.0 | 86.5 |
| 4 | 85.5 | 83.2 | 71.5 | 80.9 |
| 5 | 77.2 | 79.8 | 75.5 | 79.5 |
| Average | **82.4** | **83.2** | **79.3** | **83.4** |

#### *7.1.3.1 The effect of lighting on HR estimation*

HR estimation accuracy is reported for each lighting condition in Table 7.5, averaged across all subjects. As expected, the RGB stream was susceptible to lighting variations, generally performing best in brighter environments. The "High" and "Full" lighting provided the most accurate estimates among all levels examined, which is not surprising given that EVM enhances perceptible color changes due to changes in blood flow. Thus, a brighter scene should help capture this subtle variation. As hypothesized,

the depth appeared to be invariant to illumination changes given that the "Low" and "Full" lightings demonstrated similar accuracy.

Table 7.5: Impact of Lighting Conditions on HR Estimation Accuracy

| Stream | Accuracy | | | |
|---|---|---|---|---|
| | ***Low*** | ***Med*** | ***High*** | ***Full*** |
| RGB | 78.6 | 77.1 | 85.2 | 88.5 |
| Depth | 83.9 | 78.6 | 87.1 | 83.2 |
| NIR | 77.1 | 78.5 | 81.0 | 78.2 |
| Consensus | 81.2 | 80.5 | 86.9 | 85.1 |

As for the NIR streams, results would suggest that it is invariant to ambient light, supporting our hypothesis that this stream would be robust to low lighting. We had expected the reduction in ambient IR noise in low lighting conditions to be an advantage for NIR-based HR estimation. However, these data were recorded in an environment without significant IR noise, such as sunlight, so this effect was not observed.

The consensus estimate was the only one able to provide a consistent HR estimate with over 80% accuracy in all lighting conditions. This strongly suggests opting for the consensus estimate for HR estimation as it is highly robust to changes in lighting.

#### *7.1.3.2 The effect of subject pose on HR estimation*

At the study outset, we hypothesized that the depth stream would be most useful for HR estimation if the head were free to move subtly in the sagittal plane as a result of pulsing blood flow. The results in Table 7.6 appear to support this hypothesis, where the accuracy increases as the subject's freedom of movement also increases (from lying, to sitting, to standing). This could be due to the ballistocardiac motion having a greater impact on head oscillations, thereby revealing a coherent HR estimate. Previous studies have used motion magnification to estimate HR from depth video [159], [206]. In the case of the depth stream, using EVM to magnify changes in "colour" is equivalent to magnifying changes in the distance from the subject to the camera, which is equivalent to magnifying motion. Therefore, our results are consistent with previous studies.

Table 7.6: Impact of Subject Pose on HR Estimation Accuracy

| Stream | Accuracy | | |
|---|---|---|---|
| | ***Lying*** | ***Sitting*** | ***Standing*** |
| RGB | 77.6 | 81.0 | 88.0 |
| Depth | 77.8 | 84.5 | 87.4 |
| NIR | 83.4 | 77.0 | 74.3 |
| Consensus | 81.7 | 83.4 | 85.1 |

Our secondary hypothesis that HR estimation from the RGB and NIR stream would be invariant to the presence of movements was not supported by the data. In fact, we observe a similar trend in HR estimation accuracy from the RGB stream as we did for depth, where accuracy improved from lying to sitting to standing. This result warrants further investigation, but it is possible that the same head movement due to ballistocardiac motion causes changes in colour images as the person moves through inhomogeneous lighting. Interestingly, the NIR HR estimate appears to show a negative correlation with subject movement, where the highest accuracy is observed in the lying position.

As with lighting condition, the consensus estimate was the only one able to provide HR estimation accuracy consistently over 80% in all subject poses.

#### *7.1.3.3 Robustness of multi-modal approach to HR estimation*

In a few cases, when applying the selective passband method to individual video modalities, no modes were identified. This occurred twice with the colour stream, never with the depth stream, and 11 times with the NIR stream out of 180 videos in total. In these cases, the inconclusive modality was excluded from the multimodal selective passband consensus approach. The ability to switch dynamically between video modalities demonstrates the robustness obtained in leveraging multiple modalities for heart rate estimation. Among the 13 cases with an inconclusive video modality, we were able to obtain a consensus HR estimate from the other two streams with 94.2% accuracy. Clearly, this would have been impossible under a single modality. This again demonstrates the robustness of the multi-modal selective EVM algorithm. Future work outside the scope of this thesis could examine weighting the various modalities in different conditions. The weight of contribution could be redistributed from each modality, thereby emphasizing the modality believed to have the highest signal quality for a given environment. Additionally, other combination of modalities could be further investigated, for example with Blue or Green channel instead of Red, or with RGB channels alone. Such combinations are explored in this thesis while validating the Selective EVM approach on neonates.

## 7.2 Heart Rate Estimation on Neonates

This section uses the Selective EVM approach implemented on adult subjects to now apply it on our neonatal dataset. A few modifications to the Selective EVM technique are proposed here for application in the neonatal environment.

### 7.2.1 Selective EVM Methods

Given promising results from the Selective EVM approach implemented on adults, we propose to apply and expand this technique on neonatal data. The passbands are adjusted to reflect newborns' heart rate which are typically much higher than adults. While adult's passbands ranged from corresponding

resting HR between 35-110 bpm, neonatal resting HR can range from 90-180 bpm [28]. We select passbands exceeding and engulfing these ranges between 60 – 230 bpm and the corresponding new frequency bands are illustrated in Table 7.7.

Table 7.7: Neonatal Bandpass Filters

| Range Number | Heart Rate range (bpm) | Frequency Range (Hz) |
|---|---|---|
| 1 | 60 – 90 | 1.000 – 1.500 |
| 2 | 70 – 100 | 1.167 – 1.667 |
| 3 | 80 – 110 | 1.333 – 1.833 |
| 4 | 90 – 120 | 1.500 – 2.000 |
| 5 | 100 – 130 | 1.667 – 2.167 |
| 6 | 110 – 140 | 1.833 – 2.333 |
| 7 | 120 – 150 | 2.000 – 2.500 |
| 8 | 130 – 160 | 2.167 – 2.667 |
| 9 | 140 – 170 | 2.333 – 2.833 |
| 10 | 150 – 180 | 2.500 – 3.000 |
| 11 | 160 – 190 | 2.667 – 3.167 |
| 12 | 170 – 200 | 2.833 – 3.333 |
| 13 | 180 – 210 | 3.000 – 3.500 |
| 14 | 190 – 220 | 3.167 – 3.667 |
| 15 | 200 – 230 | 3.333 – 3.833 |

The Selective EVM approach used three streams from the camera: red channel from RGB, depth channel, near-infrared channel. The red channel was chosen following preliminary experiment that showed that the red channel carried most information for detecting blood flow in the person's face (results not shown). We explore here the three Red (R), Green (G), and Blue (B) channels with different combination with the depth (D) and near-infrared channels (I) as described in Table 7.8.

Table 7.8: Neonatal Heart Rate Estimation – Selective EVM Methods

| Method Name | Method Description |
|---|---|
| MSEVM-RDI | Multimodal Selective EVM with Red, Depth, and NIR channels |
| MSEVM-GDI | Multimodal Selective EVM with Green, Depth, and NIR channels |
| MSEVM-BDI | Multimodal Selective EVM with Blue, Depth, and NIR channels |
| USEVM-RGB | Unimodal Selective EVM with Red, Green, and Blue channels |
| MSEVM-RGBDI | Multimodal Selective EVM with Red, Green, Blue, Depth, and NIR channels |

### 7.2.2 Resting Patient HR Estimation Dataset

The adult experiment was conducted by recording the face of 5 subjects under varying positions and lighting conditions. In the NICU, such controlled data collection is not possible since the entire patient and the surrounding bedding and equipment are visible in the image. Also, we cannot reposition the patient nor adjust the lighting since our data collection study was purely observational.

For a fair comparison, we selected 10 patients in bright light, 3 patients in low light, and 1 patient undergoing phototherapy treatment. For each patient, one 30-second video clip is extracted, and four video clips are extracted for the only phototherapy patient at various times during monitoring. Further details of the dataset description is presented in Appendix E. To control the environment, we excluded clinical interventions, drastic changes in lighting, occlusions, and only included patients captured at a close distance during periods of rest to limit motion artifacts. This permits to use the entire frame as ROI, as was performed with the adult experiment, while preventing artifacts from the environment. For comparison with the adult population, we select the scenario with lying position and varying lighting among the 5 adult subjects. As gold standard measures, the ECG-based heart rate values collected from the Draeger monitor at CHEO were used to compare with the neonatal HR estimates.

### 7.2.3 Results & Discussion

Table 7.9 summarizes the performance of EVM on the neonatal patients, compared to the performance on adult subjects. For the adult-based MSEVM-RDI method, the 81.70% accuracy is comparable to the 79.77% accuracy for the neonatal population. Interestingly, among the other combinations using a single-color channel with depth and NIR, the MSEVM-BDI provided best results, followed by MSEVM-GDI and MSEVM-RDI. The USEVM-RGB unimodal combination performs the least successfully, but still rather well with 74.00% accuracy. Excluding NIR and depth streams is particularly detrimental to HR estimation accuracy during sub-optimal lighting conditions. The MSEVM-RGBDI combination including the contribution of all streams seemed to be most performant among all groups with 82.69% accuracy.

Table 7.9: Neonatal Heart Rate Estimation – Selective EVM Results

| **Lighting** | **Method** | **Population** | **Accuracy** | |
|---|---|---|---|---|
| | | | ***Avg*** | ***Std*** |
| Natural Lighting (low + high light) | MSEVM-RDI | Adult | 81.70 | 3.30 |
| | MSEVM-RDI | Neonatal | 79.77 | 13.68 |
| | MSEVM-GDI | Neonatal | 81.03 | 10.33 |
| | MSEVM-BDI | Neonatal | **82.64** | **10.64** |
| | USEVM-RGB | Neonatal | 74.00 | 17.23 |
| | MSEVM-RGBDI | Neonatal | **82.69** | **12.55** |
| Phototherapy Lighting | MSEVM-RDI | Neonatal | 89.12 | 9.42 |
| | MSEVM-GDI | Neonatal | 90.30 | 8.96 |
| | MSEVM-BDI | Neonatal | **93.89** | **6.66** |
| | USEVM-RGB | Neonatal | 87.74 | 14.00 |
| | MSEVM-RGBDI | Neonatal | 88.27 | 11.60 |

As for the phototherapy patient, all results are remarkably high (> 87%) in comparison to the performance under natural lighting. The best phototherapy combination was the MSEVM-BDI, which corroborates with the best combination under natural lighting. This strongly suggests that Blue channel provides most information among the three color channels from the enhanced pulsatile signal in neonates, and that the impact of blue-colored hue in phototherapy lighting intensifying this signal even further.

Although the MSEVM-BDI results for natural and phototherapy lighting are high in terms of accuracy, the standard deviation might be too high for this method to be used in hospitals yet. Also, the mean absolute error (MAE) for natural lighting is 25.2 ± 15.4 bpm, and 10.3 ± 11.1 bpm for phototherapy lighting. This error in HR estimate is not yet reliable enough for clinical use and further experiments with more data, patients, etc. would be required before deployment to hospitals.

Despite the promising results in neonatal HR estimation, using the entire frame can include external noise arising from fluctuating chest and stomach areas during respiration, from nurses during clinical intervention, or from moving limbs. Focusing the ROI to the face of the patient only is therefore warranted. To this end, once the face is detected, it can be automatically tracked overtime to continuously obtain a reliable ROI. Chapter 8 assembles a HR estimation pipeline that leverages bed occupancy, face detection, face tracking, and Selective EVM methods to build a robust video-based HR estimation system.

## 7.3 Heart Rate Estimation - Conclusions

This thesis presented a novel Selective EVM approach implemented and validated on an adult population before also validating on a neonatal population. This method revealed robustness to different lighting conditions and subject poses while the person is at rest. We demonstrated how comparable results are also obtained with a combination using Red, depth, and near infrared streams from both populations. Interestingly, we determined that while the Red channel is the most informative color channel for adults, the Blue channel is most useful for neonates (both in conjunction to depth and near-infrared data). This interesting finding is worth exploring with a diversity of patients undergoing phototherapy treatment, given that our CHEO dataset only included one patient. We therefore recommend leveraging a combination of Blue-Depth-NIR channels for multimodal HR estimation for future neonatal monitoring applications.

# 8 Non-Contact Neonatal Monitoring

Various non-contact monitoring studies have been reviewed in Chapter 2, and most of these monitoring approaches require a reliable ROI on which to focus analysis over time. Several of the contributions in this thesis are likely applicable to these methods, since they provide reliable ROI identification and tracking. Many research contributions described in this thesis constitute pivotal building blocks required in multiple patient monitoring pipelines. Figure 8.1 depicts several examples of non-contact monitoring pipelines enabled using research contributions from this thesis. These pipelines include non-contact heart rate (HR) or respiration rate (RR) estimation, false alarm detection, seizure/apnea detection, neonatal pain assessment, etc.

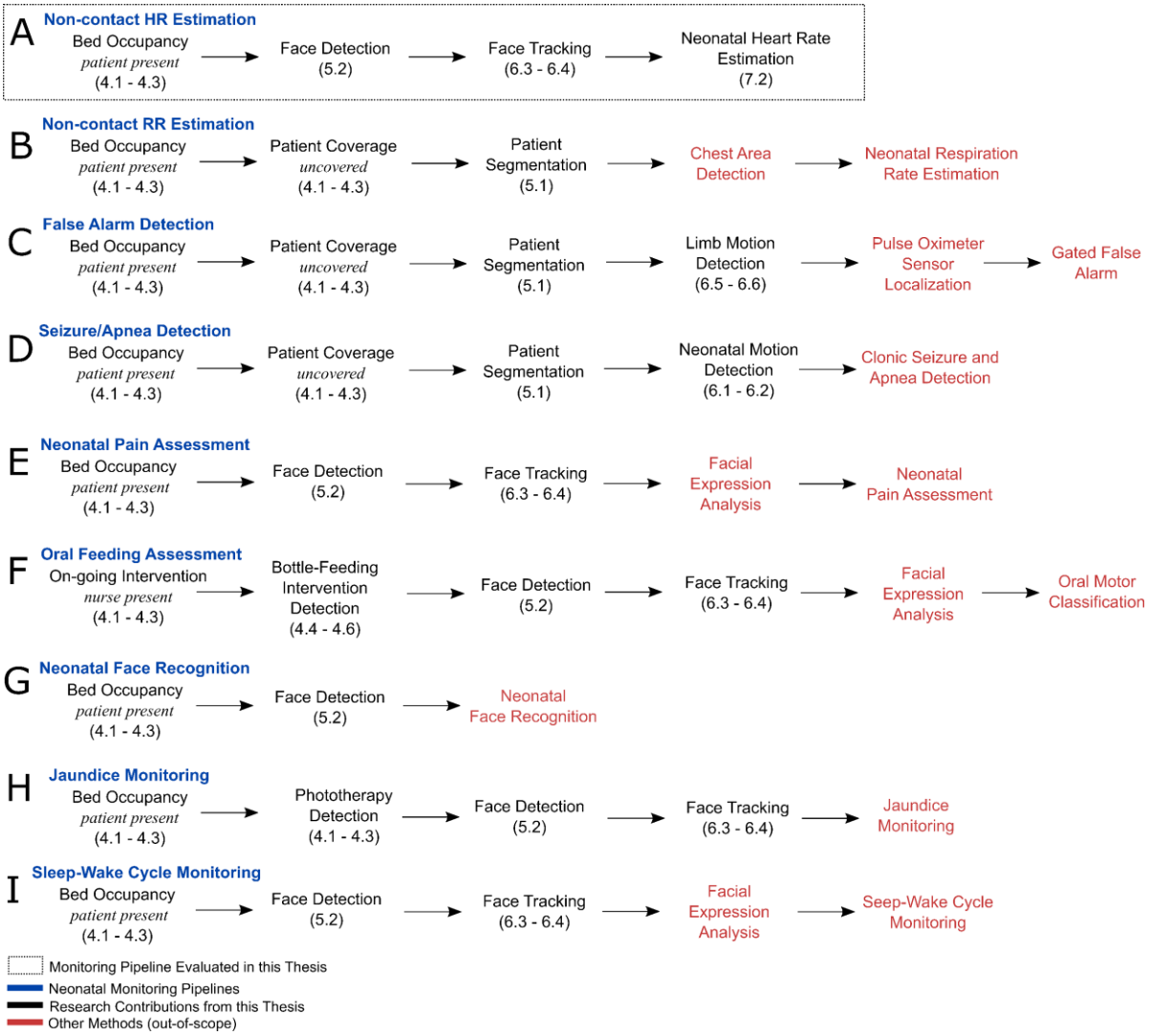



Figure 8.1: Examples of non-contact neonatal monitoring pipelines integrating research contributions from this thesis.

Among all instantiations illustrated in Figure 8.1, we here implement the "Non-contact HR Estimation" pipeline (Figure 8.1-A) to demonstrate the benefits and utility of combining multiple contributions from this thesis. To this end, in Section 8.1 we demonstrate how we can create a smart monitoring system while improving performance of HR estimation in the optimal case (i.e., patient at rest). To then address more challenging cases when HR estimation can be affected by the environment (e.g., occlusions, motions, changes in lighting conditions), a measure of pipeline uncertainty is also implemented to show how errors from each stage of the non-contact monitoring pipeline can propagate through the system (Section 8.2). From these findings, the overall uncertainty of the pipeline can be used to inform a decision support tool for clinical use.

## 8.1 Neonatal Heart Rate Monitoring

To obtain a comprehensive non-contact neonatal monitoring pipeline, multiple steps are required from analyzing the overall scene to obtaining specific information within it, as depicted in Figure 8.2.

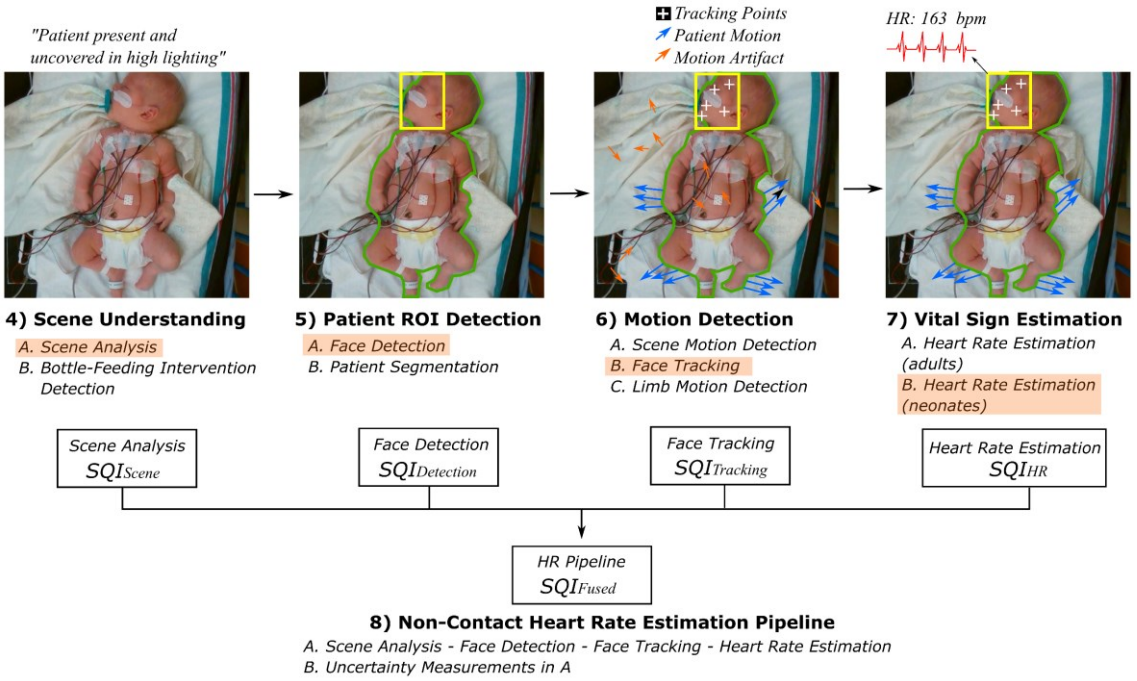



Figure 8.2: Neonatal Heart Rate Monitoring Pipeline and Uncertainty Measurements.

Our proposed non-contact heart rate monitoring pipeline comprises four stages including:

**Step 1: Bed Occupancy** (4A) - Detect if the patient is present in the bed.

**Step 2: Face Detection** (5A) - Detect the location of the patient's face to establish the ROI.

**Step 3: Face Tracking** (6B) - Track the face ROI of the patient over time.

**Step 4: HR Estimation** (7B) - Based on the ROI, estimate the patient's heart rate.

All methods and results used to create this pipeline are presented here.

### 8.1.1 Pipeline Methods

In this thesis, methods were presented in each chapter to assess the state of the art for its applicability to the neonatal environment and to advance these methods when required. The specific techniques selected to create a non-contact HR estimation pipeline are presented in Table 8.1. While monitoring applications can use different camera technologies, many rely on RGB devices (sometimes paired with another modality). For generalization, we here constrain ourselves to developing an RGB-based non-contact monitoring system. The face detection and tracking task inherently relied solely on RGB images and, although RGB-D methods are developed elsewhere in this thesis, RGB-based models are used here for determining bed occupancy and for HR estimation. Bed occupancy classification is first applied to detect whether the patient is present in the frame. In the following stage, we opt for the YOLO5Face model to perform simultaneous head orientation and face detection. Note that NICUface models could have been selected interchangeably for the face detection step; however, they do not provide reliable landmarks required for face orientation estimation since they only finetuned bounding boxes for robust face detection. Face detection is applied on the first frame to obtain the ROI; if no face is detected, the entire frame is selected as the ROI. Finally, for the face tracking method, the "tracking by detection reinforced by displacement" approach is used. This technique, presented in Section 6.3.3, depends on a selected face mask for reinitialization when tracking fails. To simplify evaluating the pipeline here, if tracking is halted, the last detected ROI is kept for the remainder of the 30 s video clip.

