Video-based infant monitoring: Recognizing appetite, pain and sleep in preterm infants

Master's Thesis

Master Artificial Intelligence - Utrecht University

27-06-2022

Author
Fabian Mijsters
7648510

Supervisor
Dr. Ronald Poppe
Second Supervisor
Dr. Albert Ali Salah

Acknowledgements

I want to start off with expressing my utmost gratitude to Dr. Ronald Poppe. He went above and beyond in supporting me during this thesis. Without his support and advice this thesis would not have reached the current quality. Secondly, I would like to thank my brother Yannick Mijsters for the interesting and motivating discussions that always resulted in a lot of extra food for thought. Additionally, I would like to thank my girlfriend Emma Jeronimus for the solid advice and for keeping up with me talking about infants for nine months. Furthermore, I would like to thank UMC Utrecht for allowing me to have the opportunity to research the most vulnerable group on this earth. I thoroughly enjoyed the cooperation with everyone at the UMC. Especially, I would like to thank Roberto Narvarro and Ilse Smits for the excellent collaboration during the planning of the recordings, figuring out how the recordings would achieve the highest level of quality, and building the annotation tool. Finally, I would like to thank my parents and my friends for their unwavering support.

Abstract

preterm infants show their internal state through cues. Nurses attempt to observe these cues as often as possible. This is not always possible due to the dynamic environment of the Neonatal intensive care unit (NICU). Missing cues can lead to misdiagnosis and an overall longer stay at the NICU. A system that can support nurses in recognizing these cues is therefore highly sought after. The overall goal of this research is to detect certain behaviors of preterm infants and use these behaviors to recognize cues.

In this study we present a rule based cue detection program. For this program we have compared three facial landmark detection and human pose estimation models. The most robust models were used to generate key-points for videos of preterm infants. These key-points serve as the foundation of the rules. In this program, medical professionals are able to describe certain behaviors in the form of a straightforward rule. A rule describes the movements of certain key-points on a preterm infant during a cue. These rules are evaluated on the key-points extracted from each frame of the videos. The detections are generated by applying a threshold to the evaluation. These cues can be used to determine the current state of an infant. Whether the infant is in pain, experiences appetite, or is in a certain sleep state. During the experiment, three medical professionals have built rules for four different cues on a training set that spanned 6 minutes. The rules were evaluated on a test set of 15 minutes. The experiment showed that medical professionals are able to build rules that can detect human annotated cues in preterm infants without any additional learning.

Contents

1 Introduction

1.1 Motivation

The infantile period of human life is the most defining period for the creation of neural pathways in the brain. Sleep, eating, and a lack of stress and pain are essential to the development of an infant's brain [56]. preterm infants spend the first weeks or days of their lives in the NICU based on their gestational age. The NICU is a dynamic place which can contribute to stress and a lack of sleep. This can have long-lasting health effects. These problems range from underdeveloped brains leading to lower capabilities or emotional and behavioral problems at school [1, 27, 19, 30, 36]. Therefore, preterm infants are at an increased risk compared to term infants. Still, a lot is unknown regarding the relation between sleep and development, specifically for preterm infants. Reducing stress, discomfort and pain while increasing the quality of sleep is therefore of the utmost importance.

Infants cannot communicate their own needs and feelings. Therefore, nurses attempt to predict the state that an infant is in. A nurse observes the infant for a limited amount of time and infers the state the infant is in from the different behaviors that the infant shows. These behaviors can be a hand moving to a mouth or movements and noises belonging to sucking for appetite [10] or different behaviors signaling pain, discomfort or sleep. These cues are used to fill out scales [29, 58, 22, 7]. An example of a scale is given in Table 1. Scales give a score to a cue. Based on the sum of these scores, the level of stress, pain, or appetite can be determined. Scales were introduced to remove subjectivity from the determination of an infant's state. Nurses monitor infants for a period before, during, and after medical procedures or feeding. Unfortunately, there are not enough nurses to monitor an infant continuously. The inconsistent and intermittent monitoring of infants can lead to the misdiagnosis of diseases which causes under- or overtreatment [68, 59, 47].

To summarize, the two main drawbacks of the current approach are the time it costs to monitor an infant and the subjectivity that accompanies individual monitoring. This research focuses on pain/discomfort, appetite, and sleep states of preterm infants. The goal of this research is to develop a continuously monitoring camera-based system to detect the aforementioned states and attempt to resolve the two main drawbacks.

1.2 Scope

The system should be able to classify the state a preterm infant is in. The cues related to an infant's state take place all over the body; facial expressions, movement of the body, and through its vital signs. Therefore, the system should be able to extract these cues from multiple modalities. Multiple modalities will be used to reduce the dependence on noise and ambiguity in a single modality and to potentially benefit from complementary information of multiple modalities. Secondly, the cues that the system detects, should be usable to determine an infant's state as shown in Figure 2.

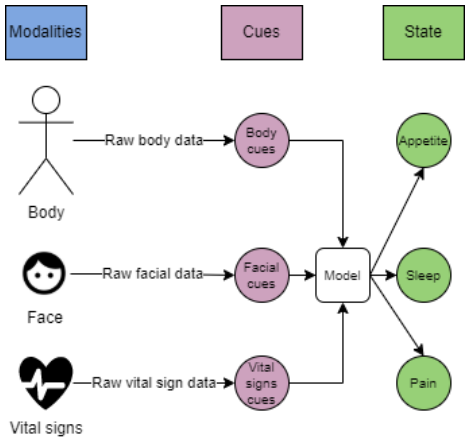


Figure 2: Schematic pipeline of the system

A pose estimation model will be used to recognize cues that occur on the body of a preterm infant. Sequences of the estimated pose can be used to determine the bodily movements of an infant. These bodily movements might consist of a hand moving towards the infant's mouth to show appetite. Fuzziness is cue that might indicate pain or stress. Facial expressions are also a telling sign of an infant's state. Frowning might be a sign of pain or appetite and slow blinking, prolonged closing of the eyes and yawning signals fatigue and sleep. Recognizing facial expressions can be done using facial landmark models. These models predict key-points on an infant's face as shown in Figure 3.

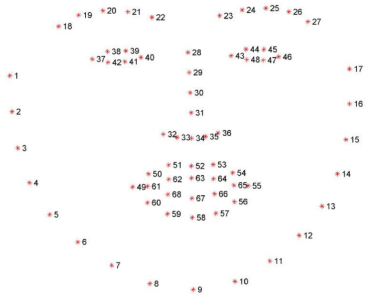


Figure 3: Facial landmarks by dlib (1 dlib [24] facial landmarks (68 key-points)

Finally, the vital signs of a preterm infant change based on its current state. Heart rate variability increases when an infant experiences pain or stress. Heart rate slows down after feeding or before sleep. [49].

The modalities can be combined using fusion. Fusion merges different modalities in different ways. Early fusion combines the modalities into one feature vector used to predict states. In contrast, late fusion combines results of modality-specific models

into one final state classification using voting ensembles [45]. The aforementioned modalities all have their own unique challenges and this is aggravated by the lack of research on automated analysis of preterm infants. The challenges will be discussed in the literature review.

1.3 Research questions

During this research, the following questions will be answered:

1.3.1 Do current publicly available facial landmarking and pose estimation models predict robust key-points?

Detecting cues of the infants is entirely reliant on the quality of the key-point detections. Applying facial landmarking and pose estimation to the infant field is a relatively small research area. There are no publicly available data-sets of infants and especially preterm infants. This is due to the privacy concerns that occur in this field. We have to determine whether or not the quality of the predictions are accurate and robust enough to be used to detect infant cues .

1.3.2 Are rules build in a rule based cue detection program able to detect cues annotated by human annotators?

To detect cues in infants, we will determine whether or not a rule based approach is able to detect cues in a low quality and quantity rich environment. We are not able to implement a state of the art action recognition model due to the low quality and low quantity of the data, . Therefore we will determine if a rule based approach model is able to strive in this environment.

1.3.3 Does the performance of cue detection rules increase if we apply majority voting to the evaluation of the rules?

We will apply majority voting to the rules created by the participants in the experiment to see whether or not combined rules outperform their individual counterparts. We will also determine whether or not low quality rules or high quality rules benefit the most from majority voting.

1.4 Thesis outline

Related literature on the different infant states will be discussed and how infants show these states in chapter 2. Then the techniques that can be used to detect these cues will be discussed and finally, the technique that will be used to combine the different modalities will be discussed.

2 Literature review

In this section we will discuss the three infant states; appetite, pain/discomfort, sleep. Firstly, we will explain each state and how to recognize it. Secondly, we divide the input video into two modalities and add vital sign data as the third modality. We will discuss each modality separately and discuss the challenges of each. Finally we will discuss how we will combine the three different modalities into one classification.

2.1 Appetite

Nutrient intake in the infantile stage of life is essential to the growth and development of infants. This is especially true for preterm infants, where proper feeding technique and nutrient intake has been linked to a shorter stay at the NICU (neonatal intensive care unit) [63]. There are two main preterm infant feeding techniques, infant-driven feeding (IDF) and traditional practitioner-driven feeding (PDF) [63]. The IDF technique is to wait for appetite cues from the infant before starting the feeding process. The PDF technique is to feed an infant based on a set schedule that can be changed to better accommodate the acting practitioner. Wellington et al.[63] researched the impact of IDF versus PDF on preterm infants. They used the PDF technique on 153 infants and IDF on 101 infants. Practitioners were instructed to give a score of the observed readiness for feeding and when this score was high enough to begin feeding. Preterm infants that showed too few signs or were not expressive enough were fed not only on cues but also according to a schedule. Preterm infants are often discharged from the NICU based on their ability to independently nipple feed. The study found that infants with a gestational age of < 28 weeks reached full nipple feeds 17 days sooner compared to PDF and were discharged nine days earlier. For infants aged 28-31 weeks, full nipple feeds were achieved eleven days sooner and were discharged nine days earlier too. For preterm infants between 32 and 34 weeks gestational age, the successful nipple feeding was, on average, achieved three days sooner and discharged three days earlier. These findings indicate that preterm infants that were fed based on the infant's own hunger cues reach a certain developmental level faster than infants that were fed on a practitioner's set schedule. Concluding, this study corroborates that accurately recognizing cues can lead to improved development.

Kirk and others [37] divide individual full nipple feeds into their behavioral parts, "Preterm infants need to be able to coordinate sucking, swallowing, and breathing, sustain alert awake behavior and preserve cardio-respiratory stability to achieve successful oral feeding." Kirk studies the effects of infant-driven feeding using a clinical pathway that was based on feeding readiness signals of the infant. The study was carried out with 53 infants with 29 in the study group and 24 in the control group. On average the study group achieved successful oral feeding between 4.5 and 6 days earlier compared to the control group. Kirk et al. [37] state that "Preterm infants may be encouraged to bottle feed before they are physiologically or behaviorally ready, subjecting them to a trial-and-error approach that can increase stress and detract from

success". Additionally, Kirk et al. [37] mention that repeated stimulation leads to an increase in oxygen use which may make the infant appear limp or rigid. This makes it harder for the infant to show future hunger cues and for practitioners to recognize them. Wellington et al. [63] studied the importance of feeding preterm infants when they show hunger signs. Kirk et al. expanded upon Wellington's findings. Kirk et al. show that recognizing these cues accurately is essential. The authors found that preterm infants are less likely to show appetite related cues when attempts are made to feed the infant while it is not hungry. All in all the aforementioned points solidify the need for accurate real-time appetite cue recognition.

2.1.1 Appetite cues

Determining when a preterm infant is hungry requires recognizing appetite cues. These cues show when an infant is hungry and, more importantly, when the infant is ready to be fed. According to Cagan [10], the primary hunger cues are fuzziness without crying, hand-to-mouth activity, rooting, and hiccups. Wellington et al. [63] also note that the muscle tone of the infant, which is the amount of tension or resistance to movement in the infant's muscles, is an important factor in determining feeding readiness. Additionally, the infant's ability to keep the pacifier is mentioned since this is an indication of the ability of an infant to suck. A study conducted on NICU nurses and their decisions to start trying nipple feeding or bottle feeding found that nurses ranked behavioral and physiologic signs higher than physical signs. The most important behavioral sign that an infant is ready for bottle feeding is, according to these nurses, nonnutritive sucking (NNS). Nonnutritive sucking is the act of sucking on objects that do not add any nutritious value such as fingers, pacifiers, or other objects. The highest-ranking physiological factor was gagging while inserting the gavage tube. The gavage tube is a small tube that is used to feed the infant. Post conceptual age (PCA) was the most important physical attribute. PCA is the number of weeks/months since conception. NNS is not always a one-on-one predictor for successful full nipple feeding or bottle feeding. When oxygen levels are checked during NNS and are proven to be stable, this can be an indication that the infant is also able to keep stable oxygen levels during oral feeding [2]. Cagan [10] concluded that the main cues pointing towards an infant's appetite are: fuzziness without crying, hand-to-mouth movement, rooting, and nonnutritive sucking. A collection of all cues divided between the three states and modalities is shown in Table 2.

2.1.2 Appetite scales

The appetite scale that Wellington et al. [63] used is shown in Table 1. The scale consists of combinations of cues that indicate the readiness of a preterm infant. The readiness score is used to determine the proper feeding technique. A preterm infant with a high readiness score (1-2) should be nipple fed. A preterm infant with a low readiness score (3-5) should be fed by gavage tube. Wellington et al. [63] use the same definition of readiness as Kirk et al. [37]. Wellington et al. [63] adds that feedings should be done every 3 hours. The technique used to feed the infant at that time is based on the readiness score.

Readiness score	Description	Feeding technique
1	Drowsy, awake or fussy prior to care. Rooting or hand to mouth. Keeps pacifier. Good tone.	Nipple feed
2	Drowsy or awake once handled. Some rooting or takes pacifier. Adequate tone	Nipple feed
3	Briefly alert with care. No hunger cues. No change in tone.	Gavage
4	Sleeping throughout care. No hunger cues. No change in tone.	Gavage
5	Needs increased oxygen with care. Apnea and or bradycardia (A/B) with care. Tachypnea over baseline with care.	Gavage

Table 1: Appetite scale [63]

2.2 Pain and discomfort

preterm infants in the NICU undergo many painful procedures [23]. Research indicates that untreated pain in infants can have short [54, 35] and long term consequences [53, 6]. Recognizing pain and stress in infants is important to ensure that adjustments can be made to the food, treatment, or care of the infant. Currently, nurses monitor infants for a short period spanning the duration of a certain procedure for example the heel prick. During this monitoring, nurses determine the state of an infant, specifically how much pain and discomfort the infant is experiencing. Nurses determine the pain level by writing down which cues the infant shows during this time. These cues are used to fill out certain scales [29]. A low pain level in an infant is considered discomfort. Discomfort can occur after feeding and can be detrimental to an infant's sleep pattern.

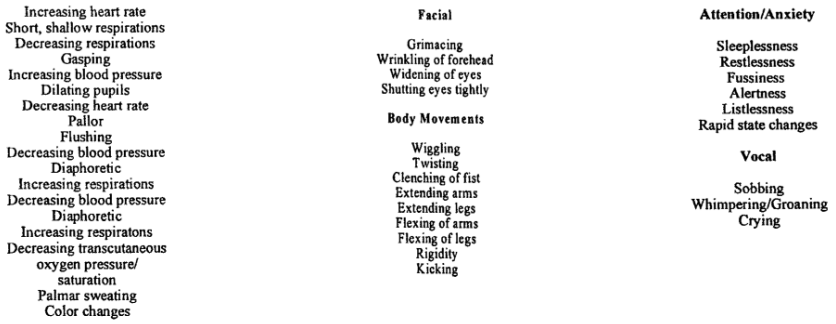


Figure 4: Infant pain and discomfort cues [31]

State	Modality	Type	Source
Appetite			
Fuzziness without crying	Body	Binary	[10, 63, 37]
Hand-to-mouth movement	Body	Binary	[10, 63, 37]
Rooting	Face	Binary	[10, 63, 37]
Nonnutritive sucking	Face	Binary	[10, 63, 37]
Pain and Discomfort			
Increasing heart rate	Vital signs	Continuous	[31]
Short shallow respirations	Vital signs	Continuous	[31]
Decreasing resporation	Vital signs	Continuous	[31]
Gasping	Vital signs | Face	Binary	[31]
Increasing blood pressure	Vital signs	Continuous	[31]
Dilating pupils	Face	Binary	[31]
Decreasing heart rate	Vital signs	Continuous	[31]
Pallor	Body | Face	Binary	[31]
Flushing	Face	Binary	[31]
Decreasing blood pressure	Vital signs	Continuous	[31]
Diaphoretic	Body | Face	Binary	[31]
Decreasing trancutaneous	Vital signs	Binary	[31]
Oxygen pressure/saturation	Vital signs	Continuous	[31]
Palmar sweating	Body	Binary	[31]
Color changes	Body | Face	Binary	[31]
Grimacing	Face	Binary	[31, 25]
Wrinkling of forehead	Face	Binary	[31, 25]
Widening of eyes	Face	Binary	[31, 25]
Shutting eyes slightly	Face	Binary	[31, 25]
Wiggling	Body	Binary	[31]
Twisting	Body	Binary	[31]
Clenching of fist	Body	Binary	[31]
Extending arms	Body	Binary	[31]
Extending legs	Body	Binary	[31]
Flexing of arms	Body	Binary	[31]
Flexing of legs	Body	Binary	[31]
Rigidity	Body	Binary	[31]
Kicking	Body	Binary	[31]
Sobbing	Face	Binary	[31]
Whimpering/groaning	Face	Binary	[31, 25]
Crying	Face	Binary	[31, 25]

State	Modality	Type	Source
Sleep			
Open and focused eyes	Face	Binary	[18]
Closed and relaxed eyelids	Face	Binary	[18]
Rapid eye movements with closed eyes	Face	Binary	[18]
REM with open eyes	Face	Binary	[18]
Gross body movement - Not sudden movement of both limbs and torso	Body	Binary	[18]
Small body movement - Not sudden movement of (part of) one limb	Body	Binary	[18]
Twitch – Short and small movement of body or body part	Body	Binary	[18]
Jitter – Rhythmic twitch of at least three cycles and involving part or all of the body	Body	Binary	[18]
Startle/Jerk - Big sudden movement of body involving at least one extremity (e.g. Moro reflex)	Body	Binary	[18]
Stretch – Increased muscle tone of body or body part	Body	Binary	[18]
Writhing - Increased muscle tone of torso, torso is elevated, while limbs don't move	Body	Binary	[18]
Frown – Wrinkling of forehead	Face	Binary	[18]
Grimace – Contraction of whole face	Face	Binary	[18]
Smile – Mouth contraction sideways	Face	Binary	[18]
Eyebrow movement – Eyebrows going up/down	Face	Binary	[18]
Closed, squinted eyes – Eyelids contracting	Face	Binary	[18]
Reflexive facial movements –Twitch/jerk/startle in face	Face	Binary	[18]
Smacking - Smacking with lips moving towards and away from each othe	Face	Binary	[18]
Sucking - Sucking as if on a pacifier	Face	Binary	[18]
Mouthing - Movements of the mouth other than smacking, sucking, yawning and smiling	Face	Binary	[18]
Yawn - Open mouth and elevated eyebrows	Face	Binary	[18]
No facial movements	Face	Binary	[18]
Sobs – Soft moaning sound / Sighs – Deep audible respiration	Face	Binary	[18]
Grunt – Louder	Face	Binary	[18]
Hiccup – Soft gasp/hiccupping like sound	Face	Binary	[18]
Coughing – Soft cough	Face	Binary	[18]
Squeal – Squealing sound	Face	Binary	[18]
Crying – High volume 'crying'	Face	Binary	[18]
MAR	Face	Ordinal	[41]
EAR	Face	Ordinal	[41]
Regular heart rate – Stable and relatively slow heart rate	Vital Signs	Continuous	[18]
Irregular heart rate – Unstable and relatively fast heart rate	Vital Signs	Continuous	[18]
Regular respiratory frequency – Stable and relatively slow respiratory frequency	Vital Signs	Continuous	[18]
Irregular respiratory frequency – Unstable and relatively fast respiratory frequency	Vital Signs	Continuous	[18]

Table 2: State cues

2.2.1 Pain and discomfort cues

Infants are not able to communicate the level of pain or discomfort they are experiencing [32]. Therefore, observing an infant and looking for cues that might indicate pain and discomfort is the only option. Howard et al. [31] outline various cues that are an indication of pain or discomfort. These cues can be seen in Figure 4. A collection of all cues divided between the three states is shown in Table 2. Looking at Figure 4 it is clear that the cues span the three modalities; behavior, physiological, and physical cues.

