Automatic Multimodal Assessment of Neonatal Pain

by

Ghada Zamzmi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Co-Major Professor: Rangachar Kasturi, Ph.D.
Co-Major Professor: Dmitry Goldgof, Ph.D.
Yu Sun, Ph.D.
Richard Gitlin, Sc.D.
Terri Ashmeade, M.D.

Date of Approval:
May 1, 2018

Keywords: Affective computing, computer vision, video and image processing, medical application, machine learning

DEDICATION

This dissertation is dedicated to my wonderful mother, lovely sister, supportive brothers, and adorable nieces and nephews. No words can describe how much I am grateful for your presence in my life. Thank you for your unconditional love and support.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the guidance and support of many people. I would like to start by expressing my utmost gratitude to my wonderful graduate advisors, Dr. Rangachar Kasturi and Dr. Dmitry Goldgof, who were always happy to discuss research ideas and provide guidance, constant encouragement, and precious professional and personal advice. Their continuous support made graduate school an incredibly rewarding experience. I want to express my appreciation to Dr. Yu Sun for his encouragement, advice, and insightful feedback. I also want to thank my committee members, Dr. Terri Ashmeade and Dr. Richard Gitlin, and the chairperson of my defense, Dr. Ismail Uysal.

I would like to thank several other faculty members from whom I have learned a lot and who have given me valuable academic and career advice. Those professors include Dr. Jarred Ligatti, Dr. Sudeep Sarkar, Dr. Shaun Canavan, Dr. Larry Hall, Dr. Ken Christensen, and Dr. Paul Rosen. My acknowledgement would be incomplete without thanking my wonderful friends for their significant support during my journey towards this degree. I especially want to thank Yan Albright, Maryam Habadi, Cagri Cetin, Jean-Baptiste Subils, Hernan Palombo, Alireza Chakeri, Parham Phoulady, Donald Ray, Hamidreza Farhidzadeh, and Atika Chaudhary. I must also thank my great labmates and collaborators: Rahul Paul, Chih-Yun Pai, MD Sirajus Salekin, Mona Fathollahi, Fillipe Souza, Dmitry Cherezov, Saeed Alahmari, Hunter Morera, Gabriel Ruiz, Ruicong Zhi, and Matthew Compton. Last but not least, I want to thank the staff of the CSE main office for their help and the entire department for their warmth and welcoming manner.

I would like to end by thanking the parents who allowed their babies to take part in this study and the entire neonatal staff at Tampa General Hospital for their help and cooperation in the data collection. This research was supported in part by USF Women's Health Collaborative Grant.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

For several decades, pediatricians used to believe that neonates do not feel pain. The American Academy of Pediatrics (AAP) recognized neonates' sense of pain in 1987. Since then, there have been many studies reporting a strong association between repeated pain exposure (under-treatment) and alterations in brain structure and function. This association has led to the increased use of anesthetic medications. However, recent studies found that the excessive use of analgesic medications (over-treatment) can cause many side effects. The current standard for assessing neonatal pain is discontinuous and suffers from inter-observer variations, which can lead to over- or under-treatment. Therefore, it is critical to address the shortcomings of the current standard and develop continuous and less subjective pain assessment tools.

This dissertation introduces an automatic and comprehensive neonatal pain assessment system. The presented system is different from the previous ones in three principal ways. First, it is specifically designed to assess pain of neonates using data captured while they are hospitalized in the Neonatal Intensive Care Units (NICU). Second, it dynamically analyzes neonatal pain as it unfolds in a particular pattern over time. Third, it combines visual, vocal, and physiological signals to create a system that continues to assess pain even when one or more signals become temporarily unavailable. The presented system has four main components. The first three components consist of novel algorithms for analyzing the visual, vocal, and physiological signals separately. The last component combines all the three signals to create a multimodal pain assessment system. The performance of the system in recognizing pain events is comparable to that of trained nurses; hence, it demonstrates the feasibility of automatic pain assessment in typical neonatal care environments.

CHAPTER 1

INTRODUCTION

1.1 Outcomes of Neonatal Pain Experience

The International Association for the Study of Pain (IASP) defines pain as, "an unpleasant sensory and emotional experience associated with actual or potential tissue damage or described in terms of such damage." McCaffery [1] describes pain as, "whatever the experiencing person says it is, existing whenever the experiencing person says it does." Unfortunately, neonates do not have the ability to communicate this experience verbally (self-evaluation) or non-verbally by writing or pointing (Visual Analog Scale). The limited ability of neonates to communicate pain and the earlier misconception about the absence of neurological substrate for the perception of pain in neonates has led pediatricians to believe, for several decades, that neonates do not feel or remember pain. Sufficient scientific studies [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14] disproved this earlier belief and reported serious short- and long-term outcomes of pain exposure in early life.