Table 8.1: Methods used in Non-contact HR Neonatal Monitoring Pipeline

| **Step** | **Method** | **Chapter** | **Modification** |
|---|---|---|---|
| Bed Occupancy | Deep Learning RGB Bed Occupancy | 4.1.2 | None |
| Face Detection | YOLO5Face | 5.1.2 | None |
| Face Tracking | Tracking by detection reinforced by displacement | 6.3.3 | If tracking fails, keep last detection for remainder of video clip. |
| HR Estimation | USEVM-RGB | 7.2.1 | None |

To validate the pipeline, the same patients used in Chapter 7.2 are used here (13 patients in natural lighting). We compare the HR estimation results from Chapter 7 to results obtained here using the complete non-contact HR monitoring pipeline.

### 8.1.2 Pipeline Results & Discussion

Using the proposed HR monitoring pipeline, HR estimation performance increased by 4.28% accuracy for patients in natural lighting, as shown in Table 8.2. This improvement shows how beneficial this smart monitoring pipeline can be by providing an informative ROI rather than using the entire frame.

The face detection step on the first frame represents an important step in the pipeline since the face must be correctly detected before it is tracked and analyzed. As mentioned above, when face detection fails, we opted to use the entire frame as a substitute, which occurred about 10% of the time. We also considered stalling the detection until a face can be detected; however, this approach would show serious limitations during prolonged challenging scenarios (e.g., during a low lighting period lasting 20 mins, the system could not proceed for this entire time).

Table 8.2: Results from Non-contact HR Neonatal Monitoring Pipeline for Resting Patients in Natural Lighting

| **Approach** | **Accuracy** | **Standard Deviation** |
|---|---|---|
| **Entire Frame (Chapter 7)** | 74.00 | 17.23 |
| **Complete HR Estimation Pipeline (Chapter 8)** | **78.28** | **12.09** |

Overall, the smart monitoring pipeline presented here is effective since the ROI is restricted to the patient's face tracked over time, as opposed to the entire frame. Given that many monitoring applications use RGB cameras, using the RGB streams for bed occupancy and HR estimation shows the utility and applicability of our system in other applications.

While the patients here were at rest, introducing occlusions, motion, or changes in lighting environment could affect the performance of HR estimation. Experiments presenting such cases are shown in the following section.

## 8.2 Uncertainty Measurements in Non-Contact HR Estimation Pipeline

Once a monitoring pipeline is in place, it is important to evaluate how potential errors generated at each step would affect the rest of the system, as illustrated in the lower part of Figure 8.2. Most studies calculate HR estimation uncertainty from only the final stage of the pipeline [28], [61]. This ignores valuable uncertainty information from the preliminary stages of the pipeline. In this section, we develop signal quality indexes (SQI) for each stage and demonstrate that the fusion of these SQI is more informative than an SQI derived strictly from the final "HR Estimation" stage of the monitoring pipeline.

A dataset is curated to include a variety of videos representing the possible errors encountered in the pipeline (8.2.1). Independent SQI values are obtained at each stage before generating a fused SQI

encompassing the entire pipeline (8.2.2). Finally, results are presented in Section 8.2.3 to demonstrate the effectiveness of our proposed fused SQI metric as a decision support tool for clinical applications.

### 8.2.1 Challenging Scenario HR Estimation Dataset

Challenges in the patient environment can affect each stage in the non-contact monitoring pipeline differently. For example, bed occupancy can be challenged by discontinuous presence of the patient in the bed; facial occlusions in one image would challenge a face detection algorithm; facial occlusion and motion (directly from the face or indirectly during body motion) would challenge a tracking method; any change in the environment when the patient is not at rest or when the lighting conditions change drastically would impact a HR estimator. To evaluate all these scenarios, a dataset encompassing 13 distinct 30-sec videos of events from 8 patients were obtained using the bedside annotation (CEA) application to automatically identify events of interest, in addition to the same resting patients used in Sections 7.2 and 8.1.

Table 8.3: Events in the Challenging Scenario HR Estimation Dataset

| **Scenario per pipeline stage** | **Bed Occupancy** | **Face Detection** | **Face Tracking** | **Heart Rate Estimation** |
|---|---|---|---|---|
| | *Challenging Factors* | | | |
| | *absence* | *absence + occlusion* | *absence + occlusion + motion* | *ALL (non-rest/ varying light)* |
| Patient out | X | X | X | X |
| Patient in | X | X | X | X |
| Facial occlusion – temporary | | X | X | X |
| Facial occlusion – continuous | | X | X | X |
| Facial motion suction | | | X | X |
| Facial motion sneeze | | | X | X |
| Facial motion yawn | | | X | X |
| Hiccup | | | X | X |
| Clinical intervention (no occlusion) | | | X | X |
| Body motion – minor | | | X | X |
| Body motion – major | | | X | X |
| High to low light | | | | X |
| Monitor alarm light | | | | X |
| Rest – natural light | | | | |
| Rest – phototherapy light | | | | |

This resulted in a total of **15 distinct scenarios, among 27 videos, and 16 patients.** Each scenario and the expected impact on each pipeline stage is summarized in Table 8.3. Further details of the dataset description are presented in Appendix E.

### 8.2.2 Uncertainty Methods

In this section, SQI are derived from each pipeline stage before combining them to obtain a fused uncertainty metric. The fused SQI is shown to be more informative of the uncertainty of the HR estimate rather than any last-stage SQI separately.

#### *8.2.2.1 SQI of Bed Occupancy*

The SQI for this step is based on the classification scores from the "present" class measured by the deep learning RGB-based bed occupancy model. The higher the "present" confidence output, the higher the confidence in the assertion that the patient is present in the bed. In a 30-sec video, the bed occupancy model is tested every second and a confidence score is extracted. The SQI for the bed occupancy step is calculated as,

$$BO_{conf} = \frac{1}{n}\sum_{i=1}^{n} BOscore_i \ , (8.1)$$

where $BOscore$ consists of the confidence score from the Bed Occupancy (BO) model from the i$^{th}$ image frame, averaged over the $n$ seconds in the video clip (in this thesis, $n$=30). Here, a lower $BO_{conf}$ value would suggest inconsistencies in bed occupancy.

#### *8.2.2.2 SQI of Face Detection*

Similar to the SQI for bed occupancy, the confidence score of the detected "face" object produced by the YOLO5Face model is used as an SQI for face detection

$$FD_{conf} = \frac{1}{n}\sum_{i=1}^{n} FDscore_i \ , \ (8.2)$$

where $FD_{conf}$ consists of the confidence score from the Face Detection (FD) model from the i$^{th}$ image frame, averaged over the $n$ seconds in the video clip. Here, a lower $FD_{conf}$ value would suggest difficulties in detecting the face of the patient due to difference occlusion factors.

Additionally, the face detection rate over the 30-sec video is measured as another SQI to differentiate scenarios when the detection is consistent, despite a low confidence score vs. the case where high confidence detections are made interspersed with lack of detection. In an image selected per second $i$, a binary operation $FDdet_i$ is applied when there is a detection (value of 1) vs not (value of 0), and averaged over all the entire video as

$$FD_{rate} = \frac{1}{n}\sum_{i=1}^{n} (FDdet_i) \qquad (8.3)$$

$$FDdet_i = \begin{cases} 1, & \text{if } \exists \text{ FDscore}_i \\ 0, & \text{otherwise} \end{cases} \qquad (8.4)$$

Finally, a third SQI is derived based on the Nose-to-Eye-Line Angle (NELA) extracted during face orientation estimation (detailed in Appendix D). For simplicity, all videos were rotated for standardized head orientation. This permits evaluation of the pipeline with SQIs from each individual step vs. a fused SQI, independently of face orientation prediction. The confidence of NELA is calculated as

$$NELA_{conf} = \begin{cases} 1, & \text{if NELA} \in [45,135] \\ \frac{\log(271-\lfloor NELA \rfloor)}{\log(136)}, & \text{if NELA} \in (135,270] \\ \frac{\log(91+\lfloor NELA \rfloor)}{\log(136)}, & \text{if NELA} \in [0,45) \\ \frac{\log(\lfloor NELA \rfloor-269)}{\log(136)}, & \text{if NELA} \in (270,360) \\ 0, & \text{otherwise} \end{cases}, \qquad (8.5)$$

where North-facing orientations exhibit best $NELA_{conf}$ (90° ± 45°) and angles approaching South are normalized log of angles (e.g., from 135° to 270°). The mapping of NELA to $NELA_{conf}$ for all angles is depicted in Figure 8.3.

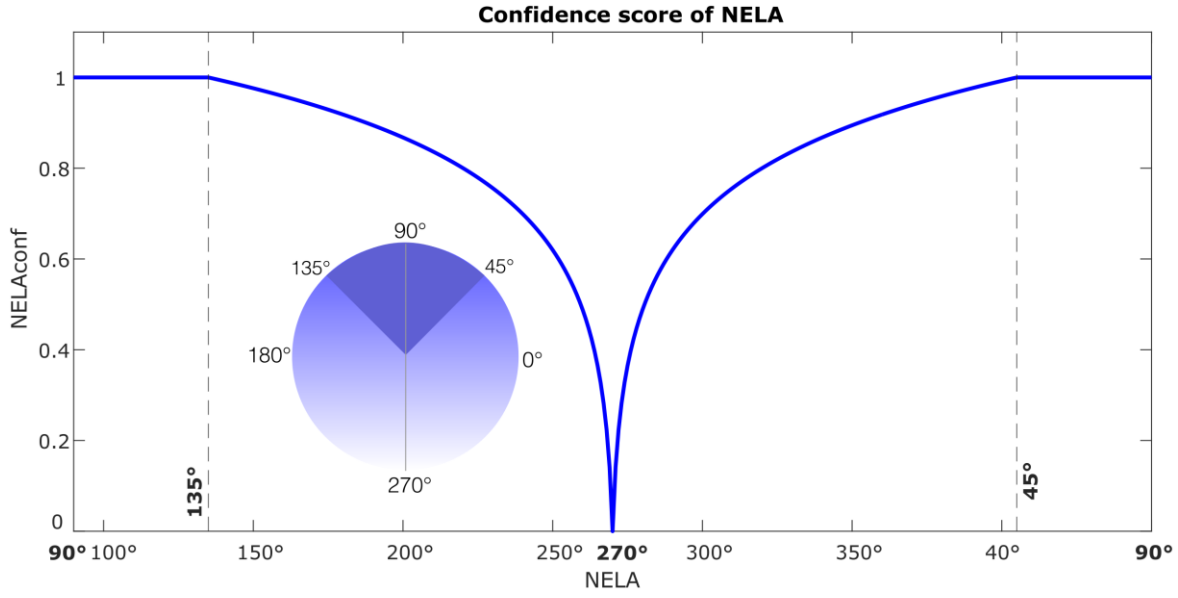



Figure 8.3: Confidence score of the NELA prediction.

#### 8.2.2.3 SQI of Face Tracking

As described in Chapter 6.3, from a detected ROI, landmarks are extracted and tracked over time. Often, during changes in the environment (e.g., during motion or occlusion), a loss in the visible landmarks is observed, and this knowledge is exploited to derive an SQI for face tracking as

$$Land_{ratio} = \frac{1}{n}\sum_{i=1}^{n}\frac{VL_i}{IVL} \quad , (8.6)$$

where $Land_{ratio}$ calculates the ratio of visible landmarks (VL) at the $i^{th}$ frame compared to the initial frame's visible landmarks (IVL), averaged over the $n$ seconds in the video clip.

#### 8.2.2.4 SQI of HR Estimation

To evaluate the quality of the HR signal, two different methods are used. The first one is based on the SQI calculations introduced by Pereira et al [61]. In their work, they compared the SQI from different ROIs extracted from thermal imaging for respiration estimation, to identify if the ROI corresponds to the respiration signal or noise. We use a similar approach using the features F1-F4 to calculate the $SQI_{Pereira}$ for HR signal estimated during rest periods vs scenarios potentially introducing noise as

$$F1 = \max_{f>3}(|C(f)| \qquad (8.7)$$

$$F2 = \frac{1}{n}\sum_{f>3}|C(f)| \qquad (8.8)$$

$$F3 = \left|\max_{f<0.1}(|C(f)| - \max_{0.1\leq f\leq 3}(|C(f)|\right| \qquad (8.9)$$

$$F4 = \left(\left|\max_{f<0.1}(|C(f)| - \max_{0.1\leq f\leq 3}(|C(f)|\right|\right) \div \left(\max\left(\max_{f<0.1}(|C(f)|\ , \max_{0.1<f<0.1}(|C(f)|\right)\right) \qquad (8.10)$$

$$SQI_{Pereira} = \begin{cases} 1 - \left[\frac{1}{2}F3 + \frac{1}{4}(F1 + F2)\right], & \text{if F4} \geq 2 \\ 1 - \frac{1}{2}(F1 + F2), & \text{otherwise} \end{cases} \quad , \qquad (8.11)$$

where the features and $SQI_{Pereira}$ is calculated for the HR signal |C(f)| from a channel C among the Red, Green, and Blue channels. This results in three distinct $SQI_{Pereira}$ values (one per channel), and an example drawn from the Red channel is depicted in Figure 8.4.

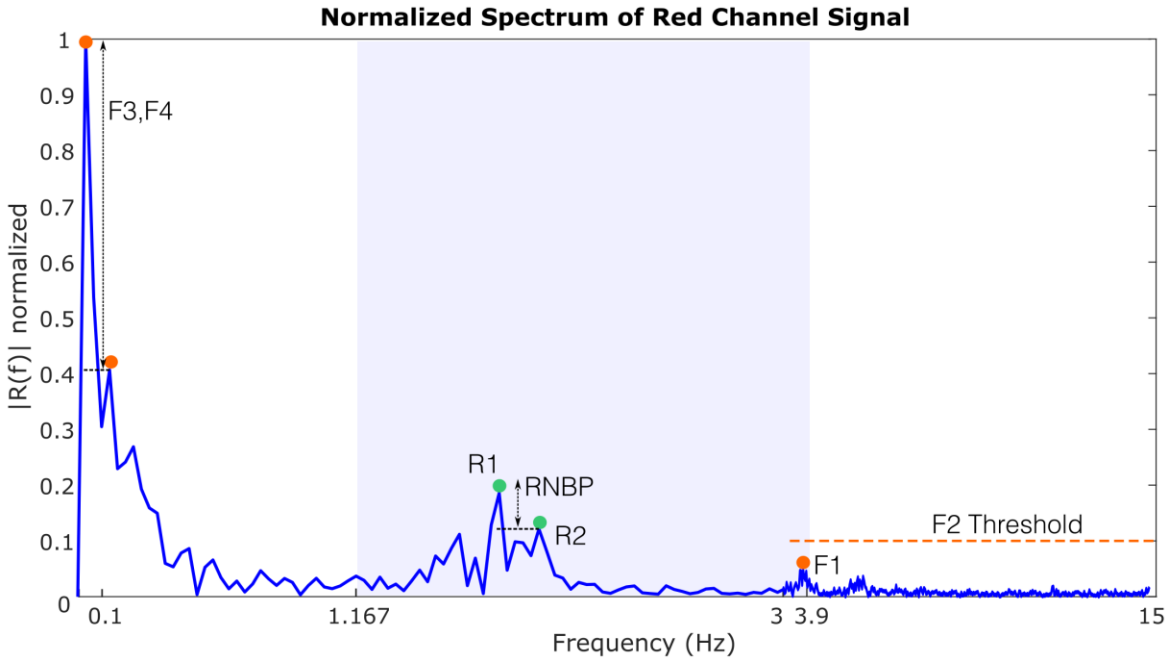



Figure 8.4: Normalized spectrum of Red Channel Signal. Features (F1-F4) for SQI calculation adapted from Pereira *et al.* [61], and features (R1-R2) for SQI measured from Ratio within Neonatal-based BandPass (RNBP). The blue-shaded area corresponds to the neonatal bandpass and the frequency axis is scaled for visualization purposes.

As a second approach, the ratio between the two maximum peaks within a neonatal passband (RNBP) corresponding to plausible HR values from newborns (1.167-3.9 Hz, 70 to 234 bpm) as

$$R1 = \max_{1.167 f > 3.9} (|C(f)| \qquad (8.12)$$

$$R2 = \max_{1.167 f > 3.9} (|C(f - R1)| \qquad (8.13)$$

$$RNBP = (|R1 - R2|) \div (\max(R1, R2)) \quad , \qquad (8.14)$$

where RNBP is calculated for the HR signal |C(f)| from a channel C among the Red, Green, and Blue channels. This results in three distinct $RNBP$ values for each channel, and an example drawn from the Red channel is depicted in Figure 8.4.

All SQIs used here at each step in the monitoring pipeline are summarized in Table 8.4

Table 8.4: Description of SQI extracted from each step in the non-contact neonatal HR monitoring pipeline

| Step | SQI | Description |
|---|---|---|
| Bed Occupancy | $\mathrm{BO}_{\mathrm{conf}}$ | Confidence score of "present" classification |
| Face Detection | $\mathrm{FD}_{\mathrm{conf}}$ | Confidence score of "face" bounding box detection |
| | $\mathrm{FD}_{\mathrm{rate}}$ | Detection rate of detected face |
| | $\mathrm{NELA}_{\mathrm{conf}}$ | Confidence score based on NELA prediction |
| Face Tracking | $\mathrm{Land}_{\mathrm{ratio}}$ | Ratio of remaining visible landmarks |
| HR Estimation | $\mathrm{Pereira}_{\mathrm{R}}$ | SQI formula from Pereira *et al.* [45] in Red channel |
| | $\mathrm{Pereira}_{\mathrm{G}}$ | SQI formula from Pereira *et al.* [45] in Green channel |
| | $\mathrm{Pereira}_{\mathrm{B}}$ | SQI formula from Pereira *et al.* [45] in Blue channel |
| | $\mathrm{RNBP}_{\mathrm{R}}$ | Ratio between two max peaks within neonatal passband in Red channel |
| | $\mathrm{RNBP}_{\mathrm{G}}$ | Ratio between two max peaks within neonatal passband in Green channel |
| | $\mathrm{RNBP}_{\mathrm{B}}$ | Ratio between two max peaks within neonatal passband in Blue channel |

#### *8.2.2.5 Fused SQI*

Leveraging all 11 stage-specific SQIs presented above and described in Table 8.4, a Gradient Boosted Regression Trees (GBRT) approach is used to combine them and arrive at a more informative fused "pipeline SQI". GBRT methods work by using an ensemble of multiple decision trees as learners in a boosting algorithm to minimize the residuals between the predicted values (Fused SQI) and the actual values (HR accuracy). During training, 500 trees are used with a learning rate of 0.01 and a tree depth of 5. A 13-fold cross-validation is employed by holding out one of the 13 challenging scenarios plus one rest event in natural light.

For each video, the computed SQI is compared to corresponding HR accuracy. A linear relationship is inferred using linear regression from iteratively reweighted least squares to reduce impact of outliers. The slope, root mean squared error (RMSE), and coefficient of determination ($R^2$) and calculated to evaluate the fit of the individual SQI vs HR accuracy.

### 8.2.3 Uncertainty Results & Discussion

Results from each SQI per step in the pipeline are illustrated in Figure 8.5, and are summarized in Table 8.5. Ideally, the SQI should follow a positive linear trend with the HR accuracy since each SQIs were derived in such a manner that a greater value would suggest a greater level of certainty in the corresponding task. In some cases, the SQIs are consistently high for most scenarios resulting in an abundance of data points in the SQI = 1 vertical line (e.g., most severe $\mathrm{BO}_{\mathrm{conf}}$, and less severe $\mathrm{FD}_{\mathrm{rate}}$, $\mathrm{NELA}_{\mathrm{conf}}$, $\mathrm{Land}_{\mathrm{ratio}}$). In other cases, the linear fit is obtained around a slope of 0 when data points are

sparsely distributed, without any pertinent trend. For visualization purposes, Figure 8.5 only shows results from the Red channel in Step 4 ($Pereira_R$ and $RNBP_R$), however, similar patterns were observed from the Green and Blue channels. Interestingly, the $Pereira_R$ metric tends to overestimate SQI values, while the $RNBP_R$ underestimates. Neither demonstrate any informative linear trend.

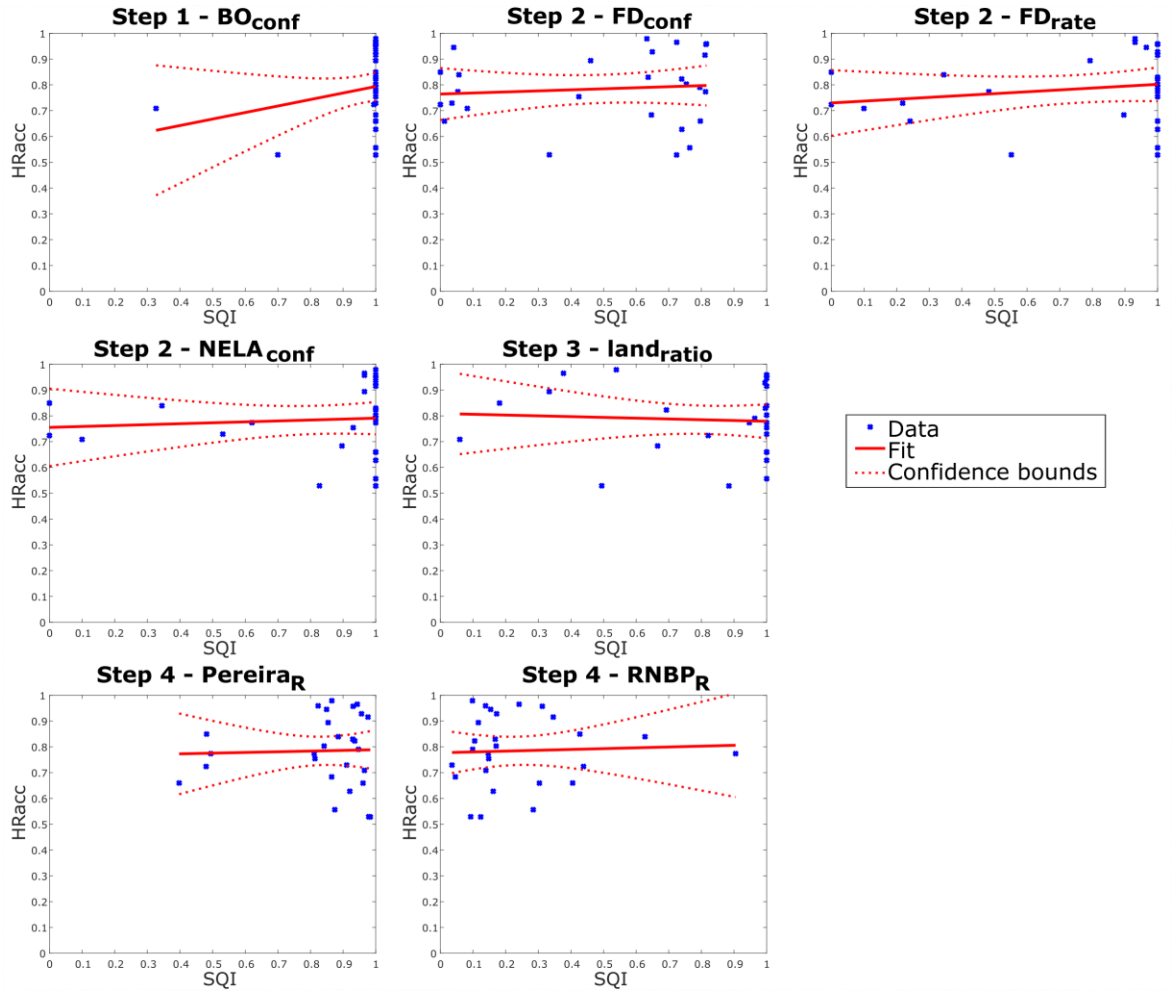



Figure 8.5: SQI plots from each step in the monitoring pipeline.