2.2.2 Pain scales

There is a wide variety of different pain scales that are updated and adapted based on new research. An example of such a scale is the Pain Assessment In Neonates(PAIN) [29] scale. When nurses use the PAIN scale, they score the behavior of the infant by looking at the breathing patterns and the facial expressions. The wide range of different pain and discomfort scales all have the same purpose and that is to quantify the amount of pain an infant is in. The difference between the scales is the cues. Certain scales focus more on body movement while other scales use the face as the primary predictor. The PAIN scale weights breathing patterns and oxygen levels the highest. Scales that quantify pain in infants focus on objective information about the infant leading to higher inter-rater reliability. Suraseranivongse et al. [55] compared three infant pain scales; cry requires O2 increased vital signs expression sleeplessness (CRIES) [38], children's and infants' postoperative pain scale (CHIPPS) [9], neonatal infant pain scale (NIPS) [40]. The scales were scored based on validity, reliability and practicality. The authors found that all the scales have excellent inter-rater reliability. CRIES [38] showed the lowest correlation with the other scales based on validity. All scales scored evenly on reliability. Based on the findings of the authors NIPS [40] was determined to be the most practical and is therefore recommended by the authors. Hudson-Barr et al. [32] compared the NIPS [40] scale to the PAIN [29] scale and showed that the PAIN scale is a valid scale with an overall correlation of 0.93.

2.3 Sleep

Sleep is essential to the development of neural pathways in an infant's brain [42]. Uninterrupted, stress-less sleep is important during the time a preterm infant spends in the NICU. Therefore, promoting sleep should be one of the most important concerns in the NICU [5]. The sleep of preterm infants can be categorized in three categories. Restless behavior is classified as active sleep (AS). preterm infants often spend 40%-60% of their total sleep time in AS [16]. Periods during sleep where the infant is less active is often called quiet sleep (QS). The switching period between different sleep states in preterm infants is called intermediate sleep (IS). The sleep state IS is also assigned when the sleep state is not clear. Apart from these three sleep states there is also the behavioral state wake [5].

The sleep of preterm infants can be disturbed in the NICU due to stress. Stress in

an infant can occur when an infant's sleep is not properly recognized and the feeding process is started, leading to increased respiration speed and other stress cues. This stress prohibits the well-needed sleep of the infant. Noise at the NICU is another reason for the lack of sleep in infants. Improper medication or misdiagnosis also leads to a lack of sleep which can have long-lasting consequences for the brain development of an infant [60]. An infant's lack of sleep leads to the underdevelopment of neural pathways in the brain. This underdevelopment can lead to emotional and behavioral problems later in life and especially in school. Behavioral and emotional problems lead to poor interactions with classmates or friends. Misbehaving in school leads to lower results, which in turn leads to lower job chances [1, 30].

2.3.1 Sleep cues

A preterm infant shows sleepiness in the following ways: The heart rate slows down, movements are reduced to a minimum and the breathing patterns become more stable and relaxed. In addition to the aforementioned clues, the eyes and mouths can be important signs of sleepiness. Increased blinking and prolonged closing of the eyelids are telling signs of fatigue. Secondly, infants show fatigue by yawning [34]. These cues can be recognized by looking at two ratios, the Eye Aspect Ratio (EAR) and the Mouth Aspect Ratio (MAR) [41]. Both ratios look at the vertical and horizontal ratios of the eye and mouth. The meaning of EAR and MAR is shown in Figure 5. Where the points around the mouth and eye correspond to the facial landmark estimation points.

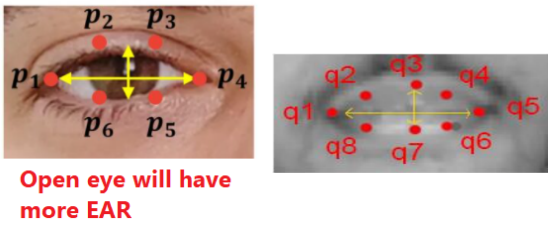


Figure 5: Eye aspect ratio left [48], Mouth aspect ratio right [41]

Traditionally EAR and MAR [41] are calculated using the outcomes of a facial landmark prediction model which will be discussed in the following section. To calculate EAR and MAR, the sum of the distances between vertical pairs is divided by the sum of the distances between the horizontal pairs to get the ratio of the vertical and horizontal spread. Finally, a threshold needs to be determined. When the MAR or EAR is below this threshold the eye or mouth is closed and otherwise, it is open. To determine the current sleep state of a preterm infant, de Groot et al. [18] show that sleep state cues occur in a wide variety of places: the eyes, the body, facial movements, sounds the infant produces, the activity level of the infant, heart rate and respiratory patterns. The detailed description of the cues is shown in Table 2. De Groot et al. [18] determined that irregular heart rate and respiratory patterns are signs of AS. Regular heart rate and respiratory patterns indicate QS.

2.3.2 Sleep scales

De groot et al. [18] created a preterm infant sleep scale: behavioral sleep stage classification for preterm infants (BeSSPI). The sleep/wake cycle of infants was divided into the four aforementioned stages: active sleep (AS), quiet sleep (QS), intermediate sleep (IS) and, wake (W). The items used in the BeSSPI scale span all three of the predefined modalities: face, body and, vital signs. Additionally, vocalizations were added. The BeSSPI is shown in Table 2 under the sleep category.

2.4 Automated measurement of behavior cues

The different cues and signs can be divided into three categories: body, face, and vital signs. The categories all have their own respective research area. Therefore these categories will all be reviewed separately. A division has been made between recognizing and classifying the state cues from the body, face and vital signs. This division ensures scalability and expandability. The body model and face model components of the system both consist of two parts. Part one is the current location of the face and body. Part two is the movements of the face and body such as frowning or moving a hand to the mouth. Representations and embeddings are needed to bridge the gap between human interpretations of these parts to a language a machine can understand.

2.4.1 Representations

Representing the position of a face can be done using facial landmarks [64]. 2D facial landmarks encode the current position of a specific area of the face as shown in Figure 3 and Figure 8. The facial landmarks representation consist of 68 or 168 key-points strategically spread out over the face. These key-points are able to capture rigid and non-rigid facial movements due to facial expression and head movements. The position of a human body can be represented with a body skeleton as

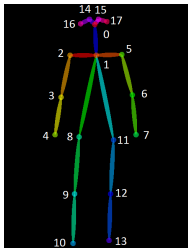


Figure 6: Body skeleton COCO [15]

shown in Figure 6. The body skeleton has the same purpose as the facial landmarks: to rigidly capture body movements. These representations translate the real world position of the body or face to a vector the system can interpret. The representations encode only the current position of the corresponding body part and do not contain information regarding the visual appearance of that part. For example the dilation

or shape of the irises is not present in the facial landmark representation. The representations do however capture most of the movement related cues discussed in Sections 2.1.1, 2.2.1 and 2.3.1. A multitude of techniques exist that can used to determine the locations of the key-points for the facial landmarks as well as for the body skeleton. The challenges that occur while trying to determine these key-points will be discussed in the following chapters. Each modality; face, body and vital signs will be discussed separately.

Expression	**Present**
Brow Bulge	True/False
Eye Squeeze	True/False
Naso-labial Furrow	True/False
Open Lips	True/False
Vertical Mouth Stretch	True/False
Horizontal Mouth Stretch	True/False
Taut Tongue	True/False
Tongue Protrusion	True/False
Chin Quiver	True/False
Lip Purse	True/False

Table 3: Neonatal Facial Coding System

2.4.2 Face

Facial coding systems are explored to determine whether facial landmarks cover enough cues to make accurate state predictions. There are a multitude of proposed facial coding systems such as the aforementioned NIPS[40] and the Neonatal Facial Coding System (NFCS) [25].

The NFCS encodes the facial action units (AU) of infants and specifically the AUs listed in Table 3. Grunau [25] et al. applied the NFCS on 42 infants during multiple procedures. The authors found that "Eye Squeeze", "Brow Bulge", "Naso-labial Furrow", and "Open Mouth" occurred in over 80% of the infants during the procedures. These AU occur in areas covered by the facial landmarks and are therefore recognizable. Over the years there have been a multitude of infant specific pain scales. Carlini et al. combined these scales into 14 regions of interest (ROI) shown in Figure 7.

The ROIs shown in Figure 7 will be the regions that need to be analyzed to be able to accurately recognize facial cues. These ROIs are essential in determining the level of pain. Low-level pain expressions can be used to determine discomfort after feeding. The more expressions from Table 3 that are noted as present, the higher the level of experienced pain in the infant.

The technique that will allow us to detect changes in the aforementioned ROIs is called 2D facial landmark detection. These key-points can be detected on sequential frames and used to detect changes in pose and expression. Wu et al. [65] provide an

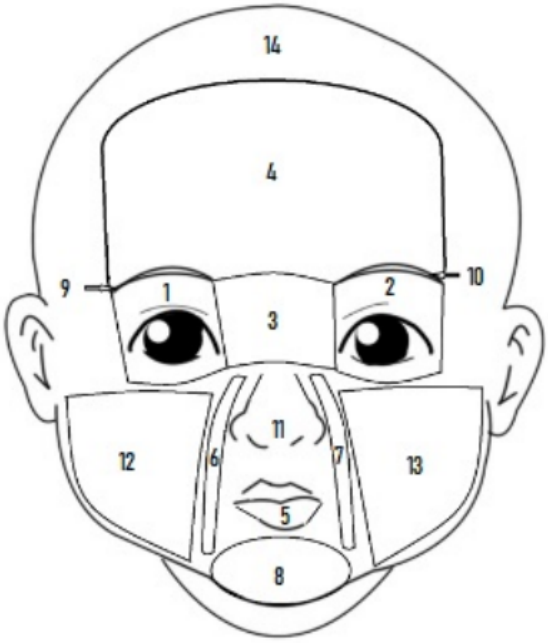


Figure 7: (1) Right and (2) Left Eye (Eye squeeze; Frown; Eyes tense; Distressed look); (3) Region between Eyebrows and (4) Forehead (Furrowed forehead; Furrowed brow; Brow bulge); (5) Mouth (Open mouth; Tense mouth; Horiz. mouth stretch; Vert. mouth stretch; Lip purse; Open lips; Taut tongue; Tongue protrusion); (6) Right and (7) Left Nasolabial Groove (Nasolabial furrow); and (8) Chin (Chin quiver). (9) Right and (10) Left Eyebrow; (11) Nose; (12) Right and (13) Left Cheek, and (14) "Other regions of the face" [12]

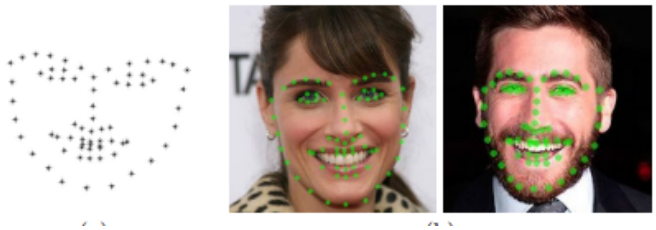


Figure 8: Facial landmark detection [65]

overview of state-of-the-art facial landmark detection techniques. First, they divide the different techniques into three categories: holistic methods, Constrained Local Model (CLM), and regression-based models. Secondly, they discuss the underlying theory and differences. Finally, they provide a comparison between the different techniques, comparing them on different datasets both in the wild and controlled,

and comparing them based on facial expressions, head poses, and occlusion. These comparisons will be combined into a list of strengths and weaknesses per technique. Infant specific facial landmark recognition is a relatively underexplored area of research. Therefore, the results of the comparisons done by Wu et al. might not be applicable to infant facial landmark detection. Datasets used to train models for facial landmark detection will contain occlusions. Especially in the pre-term infant area most, if not all, of the images contain occlusions. Infants in the NICU often are riddled with bandages and probes for food or oxygen. Therefore the focus should be on solutions that deliver accurate and robust results under occlusion. Burgos-Artizzu et al.,[8] propose a solution that attempts to predict all key-points by dividing the face into 9 segments, only looking at a single segment to predict all the points on the face, and merging these points based on a probability that a part is occluded. The proposed model outputs landmarks with a label occluded or not occluded as shown in Figure 9.

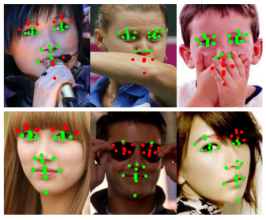


Figure 9: Facial landmark detection [8]

Yu et al.[66] propose a solution that is interesting for the NICU use case since they train multiple models to predict key-points in pre-determined occluded areas. Infants in the NICU often have the same occlusions. For example, food probes, an incubation probe, or other medical equipment. The authors propose a Consensus of Regressor (CoR) approach. Using the results of multiple regressors to determine the most likely location of a key-point and whether this key-point is occluded. The results of this model are shown in Figure 10.

As mentioned in the previous section, facial landmarks do not cover all the cues present in the face of a pre-term infant. Cues that relate to the tone or color of an infant can not be recognised using facial landmark techniques. Comparing Figure 3 and Figure 7 however, show that facial landmarks cover most of the ROIs determined by Carlini et al. [12]. Which does provide ample ability to recognise most facial cues.

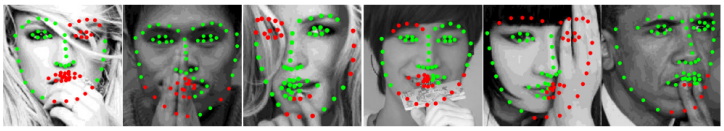


Figure 10: Facial landmark detection [8]

2.4.3 Body

Human pose estimation is a well studied subject. Its applications vary from sports analysis, gaming, to medical assistance [17]. The goal of human pose estimation is to automatically locate the position of human body parts from images or videos. Human pose estimation has to deal with pose specific challenges in pose variability and appearance variability. Apart from these specific challenges, human pose estimation has to deal with general computer vision challenges such as occlusions, truncation, and variability of image conditions [3]. A division is made between single and multi person pose estimation in the human pose estimation field. This research focuses on single person pose estimation. Single person pose estimation can be divided into two techniques, direct regression approaches and heatmap based approaches. These approaches will be discussed in the following section.

2.4.4 Single person pipeline

Wang et al. [61] explore techniques for human pose estimation using RGB-D images. RGB-D images contain two channels. The first channel is the RGB channel that contains information about the shape, color and texture of the subject. This channel is sensitive to illumination, meaning that slight changes in lighting causes colors to change. The RGB channel is used in traditional 2D CNN approaches to pose estimation. The D channel in RGB-D stands for depth. The depth channel is insensitive to lighting changes and provides additional information about the form of the subject. Additionally, depth provides information about the distance between the camera and an object which can be used to detect occlusions.

Two often used techniques for 2D pose estimation are direct regression and heatmap based pose estimation. Toshev et al. [57] propose a holistic human pose estimation as a deep neural network (DNN). The model predicts an initial pose and uses DNN-based regressors to refine the joint points by using specific higher resolution images that contain the locations of these joints. The advantage of using a regressor based model is that there is no need for a graph based model that encodes the human skeleton, as shown in Figure 6. Secondly, since the DNN uses the full image for the initial prediction, it is able to capture the full context of each body joint [57]. Heatmap based models can use underlying human priors that show the model what a human pose should look like. Chen et al. [13] present a method that uses a graphical model for human pose. In the graphical model the nodes represent the skeleton key-points as shown in Figure 6 and the edges as the pairwise relations between these nodes. The authors use a deep convolution neural network to generate heatmaps for image patches. These heatmaps show the probability of a group of pixels belonging to a skeleton key-point. Using the probabilities and the pairwise relations from the graphical model, their solution is able to generate accurate human pose estimations. Pose estimation models are most often trained on widely used publicly available datasets [17]. These datasets consist of mostly adults or children and do not include infants or pre-term infants. This causes the models that use a human prior to be biases towards adults. Therefore, pose estimations for infants is an underrepresented area of researched. The available research will be discussed in

the next section.

2.4.5 Infant specific body model

Hesse et al. [28] acknowledge the lack of research in infant pose estimation. The current field of pose estimation and body model representations of a 2D image focus on full-size body pose estimation. Therefore, Hesse et al. [28] mention the non-existence of an infant-sized body model as the first big challenge. An infant-sized body model is needed since the ratios between body part sizes differ significantly as shown in Figure 11.

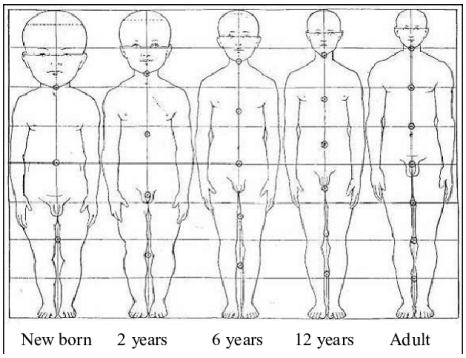


Figure 11: infant versus adult body part ratios [62]

The second challenge is the non-cooperative subjects. An infant can not be asked to strike a certain pose or to not move during recordings. The pipeline proposed by Hesse et al. [28] starts off with pre-processing the captured RGB-D images. The pre-processed images are fed into the next stage which is a registration stage. The registration stage translates the data to a common topology. A common topology is needed to learn a model. Finally, the resulting shape is predicted using Principal Component Analysis (PCA). During pre-processing the RGB-D images are transformed into a 3D point cloud. This point cloud is segmented using a simple clustering algorithm, leaving only pixels belonging to the infant and the clothing of the infant. These are classified using the color information of the RGB-D images. The resulting point cloud is transferred to the vertices of a personalized shape using the Skinned Multi-Person Linear model (SMPLb) [28]. This leaves a minimization problem on a Markov random field which is in turn solved. Leaving a clean segmentation of clothing and infant points. The authors can not match the SMPL to the base human shape since it does not generalize well to the new infant body domain [28]. Therefore, the authors use an open-source 3D character design tool to create an infant-sized body model. Further adjustments were done to the model to make it more “infant-like” including the bending of the spine and allowing the ankle to have a bigger rotation. Gradient-based optimization was used to match the points to this newly created infant body model. This process involves penalizing points that lay relatively further away from the body model compared to other points. This process

is shown in Figure 12. Personalized shapes are created for each sequence. A per-

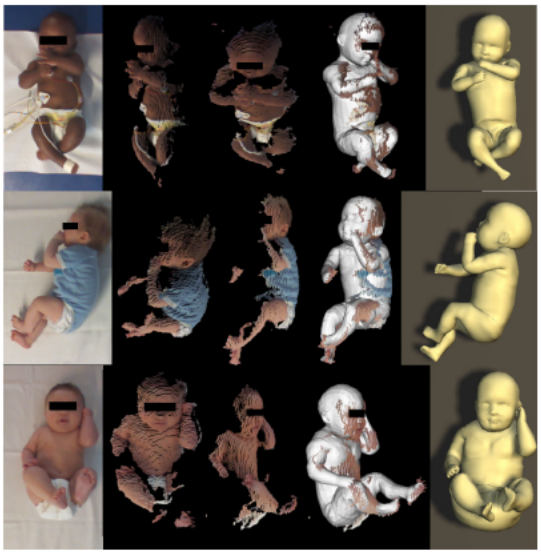


Figure 12: From left to right: RGB-D image, point cloud, point cloud second angle, point cloud with registered SMIL, and finally the rendered registration [28]

sonalized shape consists of the aforementioned steps. These personalized shapes are combined into one final infant body model using Weighted PCA (WPCA). With a weight of 1 for points belonging to clothing and a weight of 3 for points belonging to the infant. The result is a Skinned Multi-Infant Linear model (SMIL). The SMIL can be used to accurately describe an infant's body location and its current pose. Motion or movement can be determined by computing these body models for sequential frames and looking at the new locations of certain key-points on the body model. This relative movement can be used for the recognition of appetite and pain cues. According to Hesse et al., the downside of this approach is the lack of detail around the face and fingers of the infant, leaving the opportunity for missing highly specific and detailed cues, for example, squeezing of the eyes or frowning. The last unexplored modality is the vital signs modality. Vital signs and their addition to the previously discussed modalities will be discussed in the next section.