Studies have found that unexpected and repeated painful experiences during the early life period are associated with alterations in the pain sensitivity and perception [7, 8, 9, 10] (e.g., allodynia and hyperalgesia), stress-response system functioning [3, 11, 13] (e.g., high basal cortisol levels), and postnatal growth [8] (e.g., slower body weight gain and head growth). In addition, there is strong evidence that extensive pain exposure during the early life period is associated with alterations in the brain structure and function. Such alterations include changes in the cerebral white matter and subcortical grey matter [8, 10, 14], delayed corticospinal development [8, 7], changes in the number of synaptic connections and glia, as well as the degree of capillary branching that augments the blood and oxygen supply [5, 4]. These alterations can result in a variety of behavioral, developmental and learning disabilities [2, 11, 15]. A comprehensive discussion of several short-term and long-term outcomes of extensive pain exposure in early life period can be found in [15].

The recognition of adverse effects of neonatal pain has led to recommendations for increased use of analgesic medications. However, recent studies [16, 17, 18, 19] have found that the excessive use of analgesic medications such as Morphine and Fentanyl may cause several side effects. For example, Zwicker et al. [16] found that 10-fold increase in Morphine, an agent commonly used for neonatal pain management, is associated with impaired cerebellar growth in the neonatal period and poorer neurodevelopmental outcomes in early childhood period. In addition, studies demonstrate that Morphine increases apoptosis in human microglial cells [17] and neuronal like cells in neonatal rats [18]. Studies in neonatal rats have found long-term alterations in brain function and structure following exposure to Morphine [19]. Other effects of using high doses of Morphine, which include hypotension and lesser tolerance for feedings, are reported in [15]. The long-term side effects of another well-known analgesic medication (Fentanyl) are discussed in [15]. This study describes Fentanyl as an extremely potent analgesic and lists several side effects (e.g., neuroexcitation and respiratory depression) for using high doses of Fentanyl.

These findings suggest that failure to recognize and treat neonatal pain (under treatment) as well as administration of certain analgesic medications in the absence of pain (over treatment) may lead to serious outcomes such as, permanent alterations in brain structure and function. These alterations may contribute to the high incidence of neurodevelopmental disability occurring in preterm and full-term neonates. The annual cost of care related to adverse neurodevelopmental outcomes in preterm neonates alone is estimated at over 7 billion dollars [20].

1.2 Current Practice of Pain Assessment

Although neonates are incapable of articulating their pain experience, their body responds to painful stimuli in three different ways: behaviorally (e.g., facial expression, dysregulated sleep pattern, and crying), physiologically (e.g., changes in heart rate and blood pressure), and metabolically (e.g., pronounced release of catecholamines). The intensity and pattern of these responses differ across various types of pain, which are procedural, postoperative, and chronic pain [21]. Procedural pain is usually associated with a short painful stimulus such as immunization and it ends as soon as the cause of pain is removed. Postoperative pain has a clearly defined beginning point and expected end point, and it occurs after a known stimulus such as a surgical procedure. Neonates' behavioral

responses to procedural painful stimulus are usually more intense as compared to their response to postoperative pain. This can be attributed to the low physical reserves of a neonate to sustain a response and the level of sedation/analgesia. Finally, the neonatal chronic pain is defined as the persistent and ongoing pain that lasts beyond the normal three-month healing time and does not have an expected end point.

Apart from the type of pain, studies [9, 22, 23, 24, 25, 26, 27, 28] have also found strong associations between the contextual (e.g., oral sucrose) and clinical (e.g. gestational age) data, and neonates' response to pain. For example, it has been found [24, 25, 26, 28] that the neonates' age greatly affects their reaction to painful experiences. Particularly, it has been reported that neonates with low gestational age have limited ability to behaviorally communicate pain due to the underdeveloped muscles of fragile premature neonates. Differences in reaction to pain based on gender have also been reported in the literature [22, 23, 26]. Therefore, incorporating contextual and clinical data with other pain responses is necessary to refine the assessment process and obtain a context-sensitive assessment.

On average, infants receiving care in the Neonatal Intensive Care Unit (NICU) experience fourteen painful procedures per day [33, 34]. The current practice for assessing neonatal pain involves observing, by bedside caregivers, multiple behavioral (e.g., facial expression and crying) and physiological (e.g., changes in vital signs) responses of pain. At least 29 response-based pain scales have been developed to evaluate procedural and postoperative pain in neonates. Table 1.1 provides examples of validated procedural (1^{st} and 2^{nd} rows) and postoperative (3^{rd} and 4^{th} rows) pain scales [29, 30, 31, 32].

The utilization of the pediatric response-based pain scales for assessment has three main limitations. First, it relies on the caregiver's direct observation and interpretation of multiple responses, including behavioral, physiological, and metabolic responses. This practice of assessment is highly biased and is affected by several idiosyncratic factors, such as the observer's cognitive bias, identity, culture, and gender [23, 35, 36, 37]. The inter- and intra- observer variations can lead to inconsistent assessment and treatment of pain. Second, caregivers assess pain at different time intervals and are not able to provide continuous assessment of pain. The discontinuity of assessment can lead to missing pain while the neonate is left unattended; therefore, it may result in delayed intervention. Third, this practice involves a substantial time commitment and requires a large number

Table 1.1: Examples of common pediatric pain scales.