From all steps, a ~ 0 slope is obtained, with ~ 0.138 RMSE and ~ 0 $R^2$, which strongly suggests that none of these models can predict the HR accuracy based on the measured stage-specific SQI values. These results are observed qualitatively in Figure 8.5 and quantitatively in Table 8.5.

Table 8.5: Evaluation per SQI extracted from each step in the non-contact neonatal monitoring pipeline

| **Evaluation per SQI** | **Slope** | **RMSE** | $R^2$ |
|---|---|---|---|
| $BO_{conf}$ | 0.252 | 0.134 | 0.067 |
| $FD_{conf}$ | 0.041 | 0.138 | 0.010 |
| $FD_{rate}$ | 0.072 | 0.136 | 0.037 |
| $NELA_{conf}$ | 0.036 | 0.138 | 0.007 |
| $Land_{ratio}$ | -0.030 | 0.138 | 0.004 |
| $Pereira_{R}$ | 0.027 | 0.138 | 0.001 |
| $Pereira_{G}$ | -0.074 | 0.138 | 0.005 |
| $Pereira_{B}$ | 0.046 | 0.138 | 0.003 |
| $RNBP_{R}$ | 0.032 | 0.138 | 0.002 |
| $RNBP_{G}$ | 0.073 | 0.138 | 0.010 |
| $RNBP_{B}$ | 0.041 | 0.138 | 0.004 |
| **Fused** | **1.304 ± 0.052** | **0.027 ± 0.003** | **0.962 ± 0.008** |

Using a GBRT approach to intelligently fuse these SQIs, however, overcomes these issues by showing great promise in predicting HR accuracy, as demonstrated by the slope of 1.304, RMSE of 0.027, and $R^2$of 0.962. As depicted in the left side of Figure 8.6, the data points follow a positive linear trend with little standard deviation from the error distribution. These results were obtained from 13-fold cross validation testing such that the regression model was fit and evaluated on different data.

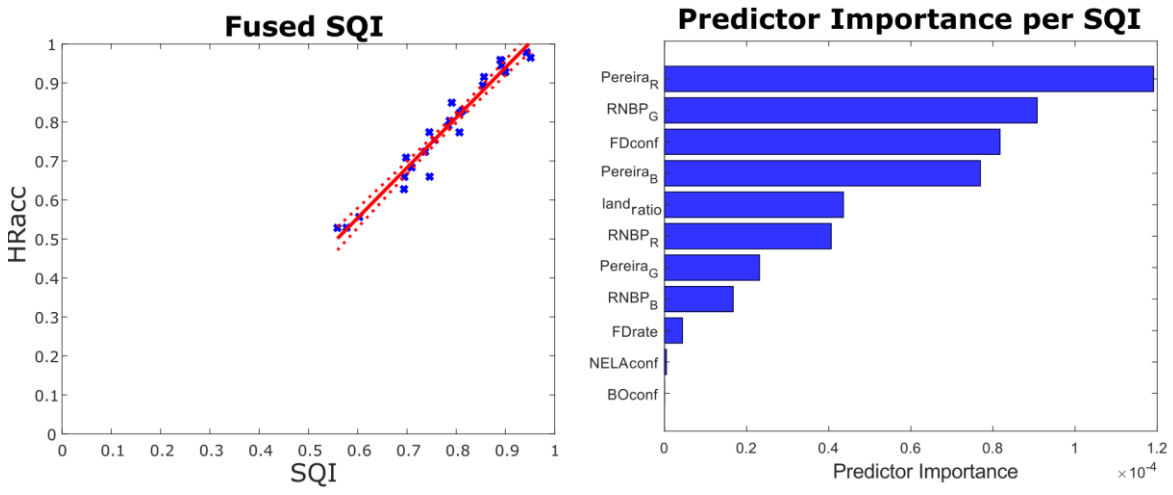



Figure 8.6: Fused SQI with predictor importance of each individual SQI.

The right side of Figure 8.6 shows the predictor importance which is estimated during training of the ensemble model. Each tree has different split of predictors during training, and the node risk $R_i$ is calculated at every node as,

$$R_i = P_i \cdot E_i \quad , \quad (8.15)$$

where $P_i$ is the probability of node $i$ and $E_i$ is the mean squared error of node $i$. Using node risk measures, predictor importance, $Imp_j$, is calculated as

$$Imp_j = \frac{1}{n}\left(R_p - \sum_{i=1}^{n} R_{c_i}\right) \quad , \quad (8.16)$$

in each regression n tree $j$ by measuring the node risk of splitting every predictor in that tree from every parent node ($R_p$) into the $n$ child nodes ($R_c$). The reported predictor importance, $Imp_{ens}$, then averages all $Imp_j$ over the ensemble of $N$ trees using the weights $W_t$ of each tree.

$$Imp_{ens} = \sum_{t=1}^{N} W_t \cdot Imp_j \quad , \quad (8.17)$$

Resulting predictor importance for our model is depicted in the right panel of Figure 8.6 where the most important predictor is the $Pereira_R$ followed by $RNBP_G$. Despite their significance, neither of these stage-specific SQI is strong predictor individually. While other vital sign monitoring studies would typically report the SQI of a signal obtained strictly from the final stage [28], [61], this thesis presents how a Fused SQI encompassing the uncertainty at each step in the monitoring pipeline can be more informative. Providing a reliable measure of vital sign estimation uncertainty is critical, if one is to make clinical decisions based on the data provided (e.g., silence an alarm or change treatment).

## 8.3 Non-Contact Neonatal Monitoring – Conclusions

This thesis has demonstrated non-contact neonatal monitoring pipelines can be created by integrating multiple research contributions presented here. More specifically, we demonstrated how a non-contact HR estimation pipeline can be implemented from a composition of bed occupancy, face detection, face tracking, and HR estimation approaches. Such pipeline demonstrated improved performance of the HR estimate compared to the results from Chapter 7 where the entire video frame was used as the ROI.

Additionally, this thesis derived multiple SQIs in each stage of the pipeline to arrive at a fused SQI that was shown to be more informative in assessing the uncertainty of final HR estimate. Using gradient boosted regression trees allowed to identify and combine the predictors providing maximal information for the Fused pipeline SQI prediction. This work showed great promise as a decision support tool, however, it might not yet be ready for deployment in hospitals. Further experimentation would require more data collected from more patients in different scenarios, in addition to careful consultations with clinicians to determine an acceptable level of model accuracy for clinical deployment.

# 9 Conclusions

This chapter summarizes the major conclusions and contributions made in this thesis. Suggestions for future work to extend this research are also provided.

## 9.1 Papers Arising from this Thesis

This thesis resulted in the following papers

1. **Y. S. Dosso**, A. Bekele, and J. R. Green, “Eulerian Magnification of Multi-Modal RGB-D Video for Heart Rate Estimation,” in Proc. of IEEE Int. Symp. Med. Meas. Appl. (MeMeA), Rome, Italy, 2018. ***Awarded the IEEE MeMeA "Women in Engineering Best Paper"*** [21]
2. **Y. S. Dosso**, A. Bekele, S. Nizami, C. Aubertin, K. Greenwood, J. Harrold, and J. R. Green, “Segmentation of patient images in the neonatal intensive care unit,” in 2018 IEEE Life Sciences Conference, LSC 2018, 2018, pp. 45–48. [22]
3. **Y. S. Dosso**, S. Aziz, S. Nizami, K. Greenwood, J. Harrold, and J. R. Green, “Neonatal Face Tracking for Non-Contact Continuous Patient Monitoring,” in 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2020, pp. 1–6. [23]
4. **Y. S. Dosso**, S. Aziz, S. Nizami, K. Greenwood, J. Harrold, and J. R. Green, “Video-Based Neonatal Motion Detection,” in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2020, vol. 2020-July, pp. 6135–6138. [24]
5. **Y. S. Dosso**, K. Greenwood, J. Harrold, and J. R. Green, "Bottle-Feeding Intervention Detection in the NICU." 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021. [25]
6. **Y. S. Dosso**, R. Selzler, K. Greenwood, J. Harrold, and J. R. Green, "RGB-D Sensor Application for Non-Contact Neonatal Monitoring." 2021 IEEE Sensors Applications Symposium (SAS). IEEE, 2021. [26]
7. **Y. S. Dosso**, K. Greenwood, J. Harrold, and J. R. Green, “RGB-D Scene Analysis in the NICU,” in Elsevier – Computers in Biology and Medicine, 2021. [27]
8. **Y. S. Dosso**, D. Kyrollos, K. Greenwood, J. Harrold, and J. R. Green, “Robust Neonatal Face Detection in Complex NICU Scenes”. IEEE Access, 2022 ***(In review)***

Throughout the period of working on this thesis research, I have also contributed and co-authored other publications from the overarching unobtrusive neonatal monitoring project.

1. A. Bekele, S. Nizami, **Y. S. Dosso**, C. Aubertin, K. Greenwood, J. Harrold, J. R. Green, "Real-time neonatal respiratory rate estimation using a pressure-sensitive mat." in 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2018. [207]

*Co-developed the data collection protocol, participated in the data collection and management steps, and edited the manuscript.*

2. S. Aziz, S. Nizami, **Y. S. Dosso**, K. Greenwood, J. Harrold, J. R. Green. "Detection of Neonatal Patient Motion Using a Pressure-Sensitive Mat." in 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2020. ***Received the IEEE MeMeA "Best Student Paper" award.*** [208]

   *Mentored the first author in the research, experimental design, and the creation and presentation of the manuscript.*

***3.*** D.G. Kyrollos, R. Hassan, **Y. S. Dosso**, J. R. Green. "Fusing Pressure-Sensitive Mat Data with Video Through Multi-Modal Registration." in IEEE 2021 International Instrumentation and Measurement Technology Conference (I2MTC), 2021. [209]

   *Mentored the first author and directly supervised the second author. Contributed to the research methodology, experimental design, and interpretation of results.*

## 9.2 Overall Results and Contributions

This thesis assessed the state of the art in a multitude of machine vision applications. While most of these are commonly implemented on an adult population, an important gap was observed when leveraging highly performant models on our neonatal population. To this end, this thesis evaluated different scenarios, typical of an NICU environment, where state of the art models would typically fail. Throughout various chapters, we identified such neonatal-specific and NICU-specific challenging scenarios as:

1. Neonatal facial features
2. Phototherapy lighting
3. Facial occlusions due to hospital equipment (mainly from ventilation support or phototherapy eye mask)
4. Complex patient pose (prone, supine, side/fetal position)
5. Varying degree of body coverage (unclothed, clothed, covered with blanket, swaddled)
6. Temporary occlusions during clinical interventions

All of these challenges can impact machine vision applications related to patient monitoring, and this thesis presents several solutions by advancing the state of the art when severe limitations were observed, often using different deep learning and image processing technologies in a complementary manner to arrive at our goal. Subsequently, this thesis demonstrated how we can compose various research contributions into a useful monitoring pipeline. An instantiation of a video-based neonatal heart rate

estimation pipeline is presented, demonstrating a direct application of research presented in this thesis, and providing a promising system as a decision support tool for future clinical use.

Table 9.1 highlights each research contribution presented in this thesis in comparison with previous neonatal monitoring studies from the literature. In some cases, research work presented here has not been accomplished by others and is marked as N/A.

Table 9.1: Thesis Contribution Compared to Other Neonatal Monitoring Studies

| **Research Work** | **This Thesis** | **Others** |
|---|---|---|
| 3A. RGB-D Camera for Video-Based Patient Monitoring | Recommended minimum camera distance<br>Open beds: 40 cm<br>Closed incubator: 30 cm | N/A |
| 4A. Scene Analysis | Accuracy per Context<br>**Intervention: 94.73 %**<br>**Occupancy: 97.64 % (98.98 % AUC)**<br>Lighting: 96.68 %<br>Phototherapy: 99.81 %<br>Coverage: 91.73 % | Accuracy per Context<br>Villarroel *et al.* [28]<br>**Intervention: 94.5 %**<br>**Occupancy: 98.8 % (98.2 % AUC) -- no interventions, patient uncovered, single bed type** |
| | Sentence Generation Accuracy<br>All 5 contexts: 73.05 %<br>All 4 contexts except coverage: 97.97 % | N/A |
| 4B. Bottle-Feeding Intervention Detection | Sensitivity: 51.51 %<br>Accuracy: 83.96 % | N/A |
| 5A. Neonatal Face Detection | Accuracy per dataset<br>**COPE: 100 % (NICUface-RF & NICUface-Y5F)**<br>NBHR: 100 % (NICUface-RF & NICUface-Y5F)<br>CHEOopt: 99.1 % (NICUface-RF)<br>CHEOch: 83.77 % (NICUface-Y5F) | Accuracy<br>Awais *et al.* [46]<br>**Patients in frontal view: 98.5 %** |
| 5B. Patient Segmentation | Jaccard<br>CNN + Post-processing image filtering (median filtering & morphological closing/opening)<br>SegNet: 82.1 % (with closing)<br>**Mask R-CNN: 83.4 % (with opening)** | Jaccard<br>Antink *et al.* [47]<br>Own encoder-decoder: 51.0%<br><br>Asano *et al.* [50]<br>**U-Net GAN: 70.4 %** |
| 6A. Neonatal Motion Detection | Precision and Recall<br><br>*Optical Flow*<br>Precision: 68.02 %<br>Recall: 90.51 % | Cenci *et al.* [1]<br>No results reported<br><br>** This work requires the input of the user to identify an ROI for motion detection. Also, patient movement are only visualized through a software, but |

| | | |
|---|---|---|
| | | no evaluation of their detection algorithm is reported. |
| 6B. Neonatal Face Tracking | Mean AUC: 0.6488 | N/A |
| 6C. Limb Motion Detection | Accuracy<br>*Right arm, left arm, right leg, left leg*<br>Mean Accuracy: 85.3 % | Cenci *et al.* [1]<br>No results reported<br><br>** This work requires the input of the user to identify an ROI for motion detection (i.e., identifying limbs). Also, patient movement are only visualized through a software, but no evaluation of their detection algorithm is reported. |
| 7B. RGB-D HR Estimation (neonates) | Mean Absolute Error<br>**Natural Light: 25.2 bpm**<br>Phototherapy Light: 10.3bpm | Villarroel *et al.* [28]<br>**Natural Light: 2.3 bpm** |
| 8A. Uncertainty Measurement in Non-Contact HR Monitoring | MSE: 0.027 ± 0.003<br>$R^2$: 0.962 $\pm$ 0.008 | N/A |

## 9.3 Future Work

This thesis assessed the state of the art in many machine vision applications pertaining to neonatal patient monitoring, and often advanced these methods when limitations were observed. A non-contact HR estimation pipeline was then proposed as the culmination of research contributions presented here. To do so, a thorough data collection process was used to acquire a unique RGB-D dataset comprising of complex NICU scenes from realistic scenarios observed in continuous monitoring. Although promising results were achieved here, future work is required to further expand this research for hospital deployment.

In this thesis, the Intel Real Sense SR300 was selected instead of the Microsoft Kinect v2 for multiple reasons discussed in Chapter 3.4.1. Newer depth-sensing technologies could, however, be used for future research in monitoring applications. The Azure Kinect should be considered since it can operate at close range and provides hardware-accelerated skeleton detection with greater resolution than the Kinect v2 (with total of 32 vertices). Recent studies have demonstrated great improvements with this device compared to its predecessors [210]. Using new devices, future work will also explore capturing RGB, depth, and near-infrared videos at full resolution, rather than downsampling one modality. In this study, the RGB channel was downsampled to 640x480 pixels, to match the depth/NIR channel resolution while remaining of high quality and reasonable file size for machine vision applications.

Our unique neonatal dataset collected at CHEO provided valuable information for the implementation of neonatal monitoring applications robust to challenging NICU scenes. In some instances, however, the acquisition of insufficient specific scenes (e.g., ongoing phototherapy or bottle-feeding intervention) resulted in the search for additional images from external sources. While this thesis presented reliable solutions to overcome this problem, other neonatal monitoring applications would require collecting a larger database of such conditions (e.g., for oral feeding assessment). Such monitoring applications are outside the scope of this thesis.

In scene understanding, all contexts were considered individually. Due to the nature of our images, various combinations between multiple contexts already exists in our dataset. Some combinatorial cases are impossible, given the hierarchy of contexts (e.g., *phototherapy* light can only be seen in *high* lighting conditions; a *covered or uncovered* patient must be *present* in the scene). Connections between the image processing and deep learning models were evaluated explicitly for some combinations ("coverage", "occupancy", and "intervention" in different "lighting") while other combinations between deep learning contexts were inherently observed due to the nature of our dataset (mix of "coverage", "occupancy", and "intervention" levels in our entire dataset before performing respective classification for each context). Future studies could investigate explicit combinatorial cases where a model incorporating cross-talk between output nodes after feature extraction layers could be explored. Predictions on each task would then inform each other, resulting in a multi-objective classification system with five binary outcomes. This method could not only improve accuracy but could also reduce prediction time by sharing computations. This present work focused on investigating which stream or fusion data model is more suitable for each context. Future work could elaborate on fusion data in combinatorial cases and investigate prediction time improvement using a cross-talk network.

Beyond the NICU environment, the scene analysis model developed here could also be used in home monitoring. Increasingly complex home monitoring devices are becoming commercially available. For instance, the Owlet technology uses video analysis to measure the length and quality of the baby's sleep while a sock can track their heart rate and oxygen levels during minimum motion [56]. A similar non-contact and contact pairing can be seen with the Nanit product [57], which tracks and analyzes the baby's sleep patterns before suggesting advice to parents to improve the baby's sleep quality. Our research could improve on these devices by tracking additional events of interest.

For bottle-feeding intervention detection, future work could transition from classification to object detection, to further analyze the entire clinical scene (e.g., within bottle-feeding event, can we identify periods of active feeding vs. pauses). Our dataset solely included intervention images where the patient and nurse can be detected in the scene, and the occurrence of the nursing bottle would distinguish between the "bottle-feeding" and "no-feeding" class. Other combinations could be investigated for

detailed analysis (e.g., patient present and bottle near the baby but absent nurse could suggest a paused feeding event). This might require more complex video analyses such as action recognition techniques. Other intervention events such as dressing, diaper change, or changing sensors will likewise require an evaluation of a sequence of images to infer context. Although, it may be difficult to find videos from outside sources to apply our data expansion technique, the approach may still be applicable since video analysis models often leverage a feature extraction step trained using individual images before concatenating frames to identify patterns in video sequences.

Depending on the hospital chart requirements, a more sophisticated sentence generator could be developed. Future work could develop deep natural language processing techniques for auto-captioning of videos and transcription in documentation forms for automated charting.

Note that for RetinaFace and YOLO5Face, lightweight models suitable for detections on embedded or mobile devices were also implemented using MobileNet-0.25 and ShuffleNet backbone models, respectively. These pretrained models were not investigated here since in our application, we are not limited in compute power, so these lightweight models are not particularly useful. Future work could however use these models for the implementation of other neonatal monitoring applications (*e.g.*, in home monitoring or in intelligent monitoring applications from smartphones).

As a proof-of-concept, this thesis presented methods for detecting neonatal motion. This work used a depth-sensing camera to capture videos, however, only color videos were analyzed. Future work beyond the scope of this thesis will investigate depth information while modifying the CNN model for feature extraction to accommodate imputing RGB-D video data. Other CNNs architectures could also be explored, given the increased number of features from depth data. The LSTM architecture could also differ by optimizing the number of layers, the number of hidden units, and other hyperparameters. Additionally, we extracted videos with length of 3-16 seconds; however, a different video length may be more suitable for different applications. Future work could evaluate the minimum video length required to provide sufficient data input. For example, in classifying interventions, a diaper change might require ten seconds to be recognized, while a feeding event might only require three seconds. This work aimed to detect motion to inform physiologic signal of motion artifacts; future work could gate alarms from the patient monitor to address this issue, thereby informing the monitoring system on signal quality.

Different methods for re-initialization could be explored in the face tracking algorithm. Tracking by detection could solely leverage a robust face detector (e.g., NICUface) to track by continuous detection of the patient's face (as opposed to only using the detector as the initial ROI). Future work could also incorporate variations in yaw and tilt by creating a 3D model of the patient's head, as opposed to the 2D mask. The depth channel from the Intel camera could therefore help in generating such model. The 5-

point face alignment could also be sufficient to capture frontal and profile frames as the patient moves about, instead of using manual frames for frontal and profile view.

Proof-of-concept is presented for limb motion detection using a semi-automated human skeleton to identify limbs. Future work could however use a pose estimation method to automate this step and obtain a dynamic definition of the limb ROI. Preliminary experiments included analysis on the depth and NIR streams from the Intel camera, which proved to be unsuccessful. Further investigation with more data from additional patients exhibiting various levels of limb motion is left for future work outside the scope of this thesis, where they could potentially aid in low-lighting environments or severe patient occlusions. Neonatal motion detection applications revealed promising results and future work will enable the deployment of such a system in the NICU as part of a motion artifact detection, quantification, and mitigation strategy. Neonatal motion could also detect clonic seizures by analyzing head and body movement patterns. These studies could leverage the detection of motion and categorization of head movements from face tracking and body movements from limb motion detection as a combined approach to evaluating overall body movements suggesting seizures or other motion-related conditions.