2.5 Vital signs

Vital signs are used to determine the progress or current state of a patient [21]. These vital signs traditionally consist of: blood pressure, temperature, pulse rate and respiratory rate. Margolius et al. [43] show that there is a difference of 6 beats per minute between an infant being awake or sleeping. Descriptive studies have shown that normal vital signs do not necessarily mean normal physiologic, physical or psychological function. Vital signs do change according to the mood of a person. Shu et al. [52] show that the moods happy, sad and normal can be classified using only

heart rate information in adults. Monrroy et al. [44] analyzed heart rate before and after feeding adult participants of the study. They found that heart rate increased, on average, 5% immediately after feeding, 12% 30 minutes after feeding, and 10% 60 minutes after feeding. These percentages are all compared to the pre-feeding baseline. The authors also found more pronounced differences in women compared to men. There is a great lack of research on vital sign analysis to predict states in infants. The available research is done on heart rate. Pados et al. [46] researched heart rate variability (HRV) during feeding to determine whether the infant experienced higher stress levels. The goal of the research was to find out if HRV is a more sensitive non-invasive indication of stress compared to the already established physiological signs; heart rate (HR), respiratory rate (RR), and oxygen saturation (SpO2) [49]. The study compared the aforementioned physiological signs during regular care PDF and a co-regulated approach to feeding infants (CoReg). CoReg attempts to make the feeding experience less stressful and thereby less straining on the infant. During CoReg the infant lays on its side with an elevated head position, ensuring minimal oral and tactile stimulation. Pados et al. found that SD12 was a strong non-linear predictor for stress. SD12 describes the interbeat variability. A Poincare plot can be used to plot the SD12 statistic. Spread out points indicate a high SD12 and clustered points indicate a low SD12. They describe a low SD12 as an indication for a high correlation between interbeat intervals and a high SD12 indicating a low correlation between interbeat intervals meaning more randomness in the interval between beats as shown in Figure 13. They found that SD12 is statistically significantly higher during normal feeding compared to CoReg indicating that normal feeding induces more stress compared to CoReg. Secondly, they found that SD12 is a robust, non-invasive, accurate measure of stress in infants.

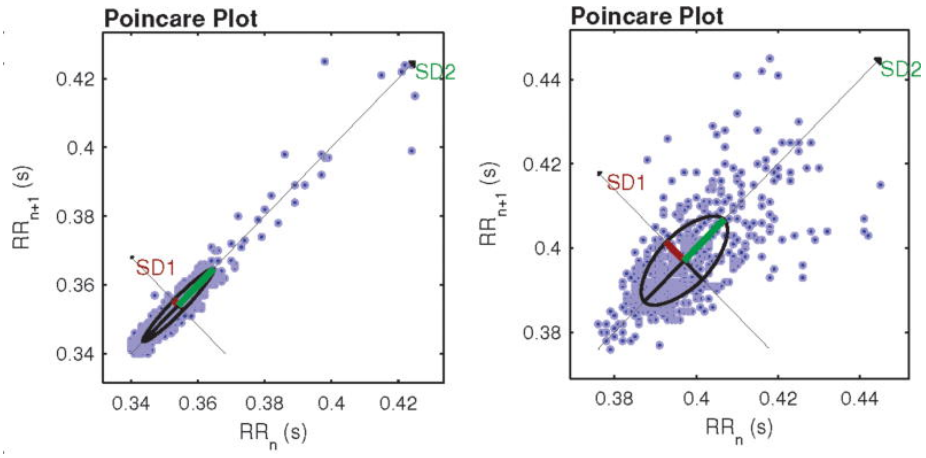


Figure 13: Differences in heart rate variability showing stress or pain: Left, low SD12 no pain or stress Right, high SD12 stress or pain [46]

2.6 Multi-modal

This research focuses on three modalities: facial cues, body movement cues, and vital sign data. These modalities need to be combined to detect the cues shown in Table 2. These cues can be used to determine whether an infant experiences pain, appetite, or fatigue(sleep) and the current sleep state. Combining these modalities is not straightforward. There are five main challenges that need to be overcome. Baltrusaitis et al. [4] outline the following five challenges. Firstly the representation, how can the data be combined to exploit the complementary and redundancy of the data. Secondly, translation, how should the data from the different modalities be translated. The translation is needed since the data is heterogeneous in nature. Thirdly, the alignment of the multiple modalities is a challenge. How can the relation between elements of the data be exploited. This can be done on a time sensitive basis, combining data points that occur at the same time but can also be done using the location of the data, combining detailed and overview information. Fourthly, how can the data be fused together to account for possible missing data or to account for the fact that some modalities might have a greater prediction power compared to others. Finally, co-learning is the concept that explores how models trained on one modality can help improve models trained on different modalities [4]. Keeping these challenges in mind, one approach could be to train three different models for each modality respectively and use some voting ensemble to come to one prediction or classification. This approach is called late fusion, where the separate models output predictions according to their modality which is later combined at the decision level [45]. As shown in Figure 14.

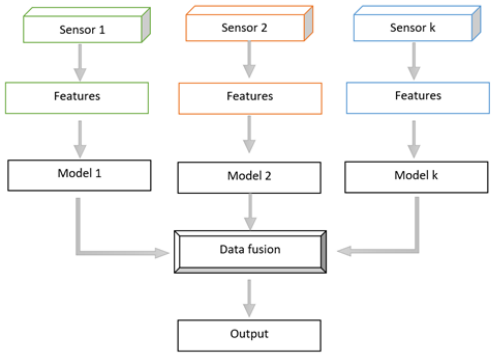


Figure 14: Late Fusion (https://medium.com/haileleol-tibebu/data-fusion-78e68e65b2d1)

Opposite of late fusion there is early fusion. Early fusion combines the inputs of each separate modality into one feature vector which is a single input value representing the data from each of the modalities. Early fusion is considerably more difficult than late fusion. Especially when the modalities are not synchronized timewise [39]. Heart rate might be supplied for every second while certain body movements might only occur once every 5 minutes. A second challenge is the importance

of each modality. A situation where an infant's heart rate increases rapidly and the heart rate variability increases might be a strong indication of pain even when the infant is not showing any of the other pain cues in its face or body. In this situation, vital sign information must be weighted higher than the other cues and should not be "drowned out" by the lack of cues from the other modalities. As shown in Figure 15.

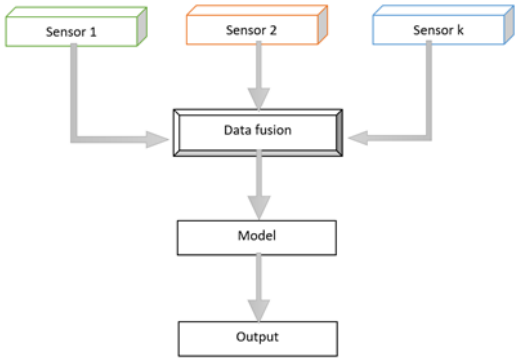


Figure 15: Early Fusion (https://medium.com/haileleol-tibebu/data-fusion-78e68e65b2d1)

Late fusion in this research can be seen as trying to predict an infant state from each modality separately, for example, it could use the body movement cues to predict a certain state, these predictions would be combined into one final prediction. This type of late fusion would reduce the usefulness of multi-modal learning. Since infants express their states through cues that span all the modalities. Thus, predicting states for each modality individually reduces the information a model has to predict a state.

Another solution could be a model equivalent of a pain or appetite scale. The scales which nurses use to determine the state of the infant all use highly similar criteria. This can be used in a model. Training three different models that output which cues were visible or shown during a certain window of time would solve two of the main challenges which is the combining of the features and the different frequencies of data collection. By using a sliding window approach where, for example, 3 minutes of footage is analyzed, the cues are detected using three deep neural networks that each predict the cues shown in the 3 minutes of video for each corresponding modality, then the window slides 1 minute over and the process is repeated. The resulting collection of cues can be combined into a single binary feature representing which cues are present and what the heart rate variability was during a window of time. These features can serve as the input for a fourth deep neural network that predicts infant states from cues and outputs which state(s) the preterm infant is experiencing. A big advantage of the previously explained solution would be the fact that the intermediate results of the three models can serve as input for traditional

infant state quantification by using scales. Secondly, the initial three cue recognition models can be easily swapped out with newer and different solutions allowing for a scalable and customizable solution.

3 Methods

This chapter consist of two parts, namely, data generation and the rule based cue detection (RBCD) program. Firstly, we will discuss the techniques that were used to generate the data for the RBCD. In the second part of this chapter we will discuss the data cleaning, parameter prediction, and the workflow of the RBCD. The RBCD program requires the location of key-points on the infants body and face. These key-points are generated by pose estimation and landmark detection models respectively. The third component required by the RBCD is vital sign data. A pose estimation model and a facial landmark detection model was used to generate the required data respectively. We will start by discussing the models that we considered for pose estimation and facial landmark detection. It is important to note that the main goal of this research is to develop the RBCD, models used for the key-point generation can be swapped when superior models are developed specifically for infants. The RBCD uses the COCO pose estimation and 168 key-point face estimation standards to allow for effortless swapping of future improved models. We have looked for a model that provided data that we could work with for the RBCD and did not try to achieve state of the art results on infants. Additionally, the RBCD can work with multiple input sizes. Therefore there will be no limitation on the output of the model whether there are 21 or 25 pose key-points or 68 to 168 facial landmarks.

3.1 Data generation and predictions

In this chapter we will start by discussing which techniques and models we used to generate the data that is used in the RBCD. Secondly, we will discuss the inner workings of the RBCD.

3.1.1 Validation

The pose estimation and facial landmark models have been manually and automatically evaluated on preterm infant with identifier 663. The recordings of infant 663 show a clear progression from the wake state through the different sleep states. Secondly, the lighting conditions are among the best of all the recordings and, additionally, the infant is minimally obstructed by tubes or fabrics for the most part of the recordings. The behaviors and cues visible in the recordings of infant 663 are representative of all the recordings. The quality of the recording however, is especially usable. The combination of the excellent lighting conditions and the lack of occlusions make it perfect for pose and facial landmark estimation.

The pre-processing of the recordings was done as follows. The recordings were rotated clockwise or counter clockwise depending on the side that the recording was

taken from. This was done to ensure that the face of the infant is upright. The videos were then padded to ensure a normal 1920x1080 resolution. Finally, every 25th frame of the recordings was extracted to serve as a validation set for the manual validation.

The validation set was used for the validation of each model. Manual validation was done on the pose estimations of each image. The manual validation consists of manually reviewing each predicted pose and determine whether it was accurate or not, to specify, all the predicted joints were compared to the actual position of these joints. If the real joints were visible and the predicted joint position was on the actual position the prediction was accepted. Real joints that are not visible due to occlusions or the viewing window were not taken into consideration during validation. The automatic validation was done on a validation set containing the first 1000 frames of a recording. The automatic validation consists of determining the change in position of each joint between frames. Ideally we would want this distance to be as small as possible. The infant is only able to move a minimal amount since the time between frames is around 33ms. Conversely, the distances should not be 0 since we do expect some minor changes. The choice was made to go with these validations since there are no ground truth labels available for the dataset.

3.1.2 Pose Estimation

For this study, multiple publicly available pose estimation models have been evaluated, namely, EfficientPose, OpenPose, and HigherHRNet. In this chapter we will evaluate these models using the metrics described in chapter 3.1.1 and discuss the results. The results of the automatic validation is shown in Table 4.

Model	Distances	Missing frames	Missing joints
EfficientPose	0.01	0	0
OpenPose	15.7	0	12.1
HigherHRNet	17.4	0	8.2

Table 4: Performance of human pose estimation n the automatic validation set

EfficientPose [24] was the first model we implemented. Implementation was straightforward and some minor changes were made to the publicly available repository [1] to allow for the input of a folder of images. EfficientPose was used to generate pose estimations of the preterm infant recordings. The model was validated on the generated validation set for manual and automatic validation. Automatic validation showed minimal changes between joint locations in sequential frames which is very promising, although in some cases the distance was 0 over multiple sequential frames which is quite unrealistic. The model achieved an average distance between frames of 0.01 as shown in Table 4. Additionally, the model did detect every joint in every frame which is not possible since the lower body of the infant is not shown in the image. The manual validation showed that the estimated poses were way off,

[1] https://github.com/daniegr/EfficientPose

(a) EfficientPose (b) OpenPose (c) HigherHRnet

Figure 16: Model comparison. Results generated on images from Salekin et al. [50, 51]

showing a folded in half body. An example result of a frame picked at random is shown in Figure 16a. These errors were present in every image in the validation set. Therefore, we decided to try a different model, namely OpenPose.

OpenPose [11] was the second model we tested for this project. OpenPose also has a publicly available repository [2]. No modifications had to be made to the available code. OpenPose showed more promising results compared to the EfficientPose model. Manual validation showed that the predicted poses were quite accurate but contained a lot of frames were the confidences of certain joints were too low to be used resulting in a large collection of frames with missing joints. OpenPose struggled with finding the location of the belly. The automatic validation confirmed these findings. On average 12.1 joints were missing per frame in the validation set as shown in Table 4. This number is reduced to 4.1 since 8 joints are not visible due to the camera viewing window. This leaves 5 joints that are visible whenever these joints are in the frame. The automatic validation also showed minimal changes in the locations of joints with high confidence scores (>0.7) over sequential frames. An example result of the OpenPose model is shown in Figure 16b. The results of the OpenPose model are quite promising but missing joints are detrimental to the RBCD. The RBCD heavily relies on the change in distances between joints over multiple frames. The quality of the output of the evaluation of a rule is severely hindered when there are multiple missing joints per frame or over multiple frames. Therefore we moved on to the next model, namely, the HigherHRNet [14] model.

The publicly available repository [3] contains a working sample of the HigherHRNet model. The publicly available code provided a framework to validate the model on

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose
[3] https://github.com/HRNet/HigherHRNet-Human-Pose-Estimation

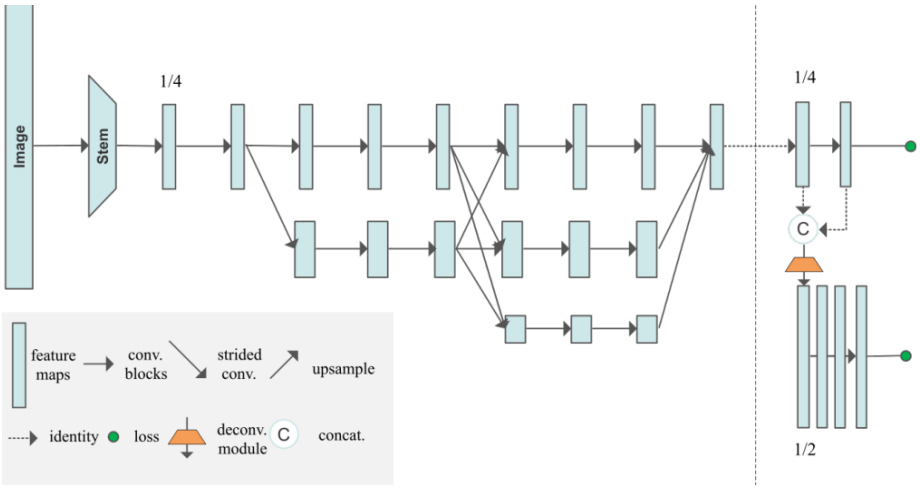


Figure 17: An illustration of HigherHRNet. The network uses HRNet as backbone, followed by one or more deconvolution modules to generate multi-resolution and high-resolution heatmaps. Multi-resolution supervision is used for training. More details are given in chapter 3.1.1. Image and caption taken from [14]

the MPII or COCO data sets. Therefore, some adjustments were made to allow for pose estimation on custom images. The automatic validation showed quite some variation in joint locations in sequential frames. As shown in Table 4, the model achieved an average distance of 17.4 which is the highest of all the models we evaluated. The model misses 8.1 joints on average which can be reduced to 0.1 since 8 joints are not visible in the images. Manual validation showed that the joints that were responsible for the largest distances between frames are joints that are not visible in the image. This can be explained since there is no threshold for minimal confidence in a joint to qualify it as a successful detection. After taking a closer look at these joints, it was clear that these joints can easily be excluded by applying a minimum threshold on the confidences. Wrongly predicted joints have a confidence of around 0.01% where the confidences of seemingly accurate joints lay around 70-80%. It is important to note that HigherHRNet is a multi person pose estimation model. Therefore, in certain frames multiple poses can be found in a single image of a single infant. The pose with the highest confidence score was used as the pose of the infant. Further post processing on the estimated poses is described in section 3.2. An example result of the model is shown in Figure 16c. HigherHRNet provided the best results out of all the models that were considered. Therefore, we picked HigherHRNet as the model that generates the poses for the RBCD. Although the output of the model is not perfect, it is good enough for the exploratory goal of this research. In a practical setting we will have to deal with noise and inaccurate measurements and we will not know the severity these issues and we might not always know before hand. The results, however, are promising enough to use this model as a base for this research.

HigherHRNet is a bottom-up model, meaning that it does not rely on a person detector that outputs a bounding box. Generally speaking, this reduces the ability to detect persons of different scales since it is not able to scale the detected bounding box to a certain size. HigherHRNet however, is created with scale sensitivity in mind. HigherHRNet uses feature pyramids to deal with scale variance. Traditional feature maps start with 1/32 resolution, the feature pyramid used by HigherHRNet starts at 1/4 resolution leading to a more detailed feature map and generates even higher resolution feature maps using deconvolution layers as shown in Figure 17.

3.1.3 Facial landmark detection

The second requirement of the RBCD is facial landmark information. This information can be used to detect cues that occur in the face of the infant, such as prolonged blinking, frowning, and yawning. Facial landmark detection is a vast field which employs a wide variety of techniques to extract the location of a certain number of key-points from a face. The number of key-points can differ significantly. Certain baselines apply a 68 key-point standard while others conform to a 168 key-point standard. The three models we are considering for this research are OpenFace, ZFace, and InsightFace. Table 5 shows the automatic validation results of the facial landmark detection models. The models will always output every key-point when the model detects a face. Therefore, this category has been left out in the table.

Model	Distances	Missing frames	Missing key-points
OpenFace	3.94	395	-
ZFace	No sequantial detections	771	-
InsightFace	43.7	53	-

Table 5: Performance of facial landmark detection on the automatic validation set

OpenFace [67] is the first model we considered for this research. OpenFace uses a rigid facial landmark detector called a Convolutional Experts Constrained Local Model (CE-CLM). This approach is mainly based on the popular use of CLM's. CLMs excel in dealing with occluded faces by using local models and constraints for each individual landmark. Recently CLMs have been outperformed by cascading regression approaches. CE-CLM is a combination of a convolutional export network (CEN) and the aforementioned constrained local model. CE-CLM uses a CEN as the local model which, according to the authors, should be better suited to deal with the complex variation in facial landmarks. OpenFace is publicly available on github[4]. The setup of OpenFace is straightforward and can be used out-of-the-box. An example of the output of the OpenFace model is shown in Figure 18a. The automatic validation of the generated facial landmarks indicate that there is little inter key-point variation between sequential frames which is positive. As shown in Table 5 the model however, does not detect a face in 395 of all the frames which is significant. The manual validation shows the same. On frames where the infant's head pose is turned to either side, the predictions are stable. The predicted key-points were determined to

[4] https://github.com/TadasBaltrusaitis/OpenFace

(a) OpenFace (b) ZFace (c) Insightface

Figure 18: Model comparison. Results generated on images from Salekin et al. [50, 51]

be inaccurate during the manual validation. Instead of predicting the key-points on the face the model predicts the key-points at the center of the screen, meaning that the model is not able to deal with significant pose changes of the infant's face. These frames could be filtered out since the pose prediction capabilities of OpenFace indicate a sideways facing face. This will however, result in too many missing frames. The missing frames shown in Table 5 do not include frames we would filter out. Therefore, we decided to move to the next model, namely, ZFace [33].

The ZFace model uses a 3D cascading regression technique. The model first attempts to predict the location of a set of landmarks and their visibility. Then the model attempts to fit a 3D model of a face to these initial markers. The method makes no assumptions about illumination and surface properties and should be able to handle a variety of poses. ZFace is publicly available on github [5]. ZFace is implemented in Matlab and required significant modifications for a usable environment. ZFace was used for facial landmark detection on the sequential and non subsequent validation set. The automatic testing showed that there are no subsequent frames to detect distance differentials in. Additionally, ZFace does not detect a face in 771 frames. The manual validation showed that the model is confused about the lower part of the infant's face. This is especially clear in Figure 18b. The figure shows that the bandage on the chin of the infant is confused for the mouth. The predictions in the validation set show that the model is uncertain about the mouth location and alternate between a correct prediction and the prediction shown in the figure. The location of key-points around the mouth are vital for the detection of pain and hunger. Therefore, we made the decision to implement a third model, namely InsightFace [20].