Scale	**Age**	**Behavioral**	**Physiological**	**Psychometric**
[29] NIPS	28-38 gestation weeks	Facial expression, crying, arms/legs movement, and arousal state	Breathing pattern	Inter-rater reliability: (r=0.92-0.97) Internal Consistency: (α= 0.87-0.95) Content validity Concurrent validity: (r=0.53-0.83)
[30] NFCS	$\geq$ 25 gestation weeks	Brow bulge Eye squeeze Nasolabial furrow Open lips Horizontal mouth Vertical mouth Lips pursed Taut tongue Chin quiver Tongue protrusion	N/A	Inter/Intra-rater reliability $>$ 0.85 Internal Consistency: (α= 0.87-0.95) Content validity Face validity Construct validity
[31] N-PASS	23-40 gestation weeks	Facial expression, behavior movements, crying/irritability, and extremities tone	Heart rate, respiratory, blood pressure, and O saturation	Inter-rater reliability: (r=0.85-0.95) Intra-rater reliability: (r=0.87) Internal consistency: (α= 0.84-0.89) Construct validity: (P$<$.0001)
[32] CRIES	32-60 gestation weeks	Facial expression, crying, and sleeping state	Requires oxygen increase and VS	Inter-rater reliability: (r=0.98) Construct and content validity

of well-trained caregivers to ensure the proper utilization of the pain scale. The substantial cost of this practice makes it infeasible in under-developed countries where medical professionals and resources are scarce. As accurate pain assessment is essential to obtain an adequate pain management, researchers are obligated to address the shortcomings of the current practice and create consistent, continuous, and inexpensive pain assessment tools to guide treatment.

1.3 Automatic Assessment of Neonatal Pain

This dissertation introduces an automatic and multimodal pain assessment system that addresses the shortcomings of the current pain assessment. Combining multiple modalities or pain responses allows for the assessment of pain during circumstances when not all pain indicators are available due to developmental stage, clinical condition or level of activity. Our automatic pain assessment system decreases the caregiver's burden of observation and documentation while providing continuous monitoring. The main contributions of this dissertation can be summarized as follows:

1. It introduces a comprehensive review of the current neonatal pain assessment methods, reviews the current challenges, and provides implications for new research (Chapter 2).

2. It presents a fully automated and multimodal approach for neonatal pain assessment that can be easily adopted and integrated into clinical environments since it uses non-invasive devices (RGB cameras) for pain monitoring. This dissertation is the first to propose an automated version of the current pediatric scales that integrates multiple behavioral and physiological pain responses for assessment.

3. It introduces a unique, well-annotated, and multidimensional neonatal database that can be used to advance the research of automatic pain assessment (Chapter 3).

4. It presents novel handcrafted and deep learning methods for analyzing different pain responses, namely facial expression, body movement, and crying sound. It also presents methods for integrating these pain responses with physiological readings to obtain a multimodal assessment system (Chapter 4).

5. It presents a unimodal and multimodal approach for pain assessment and integrates the contextual data by developing age-specific, gender-specific, race-specific, and weight-specific pain assessment models (Chapter 5).

6. It provides several implications and directions for future research (Chapter 6).

1.4 Dissertation Roadmap

The rest of this dissertation is structured as follows. Chapter 2 delivers a review of the existing automatic methods for pain assessment and summarizes the current challenges. Chapter 3 provides a detailed description of the Neonatal Pain Assessment Database (NPAD) followed by our novel algorithms for assessing pain in Chapter 4. Chapter 5 presents the experimental setup and discusses the results. Finally, Chapter 6 concludes the dissertation and presents several directions for future work.

CHAPTER 2

AUTOMATIC PAIN ASSESSMENT: STATE OF THE ART

2.1 Note to Reader

Portions of this chapter were published in IEEE Reviews in Biomedical Engineering [38]. Permission from the publisher is included in Appendix A.

2.2 Background

Only a handful of efforts have been made to analyze and assess neonatal pain using Computer Vision and Machine Learning technologies. In contrast, a rich variety of methods were proposed to assess adult's pain based on analysis of facial expression [39, 40, 41, 42, 43, 44, 43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74], head pose [58, 54, 56], or physiological data [60, 56, 61, 75]. We believe that the lack of the automatic neonatal pain recognition methods can be attributed to the limited number of annotated and publicly-available neonatal databases. To date, we are aware of only two databases, COPE [76] and YouTube videos [77], that are available upon request for research in neonatal pain. Another reason is the common belief that the algorithms designed for adults would have similar performance when applied to neonates. Contrary to this belief, we think the methods designed for assessing adults' pain would not produce similar performance and might completely fail due to two main reasons.