Considering that the multi-modal selective EVM was shown to be effective for HR estimation using color magnification, we expect it could also be used for other applications such as respiration rate (RR) estimation from EVM motion magnification [211], [212]. This could be investigated in future work for neonatal RR estimation. One main advantage of utilizing three modalities is the applicability of our algorithm in different environments. Doing so ensures that at least one or two streams provide an accurate estimate in case of failure from the other one or two.

While the Selective EVM method was suitable for estimating resting adult heart between 35-110 bpm and neonatal heart rate between 60-230 bpm, our proposed approach would need to be revisited for patients with especially low or high heart rate that would fall outside of these ranges. Additionally, our study only included patients who did not have a heart condition; our HR estimation method would therefore need to be investigated for such patients exhibiting potential fluctuations in the heart rate signal.

Comprehensively, this thesis presented great promise in bridging the gaps between the highly performing state-of-the-art adult-based methods to an understudied neonatal population, and this represents the core contribution our work presented here. By assessing the state of the art, useful information is provided in this thesis with respect to the complex NICU scenes. Reliable solutions are demonstrated across many research contributions by often advancing the state of the art, and a non-contact HR estimation pipeline is presented as a combination of multiple contributions. Doing so, this thesis advances the field of machine vision and video-based patient monitoring; future work could then expand the research presented here for future deployment of our system in clinical settings. More

investigation and experiment would need to be taken into consideration for future hospital deployment, including the investigation of clinical relevance of vital sign estimates through error grid analyses, the addition of expert opinion for a clinically acceptable performance of our machine vision models, and future fusion and registration of depth-sensing video-based data with pressure data from the PSM. All of these important applications will permit to obtain a robust and comprehensive non-contact, non-invasive, and unobtrusive neonatal monitoring pipeline to ultimately improve patient care.

# References


[1] A. Cenci, D. Liciotti, E. Frontoni, A. Mancini, and P. Zingaretti, "Non-Contact Monitoring of Preterm Infants Using RGB-D Camera," in *Volume 9: 2015 ASME/IEEE International Conference on Mechatronic and Embedded Systems and Applications*, Aug. 2015, p. V009T07A003. doi: 10.1115/DETC2015-46309.

[2] L. Cattani, D. Alinovi, G. Ferrari, R. Raheli, E. Pavlidis, C. Spagnoli, and F. Pisani, "Monitoring infants by automatic video processing: A unified approach to motion analysis," *Comput. Biol. Med.*, vol. 80, pp. 158–165, Jan. 2017, doi: 10.1016/J.COMPBIOMED.2016.11.010.

[3] H. Rehouma, R. Noumeir, P. Jouvet, W. Bouachir, and S. Essouri, "A computer vision method for respiratory monitoring in intensive care environment using RGB-D cameras," in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Nov. 2017, pp. 1–6. doi: 10.1109/IPTA.2017.8310155.

[4] S. Fernando, W. Wang, I. Kirenko, G. de Haan, S. Bambang Oetomo, H. Corporaal, and J. van Dalfsen, "Feasibility of Contactless Pulse Rate Monitoring of Neonates using Google Glass," in *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies,"* 2015, pp. 198–201. doi: 10.4108/eai.14-10-2015.2261589.

[5] J. H. Klaessens, M. van den Born, A. van der Veen, J. Sikkens-van de Kraats, F. A. van den Dungen, and R. M. Verdaasdonk, "Development of a baby friendly non-contact method for measuring vital signs: First results of clinical measurements in an open incubator at a neonatal intensive care unit," in *Advanced Biomedical and Clinical Diagnostic Systems XII*, Feb. 2014, vol. 8935, p. 89351P. doi: 10.1117/12.2038353.

[6] Y. Shi, P. Payeur, M. Frize, and E. Bariciak, "Thermal and RGB-D Imaging for Necrotizing Enterocolitis Detection," *IEEE Med. Meas. Appl. MeMeA 2020 - Conf. Proc.*, Jun. 2020, doi: 10.1109/MEMEA49120.2020.9137344.

[7] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: Why Reinventing a Face Detector," *arXiv Prepr. arXiv2105.12931*, 2021.

[8] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[10] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 2980–2988, Dec. 2017, doi: 10.1109/ICCV.2017.322.

[11] M. Villarroel, A. Guazzi, J. Jorge, S. Davis, P. Watkinson, G. Green, A. Shenvi, K. McCormick, and L. Tarassenko, "Continuous non-contact vital sign monitoring in neonatal intensive care unit," *Healthc. Technol. Lett.*, vol. 1, no. 3, pp. 87–91, Sep. 2014, doi: 10.1049/htl.2014.0077.

[12] W. Hashim, A. Al-Naji, I. A. Al-Rayahi, and M. Oudah, "Computer Vision for Jaundice Detection in Neonates Using Graphic User Interface," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1105, no. 1, p. 12076.

[13] B. Huang, W. Chen, C.-L. Lin, C.-F. Juang, Y. Xing, Y. Wang, and J. Wang, "A neonatal dataset and benchmark for non-contact neonatal heart rate monitoring based on spatio-temporal neural networks," *Eng. Appl. Artif. Intell.*, vol. 106, p. 104447, 2021.

[14] P. S. Hamilton, M. Curley, and R. Aimi, "Effect of adaptive motion-artifact reduction on QRS detection.," *Biomed. Instrum. Technol.*, vol. 34, no. 3, pp. 197–202, 2000.

[15] J. W. Salyer, "Neonatal and pediatric pulse oximetry.," *Respir. Care*, vol. 48, no. 4, pp. 386–96; discussion 397-8, Apr. 2003.

[16] O. M. Cho, H. Kim, Y. W. Lee, and I. Cho, "Clinical alarms in intensive care units: Perceived obstacles of alarm management and alarm fatigue in nurses," *Healthc. Inform. Res.*, vol. 22, no. 1, pp. 46–53, Jan. 2016, doi: 10.4258/hir.2016.22.1.46.

[17] M. T. Petterson, V. L. Begnoche, and J. M. Graybeal, "The Effect of Motion on Pulse Oximetry and Its Clinical Significance," *Anesth. Analg.*, vol. 105, no. On Line Suppl., pp. S78–S84, Dec. 2007, doi: 10.1213/01.ane.0000278134.47777.a5.

[18] K. J. Barrington, N. N. Finer, and C. A. Ryan, "Evaluation of pulse oximetry as a continuous monitoring technique in the neonatal intensive care unit," *Crit. Care Med.*, vol. 16, no. 11, pp. 1147–1153, 1988, doi: 10.1097/00003246-198811000-00013.

[19] R. C. Cartwright-Vanzant, "Medical record documentation: Legal aspects in neonatal nursing," *Newborn Infant Nurs. Rev.*, vol. 10, no. 3, pp. 134–137, Sep. 2010, doi: 10.1053/j.nainr.2010.06.008.

[20] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian Video Magnification for Revealing Subtle Changes in the World", Accessed: Dec. 05, 2017. [Online]. Available: https://people.csail.mit.edu/mrub/papers/vidmag.pdf

[21] Y. S. Dosso, A. Bekele, and J. R. Green, "Eulerian Magnification of Multi-Modal RGB-D Video for Heart Rate Estimation," 2018.

[22] Y. S. Dosso, A. Bekele, S. Nizami, C. Aubertin, K. Greenwood, J. A. Harrold, and J. R. Green, "Segmentation of patient images in

the neonatal intensive care unit," in *2018 IEEE Life Sciences Conference, LSC 2018*, Dec. 2018, pp. 45–48. doi: 10.1109/LSC.2018.8572169.

[23] Y. S. Dosso, S. Aziz, S. Nizami, K. Greenwood, J. Harrold, and J. R. . Green, "Neonatal Face Tracking for Non-Contact Continuous Patient Monitoring," in *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Jun. 2020, pp. 1–6. doi: 10.1109/MeMeA49120.2020.9137300.

[24] Y. S. Dosso, S. Aziz, S. Nizami, K. Greenwood, J. Harrold, and J. R. Green, "Video-Based Neonatal Motion Detection," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Jul. 2020, vol. 2020-July, pp. 6135–6138. doi: 10.1109/EMBC44109.2020.9175354.

[25] Y. S. Dosso, K. Greenwood, J. A. Harrold, and J. R. Green, "Bottle-Feeding Intervention Detection in the NICU," *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, vol. 2021, pp. 1814–1819, Nov. 2021, doi: 10.1109/EMBC46164.2021.9631105.

[26] Y. S. Dosso, R. Selzler, K. Greenwood, J. A. Harrold, and J. R. Green, "RGB-D sensor application for non-contact neonatal monitoring," *2021 IEEE Sensors Appl. Symp. SAS 2021 - Proc.*, Aug. 2021, doi: 10.1109/SAS51076.2021.9530044.

[27] Y. Souley Dosso, K. Greenwood, J. Harrold, and J. R. Green, "RGB-D scene analysis in the NICU," *Comput. Biol. Med.*, vol. 138, p. 104873, Nov. 2021, doi: 10.1016/J.COMPBIOMED.2021.104873.

[28] M. Villarroel, S. Chaichulee, J. Jorge, S. Davis, G. Green, C. Arteta, A. Zisserman, K. McCormick, P. Watkinson, and L. Tarassenko, "Non-contact physiological monitoring of preterm infants in the Neonatal Intensive Care Unit," *npj Digit. Med.*, vol. 2, no. 1, pp. 1–18, Dec. 2019, doi: 10.1038/s41746-019-0199-5.

[29] S. L. Rossol, J. K. Yang, C. Toney-Noland, J. Bergin, C. Basavaraju, P. Kumar, and H. C. Lee, "Non-Contact Video-Based Neonatal Respiratory Monitoring," *Children*, vol. 7, no. 10, p. 171, Oct. 2020, doi: 10.3390/children7100171.

[30] M. S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, and Y. Sun, "Multi-Channel Neural Network for Assessing Neonatal Pain from Videos," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, Aug. 2019, vol. 2019-Octob, pp. 1551–1556. doi: 10.1109/SMC.2019.8914537.

[31] S. Brahnam, L. Nanni, S. McMurtrey, A. Lumini, R. Brattin, M. Slack, and T. Barrier, "Neonatal pain detection in videos using the iCOPEvid dataset and an ensemble of descriptors extracted from Gaussian of Local Descriptors," *Appl. Comput. Informatics*, 2019, doi: 10.1016/j.aci.2019.05.003.

[32] Y. Sun, D. Kommers, W. Wang, R. Joshi, C. Shan, T. Tan, R. M. Aarts, C. Van Pul, P. Andriessen, and P. H. N. De With, "Automatic and Continuous Discomfort Detection for Premature Infants in a NICU Using Video-Based Motion Analysis," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Jul. 2019, pp. 5995–5999. doi: 10.1109/EMBC.2019.8857597.

[33] S. Orlandi, K. Raghuram, C. R. Smith, D. Mansueto, P. Church, V. Shah, M. Luther, and T. Chau, "Detection of Atypical and Typical Infant Movements using Computer-based Video Analysis," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2018, pp. 3598–3601. doi: 10.1109/EMBC.2018.8513078.

[34] F. Pisani, C. Spagnoli, E. Pavlidis, C. Facini, G. M. Kouamou Ntonfo, G. Ferrari, and R. Raheli, "Real-time automated detection of clonic seizures in newborns," *Clin. Neurophysiol.*, vol. 125, no. 8, pp. 1533–1540, 2014, doi: 10.1016/j.clinph.2013.12.119.

[35] N. B. Karayiannis, G. Tao, J. D. Frost, M. S. Wise, R. A. Hrachovy, and E. M. Mizrahi, "Automated detection of videotaped neonatal seizures based on motion segmentation methods," *Clin. Neurophysiol.*, vol. 117, no. 7, pp. 1585–1594, Jul. 2006, doi: 10.1016/j.clinph.2005.12.030.

[36] R. Weber, A. Simon, F. Poree, and G. Carrault, "Deep transfer learning for video-based detection of newborn presence in incubator," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Jul. 2020, vol. 2020-July, pp. 2147–2150. doi: 10.1109/EMBC44109.2020.9175952.

[37] J. Jorge, M. Villarroel, S. Chaichulee, A. Guazzi, S. Davis, G. Green, K. McCormick, and L. Tarassenko, "Non-Contact Monitoring of Respiration in the Neonatal Intensive Care Unit," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, May 2017, pp. 286–293. doi: 10.1109/FG.2017.44.

[38] N. Koolen, O. Decroupet, A. Dereymaeker, K. Jansen, J. Vervisch, V. Matic, B. Vanrumste, G. Naulaers, S. Van Huffel, and M. De Vos, "Automated Respiration Detection from Neonatal Video Data," *Proc. 4th Int. Conf. Pattern Recognit. Appl. Methods*, pp. 164–169, 2015, doi: 10.5220/0005187901640169.

[39] M. Villarroel, A. Guazzi, J. Jorge, S. Davis, P. Watkinson, G. Green, A. Shenvi, K. McCormick, and L. Tarassenko, "Continuous non-contact vital sign monitoring in neonatal intensive care unit," *Healthc. Technol. Lett.*, vol. 1, no. 3, pp. 87–91, Sep. 2014.

[40] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–8, 2012.

[41] D. G. Kyrollos, J. B. Tanner, K. Greenwood, J. Harrold, and J. R. Green, "Noncontact Neonatal Respiration Rate Estimation Using Machine Vision," in *2021 IEEE Sensors Applications Symposium (SAS)*, 2021, pp. 1–6.

[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[43] M. N. Mansor, S. Yaacob, M. Hariharan, S. N. Basah, S. A. Jamil, M. L. M. Khidir, M. N. Rejab, K. K. M. Y. Ibrahim, A. A. Jamil, A. K. Junoh, and others, "Jaundice in newborn monitoring using color detection method," *Procedia Eng.*, vol. 29, pp. 1631–1635, 2012.

[44] M. J. Maisels and A. F. McDonagh, "Phototherapy for Neonatal Jaundice," *N. Engl. J. Med.*, vol. 358, no. 9, pp. 920–928, Feb. 2008, doi: 10.1056/NEJMct0708376.

[45] S. Bharadwaj, H. S. Bhatt, M. Vatsa, and R. Singh, "Domain specific learning for newborn face recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 7, pp. 1630–1641, 2016.

[46] M. Awais, C. Chen, X. Long, B. Yin, A. Nawaz, S. F. Abbasi, S. Akbarzadeh, L. Tao, C. Lu, L. Wang, and others, "Novel framework: face feature selection algorithm for neonatal facial and related attributes recognition," *IEEE Access*, vol. 8, pp. 59100–59113, 2020.

[47] C. H. Antink, J. C. M. Ferreira, M. Paul, S. Lyra, K. Heimann, S. Karthik, J. Joseph, K. Jayaraman, T. Orlikowsky, M. Sivaprakasam, and S. Leonhardt, "Fast body part segmentation and tracking of neonatal video data using deep learning," *Med. Biol. Eng. Comput.*, vol. 58, no. 12, pp. 3049–3061, Dec. 2020, doi: 10.1007/S11517-020-02251-4/FIGURES/9.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016. Accessed: Aug. 27, 2021. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/

[49] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," 2015.

[50] H. Asano, E. Hirakawa, H. Hayashi, K. Hamada, Y. Asayama, M. Oohashi, A. Uchiyama, and T. Higashino, "A method for improving semantic segmentation using thermographic images in infants," *BMC Med. Imaging*, vol. 22, no. 1, pp. 1–13, Dec. 2022, doi: 10.1186/S12880-021-00730-0/FIGURES/5.

[51] J. K. Aggarwal and Q. Cai, "Human motion analysis: a review," in *Proceedings IEEE Nonrigid and Articulated Motion Workshop*, 1999, pp. 90–102. doi: 10.1109/NAMW.1997.609859.

[52] S. Nizami, A. Bekele, M. Hozayen, K. Greenwood, J. Harrold, and J. R. Green, "Measuring uncertainty during respiratory rate estimation using pressure-sensitive mats," *IEEE Trans. Instrum. Meas.*, vol. in press, 2018.

[53] S. Nizami, J. R. Green, and C. McGregor, "Implementation of artifact detection in critical care: a methodological review.," *IEEE Rev. Biomed. Eng.*, vol. 6, pp. 127–42, Jan. 2013, doi: 10.1109/RBME.2013.2243724.

[54] P. Bifulco, G. Gargiulo, M. Romano, A. Fratini, and M. Cesarelli, "Bluetooth Portable Device for Continuous ECG and Patient Motion Monitoring During Daily Life," *IFMBE Proc.*, vol. 16, pp. 369–372, 2007.

[55] S. Patel, Bor-rong Chen, T. Buckley, R. Rednic, D. McClure, D. Tarsy, L. Shih, J. Dy, M. Welsh, and P. Bonato, "Home monitoring of patients with Parkinson's disease via wearable technology and a web-based application," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, Aug. 2010, pp. 4411–4414. doi: 10.1109/IEMBS.2010.5627124.

[56] "Canada Owlet Smart Sock -Baby Heart Rate & Oxygen Monitor | Owlet Care – Owlet Canada." https://owletcare.ca/ (accessed Sep. 09, 2019).

[57] "Nanit Smart Baby Monitoring System with Breathing Wear | Nanit Canada." https://www.nanit.com/ (accessed Sep. 09, 2019).

[58] H. Rehouma, R. Noumeir, W. Bouachir, P. Jouvet, and S. Essouri, "3D imaging system for respiratory monitoring in pediatric intensive care environment," *Comput. Med. Imaging Graph.*, vol. 70, pp. 17–28, Dec. 2018, doi: 10.1016/j.compmedimag.2018.09.006.

[59] J. Gleichauf, C. Niebler, and A. Koelpin, "Automatic non-contact monitoring of the respiratory rate of neonates using a structured light camera," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Jul. 2020, vol. 2020-July, pp. 4118–4121. doi: 10.1109/EMBC44109.2020.9175948.

[60] H. E. Rice, C. L. Hollingsworth, E. Bradsher, M. E. Danko, S. M. Crosby, R. N. Goldberg, D. T. Tanaka, R. B. Knobel, and G. Brumley, "Infrared Thermal Imaging (Thermography) of the Abdomen in Extremely Low Birthweight Infants," *J. Surg. Radiol.*, pp. 82–89, 2010.

[61] C. B. Pereira, X. Yu, T. Goos, I. Reiss, T. Orlikowsky, K. Heimann, B. Venema, V. Blazek, S. Leonhardt, and D. Teichmann, "Noncontact Monitoring of Respiratory Rate in Newborn Infants Using Thermal Imaging," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 4, pp. 1105–1114, Apr. 2019, doi: 10.1109/TBME.2018.2866878.

[62] S. Ervural and M. Ceylan, "Classification of neonatal diseases with limited thermal Image data," *Multimed. Tools Appl.*, pp. 1–29, Aug. 2021, doi: 10.1007/S11042-021-11391-0/TABLES/6.

[63] V. Bach, S. Delanaud, L. Barcat, E. Bodin, P. Tourneux, and J. P. Libert, "Distal skin vasodilation in sleep preparedness, and its impact on thermal status in preterm neonates," *Sleep Med.*, vol. 60, pp. 26–30, Aug. 2019, doi: 10.1016/J.SLEEP.2018.12.026.

[64] R. B. Knobel-Dail, D. Holditch-Davis, R. Sloane, B. D. Guenther, and L. M. Katz, "Body temperature in premature infants during

the first week of life: Exploration using infrared thermal imaging," *J. Therm. Biol.*, vol. 69, pp. 118–123, Oct. 2017, doi: 10.1016/J.JTHERBIO.2017.06.005.

[65] H. X. Liu, H. X. Liu, and B. Ran, "Vision-Based Stop Sign Detection and Recognition System for Intelligent Vehicles," *Transp. Res. Rec.*, 2001.

[66] J. Zhao, S. Zhu, and X. Huang, "Real-time traffic sign detection using SURF features on FPGA," 2013. doi: 10.1109/HPEC.2013.6670350.

[67] A. Arunmozhi, S. Gotadki, J. Park, and U. Gosavi, "Stop Sign and Stop Line Detection and Distance Calculation for Autonomous Vehicle Control," in *IEEE International Conference on Electro Information Technology*, Oct. 2018, vol. 2018-May, pp. 356–361. doi: 10.1109/EIT.2018.8500268.

[68] C. Dewi, R. C. Chen, Y. T. Liu, Y. S. Liu, and L. Q. Jiang, "Taiwan Stop Sign Recognition with Customize Anchor," in *ACM International Conference Proceeding Series*, Jun. 2020, pp. 51–55. doi: 10.1145/3408066.3408078.

[69] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst. 25 (NIPS 2012)*, 2012.

[70] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable Object Detection using Deep Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2147–2154.

[71] L. Herranz, S. Jiang, and X. Li, "Scene recognition with CNNs: objects, scales and dataset bias," 2016.

[72] 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," *Int J Comput Vis*, vol. 115, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[73] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, Jul. 2018, doi: 10.1016/j.neucom.2018.03.030.

[74] T. Wang and S. Su, "Efficient Scene Layout Aware Object Detection for Traffic Surveillance," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 926–933.

[75] M. de Bruijne, "Machine learning approaches in medical image analysis: From detection to diagnosis," *Medical Image Analysis*, vol. 33. Elsevier B.V., pp. 94–97, Oct. 2016. doi: 10.1016/j.media.2016.06.032.

[76] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, J. Wei, and K. Cha, "Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography," *Med. Phys.*, vol. 43, no. 12, pp. 6654–6666, Nov. 2016, doi: 10.1118/1.4967345.

[77] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep Learning Applications in Medical Image Analysis," *IEEE Access*, vol. 6, pp. 9375–9379, Dec. 2017, doi: 10.1109/ACCESS.2017.2788044.

[78] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle, "Knowledge transfer for melanoma screening with deep learning," in *Proceedings - International Symposium on Biomedical Imaging*, Jun. 2017, pp. 297–300. doi: 10.1109/ISBI.2017.7950523.

[79] A. Romero Lopez, X. Giro-I-Nieto, J. Burdick, and O. Marques, "Skin lesion classification from dermoscopic images using deep learning techniques," in *Proceedings of the 13th IASTED International Conference on Biomedical Engineering, BioMed 2017*, Apr. 2017, pp. 49–54. doi: 10.2316/P.2017.852-053.