[5] https://github.com/AffectAnalysisGroup/AFARtoolbox

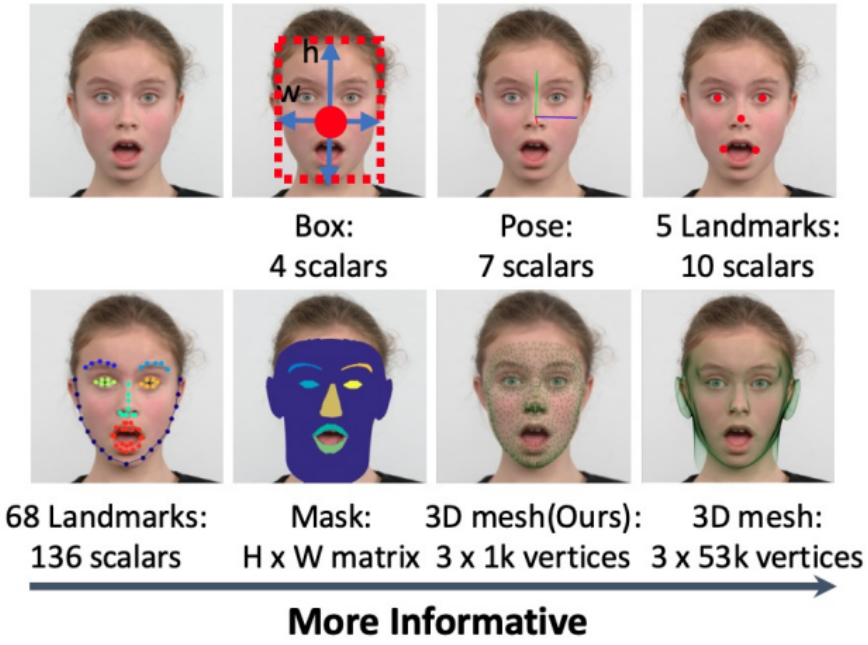


Figure 19: An illustration of the coarse to fine approach of InsightFace [20]

InsightFace uses the RetinaFace [20] model. RetinaFace applies a single shot, multi level face localization technique. This model is able to detect facial landmarks. RetinaFace uses a coarse to fine approach as shown in Figure 19. Compared to ZFace, RetinaFace attempts to first predict the locations of 5 key-points. Then expand the 5 key-points to 68 key-points and finally fit a mask to these key-points. ZFace attempts to fit a mask to an initial set of key-points thus skipping the step that adds more key-points. Multiple different techniques are used to generate these different outputs. All of these tasks aim to establish the semantic correspondence between different face images. Therefore, these tasks are combined in a unified framework where they are trained jointly to ensure that the different techniques complement each other. InsightFace showed promising results during the automatic and manual validation, especially the ability of the model to accurately determine the exact location of the eyelids and the mouth. This is shown in Figure 20. The manual validation showed that there are scenarios where only part of the face is visible, either due to obstructions or occlusions by limbs or tubes, that the model is not able to rigidly detect facial landmarks for. This led to some significant distances between subsequent frames with rapid movements as shown in Table 5. This can be managed since the model is able to achieve the best performance with regards to the eyes and mouth compared to the two other models. Additionally, Table 5 shows that InsightFace is able to detect a face more often than any other model we evaluated. Manual validation showed that the predictions around the eyes and mouth are accurate but as

shown on Figure 20 that some landmarks around obstructed areas around the outline of face can shift a bit.

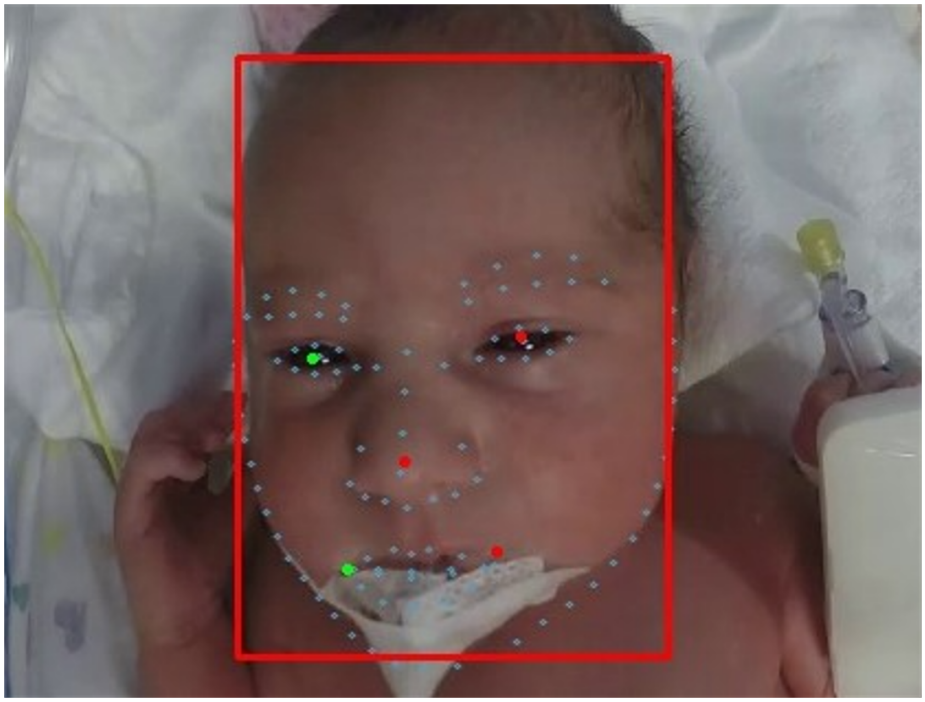


Figure 20: A close up of the InsightFace output on images from Salekin et al. [50, 51]

3.2 Rule based cue detection

3.2.1 Introduction

The goal of this research is to find out whether or not a rule based approach is able to robustly detect cues that infants portray. To detect cues we need to classify our quantitative measurement data as the corresponding cues. Since these cues are scarce, we cannot rely on a deep learning model to detect these cues since there is simply not enough data to train a model. Therefore, we decided to take a rule based approach. A rule based approach allows the experts in the field to build rules that encode their expert knowledge in cue detection.

3.2.2 Tool

To facilitate the rule based approach, we developed the Rule Based Cue Detection tool (RBCD). The RBCD allows an expert or a layman to create a rule that is able to detect cues from the landmark and pose estimation data. The tool uses three kinds of blocks to describe a rule. Namely, inputs, operators and outputs. The user is able

to connect these blocks to build rules as expansive as needed. The inputs that are available to a user are facial landmarks, pose estimation joint locations, numbers, and rules that the user has already created. The input data is generated by the models discussed in chapter 3.1. An example of available operators are the euclidean distance between points, logic operators, mathematical operators, and equality and inequality operators. The outputs correspond to the aforementioned states with a number of frames or seconds threshold. Additionally, a graph output is available which allows the user to graph the evaluation of a rule. A list of all the available inputs, operators and outputs is given in the Appendix.

A rule build in the RBCD can be seen as a acyclic tree. The RBCD iterates over all the links and blocks present in the current rule setup and generates an acyclic graph where the nodes correspond to the blocks and the paths of the graph correspond to the links between blocks. The RBCD can export these graphs as a JSON file which can be used in the evaluation script. The evaluation script traverses the acyclic graph recursively until the entire graph is filled. The program populates the input nodes with the corresponding values. When the program encounters an operator it checks whether the child nodes of the operator contain values. When this is the case the program executes the operator on the child values. When a child does not contain a value the program steps into that child to evaluate it. The rule is evaluated when the output node contains a value.

3.2.3 Inputs, operators and outputs

A rule in the RBCD consists of inputs, operators and outputs. The inputs are the coordinates of a key-point. This can be the x and y location for 2 dimensional predictions or the x, y and z coordinates for 3 dimensional predictions. The key-points generated by the facial landmark model and the pose estimation model can flawlessly be combined as long as they have the same dimensions. Numbers and the output of previously created rules are also available as inputs. Using the outputs of existing rules are especially useful while creating extensive and complex rules.

As mentioned before the RBCD supports four groups of operators. Namely, distances between points, logic operators, mathematical operators, and equality and inequality operators. Distances or the change in distances over time are useful when the user attempts to detect certain small behavioral cues. For example, whether or not a mouth has opened far enough to be classified as a yawn. Mathematical operators like addition, subtraction, multiplication and division are available to the user to manipulate or combine results of earlier operators. The aforementioned operators can be used to continue our example of a yawn detection rule. The distance between the eyelids can also be used for yawn detection. One can subtract the distance between the eyelids from the distance between the lips to get a higher value whenever the infant opened its mouth while simultaneously closings its eyes. This is a stronger indicator of a yawn compared to just looking at the mouth. The user also has the ability to use the coordinates of earlier or later frames by using the "skip frames" operator. This is useful whenever the user attempts to use the change in distances over time. Logical operators and equality operators like smaller than, big-

ger than, "and", and "or" are available to the user. The user can use these operators to only evaluate a part of the rule when the condition is true. For example, use the change in distance between the wrists and the head to determine whether or not the arm is moving. If the wrists are not visible the user can use a logical operator to instead use the distance between the elbow and the head.

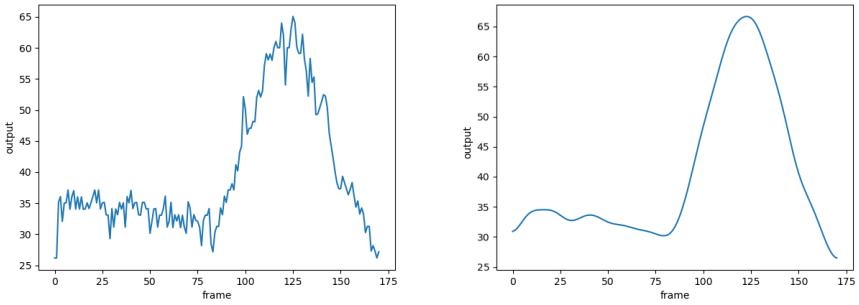
All rules end in an output. These outputs correspond to the three infant states or a user defined output. Additionally, the user can use the "graph" output. This output generates a graph of the value passed to the output node. This can be used to evaluate the effectiveness of the rule. The output nodes can also use a threshold of frames or seconds to output true, only when the rule is evaluated as true for the set number of frames or seconds.

3.2.4 Pre-processing

The raw output of the facial landmark and human pose estimation models can be quite jittery. The distance between the upper and lower lip is plotted for a fragment of a video that is annotated as yawn in Figure 21a. A yawn is generally a smooth motion. Therefore, ideally, the graph would show a smooth curve. However, this is not the case. The figure shows a jagged line which can ruin the evaluation of certain rules. For example, If the user builds a rule that determines whether or not the mouth of an infant is closed by using a certain threshold that indicates a closed mouth. Then a jagged curve around the heights of threshold might result in alternating between true and false rapidly even if the mouth of the infant is stationary for the entire time of the evaluation. The model that generates the landmarks on which the rule is evaluated predicts each frame individually, leading to predictions that can shift a few pixels each frame, which in turn leads to this jagged line. Therefore, we decided to apply a median filter with a window of 7. A median filter is able to deal with extreme noise and missing values. Additionally we apply a 1 dimensional Gaussian filter with a sigma of 5 on the output of the rule. When the rule with the filters is evaluated on the same fragment, Figure 21b is generated. Leading to a smooth curve which more accurately describes the movement of the infant. These filters are not required and the user can determine whether or not to use them based on their use case. A filter can reduce the details in slight changes in values and thus should only be used when it would not negatively impact the evaluation of a rule.

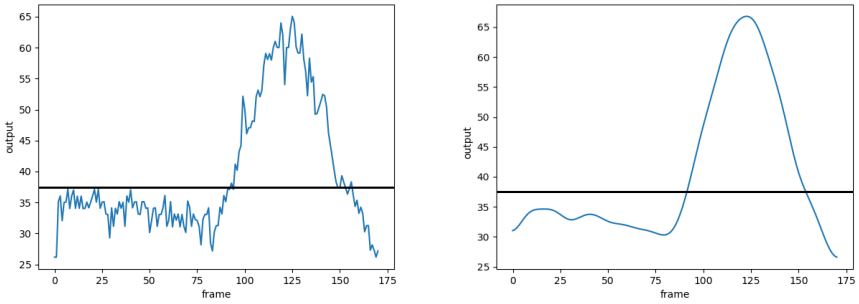
The real issue with using the non pre-processed data arises when we try to threshold the data to classify a certain state. Say we want to detect a yawn in the fragment. Ideally, we would want to establish a threshold that lies just above the resting mouth position of the infant. The same graphs as shown before are shown in Figure 22.

Figure 22a shows the non-smoothed output with a threshold at 37.5. The rule would evaluate true for 3 separate segments. The first segment is the actual yawn between frames 80 and 145. The second and third segments lie after the initial yawn between 145 and 155. The additional segments are a result of the noise generated by the model. Figure 22b shows the smoothed evaluation of the rule with a threshold at



(a) Raw rule output (b) Smoothed rule output

Figure 21: Pre-processing rule output



(a) Raw rule output with threshold at 37.5 (b) Smoothed output with threshold at 37.5

Figure 22: Pre-processing rule output with thresholds

37.5. The rule would evaluate true for only the segment that lies between 80 and 150 thus resulting in the detection of a single yawn. The exact value of a threshold is less important after filtering and any value that is able to differentiate between yawns and non-yawns should be sufficient. The only thing that will change is the start and end time of the yawn.

The second pre-processing step is to rectify the missing frames and low confidence key-points. To start off, we can see frames with low confidence scores as missing frames. Manual evaluation of the predicted key-points showed that low confidence scores are synonymous for bad predictions. We have four options when it comes to filling in missing frames. Firstly, we can decide to not try to fill in missing frames since we do not have accurate information about these frames. Secondly, we can copy the last known frame and use it to fill in the missing frames. This is a valid approach especially if only 1 or 2 frames are missing. Thirdly, we can attempt to generate the missing data ourselves. We can generate the estimated pose data by

taking the frame before the missing frames segment and the frame that follows it and interpolate values from the start frame to the end frame and evenly space the generated numbers based on the amount of missing frames. This can also be done with a median filter. This approach is especially feasible when we miss frames over the duration of a few frames or seconds. Anything longer than that will not be an accurate representation of the infant's movements since the infant could perform multiple different movements in that time frame. Finally, we can decide that there is simply not enough data to evaluate a rule upon. This option is regrettably needed in situations where we only have predictions for a small percentage of the frames.

To ensure as much customizability as possible for the user, we have decided to implement all the aforementioned options and let the user decide which option to use. The best option is often dependent on the data. The user would get a better experience evaluating rules on noisy data with a higher sigma value for the Gaussian filter. We created a script that assigns a completeness score to the input data to help the user decide which missing values option is the best for the selected data. The completeness score is a function that returns the percentage of complete frames based on the confidence threshold the user provided. The distribution of the number of subsequently missing frames will also be reported.

3.2.5 Rules

In this sub-chapter we will walk through the complete process of building and evaluating a rule that is able to detect yawns.

An example of the rule that was used to generate the graphs shown in Figures 21 and 22 is shown in Figure 23. The rule shown in Figure 23 attempts to detect yawns. A yawn can be described as an infant opening its mouth wider than a certain threshold. This rule uses the distance between the left and right part of the nose as the threshold.

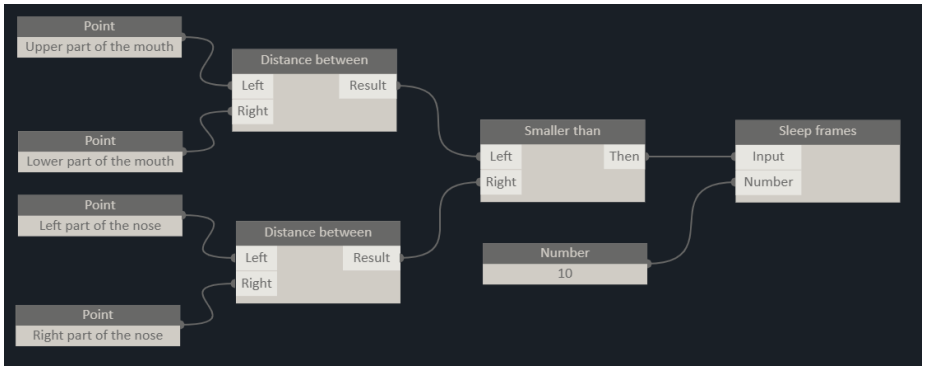


Figure 23: A rule build in the RBCD that tries to capture yawns

A more common approach to detect yawns is to calculate the Mouth Aspect Ratio (MAR) [26]. MAR encodes the ratio between the vertical and horizontal openness of

the mouth. A rule calculating and plotting MAR is shown in the appendix in Figure 39. The output of this MAR rule is a graph.

The graph shown in Figure 24 is generated by evaluating a yawn rule on a fragment of the video data that has been annotated as a yawn. The graph's y-axis shows the distance between the upper and lower part of the mouth. The graph's x-axis shows the frame numbers. The graph shows the mouth of the infant going from a resting position to fully opened and back to a resting position.

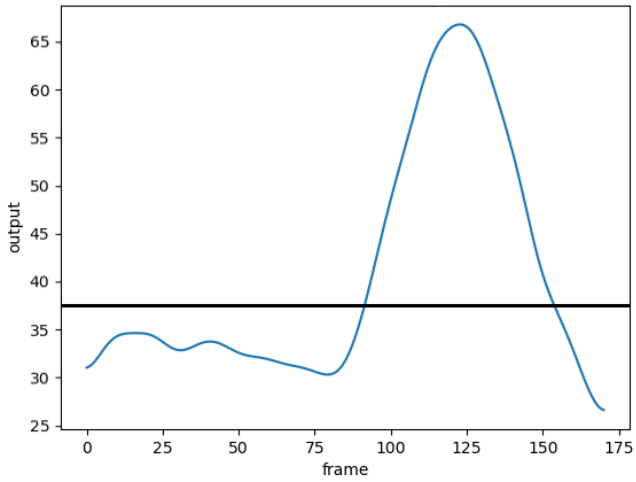


Figure 24: Output of the Yawn Rule

The data upon which the rule is evaluated does not represent a real world scenario since we would not know when a yawn would occur. To simulate a real world scenario, we evaluated the MAR rule on a 3 minute video containing multiple yawn annotations. The result of this evaluation is shown in Figure 25. The black vertical bars indicate fragments of the video that are annotated as a yawn. The graph shows that the highest peaks lie in the segments that have been annotated as yawns.

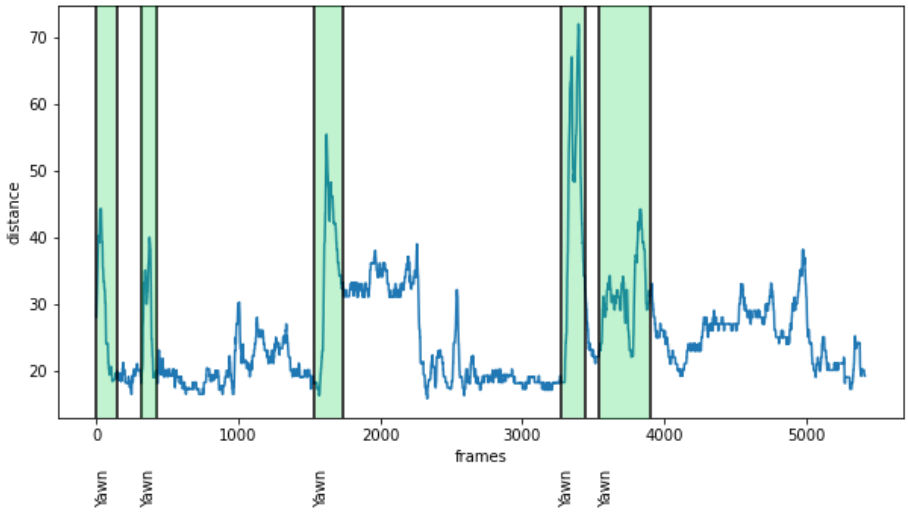


Figure 25: Yawn rule evaluated on 3 minute video

Figure 26 shows the same graph but with the 37.5 threshold as used in the previous sub-chapter. The graph shows that the five yawns that are present in this video are detected. Figure 25 also shows that there is a prolonged spike around the 2000 frames mark that is classified incorrectly as a yawn. Therefore, we can conclude that the height of the line is not only an indication of a yawn but can be an indication of other mouth related cues. The yawns share a different characteristic namely, a fast growth. The fast growth is the fast transition from a closed mouth to a fully opened mouth. This fast growth can be quantified by the derivative.