First, the pain's dynamics and facial morphology vary between infants and adults. It was reported [78] that infants' facial expressions include additional movements and units that are not present in the Facial Action Coding System (FACS) [79]. Therefore, the Neonatal FACS, also known as NFCS [78], was introduced as an extension of FACS. In addition to facial expression, neonates' sound and movement during pain have different pattern and dynamics than those of

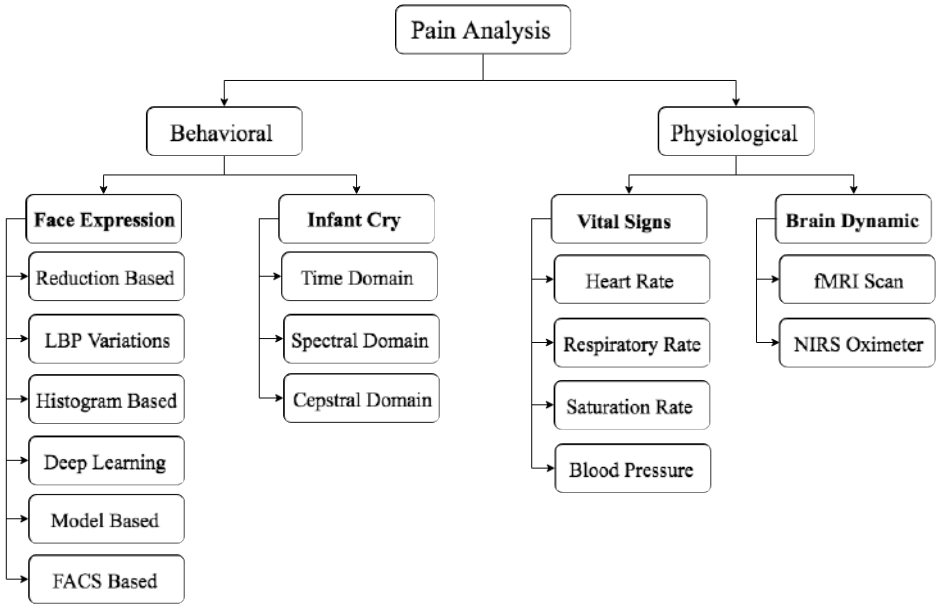


Figure 2.1: Tree diagram of the automatic feature representation methods.

adults. Second, we think the preprocessing stage (e.g., face and body tracking) is more challenging in the case of neonates because they are considered uncooperative subjects.

In this chapter, we extensively explore the current efforts for assessing neonatal pain automatically. In particular, we present a systematic review of the current methods that extract pain-relevant features from neonatal data (Section 2.3). We also categorize pain recognition into pain detection and pain intensity estimation (Section 2.4). We define pain detection as the task of detecting the presence or absence of pain and pain intensity estimation as the task of estimating the intensity of the detected pain (i.e., how much an infant is in pain). Finally, we review the pain databases that are available for research use in Section 2.5 and discuss the current limitations of automatic pain assessment in Section 2.6.

2.3 Feature Extraction and Representation Methods

Several methods were introduced to extract pain-relevant features from neonates' behavioral or physiological data. We grouped these methods into three main categories, namely feature representation of behavioral response, feature representation of physiological response, and fusion of pain responses, and divided these categories further as illustrated in Figure 2.1.

2.3.1 Behavioral Pain Responses

Feature representation of behavioral response can be defined as the task of extracting pain-relevant features from a behavioral response of pain such as facial expression or crying sound. We present next the existing methods that analyze neonatal behavioral responses to extract useful features for pain classification.

2.3.1.1 Facial Expression Analysis

Facial expression is one of the most common and specific responses of pain. Facial expression of pain is defined as the movements and distortions in facial muscles associated with a painful stimulus. The facial movements associated with pain in neonates include: deepening of the nasolabial furrow, brow lowering, narrowed eyes, vertical and horizontal mouth stretch, lip pursing, lip opening, tongue protrusion, taut tongue, and chin quiver [78].

The automatic recognition of facial expression consists of three main stages: (1) face detection and registration; (2) feature extraction; and (3) pain expression recognition. Face detection is a mature area of research and, therefore, will not be discussed further. Several methods were proposed to extract pain-relevant features from neonates' images and videos. We broadly divided these methods based on their underlying algorithms into sex groups: Feature Reduction Based methods, Local Binary Pattern (LBP) Variation Based methods, Histogram of Oriented Gradients (HOG) Based methods, Deep Learning Based methods, Model Based methods, and Facial Action Coding System [FACS]. Figure 2.1 presents a summary of these groups as a tree diagram.

2.3.1.1.1 Feature Reduction Based Methods

A simple approach to extract pain-relevant features from static images is to convert the image's pixels into a vector of $N_x \times N_y \times 1$ dimensions, where N_x and N_y represent the image's width and height. Then, feature reduction methods such as Principal Component Analysis (PCA) and Sequential Floating Forward Selection (SFFS) could be applied to reduce the vector's dimensionality.

PCA is a statistical method to reduce the dimensionality of a given feature space by identifying a small number of uncorrelated features or variables, known as principle components. Those components represent the dimensions along which the data points are mostly spread out. A detailed explanation of PCA along with its mathematical formulation can be found in [80].