[80] N. Bayramoglu and J. Heikkilä, "Transfer learning for cell nuclei classification in histopathology images," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9915 LNCS, pp. 532–539. doi: 10.1007/978-3-319-49409-8_46.

[81] L. Y. Pratt, J. Mostow, and C. A. Kamm, "Direct transfer of learned information among neural networks," in *AAAI'91 Proceedings of the ninth National conference on Artificial intelligence*, 1991, pp. 584–589.

[82] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.

[83] A. Cheevakasemsook, Y. Chapman, K. Francis, and C. Davies, "The study of nursing documentation complexities," *Int. J. Nurs. Pract.*, vol. 12, no. 6, pp. 366–374, Dec. 2006, doi: 10.1111/j.1440-172X.2006.00596.x.

[84] A. E. Carroll, P. Tarczy-Hornoch, E. O'Reilly, and D. A. Christakis, "Resident documentation discrepancies in a neonatal intensive care unit," *Pediatrics*, vol. 111, no. 5 I, pp. 976–980, May 2003, doi: 10.1542/peds.111.5.976.

[85] L. E. Moody, E. Slocumb, B. Berg, and D. Jackson, "Electronic Health Records Documentation in Nursing Electronic Health Records Documentation in Nursing: Nurses' Perceptions, Attitudes, and Preferences," *Comput. Informatics, Nurs.*, vol. 22, no. 6, pp. 337–344, 2004.

[86] T. F. Kelley, D. H. Brandon, and S. L. Docherty, "Electronic Nursing Documentation as a Strategy to Improve Quality of Patient

Care," *J. Nurs. Scholarsh.*, vol. 43, no. 2, pp. 154–162, Jun. 2011, doi: 10.1111/j.1547-5069.2011.01397.x.

[87] B. Ahirwal, M. Khadtare, and R. Mehta, "FPGA based system for Color Space Transformation RGB to YIQ and YCbCr," in *2007 International Conference on Intelligent and Advanced Systems, ICIAS 2007*, 2007, pp. 1345–1349. doi: 10.1109/ICIAS.2007.4658603.

[88] S. Jayaram, S. Schmugge, M. C. Shin, and L. V. Tsap, "Effect of colorspace transformation, the illuminance component, and color modeling on skin detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004, vol. 2. doi: 10.1109/cvpr.2004.1315248.

[89] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2015.

[90] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, J. Miao, and Q. Ye, "CoCNN: RGB-D deep fusion for stereoscopic salient object detection," *Pattern Recognit.*, vol. 104, p. 107329, Aug. 2020, doi: 10.1016/j.patcog.2020.107329.

[91] Y. Li, J. Zhang, Y. Cheng, K. Huang, and T. Tan, "Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation," in *Proceedings - International Conference on Image Processing, ICIP*, Feb. 2018, vol. 2017-September, pp. 1262–1266. doi: 10.1109/ICIP.2017.8296484.

[92] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognit.*, vol. 86, pp. 376–385, Feb. 2019, doi: 10.1016/j.patcog.2018.08.007.

[93] R. R. Hill, J. Park, and B. F. Pados, "Bottle-Feeding Challenges in Preterm-Born Infants in the First 7 Months of Life," *Glob. Pediatr. Heal.*, vol. 7, p. 2333794X2095268, Jan. 2020, doi: 10.1177/2333794X20952688.

[94] B. F. Pados, J. Park, H. Estrem, and A. Awotwi, "Assessment tools for evaluation of oral feeding in infants younger than 6 months," *Adv. Neonatal Care*, vol. 16, no. 2, pp. 143–150, Apr. 2016, doi: 10.1097/ANC.0000000000000255.

[95] C. Lau, "Development of infant oral feeding skills: what do we know?," *Am. J. Clin. Nutr.*, vol. 103, no. 2, pp. 616S-621S, Feb. 2016, doi: 10.3945/ajcn.115.109603.

[96] M. M. Palmer, K. Crawley, and I. A. Blanco, "Neonatal Oral-Motor Assessment scale: a reliability study.," *J. Perinatol.*, vol. 13, no. 1, pp. 28–35, Jan. 1993.

[97] S. M. Thoyre, C. S. Shaker, and K. F. Pridham, "The early feeding skills assessment for preterm infants.," *Neonatal network : NN*, vol. 24, no. 3. NIH Public Access, pp. 7–16, 2005. doi: 10.1891/0730-0832.24.3.7.

[98] B. F. Pados, H. H. Estrem, S. M. Thoyre, J. Park, and C. McComish, "The Neonatal Eating Assessment Tool: Development and Content Validation," *Neonatal Netw.*, vol. 36, no. 6, pp. 359–367, Nov. 2017, doi: 10.1891/0730-0832.36.6.359.

[99] J. Li, C. Shao, and X. Zhang, "Pedestrian detection & tracking from a moving vehicle," in *Proceedings of the 2009 International Conference on Image Processing, Computer Vision, and Pattern Recognition, IPCV 2009*, 2009, vol. 1, pp. 227–232. doi: 10.1109/ivs.2000.898368.

[100] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral Deep Neural Networks for Pedestrian Detection," *Br. Mach. Vis. Conf. 2016, BMVC 2016*, vol. 2016-September, pp. 73.1-73.13, Nov. 2016, Accessed: Mar. 28, 2021. [Online]. Available: http://arxiv.org/abs/1611.02644

[101] P. A. P. Ferraz, B. A. G. de Oliveira, F. M. F. Ferreira, and C. A. P. da Silva Martins, "Three-stage RGBD architecture for vehicle and pedestrian detection using convolutional neural networks and stereo vision," *IET Intell. Transp. Syst.*, vol. 14, no. 10, pp. 1319–1327, Oct. 2020, doi: 10.1049/iet-its.2019.0367.

[102] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587.

[103] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[104] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015, pp. 91–99.

[105] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 779–788, Jun. 2015, Accessed: Jul. 02, 2021. [Online]. Available: http://arxiv.org/abs/1506.02640

[106] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[107] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv*, Apr. 2018, Accessed: May 24, 2021. [Online]. Available: http://arxiv.org/abs/1804.02767

[108] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 2020, Accessed: Jul. 02, 2021. [Online]. Available: http://arxiv.org/abs/2004.10934

[109] G. Jocher, A. Stoken, A. Chaurasia, J. Borovec, NanoCode012, TaoXie, Y. Kwon, K. Michael, L. Changyu, J. Fang, A. V, Laughing, tkianai, yxNONG, P. Skalski, A. Hogan, J. Nadar, imyhxy, L. Mammana, *et al.*, "ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support." Zenodo, Oct. 2021. doi: 10.5281/zenodo.5563715.

[110] M. Tan, R. Pang, and Q. V Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10781–10790.

[111] M. Tan and Q. V Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv*, pp. 6105–6114, May 2019, Accessed: Mar. 22, 2021. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html

[112] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020, pp. 213–229.

[113] 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," *Int J Comput Vis*, vol. 88, pp. 303–338, 2010, doi: 10.1007/s11263-009-0275-4.

[114] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.

[115] H. A. Hosni Mahmoud and H. A. Mengash, "A novel technique for automated concealed face detection in surveillance videos," *Pers. Ubiquitous Comput.*, vol. 25, no. 1, pp. 129–140, Feb. 2021, doi: 10.1007/s00779-020-01419-x.

[116] T. Zhang, J. Li, W. Jia, J. Sun, and H. Yang, "Fast and robust occluded face detection in ATM surveillance," *Pattern Recognit. Lett.*, vol. 107, pp. 33–40, May 2018, doi: 10.1016/j.patrec.2017.09.011.

[117] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016.

[118] A. Zadeh, T. Baltrušaitis, and L.-P. Morency, "Convolutional Experts Constrained Local Model for Facial Landmark Detection," *arXiv*, 2017.

[119] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *2018 13th IEEE International Conference on Automatic Face \& Gesture Recognition (FG 2018)*, 2018, pp. 98–105.

[120] K. Khan, M. Mauro, and R. Leonardi, "Multi-class semantic segmentation of faces," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 827–831.

[121] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.

[122] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," 2010.

[123] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, 2011, pp. 2144–2151.

[124] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, 2016.

[125] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *arXiv*, Nov. 2015.

[126] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," Jun. 2016.

[127] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015.

[128] 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," *Healthc. Technol. Lett.*, vol. 1, no. 3, pp. 87–91, Sep. 2014, doi: 10.1049/htl.2014.0077.

[129] Shifeng Li, Huchuan Lu, and Xingqing Shao, "Human Body Segmentation via Data-Driven Graph Cut," *IEEE Trans. Cybern.*, vol. 44, no. 11, pp. 2099–2108, Nov. 2014, doi: 10.1109/TCYB.2014.2301193.

[130] C.-H. Chuang, J.-W. Hsieh, C.-C. Lee, Y.-N. Chen, and L.-W. Tsai, "Human Body Part Segmentation of Interacting People by Learning Blob Models," in *2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Jul. 2012, pp. 367–370. doi: 10.1109/IIH-MSP.2012.95.

[131] E. Borenstein and J. Malik, "Shape Guided Object Segmentation," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, 2006, vol. 1, pp. 969–976. doi: 10.1109/CVPR.2006.276.

[132] Zhe Lin and L. S. Davis, "Shape-Based Human Detection and Segmentation via Hierarchical Part-Template Matching," *IEEE Trans.*

*Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 604–618, Apr. 2010, doi: 10.1109/TPAMI.2009.204.

[133] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous Detection and Segmentation," in *Computer Vision – ECCV 2014*, 2014, pp. 297–312.

[134] 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," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[135] M. Everingham, S. M. Ali Eslami, L. Van Gool, C. K. I Williams, J. Winn, A. Zisserman, M. Everingham, S. M. A Eslami, J. Winn, L. K. Van Gool Leuven, B. L. Van Gool ETH, S. C. K I Williams, and A. Zisserman, "The PASCAL Visual Object Classes Challenge: A Retrospective," *Int J Comput Vis*, vol. 111, pp. 98–136, 2015, doi: 10.1007/s11263-014-0733-5.

[136] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv*, Sep. 2014.

[137] L. Zhang and Y. Liang, "Motion Human Detection Based on Background Subtraction," in *2010 Second International Workshop on Education Technology and Computer Science*, 2010, pp. 284–287. doi: 10.1109/ETCS.2010.440.

[138] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001, doi: 10.1109/34.910878.

[139] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1–3, pp. 185–203, Aug. 1981, doi: 10.1016/0004-3702(81)90024-2.

[140] N. J. Bauer and P. N. Pathirana, "Object focused simultaneous estimation of optical flow and state dynamics," in *2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, Dec. 2008, pp. 61–66. doi: 10.1109/ISSNIP.2008.4761963.

[141] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description." pp. 2625–2634, 2015.

[142] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating Videos to Natural Language Using Deep Recurrent Neural Networks," Dec. 2014.

[143] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[144] A. Grushin, D. D. Monner, J. A. Reggia, and A. Mishra, "Robust human action recognition via long short-term memory," 2013. doi: 10.1109/IJCNN.2013.6706797.

[145] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9907 LNCS, pp. 816–833. doi: 10.1007/978-3-319-46487-9_50.

[146] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiol. Meas.*, vol. 35, no. 5, pp. 807–831, May 2014, doi: 10.1088/0967-3334/35/5/807.

[147] L. K. Mestha, S. Kyal, Beilei Xu, L. E. Lewis, and V. Kumar, "Towards continuous monitoring of pulse rate in neonatal intensive care unit with a webcam," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2014, pp. 3817–3820. doi: 10.1109/EMBC.2014.6944455.

[148] G. R. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Proceedings - 4th IEEE Workshop on Applications of Computer Vision, WACV 1998*, 1998, vol. 1998-Octob, pp. 214–219. doi: 10.1109/ACV.1998.732882.

[149] D. Decarlo and D. Metaxas, "Optical flow constraints on deformable models with applications to face tracking," *Int. J. Comput. Vis.*, vol. 38, no. 2, pp. 99–127, Jul. 2000, doi: 10.1023/A:1008122917811.

[150] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," 2008. doi: 10.1109/CVPR.2008.4587572.

[151] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th international joint conference in Artificial Intelligence*, 1981, pp. 674–679.

[152] C. Tomasi and T. Kanade, "Detection and Tracking of Point Features," *Int. J. Comput. Vis.*, vol. 9, pp. 137--154, 1991.

[153] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, 2012, doi: 10.1109/TPAMI.2011.239.

[154] G. Chandan, A. Jain, H. Jain, and Mohana, "Real Time Object Detection and Tracking Using Deep Learning and OpenCV," in *Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA 2018*, Dec. 2018, pp. 1305–1308. doi: 10.1109/ICIRCA.2018.8597266.

[155] S. L. Bennett, R. Goubran, and F. Knoefel, "Adaptive eulerian video magnification methods to extract heart rate from thermal video,"

in *2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, May 2016, pp. 1–5. doi: 10.1109/MeMeA.2016.7533818.

[156] N. Miljkovic and D. Trifunovic, "Pulse rate assessment: Eulerian Video Magnification vs. electrocardiography recordings," in *12th Symposium on Neural Network Applications in Electrical Engineering (NEUREL)*, Nov. 2014, pp. 17–20. doi: 10.1109/NEUREL.2014.7011447.

[157] B. Aubakir, B. Nurimbetov, I. Tursynbek, and H. A. Varol, "Vital sign monitoring utilizing Eulerian video magnification and thermography," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2016, pp. 3527–3530. doi: 10.1109/EMBC.2016.7591489.

[158] N. Bernacchia, L. Scalise, L. Casacanditella, I. Ercoli, P. Marchionni, and E. P. Tomasini, "Non contact measurement of heart and respiration rates based on Kinect," in *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Jun. 2014, pp. 1–5. doi: 10.1109/MeMeA.2014.6860065.

[159] C. Yang, G. Cheung, and V. Stankovic, "Estimating heart rate via depth video motion tracking," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, Jun. 2015, pp. 1–6. doi: 10.1109/ICME.2015.7177517.

[160] X. He, R. Goubran, and F. Knoefel, "IR night vision video-based estimation of heart and respiration rates," in *2017 IEEE Sensors Applications Symposium (SAS)*, 2017, pp. 1–5. doi: 10.1109/SAS.2017.7894087.

[161] R. van Donselaar and W. Chen, "Design of a smart textile mat to study pressure distribution on multiple foam material configurations," in *Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies - ISABEL '11*, Oct. 2011, pp. 1–5.

[162] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting Pulse from Head Motions in Video," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 3430–3437. doi: 10.1109/CVPR.2013.440.

[163] J. Chen, Z. Chang, Q. Qiu, X. Li, G. Sapiro, A. Bronstein, and M. Pietikainen, "RealSense = real heart rate: Illumination invariant heart rate estimation from videos," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Dec. 2016, pp. 1–6. doi: 10.1109/IPTA.2016.7820970.

[164] J. Nikolic-Popovic and R. Goubran, "Impact of motion artifacts on video-based non-intrusive heart rate measurement," in *2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, May 2016, pp. 1–6. doi: 10.1109/MeMeA.2016.7533740.

[165] Z. Liu, J. Wu, L. Fu, Y. Majeed, Y. Feng, R. Li, and Y. Cui, "Improved Kiwifruit Detection Using Pre-Trained VGG16 with RGB and NIR Information Fusion," *IEEE Access*, vol. 8, pp. 2327–2336, 2020, doi: 10.1109/ACCESS.2019.2962513.

[166] J. Jiang, X. Feng, F. Liu, Y. Xu, and H. Huang, "Multi-Spectral RGB-NIR Image Classification Using Double-Channel CNN," *IEEE Access*, vol. 7, pp. 20607–20613, 2019, doi: 10.1109/ACCESS.2019.2896128.

[167] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7576 LNCS, no. PART 5, pp. 746–760. doi: 10.1007/978-3-642-33715-4_54.

[168] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Oct. 2015, vol. 07-12-June-2015, pp. 567–576. doi: 10.1109/CVPR.2015.7298655.

[169] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5828–5839. Accessed: Mar. 28, 2021. [Online]. Available: www.scan-net.org

[170] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "SceneNet RGB-D: 5M Photorealistic Images of Synthetic Indoor Trajectories with Ground Truth," *arXiv*, Dec. 2016, Accessed: Mar. 28, 2021. [Online]. Available: http://arxiv.org/abs/1612.05079

[171] J. Zeng, Y. Tong, Y. Huang, Q. Yan, W. Sun, J. Chen, and Y. Wang, "Deep Surface Normal Estimation with Hierarchical RGB-D Fusion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6153–6162.

[172] Intel, "Intel® RealSense™ Camera SR300 | Intel® Software." https://www.intel.com/ (accessed May 10, 2021).

[173] M. Hozayen, S. Nizami, K. Dick, and J. R. Green, "Patient Monitor Data Import Software," in *National Conference on Undergraduate Research*, 2018, p. in press.

[174] S. Nizami, A. Basharat, A. Shaukat, U. Hameed, S. A. Raza, A. Bekele, R. Giffen, and J. R. Green, "CEA: Clinical Event Annotator mHealth Application for Real-time Patient Monitoring," 2018.

[175] A. Jubran, "Pulse oximetry," *Crit. Care*, vol. 19, no. 1, p. 272, Jul. 2015, doi: 10.1186/s13054-015-0984-8.

[176] W. W. Hay, D. J. Rodden, S. M. Collins, D. L. Melara, K. A. Hale, and L. M. Fashaw, "Reliability of conventional and new pulse oximetry in neonatal patients," *J. Perinatol.*, vol. 22, no. 5, pp. 360–366, Jun. 2002, doi: 10.1038/sj.jp.7210740.

[177] T. E. Bachman, T. E. Bachman, N. P. Iyer, C. J. L. Newth, P. A. Ross, and R. G. Khemani, “Thresholds for oximetry alarms and target range in the NICU: An observational assessment based on likely oxygen tension and maturity,” *BMC Pediatr.*, vol. 20, no. 1, p. 317, Jun. 2020, doi: 10.1186/s12887-020-02225-3.

[178] N. Laroia, D. L. Phelps, and J. Roy, “Double wall versus single wall incubator for reducing heat loss in very low birth weight infants in incubators,” *Cochrane Database of Systematic Reviews*, no. 2. John Wiley and Sons Ltd, 2007. doi: 10.1002/14651858.CD004215.pub2.

[179] R. A. Koestoer, I. Roihan, and A. D. Andrianto, “Product design, prototyping, and testing of twin incubator based on the concept of grashof incubator,” in *AIP Conference Proceedings*, 2019, vol. 2062, p. 020013. doi: 10.1063/1.5086560.

[180] C. Lehmann, J. Brittelli, and M. Gouzman, “WO2014159951A1 - Portable infant incubator - Google Patents,” 2014 Accessed: May 01, 2021. [Online]. Available: https://patents.google.com/patent/WO2014159951A1/en

[181] E. F. Bell and G. R. Rios, “A Double-Walled Incubator Alters the Partition of Body Heat Loss of Premature Infants,” *Pediatr. Res*, vol. 17, pp. 135–140, 1983.

[182] E. N. Hey and L. E. Mount, “Heat Losses from Babies in Incubators,” *Arch. Dis. Childh*, p. 75, 1967, doi: 10.1136/adc.42.221.75.

[183] S. K. Biswas, M. Mahmudul, A. Mia, R. Islam, and S. Sinha, “Design of a Low cost Non Electrical Type Baby Incubator for Developing Country,” *Int. J. Sci. Eng. Res.*, vol. 7, no. 11, 2016, Accessed: May 01, 2021. [Online]. Available: http://www.ijser.org

[184] P. T. Kapen, Y. Mohamadou, F. Momo, D. K. Jauspin, and G. Anero, “An energy efficient neonatal incubator: mathematical modeling and prototyping,” *Health Technol. (Berl).*, vol. 9, no. 1, pp. 57–63, Jan. 2019, doi: 10.1007/s12553-018-0253-3.

[185] Altuglas International, “OPTICAL & TRANSMISSION CHARACTERISTICS Acrylic Sheet 2”.

[186] N. Otsu, “A Threshold Selection Method from Gray-Level Histograms,” *IEEE Trans Syst Man Cybern*, vol. SMC-9, no. 1, pp. 62–66, 1979, doi: 10.1109/tsmc.1979.4310076.

[187] “Search: phototherapy | Flickr.” https://www.flickr.com/ (accessed Mar. 21, 2021).

[188] L. Perez and J. Wang, “The Effectiveness of Data Augmentation in Image Classification using Deep Learning,” *arXiv*, Dec. 2017.

[189] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning requires rethinking generalization,” *arXiv*, Nov. 2016.

[190] I. Analyst, “How to do skin detection of an image?,” 2017. https://www.mathworks.com/matlabcentral/answers/360260-how-to-do-skin-detection-of-an-image (accessed Mar. 21, 2021).

[191] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv*, Oct. 2018.

[192] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit,” in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Jun. 2015, pp. 55–60. doi: 10.3115/v1/p14-5010.

[193] K. Papineni, S. Roukos, T. Ward, W. Zhu, and Y. Heights, “IBM Research Report Bleu : a Method for Automatic Evaluation of Machine Translation,” *Science (80-. ).*, vol. 22176, pp. 1–10, 2001, doi: 10.3115/1073083.1073135.

[194] “Search: bottle feeding baby | Flickr.” https://www.flickr.com/ (accessed Mar. 21, 2021).

[195] G. A. Miller, “WordNet: A Lexical Database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.

[196] X. Guo, W. Chen, and J. Yin, “A simple approach for unsupervised domain adaptation,” in *Proceedings - International Conference on Pattern Recognition*, Jan. 2016, vol. 0, pp. 1566–1570. doi: 10.1109/ICPR.2016.7899860.

[197] H. Wu, Y. Yan, M. K. Ng, and Q. Wu, “Domain-attention Conditional Wasserstein Distance for Multi-source Domain Adaptation,” *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 4, pp. 1–19, Jul. 2020, doi: 10.1145/3391229.

[198] S. Brahnam, L. Nanni, and R. Sexton, “Introduction to neonatal facial pain detection using common and advanced face classification techniques,” in *Advanced Computational Intelligence Paradigms in Healthcare--1*, Springer, 2007, pp. 225–253.

[199] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack, “SVM classification of neonatal facial images of pain,” in *International Workshop on Fuzzy Logic and Applications*, 2005, pp. 121–128.