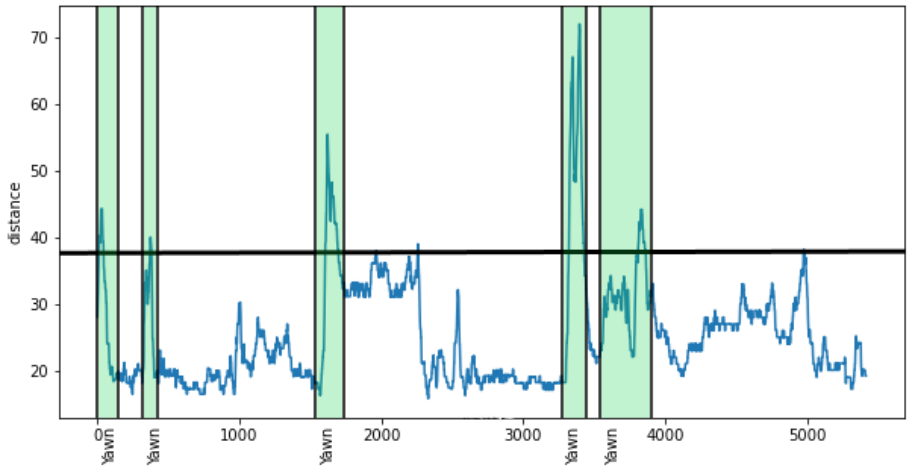


Figure 26: Yawn rule evaluated on 3 minute video with threshold at 37.5

The results of plotting the derivative of the yawn rule are shown in Figure 27. The higher values around the 2000 frames mark have disappeared. The graph of the derivative allows us to detect all yawns without detecting any non-yawns as yawns.

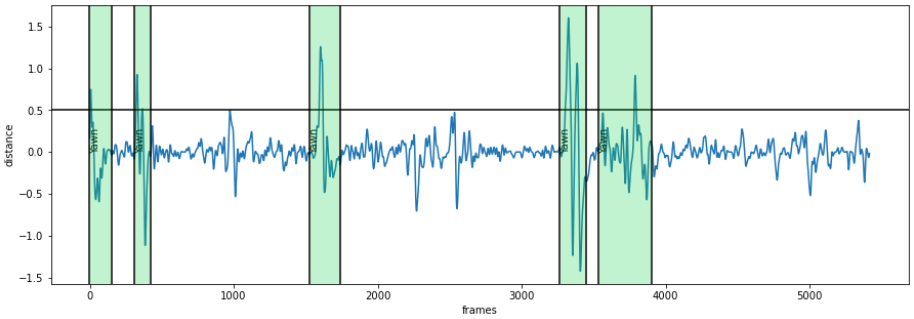


Figure 27: Derivative of the Yawn rule with a threshold at 0.5

4 Experiments

The goal of this exploratory research is to determine whether or not we are able to detect cues with a rule based approach. Therefore we want to test the ability of medical professionals to build rules in the RBCD and whether or not these rules can rigidly detect cues. Therefore, we decided to carry out an experiment that should be able to quantify whether or not the RBCD can be used for this purpose. The participants are all trained medical professionals that are familiar with the preterm infants. The participants will be required to build rules for 4 different cues. These rules will be evaluated on a 15 minute video.

4.1 Participants

The experiment will be carried out by three medical professionals. The experts are familiar with preterm infants and have been involved in annotating the data. The participants do not have any experience with building rules to detect cues. One of the participants has been involved in annotating the cues in the training and test set. The other participants have only been instructed how to annotate a video but have not been involved in any annotations at the time of the experiment.

4.2 Cues and modalities

The participants will all be required to build rules for two cues in the face modality and two cues in the body modality. The vital sign modality has been left out since there are still issues with the correct authorization to access this data. The participants will be creating rules for the yawn and frown cues in the face. Additionally, the participants will be creating rules for the head movement and arm movement body cues.

4.3 Train and test data

The participants are supplied with a 6 minute sequential video. The cues present in the 6 minute video are shown in Table 6. The train video is overlayed with both the landmark and pose estimation data. The participants have access to a video showing the facial landmarks and pose estimations. The participants will also be notified if they are looking at a certain cue in the video. The rules created by the participants will be evaluated on a 15 minute sequential test video. The cues present in the train video are shown in Table 6. The cues visible in the test video are shown in Table 7.

Facial cues		**Body Cues**	
Cues	*Count*	*Cues*	*Counter*
Eyes open and moving	70	Gross body movement	12
Smacking	7	Breathing movements	11
Yawn	6	**Arm movement**	11
Frown	6	High muscle tone	9
Mouth movement	7	**Head movement**	6
Eyebrow movement	1	Small hand movement	3
Squeezed eyes	1	Hand reflex (jerk?)	2
		Stretch	3
		High muscle tension	2
		Coughing reflex	1
		Stretch hand	1
		Arm jitter	1
		Small jerk	1
		High muscle tone	1
		Jitter - High muscle tone	1
		Arm jerk/jitter	1
		Gross body movements	1

Table 6: Cues present in the training set

Facial cues		**Body Cues**	
Cues	*Count*	*Cues*	*Counter*
Eyes open and moving	86	Gross body movement	26
Smacking	21	Breathing movements	21
Yawn	19	**Arm movement**	16
Frown	5	High muscle tone	16
Mouth movement	19	**Head movement**	6
Eyebrow movement	4	Small hand movement	4
Squeezed eyes	1	Hand reflex (jerk?)	9
		Stretch	9
		Arm jitter	1
		Jerks	8
		Arm jerk/jitter	5

Table 7: Cues present in the test set

4.4 Experiment setup

The participants are able to view the output of their rule an unlimited number of times to accurately simulate a real world environment. The participants are also allowed to make changes to their rules after they have seen the output. The participants will build the rules in the RBCD. The interface of the RBCD is shown in Figure 28. The face in the image allows the user to determine which key-points correspond

to each position in the face. The participants can add components for a rule by using the "Add an" button group. The participants can click an added component and connect it to a different component by following it up with a click on the correct side of an operator. When the participants have created a rule they can save the rule and export it in JSON format. The JSON file is then imported in a python script that evaluates the rule on the training data. The results of the evaluation are then shown to the participant which can be used to update the rule.

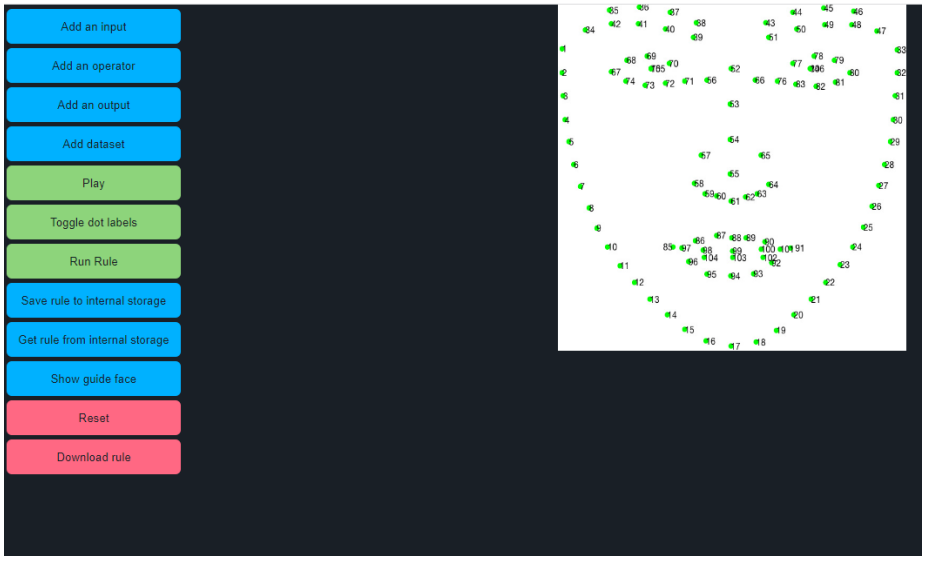


Figure 28: The interface of the RBCD

5 Results

In this section we will start by briefly describing the rules build by the participants. Secondly, we will describe the resulting evaluations on the training and test-set. Thirdly, we will conclude with describing the overlap between the number of cues that were found, not found, and wrongly identified as a cue. Finally, we will discuss whether or not majority voting outperforms their individual counterparts.

The rules built by the participants during the experiment are shown in the Appendix. The results of the evaluation on the training set are shown in Table 9 and the evaluation on the test set is shown in Table 10. The legend of both of the tables is shown in Table 8.

5.1 Participants rules

In this chapter we will illustrate the differences between the rules of each participant by describing the rules per cue.

5.1.1 Yawn

Participant 1 built a rule that should be able to detect yawns by plotting the distance between the lower side of the upper lip and the higher side of the lower lip. This results in the smallest distance between the lips. Participant 2 decided to use the distance between the left side of the upper lip and the middle of the lower lip. Participant 3 used the distance between the right side of the upper lip and the left side of the lower lip.

5.1.2 Frown

Participant 1 built a rule for the detection of frowns by calculating the distance between the left eyebrow and the upper part of the nose. This rule should classify frowning if the distance is smaller than a threshold. Participant 2 built a rule using the distance between the most right part of the right eyebrow and the lower left part of the left eye. Participant 3 summed the distance between the right eyebrow and the right eye with the distance between the lower right side of the lower lip and the upper left side of the upper lip.

5.1.3 Arm movement

Arm movement annotations consist of a general movement without specifying whether the left or the right arm moved. Therefore participants were instructed to create the same rule for the right and left arm. We will only summarize the rule created for the left arm. Participant 1 attempted to detect arm movements by determining the distance between the left wrist and the left shoulder. Participant 2 tried to detect arm movements by calculating the distance between the left wrist and the nose of the infant. Participant 3 used the distance between the left wrist and the left ear.

5.1.4 Head movement

Participant 1 attempted to detect head movements by utilizing the distance between the eyebrows. Participant 2 used the distance between the left shoulder and the infant's left cheek. Participant 3 divided the distance between the nose and the left shoulder by the distance between the left ear and the left shoulder.

5.2 Evaluation

In this chapter we will describe the evaluation of the rules described in section 5.1. The evaluations of the rules are shown in Tables 9 and 10. Secondly, we will show the overlap between the rules of the participants. The overlap between the cues that the participants detected on the training set is shown in the appendix in Table 20. The overlap between detections on the test set is shown in the appendix in Table 21. Ids are assigned to each cue to refer to them in this chapter. This chapter will be reflected upon in Chapter 6, the discussion.

Cue	-
True positives	-
False positives	False Negatives

Table 8: Legend for the results tables

Train	Yawn		Frown		Head		Arm L		Arm R	
Participant 1	3		4		1		1		5	
	0	3	13	2	0	5	1	10	16	6
Participant 2	3		4		6		11		10	
	6	3	21	2	14	0	5	0	41	1
Participant 3	3		6		1		2		10	
	0	3	20	0	0	5	1	9	35	1

Table 9: Evaluation of the experiment rules on the training set

Test	Yawn		Frown		Head		Arm L		Arm R	
Participant 1	7		3		0		0		0	
	1	12	29	2	1	6	2	16	8	16
Participant 2	19		3		6		11		12	
	14	0	37	2	74	0	15	5	185	4
Participant 3	7		5		3		0		13	
	0	12	99	0	43	3	1	16	228	3

Table 10: Evaluation of the experiment rules on the test set

5.2.1 Yawn

Participant 1 detected 3 of the 6 yawns in the training set and 7 of the 12 yawns in the test set and detected 1 yawn where there were none. Participant 2 also detected 3 of the 6 yawns in the training set but also classified 6 instances as yawns that were not annotated as yawns. Participant 2 detected all 19 yawns in the test set but also classified 14 instances as yawns that were not annotated as a yawn. Participant 3 detected 3 of the 6 yawns in the training set. Participant 3 detected 7 of the 19 yawns in the training set and detected no yawns where there were none. We can note that participant 2 detects overall more instances than the other participants. This is mainly due to the fact that the best threshold for participant 2 leads to less conservative predictions compared to participants 1 and 3.

The rules of participants 1, 2 and 3 share significant overlap and every rule depends on the distance between the lips of the infant. Therefore we also see significant overlap between the cues that are detected by each rule. All of the participants detected and missed the same yawns in the training set.

Train	Accuracy	Precision	Recall	F1-score
1	0.911	1.0	0.109	0.196
2	0.906	0.561	0.241	0.337
3	0.91	1.0	0.098	0.178
Test	Accuracy	Precision	Recall	F1-score
1	0.857	1.0	0.081	0.15
2	0.884	0.842	0.315	0.459
3	0.854	1.0	0.06	0.06

Table 11: Performance measures of the yawn rule on the training and test set

Table 11 shows the performance measures of the yawn rules built by the participants. The table shows low recall score for every rule. This is due to the fact that an annotation spans the entire yawn from opening to the closing of the mouth. When participants depend on the distance between the lips and threshold this distance, they will only detect part of the yawn annotation. The precision of the rules built by

participants 1 and 3 is remarkably high. The evaluation of the rules made by these participants do not contain any wrong predictions. However, they have missed some cues. This allows us to have confidence in a yawn prediction, but we should know that the rules might miss yawns. The rule of participant 2 makes more mistakes but does score higher on the recall measure. The evaluation of the rule makes more mistakes but is also able to detect more cues. This is also illustrated in Figure 29 and Figure 30.

Yawns 2, 3 and 6 are detected by all of the participants as shown in Tables 20 and 21. The video shows that during these yawns the infant's face is clearly visible. This results in accurate key-point predictions that follow the real movements of parts of the face. For example the key-points around the mouth follow the rapid opening of the mouth and are precisely placed upon the lower and upper side of each lip. This allows rules that rely on the distance between the lips to detect yawns. Yawn 1, 4 and 5 show the same mouth opening behavior of the mouth during a yawn. Yawn 1 starts off with the tongue of the infant obstructing the lower lip. Resulting in the lower lip key-points being predicted around the tongue which stays close to the upper lip. During yawn 4, the infant's face is pressed into the mattress and only the left part of the mouth is visible. This results in the key-points not moving during the opening of the mouth and thus showing a closed mouth while it is opening. The video of the last missed yawn in the training set shows that the infant closes its eyes fully, frowns but does not open its mouth fully. Figure 29 illustrates the overlap between the participants and the ground truth on the training data.

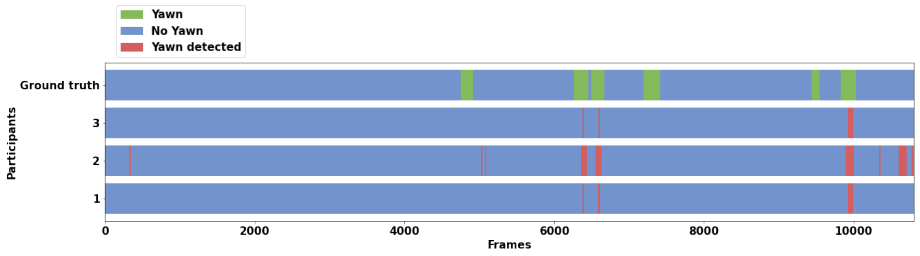


Figure 29: Participant detections overlap of the yawn cue on the training set compared to the ground truth

The detections of the yawns in the testing set alternate between yawns detected by only participant 2 and yawns detected by all of the participants. The videos where all of the participants have detected yawns all contain a yawn that is not obstructed by a hand, blanket or by the tongue. Especially the tongue obstructing the lower lip leads to bad predictions. The tongue gets confused with the lower lip which leads to the predictions of the lower lip around the tongue. This results in the predictions not showing the full open mouth but only a slight opening, since the tongue is close to the upper lip. We can also see in the recording that extreme movements of the head moving side to side leads to worse predictions than when the head remains stationary. Figure 30 illustrates the overlap between the participants and the ground truth on the test data.

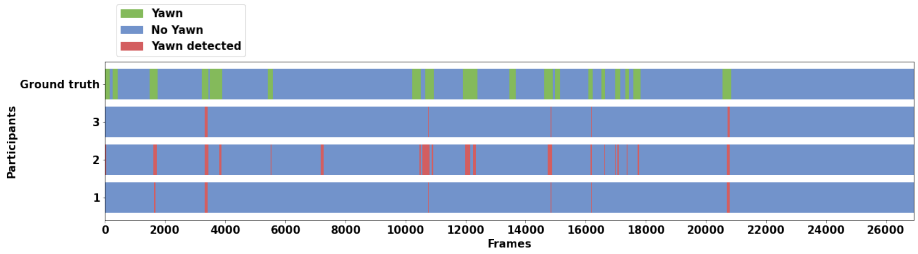


Figure 30: Participant detections overlap of the yawn cue on the test set compared to the ground truth

5.2.2 Frown

Participant 1 detected 4 of the 6 frowns but also detected 13 non-frowns as frowns in the training set. The rule evaluated on the test set found 3 of the 5 frowns but also classified 29 instances as frowns that were not annotated as a frown. Participant 2 detected 4 of the 6 frowns in the training set while obtaining 21 false positives on the training set. The test set evaluation detected 3 of the 5 frowns but also 37 false positives. Participant 3 found 6 of the 6 frowns in the training set and contained 20 false positives. The evaluation on the test set found 5 of the 5 frowns but contained 99 false positives.

Train	Accuracy	Precision	Recall	F1-score
1	0.917	0.282	0.268	0.275
2	0.898	0.218	0.285	0.247
3	0.238	0.063	0.866	0.118
Test	Accuracy	Precision	Recall	F1-score
1	0.938	0.036	0.038	0.037
2	0.914	0.047	0.09	0.061
3	0.254	0.035	0.856	0.067

Table 12: Performance measures of the frown rule on the training and test set

Table 12 shows that the rule of participant 3 performs the worst in the accuracy, precision and f1-score measures. On the recall however, the participant scores the highest. This is due to the non-conservative threshold. Participants 1 and 2, score similarly in all the measures with participant 1 scoring slightly higher on the training set and participant 2 scoring higher on the test set.

Participant 3 detected all of frowns in the training set. Participants 1 and 2 only detect frown 2, 3, 4, and 5. As mentioned before, all participants obtain more false positives than true positives. This is mainly due to the difficulty of detecting eyebrows on preterm infants. The eyebrows of preterm infants are barely visible due to the thin and light colored nature of the hairs. The key-points that correspond to the eyebrows are predicted between the eyebrows and the eyes. The inaccurate pre-

dictions also lead to challenges in detecting cues that occur around the eyebrows. The video of frowns 1, 5 and 6 show a clear downward trends of the eyebrows and a large wrinkle appears at the upper part of the nose of the infant. The video of frowns 2, 3 and 4 show only slight movements of the eyebrows and the distinct wrinkle above the nose does not appear. Every rule depends on the distances between the eyebrows and the eyes. The inaccurate detections of the key-points around the eyebrows do not result in a reliable way to detect frowns. Figure 31 illustrates the overlap between the participants and the ground truth on the training data. Figure 31 shows that participant 3 classifies movement as frowns very easily. This explains why participant 3 is able to find every cue. This could be useful if it is extremely important to find every cue and false positives do not matter. This is however, an unlikely scenario.

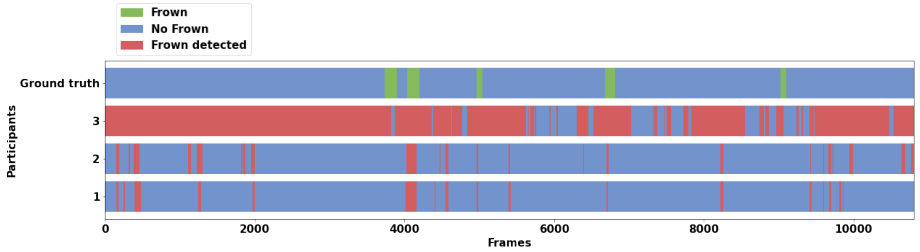


Figure 31: Participant detections overlap of the frown cue on the training set compared to the ground truth

The first frown in the test set shows the typical wrinkle at the top of the nose and a strong frown. This video however shows that the predictions around the eyebrows do not follow the actual movement of the eyebrows and thus the distance between the eyes and eyebrows does not change during the frown. The videos of frowns 2, 4 and 5 show strong frowns with accurate movements of the predictions around the eyebrows. As shown in Table 21, these are also the frowns that are detected by all of the participants. The video of Frown 3 shows that part of the face of the infant is obstructed. During the obstruction, the predictions around the eyebrows do not follow that actual movement of the eyebrows. Therefore, rules that depend on the distance between the eyebrows and the eyes have a harder time detecting these cues. Figure 32 illustrates the overlap between the participants and the ground truth on the test data. We can see that participant 3 follows the same pattern of classifying a data point as a frown as often as possible.

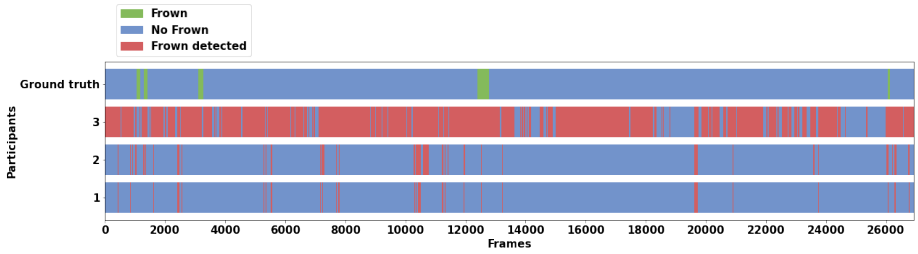


Figure 32: Participant detections overlap of the frown cue on the test set compared to the ground truth

5.2.3 Arm movement

Participant 1 detected 1 of 11 arm movements with 1 false positive on the training set. Participant 1 detected no arm movements in the test set where there were 16 but detected 2 false positives. Participant 2 detected 11 of the 11 arm movements with 5 false positives on the training set, on the test set the participant found 11 of the 16 arm movements and detected 15 false positives. Participant 3 detected 2 of the 11 arm movements with 1 false positive on the training set. The test set evaluation found 0 of the 16 arm movements with 1 false positive.