SFFS [81] is a well-known method for feature selection. Sequential Feature Selection (SFS) methods are a family of greedy search algorithms that are used to reduce an initial d dimensional feature space to a $k - dimensional$ feature subspace where $k < d$ by sequentially adding a single feature until there is no further improvement in the classifier performance. SFFS is a method to construct the best feature subset by adding to a subset, initially equal to null, a single feature that satisfies some criterion function. The difference between SFS and SFFS is that SFFS allows, according to the criterion function, to exclude the worst feature from the subset. That is, it allows to dynamically increase and decrease the features until the best subset is reached.

Brahnam et al. [82] presented one of the first studies in the machine recognition of pain. A feature reduction-based approach was proposed and applied on the Classification of Pain Expressions (COPE) database. This database consists of 204 RGB images captured for 26 Caucasian infants, half of which were girls, using Nikon D100 digital camera. The infants' age ranges from 18 hours to 3 days old and all infants were in good health. The face images of infants were taken while experiencing four different stimuli: pain stimulus during the heel lancing (60 images), rest/cry stimulus during the transportation of an infant from one crib to another (63 rest images and 18 cry images), air stimulus to the nose (23 images), and the friction stimulus, which involves receiving friction on the external lateral surface of the heel with cotton soaked in alcohol (36 images).

To extract pain relevant features, each image was rotated, cropped, converted to grayscale, and reduced to 100 x 120 pixels. The rescaled image was then concatenated into a feature vector of 12000 dimensions with values ranging from 0 to 255. To reduce the high dimensionality of this vector, PCA was applied. For classification, distance-based classifiers (i.e., PCA and Linear Discriminant Analysis [LDA]), and Support Vector Machine (SVM) were used to classify the infants' images into one of the following pairs: pain/no-pain, pain/rest, pain/cry, pain/air-puff, and pain/friction. The results showed that SVM with a polynomial kernel of degree 3 evaluated using 10-fold cross-validation achieved the best recognition rate and outperformed distance-based classifiers in classifying pain versus no-pain (88.00%), pain versus rest (94.62%), pain versus cry (80.00%), pain versus air-puff (83.33%), and pain versus friction (93.00%).

The above-discussed work is extended in [83] to include Neural Network Simultaneous Optimization Algorithm (NNSOA) for classification along with LDA, PCA, and SVM. Leave-one-subject-out cross-validation was used to evaluate NNSOA classifier instead of 10-fold cross-validation. The re-

sults showed that NNSOA has the highest average classification rate (90.20%) in classifying infants' images as pain (60 images) or no pain (144 images). SVM, PCA and LDA achieved average classification rates of 82.35%, 80.39% and 76.9%, respectively.

Instead of detecting the presence or absence of the pain expression, Golahmi et al. [84] presented a sparse kernel machine learning algorithm, known as Relevance Vector Machine (RVM), to estimate the intensity level of the detected pain expression. RVM is a Bayesian version of SVM that provides the posterior probabilities for the class memberships. In the preprocessing stage, a total of 181 images from COPE dataset were standardized using similarity transformation, cropped using 70×93 window to get the exact facial region, and converted to grayscale. Then, each image was converted into a 6510-dimensional vector by column stacking the intensity values. These vectors were used to build an RVM model, which was evaluated using leave-one-image-out cross-validation method.

For validation, the estimated pain intensity generated by the RVM algorithm (i.e., posterior probability or uncertainty of the class membership) was compared with pain intensity assessment of five expert and five non-expert examiners. To give the human examiners a prior knowledge for the assessment, two images of pain and no-pain conditions were selected for each infant and assigned a score of 0 (no-pain) and 100 (pain). Then, the human examiners were asked to provide a score that ranges from 0 to 100 for each image. The results showed moderate agreement between the assessment of expert examiners (0.47 Weighted Kappa Coefficient with 95% confidence interval of 0.37 to 0.57) and the assessment of non-expert examiners (0.46 Weighted Kappa Coefficient with 95% confidence interval of 0.36 to 0.55) as compared with the assessment of RVM. The agreement, measured using the Weighted Kappa Coefficient, between expert and non-expert examiners was 0.78 with a 95% confidence interval.

2.3.1.1.2 Local Binary Pattern Based Methods

The methods presented in this category utilize Local Binary Pattern (LBP) descriptor or its variants for analysis. Local Binary Pattern (LBP) [85] is one of the most popular texture descriptors in Computer Vision. The popularity of LBP can be attributed to its simplicity, low computational complexity, and robustness to illumination variations and alignment error [85].