[200] Z. Tang, X. Liu, H. Chen, J. Hupy, and B. Yang, “Deep learning based wildfire event object detection from 4K aerial images acquired by UAS,” *AI*, vol. 1, no. 2, pp. 166–179, 2020.

[201] X. Zhang, D. Zhang, A. Leye, A. Scott, L. Visser, Z. Ge, and P. Bonnington, “Autonomous Incident Detection on Spectrometers Using Deep Convolutional Models,” *Sensors*, vol. 22, no. 1, p. 160, 2022.

[202] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Oct. 2015, vol. 07-12-June, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

[203] C. Harris, C. Harris, and M. Stephens, "A combined corner and edge detector," *PROC. FOURTH ALVEY Vis. Conf.*, pp. 147--151, 1988.

[204] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001, vol. 1, pp. I-511-I–518. doi: 10.1109/CVPR.2001.990517.

[205] J. Shi and C. Tomasi, "Good Features to Track," in *Proceedings of 1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, 1994, pp. 593--600.

[206] A. Procházka, M. Schätz, O. Vyšata, and M. Vališ, "Microsoft Kinect Visual and Depth Sensors for Breathing and Heart Rate Analysis," *Sensors*, vol. 16, no. 12, p. 996, Jun. 2016, doi: 10.3390/s16070996.

[207] A. Bekele, S. Nizami, Y. S. Dosso, C. Aubertin, K. Greenwood, J. Harrold, and J. R. Green, "Real-time Neonatal Respiratory Rate Estimation using a Pressure-Sensitive Mat," *MeMeA 2018 - 2018 IEEE Int. Symp. Med. Meas. Appl. Proc.*, pp. 3431–3440, Aug. 2018, doi: 10.1109/MEMEA.2018.8438682.

[208] S. Aziz, Y. S. Dosso, S. Nizami, K. Greenwood, J. Harrold, and J. R. Green, "Detection of Neonatal Patient Motion Using a Pressure-Sensitive Mat," *IEEE Med. Meas. Appl. MeMeA 2020 - Conf. Proc.*, Jun. 2020, doi: 10.1109/MEMEA49120.2020.9137147.

[209] D. G. Kyrollos, R. Hassan, Y. S. Dosso, and J. R. Green, "Fusing Pressure-Sensitive Mat Data with Video through Multi-Modal Registration," *Conf. Rec. - IEEE Instrum. Meas. Technol. Conf.*, vol. 2021-May, May 2021, doi: 10.1109/I2MTC50364.2021.9459886.

[210] M. Tölgyessy, M. Dekan, and L. Chovanec, "Skeleton Tracking Accuracy and Precision Evaluation of Kinect V1, Kinect V2, and the Azure Kinect," *Appl. Sci. 2021, Vol. 11, Page 5756*, vol. 11, no. 12, p. 5756, Jun. 2021, doi: 10.3390/APP11125756.

[211] C. Liu, A. Torralba, W. T. Freeman, F. Durand, E. H. Adelson, C. Liu, A. Torralba, W. T. Freeman, F. Durand, and E. H. Adelson, "Motion magnification," in *ACM SIGGRAPH 2005 Papers on - SIGGRAPH '05*, 2005, vol. 24, no. 3, p. 519. doi: 10.1145/1186822.1073223.

[212] A. Al-Naji and J. Chahl, "Remote respiratory monitoring system based on developing motion magnification technique," *Biomed. Signal Process. Control*, vol. 29, pp. 1–10, Aug. 2016, doi: 10.1016/J.BSPC.2016.05.002.

[213] I. Yun, J. Jeung, M. Kim, Y. S. Kim, and Y. Chung, "Ultra-Low Power Wearable Infant Sleep Position Sensor," *Sensors 2020, Vol. 20, Page 61*, vol. 20, no. 1, p. 61, Dec. 2019, doi: 10.3390/S20010061.

[214] A. Hirata, N. Ito, and O. Fujiwara, "Effect of electromagnetic polarization on whole-body average SAR in infant model for far-field exposures," *Proc. 20th Int. Zurich Symp. Electromagn. Compat. EMC Zurich 2009*, pp. 325–328, 2009, doi: 10.1109/EMCZUR.2009.4783456.

[215] J. Valero De Bernabé, T. Soriano, R. Albaladejo, M. Juarranz, M. E. Calle, D. Martínez, and V. Domínguez-Rojas, "Risk factors for low birth weight: a review," *Eur. J. Obstet. Gynecol. Reprod. Biol.*, vol. 116, no. 1, pp. 3–15, Sep. 2004, doi: 10.1016/J.EJOGRB.2004.03.007.

[216] C. Eastwood-Sutherland, T. J. Gale, P. A. Dargaville, and K. Wheeler, "Non-contact respiratory monitoring in neonates," in *The 7th 2014 Biomedical Engineering International Conference*, Nov. 2014, pp. 1–5. doi: 10.1109/BMEiCON.2014.7017373.

[217] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005, vol. I, pp. 886–893. doi: 10.1109/CVPR.2005.177.

[218] T. Ojala, M. Pietikäinen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Proceedings - International Conference on Pattern Recognition*, 1994, vol. 3, pp. 582–585. doi: 10.1109/ICPR.1994.576366.

# Appendices

In addition to the preliminary experiments presented in Chapter 3, further work was done to ensure that the Intel RealSense SR300 camera is safe to use in the NICU. We must ensure that the maximum amount of power emitted by the device (Appendix A), does not affect the newborns health and growth. More specifically, the maximum power from the infrared projector potentially reaching the eye of the newborn is calculated (Appendix B). Appendix C presents preliminary work in face detection using the cascade model and deep learning Faster R-CNN, two popular models. Appendix D demonstrates a face orientation estimation method implemented using YOLO5Face upon thorough experiments presented in Chapter 5.1, and applied in Chapter 5.2 to obtain a patient segmentation dataset. Finally, Appendix E describes the further details for dataset description or obtained results presented in this thesis.

## Appendix A - Maximum power emitted by the device

These calculations present the maximum power emitted by the Intel RealSense SR300 [172] to ensure that this measure does not exceed the maximum power that can reach the newborn without causing arm to the patient's eyes, growth, and overall health.

$$s = r\theta = r_{max}\,\theta_{WorstCaseAngle} = 0.352cm \times 100mrad = 0.00352cm \times 0.1rad = 3.52x10^{-4}$$

$$P_{pupil} = I_E \times s = \frac{123.12mW}{sr} \times 3.52x10^{-4} = 0.04334mW$$

$$FOP = 64° \times 75° = 4800°^2 = 1.462sr \text{ where } 1sr = \left(\frac{180}{\pi}\right)^2 °^2$$

$$I_E = \frac{180mW}{1.462sr} = \frac{123.12mW}{sr}$$

$$P_{pupilPerLED} = \frac{P_{SupplyLimitCurrent}}{LEDs} \times P_{pupil} = \frac{0.6A}{3} \times 0.04334mW = 0.008668mW$$

$$P_{total} = P_{pupilPerLED} \times LEDs = 0.008668mW \times 3LEDs = 0.026004mW$$

s = # of steradians that the incident light exposes on the eye

r = arc radius (maximum dilated radius for an average eye = 0.352cm)

θ = arc angle (worst case angle as defined in ANSI Acceptable Exposure Limits = 100mrad)

$P_{pupil}$ = Worst case continuous incident power appearing on the user's eyes

$I_E$ = Maximum radiant intensity

FOP = Field of projection of the device (60° ± 4° vertically, 72.5° ± 2° horizontally)

Based on the calculations above, the Intel camera emits 0.03367mW in its infrared projector's entire field of projection of maximum 64° vertically by maximum 75° horizontally, and has 3 major light-emitting diodes (LEDs) consisting of

- 1 Power LED Indication: small light next to the camera light which turns on when the camera is active)
- 1 Infrared Projector: Class 1 laser compliant coded light infrared projector system
- 1 Activity LED: Green LED, illuminates when transmitting video over USB3

Considering the IR LED is significantly more powerful than the other two light sources, we estimated that a total of 180mW is generated by the device (two other LEDs negligible).

The input current of the Intel RealSense SR300 is 0.6A, r_max=352cm is the max dilated radius of an average sized eye, and Θ_worstCaseAngle = 100mrad by the American National Standards Institute, Acceptable Emissions Limits (ANSI, AEL).

To now ensure the safety of the neonates when capturing videos with the Intel camera, we estimated the amount of infrared radiations to reach the eye. The closest distance in which videos would be captured takes place with the camera mounted on the top of the insulator, above the glass covering the patient. We set this circumstance as a worst case scenario example, all other setting (i.e., Overhead warmer and crib) would have fewer emissions. The following sketch in Figure A.1 represents the incubator bed.

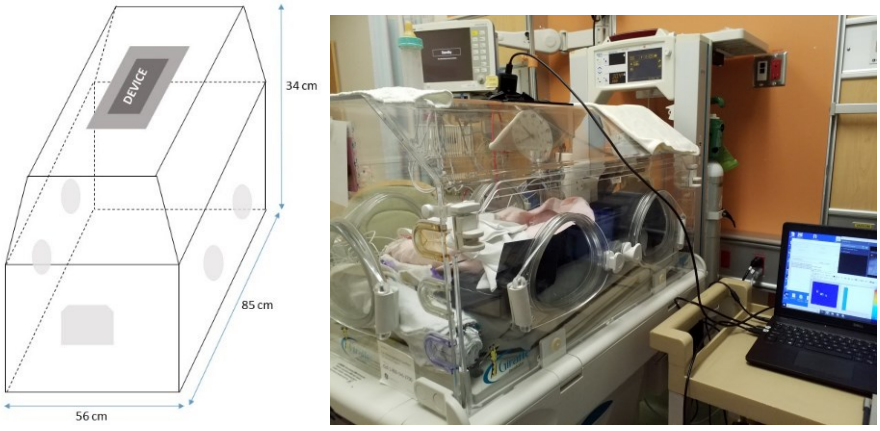



Figure A.1: Sketch of top part of incubator with a picture of a real incubator for reference.

Note: Given a standard head size for neonates as about 15 cm, the neonate placed inside the incubator is about 21 cm from the top of the incubator.

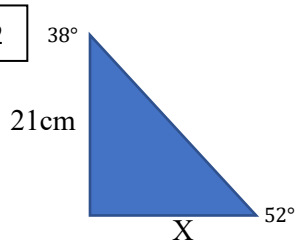
## Appendix B - Maximum power reaching the eyes

The International Commission on Non-Ionizing Radiation Protection (ICNIRP), has set a limit of 2 W/kg on the specific absorption rate (SAR), the rate at which the body absorbs radio frequency (RF) energy. They also evaluated the whole-body average SAR (WBSAR) and restricted limits of 0.4 W/kg for occupational exposure or 0.08 W/kg for public exposure. A few studies conducted on infants have reviewed these values for implementation of a wearable sensor for sleep monitoring [213] or to evaluate the SAR limit for infants of varying ages (9 month or 3 years old) [214]. For this thesis, the lowest possible limit of 0.08 W/kg is selected and calculated with 500 g as the lowest potential possible newborn weight as reported by Valero de Bernabe *et al.* [215] in the "extremely low birth weight" category.

The maximum power allowed to be emitted on neonates is $0.08\frac{W}{kg} \times 0.5\ kg = \ 0.004\ W\ (or\ 4\ mW)$. The maximum power reaching their eyes must therefore stand far below this amount.

1) "Horizontal View" from the top of the incubator
(34cm high from the bed, device's horizontal angle = 75°)

H/2

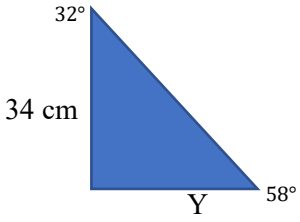



$$\frac{34cm}{sin52°} = \frac{x}{sin38°} \rightarrow x = \frac{34cm \times sin38°}{sin52°} = 26.22cm$$

$$Width\ view = 2x = 2 \times 26.22cm = 52.44cm$$

180-90-H/2

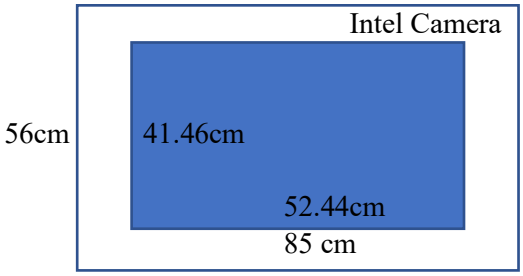
2) "Lateral View" from the top of the incubator
(34cm high from the bed, device's horizontal angle = 64°)




$$\frac{34cm}{sin58°} = \frac{y}{sin32°} \rightarrow y = \frac{34cm \times sin32°}{sin58°} = 20.73cm$$

$$Length\ view = 2y = 2 \times 20.73cm = 41.46cm$$

3) Total Field of View Incubator




$$Total\ View = 52.44cm \times 41.46cm$$
$$Incubator\ Bed = 85cm \times 56cm$$

Intel RealSense SR300's viewing surface area = $52.44cm \times 41.46cm = 2174.1624cm^2$

Eye's surface area = $\pi \times r_{max}^2 = \pi \times 0.352cm^2 = 0.3893cm^2$
(Based on the max dilated radius of an average sized eye)
Maximum exposure = $2 \times Eye's\ surface\ area = 2 \times 0.3893cm^2 = 0.7786cm^2$

% Exposure = $\frac{0.7786cm^2}{2174.1624cm^2} = 0.0003581$ or 0.03581%

Maximum power reaching the eye = %Exposure $\times P_{total}$
= 0.03581% x $0.026004mW$
= $9.312\ x\ 10^{-6}mW$

From the above calculations, the maximum exposure or infrared light emitted from the Intel RealSense SR300 pertains to 0.03581% of its entire field of view. Furthermore, the maximum power reaching the eyes of the neonate amounts to $9.312\ x\ 10^{-6}\ mW$ which is far below the maximum power allowed to be emitted on neonates (i.e., $4\ mW$). These calculations show evidence that the Intel device is safe to use for our experiment on neonates. Few studies have also used the Intel RealSense SR300 for clinical purposes in the NICU as a monitoring tool, mainly to measure respiratory patterns [216].

## Appendix C – Preliminary Work on Neonatal Face Detection

A popular face detection technique was introduced by Viola and Jones in 2001 where they used Haar-like features and a cascade classifier, thereby often subsequently referred to as the Haar-Cascade model [204]. Haar features were proposed by Alfred Haar in 1909, where he utilized different convolutional kernels to represent edge and line features of different size and change in gradient. For example, the difference in intensities from the eyes and the bridge between them can be represented by a three-rectangle line feature with a lighter rectangle in the middle. Viola and Jones thereby exploited this utility by creating integral images to evaluate Haar features more rapidly and efficiently. Integral images are obtained by incrementally implementing the sum of areas across an image. This computer vision approach was successful for detecting the face of upright standing adults from frontal view. However, other viewpoints, minor occlusions, and lighting variations can severely impact this face detection algorithm. More complex techniques, including deep neural networks, have recently been implemented to overcome this issue.

Some of the most famous face detection algorithms such as OpenCV's Haar-Cascade introduced by Viola and Jones [204], using Haar features and cascade classifiers, failed to detect the face of any patients from our dataset. The training dataset comprising primarily of adult population would strongly explain those results. This algorithm has been developed using frontal or profile views of upright standing adults, which differ significantly from our images of newborns exhibiting a range of poses (e.g., supine, prone, or on their side). A method suitable for detecting faces of newborns is warranted. Specifically, in clinical setting, multiple challenges often occur and must be considered to consistently detect the patient's face for non-contact continuous monitoring. This section presents preliminary results obtained at the beginning stages of data collection.

The following sections present the implementation of a computer vision Cascade model and a deep learning Faster R-CNN model. Both methods are evaluated using AP50 metric (as discussed in Chapter 5.1.3).

### Neonatal Face Detection using Cascade Model

From a computer vision approach, we can use a cascade model with a set of features to identify those corresponding to eyes, a nose, and a mouth. Cascade classifiers are composed of multiple stages at which each of them comprises of an ensemble of weak learners. A sliding window is projected across the image with a set of features, and a prediction is acquired at each window until the entire image is processed. In a subsequent stage, another sliding window is used, and new predictions are made based on the previous stage by boosting procedure. If the object is detected within the window, it is reported as a positive, and as a negative otherwise. At each stage, negative samples are rejected from the dataset as fast as possible

thereby decreasing the false negative rate and increasing the true positive rate. This also makes the detection task harder for subsequent stages to focus on weaker learners. Eventually, after a series of cascading stages, the number of sub-windows will be significantly reduced, with remaining regions encompassing the object of interest.

Three different feature types are explored here to implement the cascade network. First, the traditional Haar features utilizing different convolutional kernels to obtain edges and line features, as presented in Chapter 2.4.2 [204].

Second, Histogram of Oriented Gradients (HOG) features comprise of the distribution of the gradient direction to obtain the most useful information residing at edge points where the gradient is high [217]. To this end, HOG features represent an image by converting it to a feature vector containing all important texture information from that image, thereby informing the delineation of objects in the scene.

Lastly, Local Binary Patterns (LBP) can use circular-like pixel neighborhood to extract texture information in the image [218]. A 3x3 pixel neighbourhood is thresholded using the center value as the thresholding point. A new binary value is set as 1 if it is greater than the threshold, and 0 otherwise. All binary values are concatenated into an 8-digit binary number, converted to a decimal value, and set as a new pixel intensity for the central pixel.

Each of these three feature descriptors are implemented individually in the cascade model for comparison. All cascade models are run for 20 stages with 0.5 false positive rate.

**Neonatal Face Detection using Faster R-CNN model**

In comparison with the cascade model requiring explicit feature extraction, a deep learning model can learn such features automatically during training. We explore the state-of-the-art Faster R-CNN network given the remarkable results in the object detection task [104].

Each input image is fed through a CNN (selected VGG-16) consisting of a sequence of convolutional and max pooling layers to obtain a pertinent object features represented as a feature map. Simultaneously, this feature map is used as input for the RPN, where a 3x3 sliding window scans the image using anchor boxes at each window. These anchors are represented by boxes of different sizes (area of 128x128, 256x256, and 512x512) and different aspect ratios (1:1,1:2, and 2:1), resulting in a total of nine anchors per window. One additional convolutional layer forms the box-location layer, which encodes the proposed location of the box as represented by a four-tuple (r,c,h,w), where (r,c) indicates the top-left corner and (h,w) forms the height and width of the bounding box for the proposed region. Two more fully-connected layers are added; a box-classification layer performing a scored binary classification for the proposed box either belonging to the object class or the background, and a box-regression layer, providing the regressed bounds for the previously proposed box locations. All scores

are then ranked, and the N top-ranked scores are selected as region proposals. These regions are later pooled with extracted features, thus creating a ROI pooling layer. The network then utilizes a sequence of fully connected layers to process the ROI features, before branching out into two output layers. One softmax layer is utilized for predicting the class of the object. One linear layer predicts the location of the bounding box pertaining to the object, and is represented as four-tuple (R,C,H,W), where (R,C) indicates the top-left corner and (H,W) forms the height and width of the bounding box for the object of interest.

The Faster R-CNN model can be considered as the combination of a Fast R-CNN and RPN network working together to detect region proposals and classifying them. The Faster R-CNN model is trained over the following four different stages:

1. Training a Region Proposal Network (RPN)
2. Training a Fast R-CNN Network using the RPN from step 1
3. Re-training RPN using weight sharing with Fast R-CNN
4. Re-training Fast R-CNN using updated RPN

In sequence, the Fast R-CNN and RPN network can be trained over these four simple steps, only final results are reported. The model was trained over 10 epochs using an initial learning rate of 1e-6 and stochastic gradient descent with momentum (SGDM).

**Neonatal Face Detection Dataset**

From video recordings, 1 image was extracted per 30 second of video. A total of 1338 images were used from 9 patients. Images were selected as "positive" if the face of the was clearly visible, under various lighting conditions and with possible minor facial occlusions from beddings or hospital equipment (e.g., nasogastric tubing tape). If the patient's face is not visible (e.g., major occlusion, extremely low or no lighting, or patient facing backwards) or the patient is absent from the bed, they were selected as "negative" images. To prevent from overfitting, redundant images were excluded, resulting in 1104 positive imaged and 234 negative images. All positive images were standardized to have the face of the patient oriented North, and the "face" object was annotated by drawing a bounding box over the visible facial areas.

Some pre-processing image augmentation techniques were applied using a vertical mirroring on positive images to capture both profiles when the patient is lying on their side. For negative images, vertical mirroring, horizontal mirroring, and 180 degrees rotation were implemented.

**Face Detection Results**

Table A.1: Neonatal Face Detection Results

| Model | AP50 |
|---|---|
| Haar-Cascade | 21.88 % |
| HOG-Cascade | 38.05 % |
| LBP-Cascade | 44.73 % |
| Faster R-CNN | 61.48 % |

Among the results from the cascade models, the LBP-Cascade worked best, as demonstrated by the greater mAP performance in Table A.1. All presented features (Haar, HOG, and LBP) can be error-prone due to vision-based challenges during lighting variations or facial occlusions. Since these features rely heavily on the pixel intensities in the image and visual appearance of the object, variations occurring from NG tube tapings in some patients could severely impact the model's predictions. Among all, LBP features are more robust to these challenges by considering a circular-like rather than a linear-like neighborhood (Haar), and less impacted by higher gradient differences due to minor occlusions (HOG).

In contrast, the deep learning Faster R-CNN outperformed the cascade models with an AP50 difference of 16.75 % compared to the best performing LBP-Cascade model. The deep learning model learns features from the feature extraction step in the CNN block using VGG-16, thereby being even more robust to illuminations changes and minor occlusions.