Train	Accuracy	Precision	Recall	F1-score
1	0.847	0.45	0.041	0.076
2	0.491	0.202	0.796	0.322
3	0.851	0.603	0.057	0.105
Test	Accuracy	Precision	Recall	F1-score
1	0.928	0.0	0.0	0.0
2	0.604	0.082	0.447	0.139
3	0.927	0.0	0.0	0.0

Table 13: Performance measures of the arm movement rule on the training and test set

Table 13 shows the performance measures of all the rules of the participants. Participants 1 and 3 score similarly on the training and test set. This is also shown in Figures 33 and 34. Participant 2 achieves the highest recall and f1-score out of all of the participants. Figure 33 illustrates that this is due to the large amount of positive predictions made by the rule. The performance measures of participants 1 and 3 on the test set show that the rules do not generalize well to the unseen data.

The arm movements can be categorized into three main categories. The first category is the wild swinging of the arms. Cue 1 and 10 belong in this category. The second category is very subtle movements. Cue 2, 3, 7, 8, and 9 belong in this category. The last category are movements that start from or end outside of the viewing angle of the camera. Cue 4,5 and 6 belong in this category. The participants have been

instructed to create a rule that should be able to detect movements. As described in chapter 5.1.3, the rules of the participants describe the distance between the arms and an anchor point. The anchor point can, for example, be a shoulder or the nose. When rules use these distances, it is non trivial to detect arms that move towards the anchor point. These movements would result in a downwards curve during the annotation. Downwards curve are not detectable by a threshold that detects upwards curves. Figure 33 illustrates the overlap between the participants and the ground truth on the training data. The figure shows that participant 2 is not conservative with arm movement predictions.

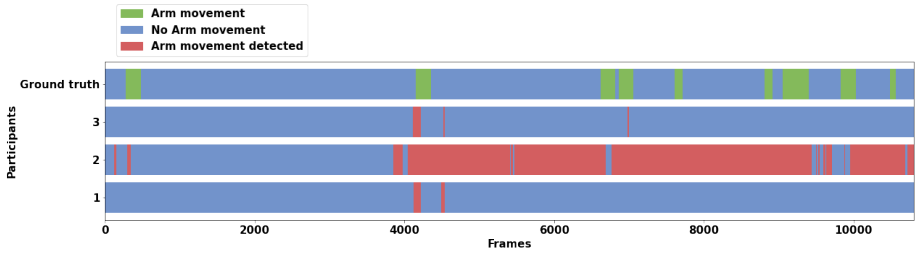


Figure 33: Participant detections overlap of the arm movement cue on the training set compared to the ground truth

The first arm movement in the test set is the movement of the right arm which starts underneath the blanket and ends above the blanket, and a small twitch in the left arm of the infant. The second arm movement is fully covered under the right blanket with only the fingers of the right arm showing that the arm has moved. Arm movement 3, 4 and 5 all occur with the left elbow outside of the viewing angle of the camera. The model however is able to determine the position of the elbow. Arm movement 6 shows that only the elbow moves while the shoulder and wrist remain stationary. The videos that show arm movements 7, 8 and 9 show movements of the elbow while the hand obstructs the view of the elbow due to the fact that it is pointed directly at the camera. Cues 10 and 11 show the infant stretching its left arm fully outside of the camera view box therefore, these cue are not detectable. The remaining videos (cues 12-16) show the left arm of the infant moving away from the face in a slight downwards direction. These videos show that the arm movement is very subtle and only moves slightly. Figure 34 illustrates the overlap between the participants and the ground truth on the test data.

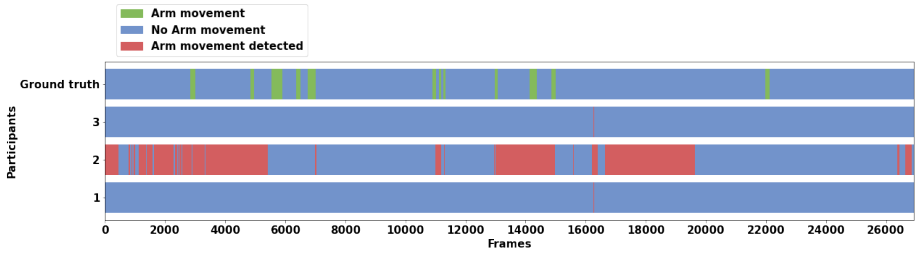


Figure 34: Participant detections overlap of the arm movement cue on the test set compared to the ground truth

5.2.4 Head movement

The head movement evaluations showed that participant 1 found 1 of the 6 head movements in the training set and 0 of the 6 head movements in the test set with 1 false positive. Participant 2 detected 6 of the 6 head movements with 14 false positives in the training set and detected 6 of the 6 head movements in the test set with 74 false positives. Finally, participant 3 detected 1 of the 6 head movements in the training set and 3 of the 6 head movements in the test set with 43 false positives.

Train	Accuracy	Precision	Recall	F1-score
1	0.924	0.929	0.016	0.031
2	0.424	0.083	0.649	0.148
3	0.924	1.0	0.017	0.033
Test	Accuracy	Precision	Recall	F1-score
1	0.951	0.0	0.0	0.0
2	0.535	0.087	0.912	0.16
3	0.757	0.033	0.142	0.054

Table 14: Performance measures of the head movement rule on the training and test set

Table 14 shows a large difference between the performance on the training and test set. Additionally, The rules of participants 1 and 3 score the highest on the training set in the precision and accuracy metrics. The rule of participant 3 performs the best on the test set. Table 13 shows that participant 2 scores the highest on the test set. The arm movement rule of participant 2 and the head movement rule of participant 3 both use a low threshold which shows to lead to higher scores on the test set.

The video of the first head movement annotation shows the upwards movement of the head of the infant while the rest of the body moves in the same direction. The infants left shoulder lifts off the mattress in a shocking fashion. Participant 2 is the only participant that detected this movement. The participant used the distance between the head and the left shoulder. Therefore this participant is able to detect the

movement of the head. The video of the second head movement shows the infant coughing and waving its arms wildly. The quality of the predictions suffer under these movements. Facial landmark predictions disappear for more than half of the frames during the annotation. The videos of the annotations of head movement 3,4, and 5 show very slight movements of the head compared to video 1 and 2. The infant moves its head from fully to the right to 1 centimeter towards the middle. This leads to the need for more sensitive rules. Video 6 shows the infant yawning. At the end of the yawn the infant extends its head upwards towards the left side. Figure 35 illustrates the overlap between the participants and the ground truth on the training data. The figure shows that participant 2 does not predict head movements conservatively.

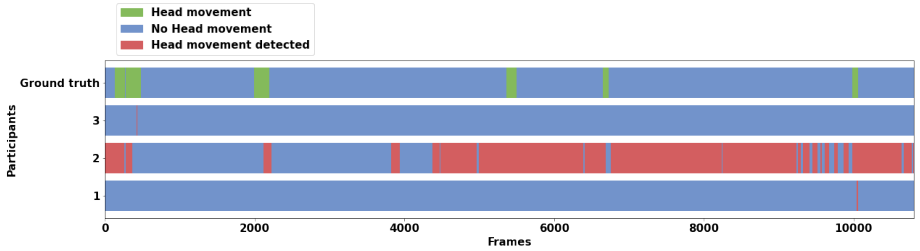


Figure 35: Participant detections overlap of the head movement cue on the training set compared to the ground truth

The first head movement in the test set is during a yawn. Where the head moves upwards a small amount in a stretching fashion. The second head movement is during a stretch and the remaining head movements are slight left to right motions of the head. We can see in Table 21 that head movements 1,2 and 6 are picked up by the rules of participants 2 and 3. The remaining movements are only picked up by participant 2. The rule of participant 1 focuses on the distance between the eyebrows. The nature of a head movement is such that the entire head should move. During the movement of the head the key-points inside the face are not able to detect this movement. Since every point is affected in the same way by the general movement for each point in the face. Figure 36 illustrates the overlap between the participants and the ground truth on the test data.

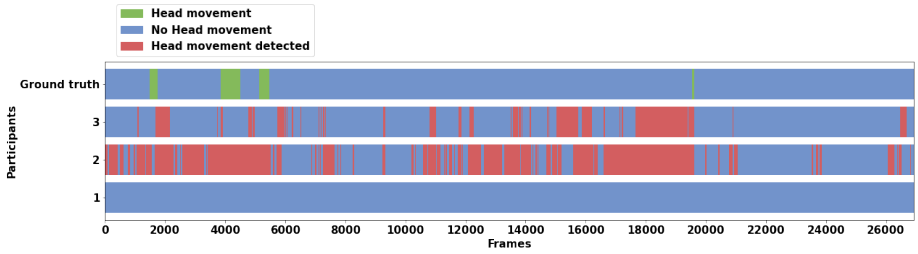


Figure 36: Participant detections overlap of the head movement cue on the test set compared to the ground truth

5.3 Inter rule reliability

Reliability and predictability of similar rules are essential to the usability and practicality of a rule based system. Therefore, we determined the inter rater reliability between the rules of the different participants. The inter rater reliability per rule per participant on the training data is shown in Table 15. The inter rater raliability per rule per participant on the test data is shown in Table 16. The tables show that similar rules receive a high inter rule reliability (IRR) score. As described in chapter 5.1.1, the yawn rule of participant 1 and 3 are highly similar. We expect that the evaluation of these rules also score a high IRR. Tables 15 and 16 show that the rules of these participants score an IRR of 0.9987 and 0.9965 respectively. The head movement rule from participant 1 is very dissimilar to the rules of participant 2 and 3. The IRR score between participant 1 and 2, and 1 and 3 is 0.4 on the training data and 0.5 on the test data. Thus similar rules lead to a high IRR and dissimilar rules lead to a low IRR. This allows us to determine that the rules build in the RBCD are reliable and predictable.

Rules	1-2	1-3	2-3
Yawns	0.9681	0.9987	0.967
Frowns	0.9568	0.2508	0.2719
Left arm movement	0.415	0.994	0.4155
Right arm movement	0.6254	0.611	0.9479
Head movement	0.3993	0.9974	0.4019

Table 15: Inter rule reliability on the training set

Rules	1-2	1-3	2-3
Yawns	0.9545	0.9965	0.9512
Frowns	0.9712	0.2646	0.2921
Left arm movement	0.6122	0.9991	0.6123
Right arm movement	0.5065	0.5219	0.901
Head movement	0.4947	0.7918	0.6321

Table 16: Inter rule reliability on the test set

5.4 Majority voting

In this section we will attempt to increase the performance of the rules by implementing majority voting. Majority voting should increase the confidence in the predictions. False positives that are only found by a single participant will not be present in the majority predictions. Additionally, detections that are made by multiple participants will be present and therefore we would theoretically have more confidence in the predictions. The downside of majority voting is that if we have one participant whose rule is significantly better than the other two. We would lose the accurate predictions of this participant if they are not present in the under performing rules. The results of implementing majority voting on the training data is shown in Table 17. The legend shown in Table 8 applies here too. The results of majority voting on the test data is shown in Table 18. When we compare Table 18 with Table 10 we can see that the amount of false positives increases in some cases. This is due to the fact that combining rules that contain false positives that span a large number of frames with rules that contain false positives that span a small number of frames leads to an increase in false positives.

Yawn		Frown		Arm L		Arm R		Head	
3		4		2		10		1	
0	3	27	2	1	9	41	1	0	5

Table 17: Majority voting on the training set

Yawn		Frown		Arm L		Arm R		Head	
7		3		0		12		3	
1	12	34	2	2	16	214	4	39	3

Table 18: Majority voting on the test set

Table 17 shows that the yawn cue finds the same 3 cues as the participants and finds no false positives. The frown cue shows that 4 out of the 6 cues are found. These cues are found by participants 1 and 2 as well. The false positives however, increase from 14 and 21 respectively to 25. The arm cues perform similarly to participant 3. The head cue performs similarly to participant 3 as well. The figures in section 5.2 show that the frown, arm movement, and head movement cues contain predictions

by a participant that is not conservative. Majority voting is able to deal with these large sections that are classified as a cue. Figure 37 is the same as Figure 31 with the addition of the majority voting results. The figure shows that the large sections predicted by participant 3 are not present in the majority voting. The impact of the lesser quality of the rule of participant 3 is still present. Every false positive of different participants that occur during the large sections automatically achieve a majority. We can conclude that majority voting allows us to have a higher confidence in the predictions if we have no prior knowledge of the quality of the rules.

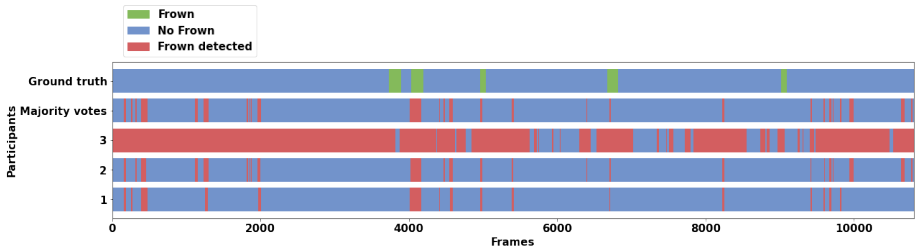


Figure 37: Participant detections overlap of the frown cue on the training set compared to the ground truth, additionally containing the majority voting results

Train	Accuracy	Precision	Recall	F1-score
Yawn	0.912	1.0	0.11	0.198
Frown	0.887	0.192	0.287	0.23
Arm	0.848	0.505	0.057	0.103
Head	0.578	0.056	0.279	0.093
Test	Accuracy	Precision	Recall	F1-score
Yawn	0.857	1.0	0.082	0.151
Frown	0.913	0.046	0.09	0.061
Arm	0.927	0.0	0.0	0.0
Head	0.52	0.011	0.097	0.019

Table 19: Performance measures of majority voting for each cue on the training and test set

Table 19 shows the performance of majority voting on the training and test set. The performance on the yawn cue achieves the same high precision as participants 1 and 3 as shown in Table 11. Additionally the accuracy, recall, and f1-score increases. Participant 2 outperforms majority voting on these categories but, as discussed in section 5.2.1, makes more mistakes and thus achieves a lower precision score. We can conclude that majority voting outperforms the rules on their own for the yawn cue.

The performance of majority voting on the frown, arm movement. and head movement rule shows the impact of the low quality rules. Majority voting is only able to significantly outperform the accuracy and precision measures of the rules that uti-

lize a low threshold. As discussed in section 5.2, rules that utilize low thresholds lead to a large number of false positives leading to a low precision score. Majority voting is able to limit the size and amount of false positives.

We can conclude that majority voting outperforms individual high quality rules, by reducing the false positives found only by a single rule. Additionally, majority voting is able to outperform rules that utilize a low threshold by combining these rules with higher quality rules.

5.5 Experiment conclusions

We can conclude that current facial landmarking and pose estimation models are often not quite good enough to accurately and robustly predict key-points on preterm infants especially during occlusions. As mentioned in chapter 5.3, the most common reason that rules do not lead to accurate predictions is the lacking quality of the key-points. The rules build using the RBCD, can lead to rules that are able to detect cues. Specifically, the participants are able to detect the yawn cue reliably as shown in Table 9 and 10. Therefore, we can conclude that rules build in the RBCD are able to detect cues whenever the rule is of sufficient quality and the key-points are accurate during the entire duration of the cue.

5.5.1 Face and body modality

We have seen in section 5.2 that facial cues with accurate key-points are detectable by rules build in the RBCD. The rules for the yawn cue achieves the best scores on the training and the test set. The frowns cues could be solved in the same manner as the yawn cues, by determining the distance between the eyebrows and the eyes. These rules however, do not perform equally well. This is mainly due to the fact that the facial landmark detection model struggles with the close to invisible eyebrows that are present on the infant. The main problem with the body cues is that the participants have not found the right way to detect movement. The issues with the current quality of the movement rules is described in section 6.3. Secondly, the recordings contain a lot of obstructions when it comes to the body parts of the infants. The lower part of the infant is never visible and the key-points on the right arm are only visible for 10% of the video. The obstructions play a lesser role in the key-points on the face. This reduces the inaccuracy of the facial landmarks. This leads to a curve that contains less noise that is caused by the key-points shifting significantly. This in turn allows for the more accurate prediction of the best threshold for a certain rule.

5.5.2 Specific cues and general cues

We have chosen two specific and two general cues to use in the experiment. The specific cues are the yawn and frown cues. The general cues are the arm movement and the mouth movement cues. We can see that the participants are generally on the same line when it comes to the different cues with some outliers per cue. We attempted to determine whether or not participants would perform significantly bet-

ter on specific or general cues. However, due to the limited quality of the rules for the general cues, we cannot with confidence conclude that participants would perform better on either category. However, it does appear that specific cues perform better than general cues. This is due to the higher quality of annotations for these cues. This will be discussed in section 6.2.

6 Discussion

During this research we have acquired insights into the current issues in the preterm infant monitoring field. The most significant issues will be discussed in this chapter. Secondly, we will reflect on this research. Finally, we will propose possible improvements to this research and the RBCD.

6.1 Automatic rule generation

During this research we have researched a rule based approach to cue detection from key-points. We have determined that the rules created by the participants are able to detect cues whenever the key-points are accurate. One advantage of the rule based approach is that the rules are explainable. Therefore, we could attempt to automatically create rules without human interference. There are a few challenges that need to be overcome before we can attempt this. The challenges will be discussed in the following chapters 6.1.1 and 6.1.2.

Automatic rule generation can be achieved in a wide variety of ways. We would require a large dataset with annotations of a large variety of preterm infants of different gestational ages. This is required to ensure that rules generalize well over different age groups. Secondly, we have to decide whether or not we would want to incorporate expert knowledge into a rule. The advantage of incorporating expert knowledge is that we will only have a few combinations that need to be evaluated. Without expert knowledge we would try all of the combinations between key-points. The disadvantage of using expert knowledge is that current biases that influence certain cues would not be present during the evaluation of the generated rules.

Automatic rule generation can also be done under the current circumstances. This will however be less ideal than the aforementioned implementation. We need to determine the current identifiability of cues. This can be done by creating rules for each cue present in the annotations and evaluating these rules on the data. This would supply us with the data needed to determine how well a cue can be detected. Then we would pick the most identifiable cues and apply a brute force approach to rule creation, where every combination of key-points and operators will be used to create rules. The rules would then be evaluated and compared to the original results of the human made rules to determine whether or not an automatic approach is viable.

6.1.1 Data quality

The quality of the data provides us with a challenge. Recordings of infants in the NICU often contain significant occlusions. Tubes, fabrics or bandages are placed on and around the infant. This reduces the robustness and quality of key-point predictions. Therefore, we are in need of models that are able to cope with these occlusions. This can be done by pre-processing the data to deal with missing detections and detecting inaccurate predictions based on the previous and next detections. A second approach would be to use a model that is able to deal with these occlusions or that provides key-point specific confidence values to provide information about which key-points can not be relied upon.

6.1.2 Data quantity

Currently, we only have access to one recording of 45 minutes of a single infant. This is due to the fact that the other recordings did not meet the quality standards or contain significant occlusions. Infants that require additional oxygen will have most, if not all, of their face obstructed. The facial landmarking models we have evaluated can not accurately predict key-points with these significant obstructions. If we were to attempt to create rules automatically, we would need a large amount of high quality data. Secondly, recording infants in the NICU is not trivial. Recordings need to be planned in accordance with the nurses instructions, to not hinder any treatment that the infants receive. Additionally, the parents need to sign a contract to allow recordings to be made of the infant. Furthermore, there is no strict schedule or time that the infant has to spend on the NICU. This leads to planned recordings being cancelled when the infant is allowed to leave the NICU. All of these issues hinder large amount of data collection.

6.2 Annotations

During the experiment the participants noted that there were some inconsistencies in the annotations. For example, certain yawns were also annotated as a head movement. These yawns, are yawns where the infant moves its head either by stretching or by moving from side to side. There was no consistency in the head annotations during yawns. Certain yawns with large movement contained a head movement annotation while other with the same degree of movement did not. This led to rules picking up head movements that fit inside of the definition of a head movement but are not annotated as such.