The basic LBP describes the image's texture by comparing the gray value of a central pixel X with the gray values of its P neighbors within a predefined circle of radius R and considering the

output of the comparison as a binary number. For example, the value of a neighbor pixel would be one if the value of that pixel is greater than the central pixel value and zero otherwise. These binary values are then encoded to form local binary patterns that are converted into decimal and accumulated into a discrete histogram. Local Ternary Pattern (LTP) [86] is an extension of LBP. The main difference between LBP and LTP is that the difference between the central pixel X and its neighbor P is represented by a 3-valued function (i.e., ternary values of 0, 1, and -1 instead of binary). An Elongated Binary Pattern (ELBP) [87] and Elongated Ternary Pattern (ELTP) [88] are variants of LBP and LTP that use an elliptic neighborhood window instead of a circular window. As discussed in [87], the elliptic neighborhood window captures the anisotropic structure of facial images more effectively.

Nanni et al. [88] presented a texture-based method to detect facial expressions of pain. In the preprocessing stage, the images of COPE database were resized, aligned, cropped to obtain the exact facial region, and divided into blocks or cells of 25×25 dimensions. Then, LBP, LTB, ELTP, and ELBP texture descriptors were applied to these cells to extract pain-relevant features. To select the most discriminate cells, SFFS feature selection algorithm was applied to a training set using the leave-one-out-cross-validation protocol. For classification of neonates' images as pain or no-pain, an ensemble of Radial Basis SVMs was built and evaluated on a testing set. The results showed that ELTP texture descriptor achieved the highest (approx. 0.93) Area Under the Curve of Receiver Operating Characteristic curve (AUC of ROC) as compared to other texture descriptors. It also showed that pain expression affects sub regions of the face and thus dividing the whole image into cells can improve the performance.

Similarly, Mansor et al. [89] presented LBP-based method that is robust to different level of illuminations. This work modified COPE database by altering the original images and adding different levels of illuminations. Then, Multi Scale Retinex (MSR) image filter was applied to remove illumination followed by LBP for feature extraction. The extracted texture features were used to train an unsupervised Gaussian classifier and supervised Nearest Mean classifier. The highest average accuracy (83%) of the proposed method was achieved by the Gaussian classifier.

Celona and Manoni [90] applied a uniform LBP descriptor (P = 8, R = 1, and 59-bins) to static images of COPE database after dividing the face region into 25 (5×5) non-overlapping regions. To retain the color information, the LBP histogram was computed for each color channel of each

region. Then, the texture features extracted for each color of each region were concatenated into a single feature vector that has 4425 dimensions (59 bins $\times$ 25 regions $\times$3 channels). This feature vector was reduced to 175 dimensions using Principal Component Analysis (PCA) followed by L2 normalization. In the final stage, SVM was trained to classify the facial images into pain or no pain. The trained classifier achieved 77.52% average accuracy, using the leave-one-subject-out cross validation protocol.

2.3.1.1.3 Histogram of Oriented Gradients Methods

The Histogram of Oriented Gradients (HOG) [91] is a feature descriptor that works by dividing a given image into connected local regions or cells and counts the occurrences of gradient orientation in each cell. Celona and Manoni [90] applied HOG descriptor to static images of COPE database to classify them as pain images or no-pain images. The presented method used 2×2 blocks of 8×8 pixel cells with an overlap of half the block and histograms of 9 bins evenly spread from 0 to 180 degrees. Applying this descriptor to 224×224 gray-scale image generates 26244-dimensions feature vector (729 regions $\times$ 4 blocks $\times$ 9 bins). This feature vector was reduced to 175 dimensions using Principal Component Analysis (PCA) followed by L2 normalization. Using the features extracted by HOG descriptor with SVM achieved 81.75% average accuracy (i.e., accuracies averaged across 26 subjects).

The works presented above utilize traditional handcrafted features such as LBP and HOG for classification. Recently, deep feature extracted by Convolutional Neural Networks (CNNs) showed good performance in several classification tasks. We present next the existing works that utilize deep feature representation for pain classification.

2.3.1.1.4 Deep Learning Based Methods

The methods under this category utilize Convolutional Neural Networks (CNNs) to extract features, at multiple levels of abstraction, directly from the source data as opposed to the handcrafted methods which are designed beforehand to extract a chosen set of features. Recently, CNNs has been successfully applied to images for pain classification tasks.

For example, Celona and Manoni [90] applied transfer learning method to static images of COPE database to classify these images as pain or no pain images. In particular, they used deep

features extracted by a pre-trained CNN (VGG-Face) to train Support Vector Machine (SVM) model. Testing the trained model on unseen data (i.e., leave-one-subject-out cross validation) achieved 82.42% average accuracy. Combining the extracted deep features with the handcrafted features (LBP) improved the pain classification and achieved an average accuracy of 83.78%.

In a different population, Neural Networks were used to extract deep features for recognizing pain of the adult subjects who were suffering from shoulder pain (UNBC-McMaster database [43]). We refer the interested reader to DeepFaceLIFT [72] and DeepPain [73] projects for further description and discussion.

The main limitation of the methods presented so far is the use of static images taken at a specific time for training and testing. Because facial expressions are dynamic events that unfold in a particular pattern over time, it is important to take the temporal information into account when developing facial expression recognition methods. Besides the dynamic nature of facial expression, occlusion, which is known to be common in clinical environments, could not be handled statically. Therefore, developing dynamic pain recognition methods is required to obtain accurate results. We present next several methods to analyze facial expression of pain dynamically in video sequences.