## Appendix D – Neonatal Face Orientation Estimation

### Face Orientation Estimation - Method

In many neonatal monitoring applications, a preprocessing step is required where images are rotated to standardize the orientation of the face. Having the patient's face oriented North facilitates the face detection and alignment task, and this rotation step is often performed manually (laborious) or programmatically through trial and error by rotating the image at 90º increments until a face is detected (unreliable if a face is detected at non-North direction without providing confidence from this orientation). We therefore propose a face orientation estimation approach where the image is rotated in four 90º increments. The "North" face orientation is predicted as the direction that produces the most confident bounding box, the most coherent facial landmark positions, or both:

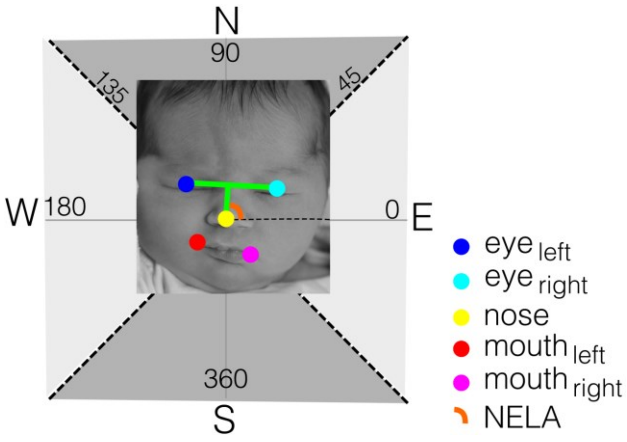



Figure A.2: Face Orientation Estimation based on Facial Landmark Position.

***Face Orientation Estimation based on face box confidence score*:** From the detected bounding boxes in all four directions, we predict that the North-facing orientation should have the highest confidence score.

***Face Orientation Estimation based on facial landmark position:*** From the detected 5-point facial landmarks in all four directions, we predict that the North-facing orientation should have landmarks positioned in manner where the nose is below the eyes. We measure the Nose-to-Eye-Line Angle (NELA) which measures the angle of a line that originates at the nose landmark and intersects the inter-ocular line at 90º. A North-facing orientation would result in a NELA of 90º ($\pm$ 45º to account for minor pose variations). This method is illustrated in Figure A.2.

***Face Orientation Estimation based on complete facial detection*:** Given the strength and weaknesses of each technique presented above, a final and more comprehensive face orientation estimation approach is presented by leveraging both the face box confidence scores and the facial landmark positions. The

North orientation is determined by selecting the detection with the highest confidence score that also has a valid NELA.

**Face Orientation Estimation - Results**

Since the face detection algorithms performed best on the COPE dataset, we use it with one of the best performing pretrained models, YOLO5Face, to evaluate our face orientation estimation approach. The COPE dataset and its annotations are artificially rotated at 90º, 180º, and 270º to create sets of images with the face oriented North, West, South, and East. As observed in Table A.2 and Figure A.3 the face orientation estimation approach based on the confidence scores alone predicts "North" as the North-facing face orientation 80.88% of the time, and "West" or "East" otherwise. It never predicts "South". The face orientation estimation based solely on landmark positions performed less accurately (50.97% precision for "North") since the South orientation is heavily misclassified as North. Compared to the confidence score based approach, this NELA method can detect other orientations (West, South, East) based on the NELA (at 180º, 270º, 360º/0º, respectively). However, if no face is detected, it cannot make a prediction (predicts *none*), while the confidence score approach is unaffected by this limitation.

Table A.2: Face Orientation Estimation (COPE + YOLO5Face) using face detection confidence score (conf), NELA, or both

| Orientation | AP30 | Conf | NELA | Conf+NELA |
|---|---|---|---|---|
| North | **100** | **80.88** | 50.97 | **99.45** |
| West | 96.09 | --- | 87.43 | --- |
| South | 93.08 | --- | 72.86 | --- |
| East | 92.67 | --- | **91.20** | --- |

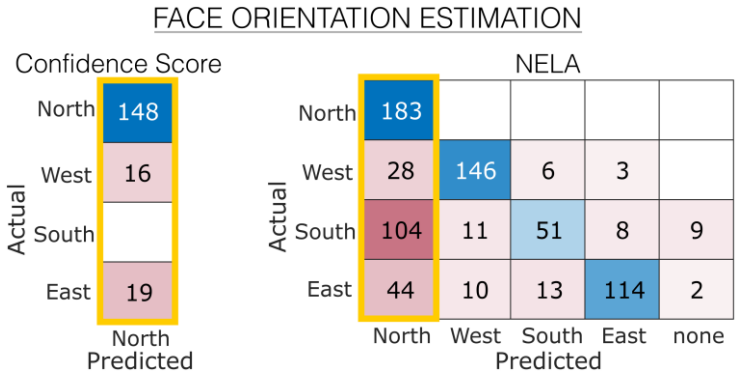



Figure A.3: Face Orientation Predictions from YOLO5Face Confidence Scores and NELA with COPE dataset

The fused approach leverages strengths from both face confidence scores and NELA, and outperforms the individual approaches with 99.45% accuracy.

In ideal cases depicted in Figure A.4-A, the score-based and NELA-based methods are both effective. With more variations in facial expression, the NELA-based technique can be affected by the localization

\of each landmark, as demonstrated in A.4-B. In many cases, it forces the position of the landmarks to face North, no matter the actual orientation. In other cases, it predicts “West” or “East” appropriately, but the “South” orientation is often predicted to be “North”. With patient occlusions, the score-based technique can be affected by a reduced face detected confidence, as demonstrated by Figure A.4-C. The combination of both confidence and NELA led to the best performance.

It is important to note that even if the detector finds a face in all orientations, facial alignment might be unreliable, and therefore not suitable for facial expression analyses. The top panel of Figure A.4-D demonstrates this, where the highest confidence score is properly detected from the North orientation compared to South. As for the landmarks, the North orientation produces a correct NELA at 87º, while the South produces an incorrect NELA at 78º given that the location of eyes- and mouth-landmarks are incorrect. Our proposed face orientation estimation approach is simple but reliable when assuming that only one face is present in the image. Additional faces might affect the desired patient’s detection confidence score, as seen in the bottom panel of Figure A.4-D, where the detector found the face of the clinical staff from their ID. Despite this interesting detection, the patient’s North-facing orientation still produced the highest scoring confidence and our proposed face orientation estimation method remains robust to the incorrect ID detection.

Standardizing the face orientation is an important preprocessing step in neonatal monitoring applications since it is often performed manually or as a trial-and-error approach. It is important to note that the South-facing orientation might be irrelevant for most adult-based detectors leveraging mostly upright standing or lying adults; however, in neonatal monitoring this direction is important since the patient could be repositioned in the bed, especially in cribs. The face orientation estimation model proposed here could therefore be of value to these studies.

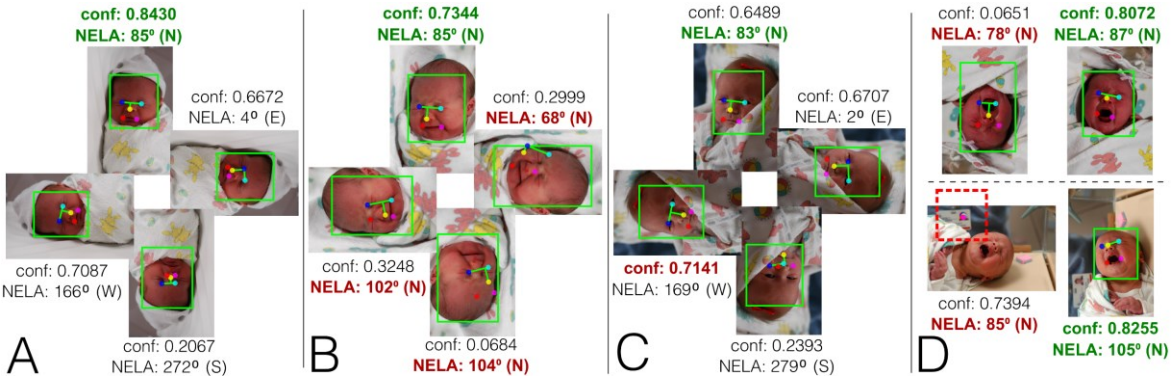



Figure A.4: Sample face orientation predictions with face detection confidence score (conf) and NELA.

## Appendix E – Detailed dataset descriptions used in this thesis

### E.1: Scene Analysis & Bottle-Feeding images per context across all patients

Table A.3: Dataset breakdown from contexts utilized for Scene Analysis and Bottle-Feeding Intervention Detection

| PT | NUM IMGS | CONTEXTS | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Lighting | | Phototherapy ("High") | | Intervention | | Feeding ("Nurse") | | Occupancy | | Coverage ("Present") | |
| | | *Low* | *High* | *Ther* | *Nat* | *Nurse* | *Noone* | *Bottle* | *NoFeed* | *Abs* | *Pres* | *Uncov* | *Cov* |
| 1 | 624 | 352 | 272 | 0 | 272 | 67 | 557 | 0 | 67 | 0 | 624 | 59 | 565 |
| 2 | 628 | 0 | 628 | 0 | 628 | 47 | 581 | 0 | 47 | 55 | 573 | 21 | 552 |
| 5 | 125 | 1 | 124 | 0 | 124 | 2 | 123 | 1 | 1 | 0 | 125 | 1 | 124 |
| 6 | 533 | 167 | 366 | 0 | 366 | 41 | 492 | 0 | 41 | 119 | 414 | 36 | 378 |
| 8 | 720 | 0 | 720 | 0 | 720 | 36 | 684 | 6 | 30 | 142 | 578 | 13 | 565 |
| 9 | 539 | 0 | 539 | 0 | 539 | 66 | 473 | 0 | 66 | 0 | 539 | 28 | 511 |
| 10 | 508 | 291 | 217 | 0 | 217 | 40 | 468 | 0 | 40 | 0 | 508 | 16 | 492 |
| 11 | 535 | 0 | 535 | 0 | 535 | 27 | 508 | 0 | 27 | 95 | 440 | 22 | 418 |
| 13 | 585 | 163 | 422 | 0 | 422 | 114 | 471 | 0 | 114 | 2 | 583 | 210 | 373 |
| 14 | 705 | 0 | 705 | 0 | 705 | 69 | 636 | 15 | 54 | 18 | 687 | 220 | 467 |
| 15 | 650 | 0 | 650 | 0 | 650 | 21 | 629 | 0 | 21 | 71 | 579 | 24 | 555 |
| 16 | 615 | 0 | 615 | 0 | 615 | 111 | 504 | 0 | 111 | 1 | 614 | 46 | 568 |
| 17 | 549 | 80 | 469 | 0 | 469 | 43 | 506 | 16 | 27 | 24 | 525 | 29 | 496 |
| 18 | 638 | 380 | 258 | 0 | 258 | 39 | 599 | 0 | 39 | 122 | 516 | 88 | 428 |
| 19 | 424 | 0 | 424 | 424 | 0 | 29 | 395 | 0 | 29 | 0 | 424 | 424 | 0 |
| 21 | 507 | 2 | 505 | 0 | 505 | 11 | 496 | 0 | 11 | 1 | 506 | 12 | 494 |
| 22 | 508 | 0 | 508 | 0 | 508 | 116 | 392 | 0 | 116 | 1 | 507 | 20 | 487 |
| 23 | 508 | 3 | 505 | 0 | 505 | 32 | 476 | 0 | 32 | 0 | 508 | 14 | 494 |
| 24 | 652 | 53 | 599 | 0 | 599 | 16 | 636 | 0 | 16 | 0 | 652 | 11 | 641 |
| 25 | 662 | 50 | 612 | 0 | 612 | 20 | 642 | 0 | 20 | 1 | 661 | 428 | 233 |
| 26 | 455 | 1 | 454 | 0 | 454 | 59 | 396 | 18 | 41 | 0 | 455 | 14 | 441 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 27 | 501 | 0 | 501 | 0 | 501 | 37 | 464 | 0 | 37 | 77 | 424 | 34 | 390 |
| 28 | 691 | 1 | 690 | 0 | 690 | 41 | 650 | 0 | 41 | 85 | 606 | 42 | 564 |
| 29 | 542 | 16 | 526 | 0 | 526 | 76 | 466 | 17 | 59 | 61 | 481 | 28 | 453 |
| 30 | 474 | 0 | 474 | 0 | 474 | 48 | 426 | 0 | 48 | 67 | 407 | 21 | 386 |
| 31 | 503 | 19 | 484 | 0 | 484 | 75 | 428 | 0 | 75 | 0 | 503 | 65 | 438 |
| 32 | 448 | 0 | 448 | 0 | 448 | 17 | 431 | 0 | 17 | 0 | 448 | 5 | 443 |
| 33 | 619 | 201 | 418 | 0 | 418 | 27 | 592 | 0 | 27 | 162 | 457 | 27 | 430 |
| 34 | 506 | 1 | 505 | 0 | 505 | 1 | 505 | 0 | 1 | 0 | 506 | 506 | 0 |
| **Total** | **15954** | **1781** | **14173** | **424** | **13749** | **1328** | **14626** | **73** | **1255** | **1104** | **14850** | **2464** | **12386** |

| | |
|---|---|
| **Avg/pt** | **550** |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Total Num Pts** | **17** | **29** | **1** | **28** | **29** | **29** | **5** | **29** | **18** | **29** | **29** | **27** |

### E.2: Scene analysis train/test cross-validation folds with results

The following patients were used for 5-fold cross-validation of the contexts “Intervention”, “Occupancy”, and “Coverage”.

Fold 1: Patients 1, 2, 5,6,8

Fold 2: Patients 9, 10, 11, 13, 14

Fold 3: Patients 15, 16, 17, 21, 22

Fold 4: Patients 23, 24, 25, 26, 27, 28

Fold 5: Patients 29, 30, 31, 32, 33, 34

**Table A.4: Intervention Context – Performance per cross-validation fold**

| Rep1 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 123 | 7 | 70 | 2430 | 81 | 3 | 112 | 2434 | 131 | 49 | 62 | 2388 | 122 | 8 | 71 | 2429 |
| 2 | 279 | 133 | 37 | 2423 | 260 | 576 | 56 | 1980 | 279 | 472 | 37 | 2084 | 282 | 132 | 34 | 2424 |
| 3 | 238 | 9 | 64 | 2518 | 228 | 318 | 74 | 2209 | 239 | 115 | 63 | 2412 | 255 | 22 | 47 | 2505 |
| 4 | 183 | 145 | 22 | 3119 | 148 | 544 | 57 | 2720 | 189 | 111 | 16 | 3153 | 183 | 78 | 22 | 3186 |
| 5 | 227 | 318 | 17 | 2530 | 118 | 18 | 126 | 2830 | 221 | 82 | 23 | 2766 | 230 | 262 | 14 | 2586 |
| **TOT** | **1050** | **612** | **210** | **13020** | **835** | **1459** | **425** | **12173** | **1059** | **829** | **201** | **12803** | **1072** | **502** | **188** | **13130** |

| Rep2 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 124 | 7 | 69 | 2430 | 82 | 3 | 111 | 2434 | 130 | 53 | 63 | 2384 | 122 | 8 | 71 | 2429 |
| 2 | 279 | 132 | 37 | 2424 | 258 | 585 | 58 | 1971 | 280 | 478 | 36 | 2078 | 281 | 128 | 35 | 2428 |
| 3 | 238 | 10 | 64 | 2517 | 228 | 320 | 74 | 2207 | 239 | 115 | 63 | 2412 | 281 | 128 | 35 | 2428 |
| 4 | 182 | 142 | 23 | 3122 | 149 | 552 | 56 | 2712 | 189 | 117 | 16 | 3147 | 182 | 141 | 23 | 3123 |
| 5 | 229 | 282 | 15 | 2566 | 117 | 17 | 127 | 2831 | 207 | 58 | 37 | 2790 | 225 | 292 | 19 | 2556 |
| **TOT** | **1052** | **573** | **208** | **13059** | **834** | **1477** | **426** | **12155** | **1045** | **821** | **215** | **12811** | **1091** | **697** | **183** | **12964** |

| Rep3 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 124 | 7 | 69 | 2430 | 81 | 3 | 112 | 2434 | 131 | 53 | 62 | 2384 | 121 | 8 | 72 | 2429 |
| 2 | 279 | 133 | 37 | 2423 | 259 | 571 | 57 | 1985 | 278 | 475 | 38 | 2081 | 282 | 131 | 34 | 2425 |
| 3 | 238 | 9 | 64 | 2518 | 228 | 321 | 74 | 2206 | 240 | 115 | 62 | 2412 | 243 | 11 | 59 | 2516 |
| 4 | 182 | 143 | 23 | 3121 | 149 | 550 | 56 | 2714 | 189 | 115 | 16 | 3149 | 182 | 142 | 23 | 3122 |
| 5 | 229 | 277 | 15 | 2571 | 117 | 17 | 127 | 2831 | 206 | 59 | 38 | 2789 | 225 | 287 | 19 | 2561 |
| **TOT** | **1052** | **569** | **208** | **13063** | **834** | **1462** | **426** | **12170** | **1044** | **817** | **216** | **12815** | **1053** | **579** | **207** | **13053** |

| Rep4 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 123 | 7 | 70 | 2430 | 80 | 3 | 113 | 2434 | 130 | 51 | 63 | 2386 | 121 | 8 | 72 | 2429 |
| 2 | 279 | 128 | 37 | 2428 | 259 | 583 | 57 | 1973 | 279 | 472 | 37 | 2084 | 283 | 127 | 33 | 2429 |
| 3 | 238 | 10 | 64 | 2517 | 227 | 317 | 75 | 2210 | 240 | 115 | 62 | 2412 | 242 | 10 | 60 | 2517 |
| 4 | 182 | 142 | 23 | 3122 | 147 | 552 | 58 | 2712 | 189 | 115 | 16 | 3149 | 182 | 141 | 23 | 3123 |
| 5 | 229 | 272 | 15 | 2576 | 117 | 18 | 127 | 2830 | 208 | 60 | 36 | 2788 | 225 | 291 | 19 | 2557 |
| **TOT** | **1051** | **559** | **209** | **13073** | **830** | **1473** | **430** | **12159** | **1046** | **813** | **214** | **12819** | **1053** | **577** | **207** | **13055** |

| Rep5 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 123 | 7 | 70 | 2430 | 81 | 3 | 112 | 2434 | 130 | 53 | 63 | 2384 | 120 | 8 | 73 | 2429 |
| 2 | 280 | 134 | 36 | 2422 | 256 | 562 | 60 | 1994 | 279 | 475 | 37 | 2081 | 282 | 135 | 34 | 2421 |
| 3 | 238 | 9 | 64 | 2518 | 228 | 324 | 74 | 2203 | 239 | 115 | 63 | 2412 | 242 | 10 | 60 | 2517 |
| 4 | 182 | 146 | 23 | 3118 | 150 | 550 | 55 | 2714 | 189 | 113 | 16 | 3151 | 182 | 139 | 23 | 3125 |
| 5 | 229 | 279 | 15 | 2569 | 117 | 18 | 127 | 2830 | 208 | 58 | 36 | 2790 | 225 | 288 | 19 | 2560 |
| **TOT** | **1052** | **575** | **208** | **13057** | **832** | **1457** | **428** | **12175** | **1045** | **814** | **215** | **12818** | **1051** | **580** | **209** | **13052** |

**INTERVENTION – Average Model Performance**

| RGB | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.8333 | 0.9551 | 0.6318 | 0.9448 | 0.7187 | 0.6968 |
| 2 | 0.8349 | 0.9580 | 0.6474 | 0.9476 | 0.7293 | 0.7077 |
| 3 | 0.8349 | 0.9583 | 0.6490 | 0.9478 | 0.7303 | 0.7088 |
| 4 | 0.8341 | 0.9590 | 0.6528 | 0.9484 | 0.7324 | 0.7108 |
| 5 | 0.8349 | 0.9578 | 0.6466 | 0.9474 | 0.7288 | 0.7072 |
| Average | **0.8344** | **0.9576** | **0.6455** | **0.9472** | **0.7279** | **0.7063** |
| StdDev | **0.0007** | **0.0015** | **0.0080** | **0.0014** | **0.0053** | **0.0054** |

| D | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.6627 | 0.8930 | 0.3640 | 0.8735 | 0.4699 | 0.4284 |
| 2 | 0.6619 | 0.8917 | 0.3609 | 0.8722 | 0.4671 | 0.4255 |
| 3 | 0.6619 | 0.8928 | 0.3632 | 0.8732 | 0.4691 | 0.4275 |
| 4 | 0.6587 | 0.8919 | 0.3604 | 0.8722 | 0.4659 | 0.4239 |
| 5 | 0.6603 | 0.8931 | 0.3635 | 0.8734 | 0.4689 | 0.4270 |
| Average | **0.6611** | **0.8925** | **0.3624** | **0.8729** | **0.4682** | **0.4264** |
| StdDev | **0.0016** | **0.0007** | **0.0016** | **0.0006** | **0.0016** | **0.0018** |

| RGBD3 | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.8405 | 0.9392 | 0.5609 | 0.9308 | 0.6728 | 0.6521 |
| 2 | 0.8294 | 0.9398 | 0.5600 | 0.9304 | 0.6686 | 0.6466 |
| 3 | 0.8286 | 0.9401 | 0.5610 | 0.9306 | 0.6690 | 0.6469 |
| 4 | 0.8302 | 0.9404 | 0.5627 | 0.9310 | 0.6707 | 0.6488 |
| 5 | 0.8294 | 0.9403 | 0.5621 | 0.9309 | 0.6701 | 0.6480 |
| Average | **0.8316** | **0.9399** | **0.5613** | **0.9308** | **0.6702** | **0.6485** |
| StdDev | **0.0050** | **0.0005** | **0.0011** | **0.0002** | **0.0017** | **0.0022** |

| RGBD4 | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.8508 | 0.9632 | 0.6811 | 0.9537 | 0.7565 | 0.7368 |
| 2 | 0.8564 | 0.9490 | 0.6102 | 0.9411 | 0.7126 | 0.6930 |
| 3 | 0.8357 | 0.9575 | 0.6452 | 0.9472 | 0.7282 | 0.7067 |
| 4 | 0.8357 | 0.9577 | 0.6460 | 0.9474 | 0.7287 | 0.7072 |
| 5 | 0.8341 | 0.9575 | 0.6444 | 0.9470 | 0.7271 | 0.7054 |
| Average | **0.8425** | **0.9570** | **0.6454** | **0.9473** | **0.7306** | **0.7098** |
| StdDev | **0.0103** | **0.0051** | **0.0251** | **0.0045** | **0.0160** | **0.0162** |