This scenario also occurs in the other cues. Yawns are defined as a mouth opening, the eyes closing, and a frown or stretch of the forehead. The yawn annotations however, do not always show each of these behaviors during the yawn annotation. For example, the first yawn annotation in the second video shows that the infant opens its mouth. Before the mouth is fully closed the infant frowns and closes its eyes. The annotation during this yawn stops before the infant closes its mouth fully and before the frown and the closing of the eyes starts. This led to issues when rules depend on each component of the yawn. Therefore, the annotations are in need of a stricter

definition. The stricter definition should be followed consistently. A new type of annotation should be made, whenever the infant shows a cue that does not conform to the original definition.

6.2.1 Human machine annotations

As mentioned in chapter 6.2, the current annotation workflow should be adapted to produce higher quality annotations. We can use the RBCD to update the definition of an annotation. Annotators should start with the current annotation workflow. Watching the video and annotating each fragment with the cue that is visible in that fragment. When the entire video is annotated, the annotators should build a rule according to their definition of a cue in the RBCD. The annotators can then evaluate this rule on the video. The evaluation of the rule should show all annotated instances of the cue. The rule should be updated when the evaluation does not show a certain annotation. The annotator should watch fragments of the video where the evaluations show a cue that is not annotated. The annotators should then decide whether or not they have missed the annotation of this cue or whether or not an exception to the rule should be added. Thus, all outliers will be accounted for.

We can use the RBCD to deepen our understanding and definition of infantile state cues. The subjectivity problem mentioned in chapter 1.1, is resolvable by combining human annotations and machine predictions. Humans are able to interpret the wide variety of behaviors an infant uses to show a cue. While machines are able to detect each instance of these behaviors. Humans can in turn deepen their definition and interpretations about these cues by analyzing each detected instance. Thus, furthering our knowledge related to the behaviors infants use to express their internal state.

6.2.2 Sub-annotations

A secondary approach to the challenges that occur in the annotations, is to use lower level annotations. Instead of a yawn annotation consisting of multiple parts, only annotate each component. For example, annotating each prolonged closing of the eyes, each stretch of the forehead, and each opening of the mouth. Then rules should be created for each component.The overlap between evaluation of these rules should be used to go from sub-annotations to annotations. Ideally these sub-annotations would be specific. For example, there would be multiple sub-annotations for a mouth opening based on the speed at which the mouth opens and the distance that the mouth is open. This approach reduces the subjectivity that hinders the current annotation quality.

6.3 The RBCD in a medical environment

A big advantage of the rule based approach is the explainability it provides. The rules created in the RBCD are easily understandable by humans. This is especially an advantage in the medical field, where a model's decisions that are straightforward to interpret, can support medical professionals in making diagnoses and assist in

treatment related decision making. As mentioned in chapter 1.1, infants are only actively monitored for a short period. Which can lead to misdiagnosis and allows nurses to miss pain cues. Therefore a system that is able to actively monitor infants can reduce the frequency and impact of missed cues. A system that is able to actively monitor an infant is described in the next paragraph.

Infants in the NICU should be actively monitored by a camera system. Key-points for the video should be automatically predicted in batches. Rules that attempt to detect cues should be evaluated whenever a batch of key-points is available. A dashboard needs to be provided where the nurses can view the detected cues. This dashboard should allow the nurses to see a compact overview of all the cues that were found for a certain infant. The dashboard should also allow the nurses to view a fragment of the video that contains a certain cue detection to make sure that the rules detect the right cues. Additionally, alerts should be added that can notify nurses when the infants shows a certain cue or combinations of cues. For example, this can be used to alert nurses to a sudden onset of pain or during appetite.

6.4 Rule based cue detection

The experiments have shown that the participants do not always use the most accurate ways to detect certain cues. This is not always due to the subjective definition of a cue but can also be due to not knowing how to represent a certain behavior in a cue. For example, all 3 of the participants, as mentioned in chapter 5.1.3, created rules that use the distance between a point on the arm and an anchor point in or on the face. Evaluating such a rule would lead to a graph with high curves when the arm moves away from the anchor point, and low curves when the arm moves towards the anchor point. Detecting movements from such a curve by using a threshold is not trivial. A better solution would be to determine the derivative. This can be done by calculating the change in distance between sequential frames. A derivative that is close to 0 would indicate that the arm is not moving while a high derivative would indicate that the arm is moving. Secondly, the absolute of the derivative should be taken to handle the arm moving towards the anchor point. Another challenge that arises on the arm rule of participant 1 is that the anchor point is on the left shoulder. This results in the infant being able to move its arm without the distance between the wrist and the left shoulder changing.

We can not rely on medical professionals knowing the best ways to detect certain behaviors. Therefore, additional inputs and operators should be added to the RBCD that can assist users in building higher quality rules. Especially, an operator that calculates the derivative and inputs that provide predetermined anchor points can improve the experience of the user and additionally aid the quality of the rules. Furthermore, instructions should be provided to the users on how to calculate commonly used metrics and the best ways to detect certain behaviors like movement and repetitive behaviors.

6.4.1 Parameter prediction

The RBCD allows users to automatically predict the best threshold for certain behaviors. The users should supply annotations that correspond to the behavior that should be detected. The RBCD is capable of calculating the F1-score, accuracy, precision, and recall metrics. The user is able to decide which measure should be used to optimize the threshold. The threshold is calculated by equally spreading a 1000 values between the minimum and maximum value in the evaluation. Then calculating the selected optimization measure and finally, returning the threshold that led to the highest optimization measure. The quality of the rule can be determined by calculating the best threshold on the training part of the annotations and evaluating this rule on the test part of the annotations.

Additionally, the RBCD would benefit from a multi parameter prediction capability. As mentioned in chapter 6.2, cues can consist of multiple components. The RBCD should be able to determine the best way to combine these different components. This can be done by detecting patterns in the evaluations of the different components during the cue annotations. A pattern present during yawn annotations is that, the distance between the lips increases and the distance between the eyelids decreases. Therefore, subtracting the distance between the eyelids from the distance between the lips leads to a curve that should contain spikes during times where the mouth is open while the infants eyes are closed. Which is an indication of a yawn. The time it takes to create high quality rules would decrease substantially if the RBCD is able to determine the best way to combine these components of cues.

6.4.2 Additional modalities

In this research we have utilized the face and body modalities. Additional modalities might increase the performance of certain rules. As discussed in section 2.5, the addition of the vital signs modality might improve the performance of rules. We have discussed in section 2.3.1 that vital signs can be a predictor for certain sleep states. Vital signs can be used to limit the amount of false positives by only predicting a certain cue if the vital sign data matches a certain condition.

The base rates of cues should also be considered as an additional modality. We can increase or decrease the sensitivity of a rule by detecting patterns in the occurrences of yawn cues. If yawns are more likely to occur in close proximity to other yawns then we could lower our detection threshold once a yawn has been detected. We could re-iterate over the data with the lower threshold before a yawn has been detected to ensure that it is the first yawn in the group. This can also be done on a higher level for each cue. By determining which cues often occur together, we can increase or decrease the sensitivity of certain rules in accordance with the chance that the corresponding cue might occur.

6.5 Research questions

In this section, we will reflect on the 3 research questions that were introduced in chapter 1.

6.5.1 Do current publicly available facial landmarking and pose estimation models predict robust key-points?

During this research we have compared three facial landmarking and pose estimation models. The best performing models are InsightFace and HigherHRNet respectively. Evaluating these models was not trivial. We did not have access to a key-point annotated dataset of preterm infants. Therefore, validation had to be done manually and automatically with no ground truth labels. We have seen in chapter 5, that low quality key-points are detrimental to the cue detection rules. These models have proven themselves to be reliable during recordings that do not contain any occlusions. During occlusions the reliability of the models decreases. HigherHRNet does not depend on each body part being visible. This is a big advantage since preterm infants are often partly occluded, either by blankets or medical equipment. InsightFace is able to provide robust predictions of parts of the face that are not occluded. During occlusion events, as mentioned in section 5.2.1, the reliability of the key-points decreases. Therefore, we must give a two sided answer to this question. On the one hand the models perform remarkably well on the preterm infant field. On the other hand, they do suffer from occlusions which decreases the reliability and robustness of the predictions.

The experiment has shown that rules build in the RBCD can detect cues that preterm infants show. These rules rely on the predicted key-points. Therefore, we can conclude that the models are able to predict robust enough key-points during high quality non-occluded recordings. Conversely, the experiment has shown that cues that occur on particularly challenging areas, such as the eyebrows of a preterm infant, are more challenging to detect. The eyebrows of an infant are nearly invisible which causes the model to shift its predictions for the eyebrows significantly between each frame. All in all, the rules are able to rely on the key-points generated by the models whenever unreliable data is discarded. Therefore, medical professionals should only rely on the evaluation of the rules when the data has been extensively checked. Additionally, the RBCD intentionally uses well-known standards to allow for the implementing of new and improved models.

6.5.2 Are rules build in a rule based cue detection program able to detect cues annotated by human annotators?

During the experiment we have found that the rules build in the RBCD are able to detect cues annotated by human annotators. The amount of cues detected by each participant and why they were or were not detected is extensively described in chapter 5. We can conclude that a rule based approach is a valid approach when we only have access to low quality and low quantity data. The rules are able to detect human annotated cues during times of little to no occlusions and when the rules are of suf-

ficient quality. Section 6.4 proposes improvements to the RBCD to reduce the effort it takes to create high quality rules. Additionally, section 6.2 proposes improvements to the workflow of the human made annotations. Furthermore, the aforementioned section explains different types of annotations that would improve the quality of the evaluations of the rules even further.

6.5.3 Does the performance of cue detection rules increase if we apply majority voting to the evaluation of the rules?

In section 5.4 we have explored whether or not the performance of the rules built by the participants increases when we apply majority. We have seen that the number of true positives does not increase. This is however expected behavior since majority voting does not add any information to the evaluation but improves its reliability. Figure 37 shows that majority voting removes large sections of false positives. All in all, majority voting decreases the amount of frames that are false positives. Thus increasing the confidence we can have in the evaluation of rules. Ideally, we would test majority voting on a larger set of rules since the impact of low quality rules is currently a significant issue. Additionally, applying majority voting to rules that depend on distinct components of a cue might improve the results even further.

6.6 Future Applications

The RBCD program we have developed during this research opens the door to a wide variety of different applications. Firstly, the RBCD can be applied in developing a new type of annotation-model. The RBCD is able to efficiently detect behaviors in large amounts of data. The detected behaviors can be used to identify groups of behaviors that often occur in parallel. This results in sub-groups of the same annotation. For example, the RBCD could be used to detect yawns in preterm infants. Infants can yawn in a wide variety of ways. The RBCD can be used to detect patterns in the yawns and create sub-categories for each type. The sub-categories can be utilized as more accurate annotations that provide a more detailed description of a behavior.

The RBCD can also be used to classify large amounts of unlabeled data. The RBCD is able to create a queryable dataset, by creating rules that are able to detect specific movements or behaviors. For example, this can be used to count the number of throw ins, passes, and free kicks occurring in a soccer match. An additional advantage of the system is that it is able to detect behaviors without having to be trained. This can be used to label or classify the large amount of unlabeled publicly available datasets. The detected behaviors in both of the examples can be used to index the data, thus resulting in a dataset that can be queried by the behaviors present in them.

The application of the RBCD inside the medical field can be expanded. Currently, we have applied the RBCD only to the NICU environment but more research needs to be done into other applications. For example, the progress of a patient's rehabilitation could objectively be measured by rules created in the RBCD. Additionally, the

RBCD could be used to identify early cues of diseases or conditions that show in patients that are admitted to ICU. This would allow the hospital staff to preemptively react to an unforeseen situation.

Finally, the RBCD could play a supporting role in new traditional research. Constructing databases that contain information about every behavior a preterm infant showed in the NICU could lead to new insights in patterns that are present in preterm infants that result in certain diseases or conditions later in life. Additionally, the database can be used to find patterns in the success or failure of treatments.

References

[1] Cornelieke Sandrine Hanan Aarnoudse-Moens, Nynke Weisglas-Kuperus, Johannes Bernard van Goudoever, and Jaap Oosterlaan. Meta-analysis of neurobehavioral outcomes in very preterm and/or very low birth weight children. *Pediatrics*, 124(2):717–728, August 2009.

[2] M Ahmadpour-Kacho, Y Zahed Pasha, Z Hahdinejad, and S Khafri. The effect of non-nutritive sucking on transcutaneous oxygen saturation in neonates under the nasal continuous positive airway pressure (CPAP). *International Journal of Pediatrics*, 5(3):4511–4519, 2017.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[5] A Bik, C Sam, ER de Groot, SSM Visser, X Wang, ML Tataranno, MJNL Benders, A van den Hoogen, and J Dudink. A scoping review of behavioral sleep stage classification methods for preterm infants. *Sleep Medicine*, 2022.

[6] Susanne Brummelte, Ruth E Grunau, Vann Chau, Kenneth J Poskitt, Rollin Brant, Jillian Vinall, Ayala Gover, Anne R Synnes, and Steven P Miller. Procedural pain and brain development in premature newborns. *Annals of neurology*, 71(3):385–396, 2012.

[7] Mark Buchholz, Helen W Karl, Maureen Pomietto, and Anne Lynn. Pain scores in infants: a modified infant pain scale versus visual analogue. *J. Pain Symptom Manage.*, 15(2):117–124, February 1998.

[8] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollar. Robust face landmark estimation under occlusion. In *2013 IEEE International Conference on Computer Vision*. IEEE, December 2013.

[9] W Büttner and W Finke. Analysis of behavioural and physiological parameters for the assessment of postoperative analgesic demand in newborns, infants and young children: a comprehensive report on seven consecutive studies. *Pediatric Anesthesia*, 10(3):303–318, 2000.

[10] J Cagan. Feeding readiness behavior in preterm infants. *Neonatal Network*, 14(2), 1995.

[11] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[12] Lucas Pereira Carlini, Fernanda Goyo Tamanaka, Juliana C A Soares, Giselle V T Silva, Tatiany M Heideirich, Rita C X Balda, Marina C M Barros, Ruth Guinsburg, and Carlos Eduardo Thomaz. Neonatal pain scales and human visual perception: An exploratory analysis based on facial expression recognition and eye-tracking. In *Pattern Recognition. ICPR International Workshops and Challenges*, Lecture notes in computer science, pages 62–76. Springer International Publishing, Cham, 2021.

[13] Xianjie Chen and Alan L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. *CoRR*, abs/1407.3399, 2014.

[14] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020.

[15] CMU-Perceptual-Computing-Lab. Coco body skeleton · cmu-perceptual-computing-lab/openpose.

[16] Lilia Curzi-Dascalova, Patricio Peirano, and Françoise Morel-Kahn. Development of sleep states in normal premature and full-term newborns. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 21(5):431–444, 1988.

[17] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019.

[18] ER de Groot, A Bik, C Sam, X Wang, RA Shellhaas, T Austin, ML Tataranno, MJNL Benders, A van den Hoogen, and J Dudink. Creating an optimal observational sleep stage classification system for very and extremely preterm infants. *Sleep Medicine*, 2022.

[19] Femke de Jong, Michael C Monuteaux, Ruurd M van Elburg, Matthew W Gillman, and Mandy B Belfort. Systematic review and meta-analysis of preterm birth and later systolic blood pressure. *Hypertension*, 59(2):226–234, February 2012.

[20] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[21] David Evans, Brent Hodgkinson, and Judith Berry. Vital signs in hospital patients: a systematic review. *International journal of nursing studies*, 38(6):643–650, 2001.

[22] Ana-Maria Gallo. The fifth vital sign: implementation of the neonatal infant pain scale. *J. Obstet. Gynecol. Neonatal Nurs.*, 32(2):199–206, March 2003.

[23] Sharyn Gibbins, Bonnie J Stevens, Janet Yamada, Kimberley Dionne, Marsha Campbell-Yeo, Grace Lee, Kim Caddell, Céleste Johnston, and Anna Taddio. Validation of the premature infant pain profile-revised (pipp-r). *Early human development*, 90(4):189–193, 2014.

[24] Daniel Groos, Heri Ramampiaro, and Espen AF Ihlen. Efficientpose: Scalable single-person pose estimation. *Applied intelligence*, 51(4):2518–2533, 2021.

[25] Ruth Eckstein Grunau, Tim Oberlander, Liisa Holsti, and Michael F Whitfield. Bedside application of the neonatal facial coding system in pain assessment of premature infants. *Pain*, 76(3):277–286, 1998.

[26] Isha Gupta, Novesh Garg, Apoorva Aggarwal, Nitin Nepalia, and Bindu Verma. Real-time driver's drowsiness monitoring based on dynamically varying threshold. In *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–6. IEEE, 2018.

[27] Maureen Hack, Daniel J Flannery, Mark Schluchter, Lydia Cartar, Elaine Borawski, and Nancy Klein. Outcomes in young adulthood for very-low-birth-weight infants. *N. Engl. J. Med.*, 346(3):149–157, January 2002.

[28] Nikolas Hesse, Sergi Pujades, Michael J Black, Michael Arens, Ulrich G Hofmann, and A Sebastian Schroeder. Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2540–2551, October 2020.

[29] Carrie L Hicks, Carl L von Baeyer, Pamela A Spafford, Inez van Korlaar, and Belinda Goodenough. The faces pain Scale-Revised: toward a common metric in pediatric pain measurement. *Pain*, 93(2):173–183, August 2001.

[30] Jorijn Hornman, Andrea F de Winter, Jorien M Kerstjens, Arend F Bos, and Sijmen A Reijneveld. Emotional and behavioral problems of preterm and full-term children at school entry. *Pediatrics*, 137(5), May 2016.

[31] V A Howard and F W Thurber. The interpretation of infant pain: physiological and behavioral indicators used by NICU nurses. *J. Pediatr. Nurs.*, 13(3):164–174, June 1998.

[32] Diane Hudson-Barr, Beverly Capper-Michel, Sally Lambert, Tonya Mizell Palermo, Kristen Morbeto, and Stephanie Lombardo. Validation of the pain assessment in neonates (PAIN) scale with the neonatal infant pain scale (NIPS). *Neonatal Netw.*, 21(6):15–21, September 2002.

[33] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3d face alignment from 2d video for real-time use. *Image and Vision Computing*, 58:13–24, 2017.

[34] Yingyu Ji, Shigang Wang, Yang Lu, Jian Wei, and Yan Zhao. Eye and mouth state detection algorithm based on contour feature extraction. *J. Electron. Imaging*, 27(05):1, February 2018.

[35] Céleste Johnston, Keith J Barrington, Anna Taddio, Ricardo Carbajal, and Françoise Filion. Pain in canadian nicus: have we improved over the past 12 years? *The Clinical journal of pain*, 27(3):225–232, 2011.

[36] C O Kerr-Wilson, D F Mackay, G C S Smith, and J P Pell. Meta-analysis of the association between preterm delivery and intelligence. *J. Public Health (Oxf.)*, 34(2):209–216, June 2012.

[37] A T Kirk, S C Alder, and J D King. Cue-based oral feeding clinical pathway results in earlier attainment of full oral feeding in premature infants. *J. Perinatol.*, 27(9):572–578, September 2007.

[38] Susan W Krechel and Judy Bildner. Cries: a new neonatal postoperative pain measurement score. initial testing of validity and reliability. *Pediatric Anesthesia*, 5(1):53–61, 1995.

[39] Dana Lahat, Tulay Adali, and Christian Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proc. IEEE Inst. Electr. Electron. Eng.*, 103(9):1449–1477, September 2015.

[40] Jocelyn Lawrence, Denise Alcock, Patrick McGrath, J Kay, S Brock MacMurray, and C Dulberg. The development of a tool to assess neonatal pain. *Neonatal network: NN*, 12(6):59–66, 1993.

[41] Xiaofeng Li, Jiahao Xia, Libo Cao, Guanjun Zhang, and Xiexing Feng. Driver fatigue detection based on convolutional neural network and face alignment for edge computing device. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 235(10-11):2699–2711, 2021.

[42] Susan M Ludington-Hoe, Mark W Johnson, Kathy Morgan, Tina Lewis, Judy Gutman, P David Wilson, and Mark S Scher. Neurophysiologic assessment of neonatal sleep organization: preliminary results of a randomized, controlled trial of skin contact with preterm infants. *Pediatrics*, 117(5):e909–23, May 2006.

[43] Francine R Margolius, Nancee V Sneed, and Ann D Hollerbach. Accuracy of apical pulse rate measurements in young children. *Nursing research*, 40(6):378, 1991.

[44] Hugo Monrroy, Giulio Borghi, Teodora Pribic, Carmen Galan, Adoracion Nieto, Nuria Amigo, Anna Accarino, Xavier Correig, and Fernando Azpiroz. Biological response to meal ingestion: Gender differences. *Nutrients*, 11(3):702, March 2019.