2.3.1.1.5 Model Based Methods

The basic concept of model-based algorithms is to search for the optimal parameters of an object model that best match the model and the input image. Active Appearance Model (AAM) is a well-known model-based algorithm that uses appearance (i.e., combination of texture and shape) for matching a model image to a new image. It is one of the most commonly used algorithms in various applications such as face recognition [92], facial expression recognition [93], and medical image analysis [94]. To fit an AMM model to a facial image, the error between the representative model and the input image should be minimized (i.e., a non-linear optimization problem).

Fotiadou et al. [95] used AAM for detecting infants' pain expression during acute painful procedure. The authors adopted the method proposed in [40] to analyze adults' pain expression. The database utilized in this work consists of facial expression data for 10 infants hospitalized in the NICU at a local hospital in Veldhoven, the Netherlands. Infants were recorded in four states, namely heel lancing (i.e., acute procedure), diaper change, hunger, and resting/sleeping. All videos were recorded under unconstrained lighting conditions.

For each video, AAM tracker was applied to track facial landmark points through the video frames. Then, three features were extracted from the tracked face based on AAM parameters. Specifically, SPTS (similarity-normalized shape), SAPP (similarity-normalized appearance), and CAPP (canonical normalized appearance) were extracted. SPTS feature vector contains the coordinates of the landmark points after removing all rigid geometric variations; SAPP vector represents the appearance after removing rigid geometric variations and scale; and CAPP represents the appearance after removing all the non-rigid shape variation.

A total of 15 videos for 8 infants were used to build the automated discomfort detection system. The videos of the remaining two infants were excluded from further analysis since these videos include severe occlusion caused by large face rotation or moving hands. The proposed system classified infants' facial expression as discomfort or comfort using the extracted features and an SVM classifier. To evaluate the classifier's performance, leave-one-subject-out cross validation was performed. The result (0.98 AUC) showed that the proposed system can detect discomfort accurately.

This work has three main limitations. First, the emotional states for each class was not clearly specified. For example, it was not specified clearly if the discomfort class contains only the heel puncture or if it contains heel puncture as well as wet diaper and hunger. We believe that the latter two states are different than pain and, thus, should be treated separately. Second, all the experiments were carried out using a person-specific AAM that is constructed specifically for each infant; this can lead to scalability issues in practice. Third, the proposed method requires further investigation on a larger dataset since it was evaluated on a small dataset (8 subjects).

2.3.1.1.6 Facial Action Coding System

FACS [79] is a comprehensive system that uses a set of numeric codes to describe the movements of facial muscles for all observable facial expressions. FACS's numeric codes, which represent the facial muscles' movements, are known as Action Units (AUs). Neonatal Facial Coding System (NFCS) [78] is an extension of FACS designed specifically to observe infants' pain-relevant facial movements (see Table 1.1).

The vast majority of the methods in the field of automatic facial expressions recognition use FACS to detect facial expressions. However, we are not aware of any FACS-based method that is

designed specifically to detect infants' facial expression of pain. In different population, Sikka. et al. [49] presented a FACS-based method to describe children's facial expressions of pain. The proposed method was applied to video sequences of 50 children recorded during ongoing and transient pain conditions. A total of 35 subjects were Hispanic, 9 non-Hispanic white, 5 Asian, and 1 Native American; the children' age ranges from 5 to 18 years old and 35% of the them were boys. The data were collected over three visits: 1) within 24 hours of appendectomy surgery; 2) one day after the first visit; and 3) at a follow-up visit. The transient pain was triggered by manually pressing the surgical site for 2-10 seconds. At each visit, facial expressions of the children were recorded using Canon VIXIA-HF-G10 video camera placed in an upright position. Along with the video recording, self-reported rating by the children and by-proxy rating by both a parent and a nurse were collected to get the ground truth labels.

To extract useful features from the recorded videos, the Computer Expression Recognition Toolbox (CERT) [96] was used to detect several AUs. A feature selection method was then applied to select fourteen representative AUs (e.g., AU4 brow lower, AU7 lid tighten, and AU27 mouth open) and different statistics (e.g., the mean, 75th percentile, and 25th percentile) were computed for each of these AUs to form the feature vectors. The extracted features were used to build a logistic regression model evaluated using 10-fold cross validation. The binary classification of facial expression as pain or no pain achieved good-to-excellent accuracy with 0.84- 0.94 AUC for both ongoing and transient pain. The main limitation of this work is the restricted light and motion condition. The presented algorithm requires moderate lighting and motion, which might be difficult to accomplish in clinical settings such as the NICU.