**Table A.5: Occupancy Context – Performance per cross-validation fold**

| **Rep1** | **RGB** | | | | **D** | | | | **RGBD3** | | | | **RGBD4** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 274 | 5 | 42 | 2309 | 316 | 84 | 0 | 2230 | 269 | 45 | 47 | 2269 | 295 | 4 | 21 | 2310 |
| 2 | 110 | 169 | 5 | 2588 | 111 | 72 | 4 | 2685 | 107 | 99 | 8 | 2658 | 109 | 160 | 6 | 2597 |
| 3 | 85 | 0 | 13 | 2731 | 77 | 2 | 21 | 2729 | 91 | 5 | 7 | 2726 | 88 | 1 | 10 | 2730 |
| 4 | 123 | 53 | 40 | 3253 | 158 | 110 | 5 | 3196 | 109 | 5 | 54 | 3301 | 88 | 1 | 10 | 2730 |
| 5 | 231 | 30 | 59 | 2772 | 97 | 18 | 193 | 2784 | 261 | 17 | 29 | 2785 | 231 | 23 | 59 | 2779 |
| **TOT** | **823** | **257** | **159** | **13653** | **759** | **286** | **223** | **13624** | **837** | **171** | **145** | **13739** | **811** | **189** | **106** | **13146** |

| **Rep2** | **RGB** | | | | **D** | | | | **RGBD3** | | | | **RGBD4** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 265 | 5 | 51 | 2309 | 316 | 141 | 0 | 2173 | 279 | 73 | 37 | 2241 | 264 | 9 | 52 | 2305 |
| 2 | 110 | 180 | 5 | 2577 | 111 | 64 | 4 | 2693 | 113 | 149 | 2 | 2608 | 107 | 158 | 8 | 2599 |
| 3 | 88 | 1 | 10 | 2730 | 80 | 3 | 18 | 2728 | 90 | 4 | 8 | 2727 | 87 | 1 | 11 | 2730 |
| 4 | 128 | 55 | 35 | 3251 | 158 | 110 | 5 | 3196 | 118 | 4 | 45 | 3302 | 132 | 68 | 31 | 3238 |
| 5 | 226 | 33 | 64 | 2769 | 92 | 10 | 198 | 2792 | 247 | 12 | 43 | 2790 | 230 | 22 | 60 | 2780 |
| **TOT** | **817** | **274** | **165** | **13636** | **757** | **328** | **225** | **13582** | **847** | **242** | **135** | **13668** | **820** | **258** | **162** | **13652** |

| **Rep3** | **RGB** | | | | **D** | | | | **RGBD3** | | | | **RGBD4** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 268 | 2 | 48 | 2312 | 316 | 84 | 0 | 2230 | 267 | 48 | 49 | 2266 | 269 | 1 | 47 | 2313 |
| 2 | 109 | 161 | 6 | 2596 | 110 | 58 | 5 | 2699 | 114 | 177 | 1 | 2580 | 106 | 128 | 9 | 2629 |
| 3 | 88 | 1 | 10 | 2730 | 77 | 4 | 21 | 2727 | 90 | 4 | 8 | 2727 | 89 | 1 | 9 | 2730 |
| 4 | 135 | 44 | 28 | 3262 | 155 | 35 | 8 | 3271 | 120 | 6 | 43 | 3300 | 139 | 41 | 24 | 3265 |
| 5 | 226 | 37 | 64 | 2765 | 105 | 9 | 185 | 2793 | 247 | 11 | 43 | 2791 | 234 | 25 | 56 | 2777 |
| **TOT** | **826** | **245** | **156** | **13665** | **763** | **190** | **219** | **13720** | **838** | **246** | **144** | **13664** | **837** | **196** | **145** | **13714** |

| Rep4 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 292 | 7 | 24 | 2307 | 316 | 72 | 0 | 2242 | 286 | 86 | 30 | 2228 | 286 | 10 | 30 | 2304 |
| 2 | 111 | 169 | 4 | 2588 | 109 | 61 | 6 | 2696 | 105 | 76 | 10 | 2681 | 107 | 152 | 8 | 2605 |
| 3 | 90 | 1 | 8 | 2730 | 77 | 3 | 21 | 2728 | 90 | 7 | 8 | 2724 | 88 | 1 | 10 | 2730 |
| 4 | 138 | 40 | 25 | 3266 | 158 | 123 | 5 | 3183 | 114 | 5 | 49 | 3301 | 129 | 67 | 34 | 3239 |
| 5 | 224 | 36 | 66 | 2766 | 95 | 9 | 195 | 2793 | 252 | 15 | 38 | 2787 | 224 | 24 | 66 | 2778 |
| **TOT** | 292 | 7 | 24 | 2307 | **755** | **268** | **227** | **13642** | **847** | **189** | **135** | **13721** | **834** | **254** | **148** | **13656** |

| Rep5 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 281 | 9 | 35 | 2305 | 316 | 114 | 0 | 2200 | 278 | 62 | 38 | 2252 | 291 | 14 | 25 | 2300 |
| 2 | 106 | 135 | 9 | 2522 | 109 | 47 | 6 | 2710 | 111 | 123 | 4 | 2634 | 108 | 176 | 7 | 2581 |
| 3 | 87 | 0 | 11 | 2731 | 76 | 2 | 22 | 2729 | 90 | 6 | 8 | 2725 | 88 | 1 | 10 | 2730 |
| 4 | 124 | 57 | 39 | 3249 | 157 | 87 | 6 | 3219 | 112 | 2 | 51 | 3304 | 130 | 41 | 33 | 3265 |
| 5 | 228 | 40 | 62 | 2762 | 94 | 11 | 196 | 2791 | 245 | 8 | 45 | 2794 | 234 | 24 | 56 | 2778 |
| **TOT** | **826** | **241** | **156** | **13569** | **752** | **261** | **230** | **13649** | **836** | **201** | **146** | **13709** | **851** | **256** | **131** | **13654** |

**OCCUPANCY – Average Model Performance**

| RGB | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.8381 | 0.9815 | 0.7620 | 0.9721 | 0.7983 | 0.7843 |
| 2 | 0.8320 | 0.9803 | 0.7489 | 0.9705 | 0.7882 | 0.7737 |
| 3 | 0.8411 | 0.9824 | 0.7712 | 0.9731 | 0.8047 | 0.7911 |
| 4 | 0.8707 | 0.9818 | 0.7717 | 0.9745 | 0.8182 | 0.8062 |
| 5 | 0.8411 | 0.9825 | 0.7741 | 0.9732 | 0.8062 | 0.7926 |
| Average | **0.8446** | **0.9817** | **0.7656** | **0.9727** | **0.8031** | **0.7896** |
| StdDev | **0.0150** | **0.0009** | **0.0104** | **0.0015** | **0.0110** | **0.0119** |

| D | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.7729 | 0.9794 | 0.7263 | 0.9658 | 0.7489 | 0.7310 |
| 2 | 0.7709 | 0.9764 | 0.6977 | 0.9629 | 0.7325 | 0.7136 |
| 3 | 0.7770 | 0.9863 | 0.8006 | 0.9725 | 0.7886 | 0.7740 |
| 4 | 0.7688 | 0.9807 | 0.7380 | 0.9668 | 0.7531 | 0.7355 |
| 5 | 0.7658 | 0.9812 | 0.7423 | 0.9670 | 0.7539 | 0.7363 |
| Average | **0.7711** | **0.9808** | **0.7410** | **0.9670** | **0.7554** | **0.7381** |
| StdDev | **0.0042** | **0.0036** | **0.0376** | **0.0035** | **0.0205** | **0.0221** |

| RGBD3 | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.8523 | 0.9877 | 0.8304 | 0.9788 | 0.8412 | 0.8299 |
| 2 | 0.8625 | 0.9826 | 0.7778 | 0.9747 | 0.8180 | 0.8056 |
| 3 | 0.8534 | 0.9823 | 0.7731 | 0.9738 | 0.8112 | 0.7983 |
| 4 | 0.8625 | 0.9864 | 0.8176 | 0.9782 | 0.8394 | 0.8281 |
| 5 | 0.8513 | 0.9855 | 0.8062 | 0.9767 | 0.8281 | 0.8160 |
| Average | **0.8564** | **0.9849** | **0.8010** | **0.9764** | **0.8276** | **0.8156** |
| StdDev | **0.0056** | **0.0024** | **0.0249** | **0.0022** | **0.0131** | **0.0138** |

| RGBD4 | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.8844 | 0.9858 | 0.8110 | 0.9793 | 0.8461 | 0.8359 |
| 2 | 0.8350 | 0.9815 | 0.7607 | 0.9718 | 0.7961 | 0.7820 |
| 3 | 0.8523 | 0.9859 | 0.8103 | 0.9771 | 0.8308 | 0.8188 |
| 4 | 0.8493 | 0.9817 | 0.7665 | 0.9730 | 0.8058 | 0.7925 |
| 5 | 0.8666 | 0.9816 | 0.7687 | 0.9740 | 0.8147 | 0.8025 |
| Average | **0.8575** | **0.9833** | **0.7834** | **0.9750** | **0.8187** | **0.8063** |
| StdDev | **0.0187** | **0.0023** | **0.0250** | **0.0031** | **0.0199** | **0.0214** |

**Table A.6: Coverage Context – Performance per cross-validation fold**

| Rep1 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 82 | 22 | 48 | 2162 | 108 | 324 | 22 | 1860 | 78 | 371 | 52 | 1813 | 76 | 18 | 54 | 2166 |
| 2 | 400 | 210 | 96 | 2051 | 312 | 448 | 184 | 1813 | 337 | 380 | 159 | 1881 | 399 | 247 | 97 | 2014 |
| 3 | 102 | 376 | 29 | 2224 | 114 | 676 | 17 | 1924 | 119 | 571 | 12 | 2029 | 101 | 359 | 30 | 2241 |
| 4 | 526 | 131 | 17 | 2632 | 526 | 440 | 17 | 2323 | 410 | 325 | 133 | 2438 | 519 | 116 | 24 | 2647 |
| 5 | 631 | 193 | 21 | 1957 | 165 | 225 | 487 | 1925 | 614 | 290 | 38 | 1860 | 635 | 226 | 17 | 1924 |
| **TOT** | **1741** | **932** | **211** | **11026** | **1225** | **2113** | **727** | **9845** | **1558** | **1937** | **394** | **10021** | **1730** | **966** | **222** | **10992** |

| Rep2 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 77 | 12 | 53 | 2172 | 105 | 294 | 25 | 1890 | 81 | 381 | 49 | 1803 | 78 | 20 | 52 | 2164 |
| 2 | 401 | 220 | 95 | 2041 | 297 | 393 | 199 | 1868 | 337 | 386 | 159 | 1875 | 393 | 208 | 103 | 2053 |
| 3 | 101 | 339 | 30 | 2261 | 115 | 665 | 16 | 1935 | 116 | 482 | 15 | 2118 | 101 | 381 | 30 | 2219 |
| 4 | 497 | 107 | 46 | 2656 | 530 | 394 | 13 | 2369 | 395 | 286 | 148 | 2477 | 519 | 110 | 24 | 2653 |
| 5 | 631 | 216 | 21 | 1934 | 149 | 208 | 503 | 1942 | 599 | 287 | 53 | 1863 | 619 | 217 | 33 | 1933 |
| **TOT** | **1707** | **894** | **245** | **11064** | **1196** | **1954** | **756** | **10004** | **1528** | **1822** | **424** | **10136** | **1710** | **936** | **242** | **11022** |

| Rep3 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 80 | 22 | 50 | 2162 | 108 | 315 | 22 | 1869 | 78 | 379 | 52 | 1805 | 81 | 19 | 49 | 2165 |
| 2 | 405 | 244 | 91 | 2017 | 320 | 449 | 176 | 1812 | 322 | 319 | 174 | 1942 | 390 | 209 | 106 | 2052 |
| 3 | 103 | 388 | 28 | 2212 | 117 | 692 | 14 | 1908 | 121 | 579 | 10 | 2021 | 101 | 372 | 30 | 2228 |
| 4 | 517 | 121 | 26 | 2642 | 524 | 322 | 19 | 2441 | 453 | 374 | 90 | 2389 | 510 | 104 | 33 | 2659 |
| 5 | 629 | 185 | 23 | 1965 | 150 | 197 | 502 | 1953 | 609 | 275 | 43 | 1875 | 630 | 226 | 22 | 1924 |
| **TOT** | **1734** | **960** | **218** | **10998** | **1219** | **1975** | **733** | **9983** | **1583** | **1926** | **369** | **10032** | **1712** | **930** | **240** | **11028** |

| Rep4 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 82 | 18 | 48 | 2166 | 105 | 293 | 25 | 1891 | 80 | 377 | 50 | 1807 | 80 | 20 | 50 | 2164 |
| 2 | 398 | 210 | 98 | 2051 | 324 | 492 | 172 | 1769 | 354 | 389 | 142 | 1872 | 394 | 232 | 102 | 2029 |
| 3 | 102 | 372 | 29 | 2228 | 116 | 676 | 15 | 1924 | 121 | 642 | 10 | 1958 | 101 | 399 | 30 | 2201 |
| 4 | 516 | 124 | 27 | 2639 | 527 | 397 | 16 | 2366 | 449 | 424 | 94 | 2339 | 517 | 127 | 26 | 2636 |
| 5 | 625 | 198 | 27 | 1952 | 161 | 214 | 491 | 1936 | 608 | 276 | 44 | 1874 | 630 | 226 | 22 | 1924 |
| **TOT** | **1723** | **922** | **229** | **11036** | **1233** | **2072** | **719** | **9886** | **1612** | **2108** | **340** | **9850** | **1722** | **1004** | **230** | **10954** |

| Rep5 | RGB | | | | D | | | | RGBD3 | | | | RGBD4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Folds | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN | TP | FP | FN | TN |
| 1 | 73 | 19 | 57 | 2165 | 114 | 337 | 16 | 1847 | 78 | 354 | 52 | 1830 | 81 | 22 | 49 | 2162 |
| 2 | 402 | 222 | 94 | 2039 | 315 | 457 | 181 | 1804 | 352 | 388 | 144 | 1873 | 403 | 249 | 93 | 2012 |
| 3 | 101 | 355 | 30 | 2245 | 114 | 670 | 17 | 1930 | 121 | 621 | 10 | 1979 | 103 | 427 | 28 | 2173 |
| 4 | 497 | 79 | 46 | 2684 | 527 | 369 | 16 | 2394 | 429 | 324 | 114 | 2439 | 506 | 104 | 37 | 2659 |
| 5 | 632 | 221 | 20 | 1929 | 147 | 203 | 505 | 1947 | 604 | 273 | 48 | 1877 | 636 | 240 | 16 | 1910 |
| **TOT** | **1705** | **896** | **247** | **11062** | **1217** | **2036** | **735** | **9922** | **1584** | **1960** | **368** | **9998** | **1729** | **1042** | **223** | **10916** |

**COVERAGE – Average Model Performance**

| RGB | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.8919 | 0.9221 | 0.6513 | 0.9178 | 0.7529 | 0.7175 |
| 2 | 0.8745 | 0.9252 | 0.6563 | 0.9181 | 0.7498 | 0.7124 |
| 3 | 0.8883 | 0.9197 | 0.6437 | 0.9153 | 0.7464 | 0.7102 |
| 4 | 0.8827 | 0.9229 | 0.6514 | 0.9173 | 0.7496 | 0.7130 |
| 5 | 0.8735 | 0.9251 | 0.6555 | 0.9178 | 0.7490 | 0.7113 |
| Average | **0.8822** | **0.9230** | **0.6516** | **0.9173** | **0.7495** | **0.7129** |
| StdDev | **0.0082** | **0.0023** | **0.0050** | **0.0011** | **0.0023** | **0.0028** |

| D | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.6276 | 0.8233 | 0.3670 | 0.7958 | 0.4631 | 0.3667 |
| 2 | 0.6127 | 0.8366 | 0.3797 | 0.8052 | 0.4688 | 0.3729 |
| 3 | 0.6245 | 0.8348 | 0.3817 | 0.8053 | 0.4738 | 0.3793 |
| 4 | 0.6317 | 0.8267 | 0.3731 | 0.7994 | 0.4691 | 0.3741 |
| 5 | 0.6235 | 0.8297 | 0.3741 | 0.8008 | 0.4676 | 0.3719 |
| Average | **0.6240** | **0.8302** | **0.3751** | **0.8013** | **0.4685** | **0.3730** |
| StdDev | **0.0071** | **0.0055** | **0.0058** | **0.0040** | **0.0038** | **0.0045** |

| RGBD3 | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.7982 | 0.8380 | 0.4458 | 0.8324 | 0.5721 | 0.5094 |
| 2 | 0.7828 | 0.8476 | 0.4561 | 0.8385 | 0.5764 | 0.5121 |
| 3 | 0.8110 | 0.8389 | 0.4511 | 0.8350 | 0.5797 | 0.5197 |
| 4 | 0.8258 | 0.8237 | 0.4333 | 0.8240 | 0.5684 | 0.5097 |
| 5 | 0.8115 | 0.8361 | 0.4470 | 0.8326 | 0.5764 | 0.5162 |
| Average | **0.8058** | **0.8369** | **0.4467** | **0.8325** | **0.5746** | **0.5134** |
| StdDev | **0.0162** | **0.0086** | **0.0085** | **0.0054** | **0.0044** | **0.0044** |

| RGBD4 | | | | | | |
|---|---|---|---|---|---|---|
| Rep | Sens | Spec | Prec | Acc | Fscore | MCC |
| 1 | 0.8863 | 0.9192 | 0.6417 | 0.9146 | 0.7444 | 0.7078 |
| 2 | 0.8760 | 0.9217 | 0.6463 | 0.9153 | 0.7438 | 0.7060 |
| 3 | 0.8770 | 0.9222 | 0.6480 | 0.9159 | 0.7453 | 0.7077 |
| 4 | 0.8822 | 0.9160 | 0.6317 | 0.9113 | 0.7362 | 0.6984 |
| 5 | 0.8858 | 0.9129 | 0.6240 | 0.9091 | 0.7322 | 0.6945 |
| Average | **0.8815** | **0.9184** | **0.6383** | **0.9132** | **0.7404** | **0.7029** |
| StdDev | **0.0048** | **0.0040** | **0.0102** | **0.0029** | **0.0059** | **0.0061** |

### E.3: Face detection train and test folds for finetuning NICUface-RF and NICUface-Y5F

Table A.7: Patient ID per fold in CHEOch (Unique data absent from CHEOopt).

| CHEOch_F1 | CHEOch_F2 | CHEOch_F3 |
|---|---|---|
| 1 | 22 | 10 |
| 5 | 23 | 25 |
| 14 | 26 | 30 |
| 16 | 32 | 29 |
| 19 | 35 | 31 |
| 21 | 38 | |

Table A.8: Training set partition across patients with Train/Val split across datasets

| Dataset | Stage | Patient Partition | NumPatients | NumImages |
|---|---|---|---|---|
| COPE | Train | S04 – S33 | 20 | 149 |
| | Val | S34 – S41 | 7 | 34 |
| NBHR | Train | Start – 20201014100636_1.0 | ? | 453 |
| | Val | 202010141000943_1.0 – end | ? | 112 |
| CHEOopt | Train | Rest… | 12 | 72 |
| | Val | Pt 6, 8, 28, 34 | 4 | 39 |
| CHEOch_F1 | Train | Pt 1, 5, 14, 16, 19 | 5 | 149 |
| | Val | Pt 21 | 1 | 31 |
| CHEOch_F2 | Train | Pt 22 23, 26, 32, 38 | 5 | 226 |
| | Val | Pt 35 | 1 | 52 |
| CHEOch_F3 | Train | Pt 10, 30, 29, 31 | 4 | 345 |
| | Val | Pt 25 | 1 | 68 |
| CHEOch_rest | Train | Rest… | 12 (same as CHEOopt) | 812 |
| | Val | Pt 6, 8, 28, 34 | 4 (same as CHEOopt) | 172 |

This Train/Validation partition was used for finetuning NICUface-Y5F, and all data from Train and Val set were used directly during training when finetuning NICUface-RF.

## E.7: Neonatal HR estimation detailed dataset description used in Chapter 7.2.2 and Chapter 8.2.1

Table A.9: Resting Patients & Challenging Scenario HR Estimation Dataset

| Scenario | Patient ID | Filename | Start Time | End Time |
|---|---|---|---|---|
| Rest – bright light | 1 | Part_2 | 19:29:22 | 19:29:51 |
| | 2 | Part_3 | 17:44:10 | 17:44:39 |
| | 6 | Part_47 | 19:58:07 | 19:58:36 |
| | 11 | Part_12 | 18:40:52 | 18:41:21 |
| | 13 | Part_217 | 18:23:21 | 18:23:50 |
| | 15 | Part_52 | 13:30:03 | 13:30:32 |
| | 28 | Part_4 | 12:19:17 | 12:19:46 |
| | 34 | Part_56 | 16:39:55 | 16:40:24 |
| | 36 | Part_53 | 12:46:57 | 12:47:26 |
| | 37 | Part_26 | 13:39:36 | 13:40:05 |
| Rest – low light | 1 | Part_51 | 19:53:52 | 19:54:21 |
| | 18 | 2_Part_20 | 19:56:28 | 19:56:57 |
| | 33 | Part_234 | 17:55:28 | 17:55:57 |
| Rest – phototherapy light | 19 | Part_34 | 18:57:08 | 18:57:37 |
| | | Part_40 | 19:00:08 | 19:00:37 |
| | | Part_41 | 19:00:38 | 19:01:07 |
| | | Part_74 | 19:17:08 | 19:17:37 |
| Patient out | 8 | Part_238 | 12:24:45 | 12:25:14 |
| Patient in | 28 | Part_66 | 12:50:17 | 12:50:46 |
| Facial occlusion – temporary | 1 | Part_8 | 19:32:22 | 19:32:51 |
| Facial occlusion – continuous | 16 | 1_Part_230 | 15:25:46 | 15:26:15 |
| Facial motion suction | 6 | Part_25 | 19:47:07 | 19:47:36 |
| Facial motion sneeze | 8 | Part_212 | 12:30:47 | 12:31:16 |
| Facial motion yawn | 36 | Part_89 | 13:04:57 | 13:05:26 |
| Hiccup | 8 | Part_202 | 12:06:45 | 12:07:14 |
| Clinical intervention (no occlusion) | 28 | Part_27 | 13:30:47 | 12:31:16 |
| Body motion – minor | 2 | Part_103 | 18:34:10 | 18:34:39 |
| Body motion – major | 14 | Part_243 | 19:32:56 | 19:33:25 |
| High to low light | 1 | Part_49 | 19:52:52 | 19:53:21 |
| Monitor alarm light | 1 | Part_74 | 20:04:52 | 20:05:21 |