[45] Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Majority vote of diverse classifiers for late fusion. April 2014.

[46] Britt F Pados, Suzanne M Thoyre, George J Knafl, and William Brant Nix. Heart rate variability as a feeding intervention outcome measure in the preterm infant. *Adv. Neonatal Care*, 17(5):E10–E20, October 2017.

[47] Gayle Giboney Page. Are there long-term consequences of pain in newborn or very young infants? *J. Perinat. Educ.*, 13(3):10–17, 2004.

[48] Dhruv Pandey. Eye aspect ratio(ear) and drowsiness detector using dlib, Apr 2021.

[49] Jinhee Park, Suzanne Thoyre, George J Knafl, Eric A Hodges, and William B Nix. Efficacy of semielevated side-lying positioning during bottle-feeding of very preterm infants. *J. Perinat. Neonatal Nurs.*, 28(1):69–79, January 2014.

[50] Md Sirajus Salekin, Ghada Zamzmi, Dmitry Goldgof, Rangachar Kasturi, Thao Ho, and Yu Sun. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Computers in Biology and Medicine*, 129:104150, 2021.

[51] Md Sirajus Salekin, Ghada Zamzmi, Jacqueline Hausmann, Dmitry Goldgof, Rangachar Kasturi, Marcia Kneusel, Terri Ashmeade, Thao Ho, and Yu Sun. Multimodal neonatal procedural and postoperative pain assessment dataset. *Data in Brief*, 35:106796, 2021.

[52] Lin Shu, Yang Yu, Wenzhuo Chen, Haoqiang Hua, Qin Li, Jianxiu Jin, and Xiangmin Xu. Wearable emotion recognition using heart rate data from a smart bracelet. *Sensors (Basel)*, 20(3):718, January 2020.

[53] Rebeccah Slater, Lorenzo Fabrizi, Alan Worley, Judith Meek, Stewart Boyd, and Maria Fitzgerald. Premature infants display increased noxious-evoked neuronal activity in the brain compared to healthy age-matched term-born infants. *Neuroimage*, 52(2):583–589, 2010.

[54] Bonnie J Stevens, Laura K Abbott, Janet Yamada, Denise Harrison, Jennifer Stinson, Anna Taddio, Melanie Barwick, Margot Latimer, Shannon D Scott, Judith Rashotte, et al. Epidemiology and management of painful procedures in children in canadian hospitals. *Cmaj*, 183(7):E403–E410, 2011.

[55] S Suraseranivongse, R Kaosaard, P Intakong, S Pornsiriprasert, Y Karnchana, J Kaopinpruck, and K Sangjeen. A comparison of postoperative pain scales in neonates. *BJA: British Journal of Anaesthesia*, 97(4):540–544, 2006.

[56] Maria Luisa Tataranno. *Early biomarkers of brain development in preterm infants*. PhD thesis, Utrecht University, 2018.

[57] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[58] James W Varni, Christine A Limbers, Katie Neighbors, Kris Schulz, Judith E C Lieu, Robert W Heffer, Krista Tuzinkiewicz, Rita Mangione-Smith, Jerry J Zimmerman, and Estella M Alonso. The PedsQL™ infant scales: feasibility, internal consistency reliability, and validity in healthy and ill infants. *Qual. Life Res.*, 20(1):45–55, February 2011.

[59] Jillian Vinall, Steven P Miller, Vann Chau, Susanne Brummelte, Anne R Synnes, and Ruth E Grunau. Neonatal pain in relation to postnatal growth in infants born very preterm. *Pain*, 153(7):1374–1381, July 2012.

[60] Salimah R Walani. Global burden of preterm birth. *Int. J. Gynaecol. Obstet.*, 150(1):31–33, July 2020.

[61] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. Rgb-d-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding*, 171:118–139, 2018.

[62] Sebastian Weber, Heiko Johannsen, and Volker Schindler. Protection for the smallest occupant – status quo and potentials concerning the development of child restraint systems. 01 2022.

[63] Abigail Wellington and Jeffrey M Perlman. Infant-driven feeding in premature infants: a quality improvement project. *Arch. Dis. Child. Fetal Neonatal Ed.*, 100(6):F495–500, November 2015.

[64] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2):115–142, 2019.

[65] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *Int. J. Comput. Vis.*, 127(2):115–142, February 2019.

[66] Xiang Yu, Zhe Lin, Jonathan Brandt, and Dimitris N Metaxas. Consensus of regression for occlusion-robust facial feature localization. In *Computer Vision – ECCV 2014*, Lecture notes in computer science, pages 105–118. Springer International Publishing, Cham, 2014.

[67] Amir Zadeh, Tadas Baltrusaitis, and Louis-Philippe Morency. Deep constrained local models for facial landmark detection. *CoRR*, abs/1611.08657, 2016.

[68] Jill G Zwicker, Steven P Miller, Ruth E Grunau, Vann Chau, Rollin Brant, Colin Studholme, Mengyuan Liu, Anne Synnes, Kenneth J Poskitt, Mikaela L Stiver, and Emily W Y Tam. Smaller cerebellar growth and poorer neurodevelopmental outcomes in very preterm infants exposed to neonatal morphine. *J. Pediatr.*, 172:81–87.e2, May 2016.

7 Appendix

	Arm		**Frowns**		**Yawns**		**Head**	
Cue ID	*Found*	*Missed*	*Found*	*Missed*	*Found*	*Missed*	*Found*	*Missed*
1	2	1,3	3	1,2	-	1,2,3	2	1,3
2	1,2,3	-	1,2,3	-	1,2,3	-	2,3	1
3	2	1,3	1,2,3	-	1,2,3	-	2	1,3
4	2	1,3	1,2,3	-	-	1,2,3	2	1,3
5	2	1,3	1,2,3	-	-	1,2,3	2	1,3
6	2,3	1	3	1,2	1,2,3	-	1,2	3
7	2	1,3						
8	2	1,3						
9	2	1,3						
10	2	1,3						
11	-	1,2,3						

Table 20: Detection overlap on the training set

	Yawn		**Arm**		**Head**		**Frown**	
Cue ID	*Found*	*Missed*	*Found*	*Missed*	*Found*	*Missed*	*Found*	*Missed*
1	2	1,3	2	1,3	2,3	1	3	1,2
2	2	1,3	-	1,2,3	2,3	1	1,2,3	-
3	1,2,3	-	2	1,3	2	1,3	3	1,2
4	1,2,3	-	2	1,3	2	1,3	1,2,3	-
5	2	1,3	2	1,3	2	1,3	1,2,3	-
6	2	1,3	-	1,2,3	2,3	1		
7	2	1,3	2	1,3				
8	1,2,3	-	2	1,3				
9	1,2,3	-	2	1,3				
10	2	1,3	-	1,2,3				
11	2	1,3	-	1,2,3				
12	1,2,3	-	2	1,3				
13	2	1,3	2	1,3				
14	1,2,3	-	2	1,3				
15	2	1,3	2	1,3				
16	2	1,3	2	1,3				
17	2	1,3						
18	2	1,3						
19	1,2,3	-						

Table 21: Detection overlap on the test set

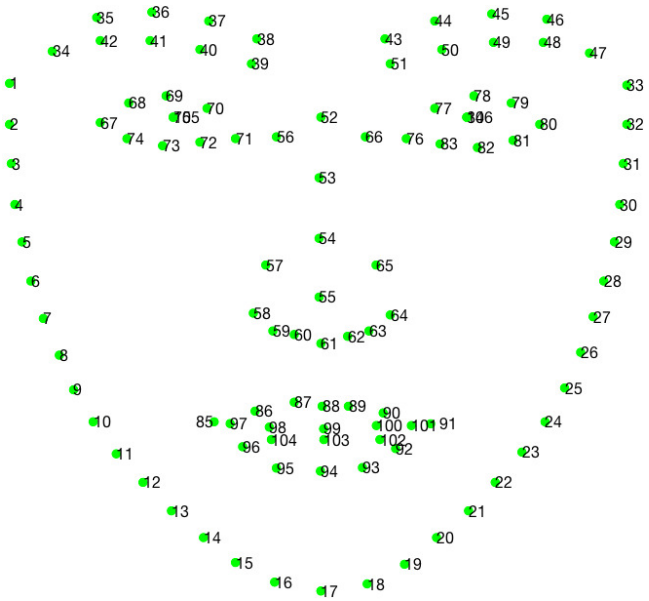


Figure 38: The legend for each image of a rule that uses facial key-points as inputs

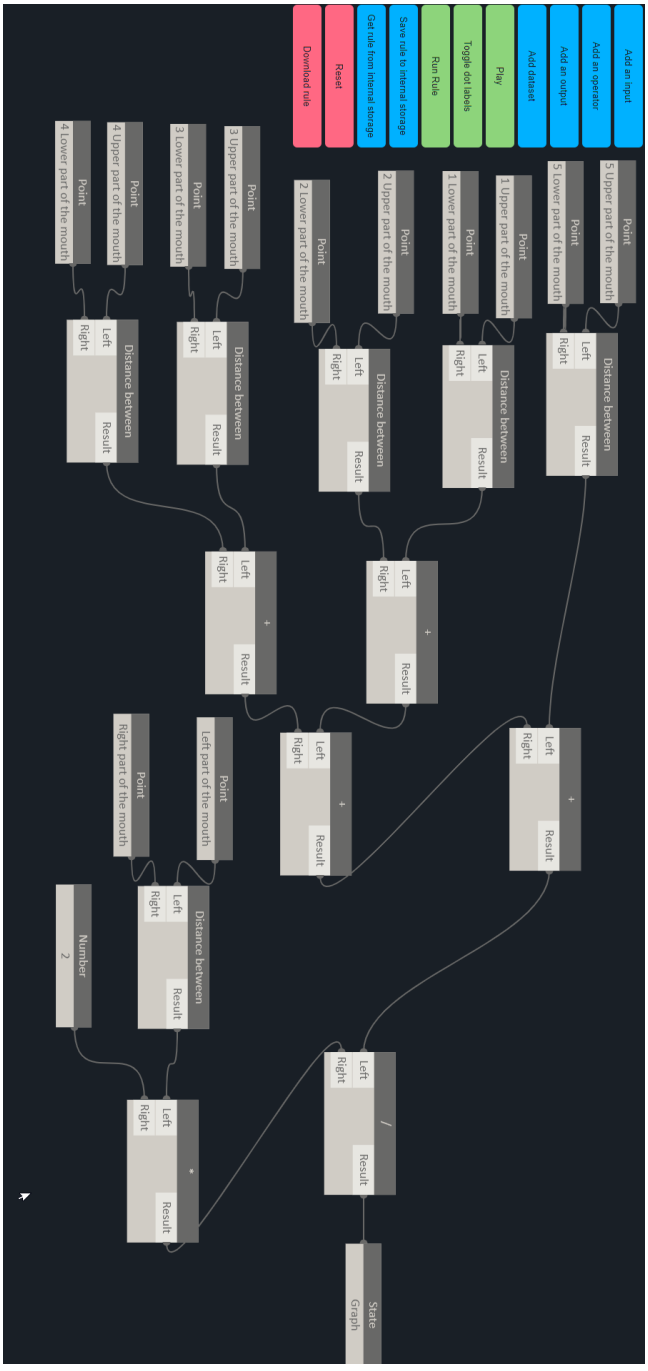


Figure 39: A rule build in the RBCD calculates and plots MAR

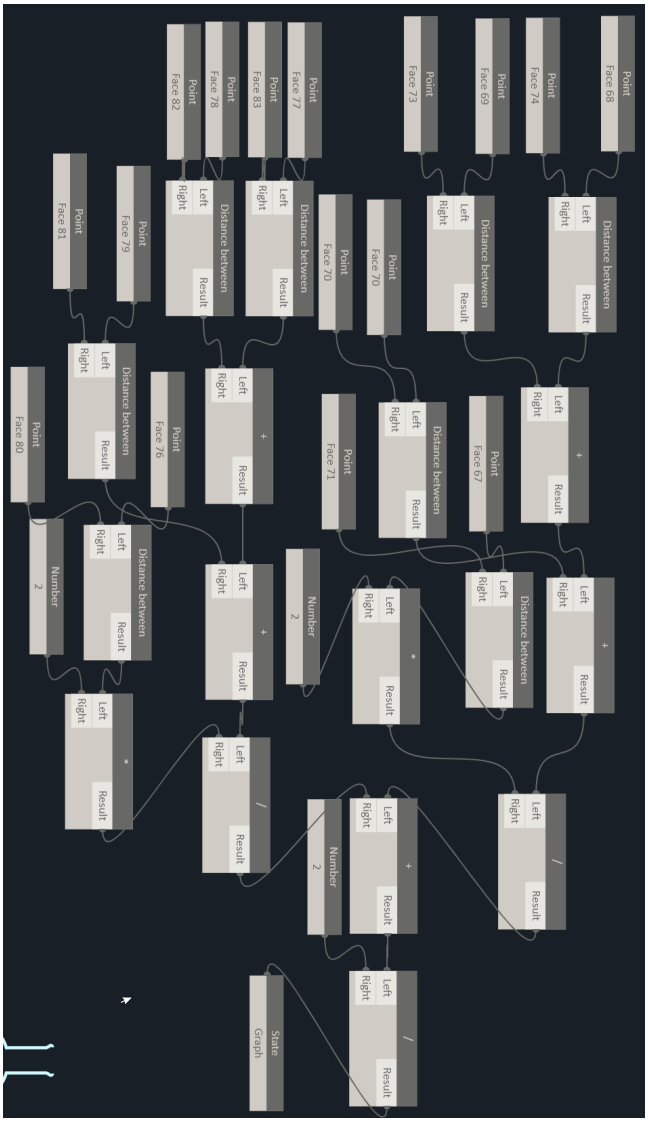


Figure 40: A rule build in the RBCD calculates and plots EAR

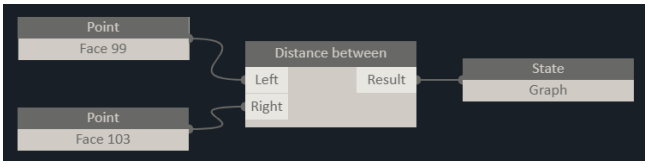


Figure 41: Participant 1 Yawn rule

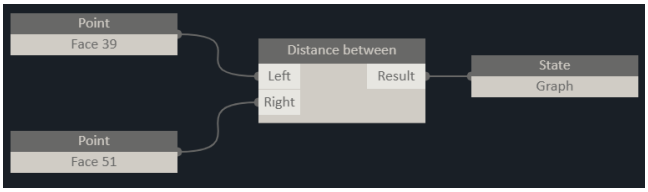


Figure 42: Participant 1 Frown rule

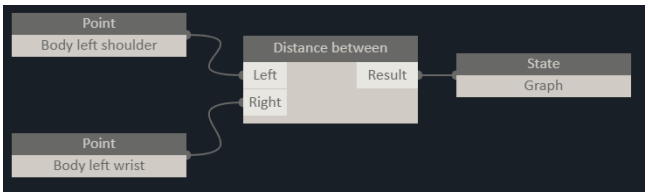


Figure 43: Participant 1 left arm rule

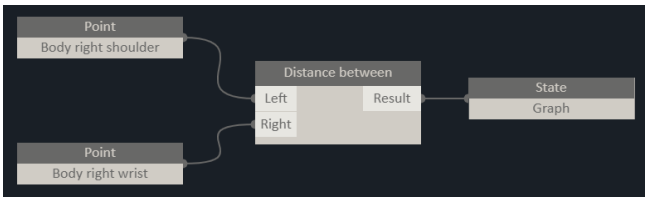


Figure 44: Participant 1 right arm rule

Figure 45: Participant 1 head movement rule

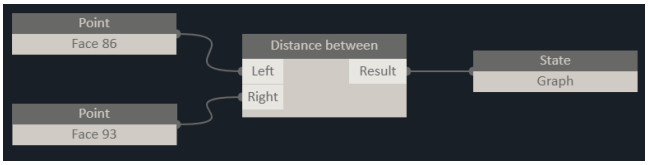


Figure 46: Participant 2 Yawn rule

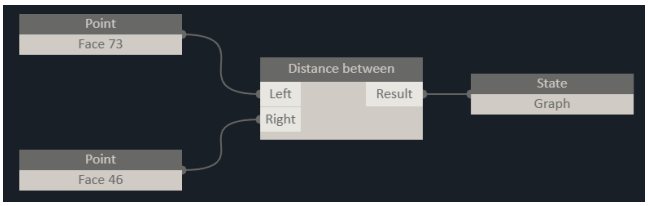


Figure 47: Participant 2 Frown rule

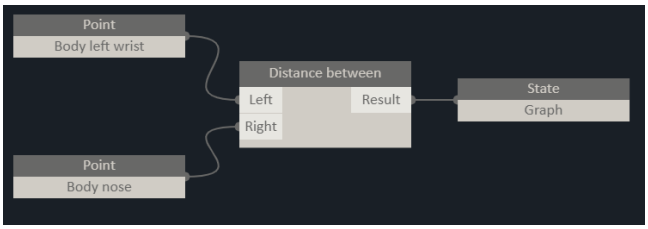


Figure 48: Participant 2 left arm rule

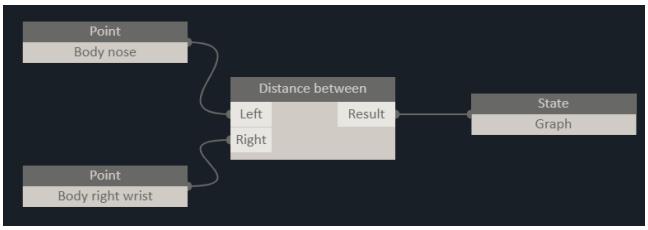


Figure 49: Participant 2 right arm rule

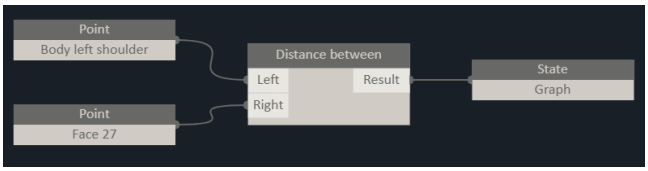


Figure 50: Participant 2 head movement rule

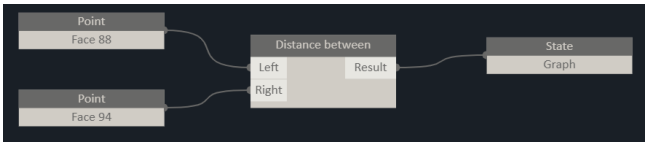


Figure 51: Participant 3 Yawn rule

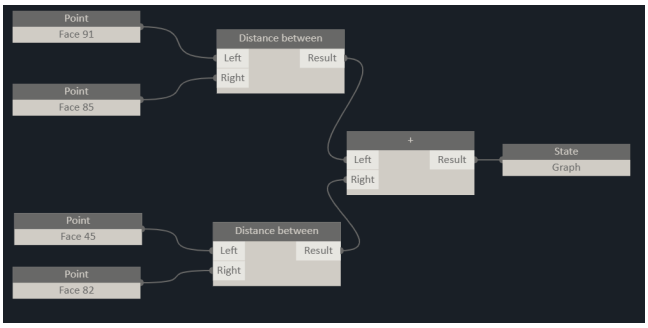


Figure 52: Participant 3 Frown rule

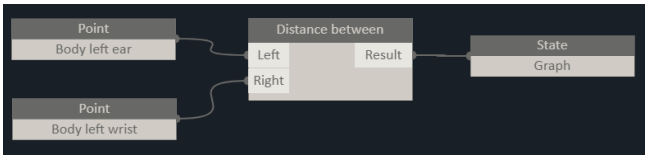


Figure 53: Participant 3 left arm rule

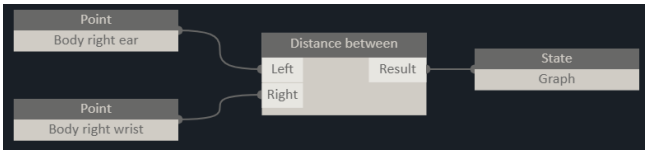


Figure 54: Participant 3 right arm rule

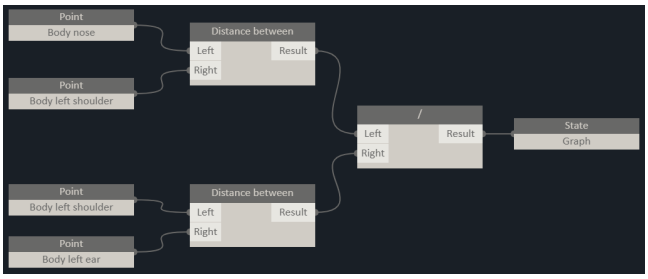


Figure 55: Participant 3 head movement rule