The main challenge of FACS-based methods is the extensive time required for labeling AUs in each video frame to get the ground truth. It has been reported [62] that a human expert needs around three hours to code one minute of a video sequence. One-way to reduce the cost of labeling is to automatically detect AUs in each frame and use them as labels. Automatic detection of facial action units in real-world conditions is a challenging area of research that is not directly relevant to this review and, thus will not be discussed further. Those who are interested in the automatic detection of AUs are referred to [97, 98] for more information.

Before we conclude this section, we would like to note that COPE is the only database of the above-presented works that is available for research in automatic pain assessment, as confirmed by

the authors through email. Similarly, the code of [88] is the only code of the above-presented works that is available, per request, for research use.

2.3.1.2 Infant Cry Analysis

Infant cry is a sign of discomfort, hunger, or pain. It conveys information that helps caregivers to assess the infant's emotional state and react appropriately. Crying analysis can be divided into two main stages: (1) signal processing stage, which includes preprocessing the signal and extracting representative features; and (2) the classification stage. We classified the existing methods of signal processing stage into: (1) Time Domain methods; (2) Frequency Domain methods; and (3) Cepstral Domain methods (see Figure 2.1).

2.3.1.2.1 Time Domain Analysis

Time Domain analysis is the analysis of a signal with respect to time (i.e., the variation of a signal's amplitude over time). Examples of Time Domain features that are commonly used for infants' sound analysis are energy, amplitude, and pause duration.

Vempada et al. [99] presented a Time Domain method to detect discomfort-relevant cries. The proposed method was evaluated on a dataset consisting of 120 cry corpuses collected during pain (30 corpuses), hunger (60 corpuses), and wet-diaper (30 corpuses). The paper does not provide information about the stimulus that triggered the pain state nor the data collection procedure. The infants' age ranged from 12- 40 weeks old and all corpuses were recorded using a Sony digital recorder with sampling rate of 44.1 kHz. In the feature extraction stage, two features were calculated: 1) Short-time energy (STE), which is the average of the square of the sample values in a suitable window; and 2) Pause duration within the crying segment. Part of these features were used to build SVM and the remaining were used to evaluate its performance. The recognition performance of pain cry, hunger cry, and wet-diaper cry were 83.33%, 27.78%, and 61.11% respectively. The average recognition rate was 57.41%.

In a different application, Time Domain methods were proposed to analyze infant cry for the purpose of diagnosing a specific disease. The interested reader is referred to [100] for further discussion.

2.3.1.2.2 Frequency Domain Analysis

Frequency Domain shows the distribution of the signal within specific ranges of frequencies. The fundamental frequency (F_0) is a well-known Frequency Domain property that represents the lowest frequency of a periodic signal. According to [101], infant cries can be classified based on the fundamental frequency into:

1. Phonated cries that have a smooth and harmonic structure with a fundamental frequency range of 400 to 500 Hz.
2. Dysphonated cries that have less harmonic structure compared to phonated cries.
3. Hyperphonated cries with an abrupt and upward shift in pitch (up to 2000Hz). This class of cries is associated with a painful stimulus.

Phonated, dysphonated, and hyperphonated fundamental frequency of neonatal sounds can be estimated using different methods presented in [101, 102, 103]. Pal et al. [104] used the Harmonic Product Spectrum (HPS) method to extract the fundamental frequency (F_0) method along with the first three formants (i.e., F_1, F_2, and F_3) from crying signals of infants recorded during several emotional states (i.e., pain, hunger, fear, sadness, and anger). The paper does not provide any information about the database (e.g., number of subjects, age range, and etc). Moreover, no information was given about the data collection procedure and the stimuli that triggered those emotional states. After extracting the features, k-means algorithm was applied to find the optimal parameters that maximize the separation between features of different types of cry. Combining F_0, F_1 and F_2 produced the best clustering and achieved an accuracy of 91% for pain, 72% for hunger, 71% for fear, 79% for sadness, and 58% for anger. The high accuracy of pain cry can be attributed to the fact that this type of cry has a distinctive and higher fundamental frequency as compared to other types of cries.

In [105], Fuller and Horii presented a Frequency Domain method to analyze four types of infant cry: pain, hunger, fussy, and cooing. The utilized database consists of vocal samples collected from 41 healthy neonates (2-6 months old). Pain cry samples (42 samples) were recorded during a routine intramuscular immunization. Hunger cry samples (16 samples) were recorded prior to infants' usual feeding time. Fussy cry samples (28 samples) were recorded during the naptime in infants that

were identified as tired. An infant's response to the mother's soft sound and fondling represented the cooing state samples (23 samples). In the preprocessing stage, the collected samples were divided into multiple time segments, with 512 data points length, that receive Hamming weighting before computing the fast Fourier transform. Then, the mean value of the spectral energy levels was computed for each vocal sample and used to perform ANOVA statistical analysis. The result showed that there is a significant difference between the cooing sound and the other cries (pain, hunger, and fussy). It also showed that the spectral characteristics of pain-induced cry is quantitatively different than the other cries (hunger and fussy). Particularly, the spectral energy distribution of pain cry has significantly less difference between the amplitude of the various frequency locations and maximal amplitude than other cries (hunger and fussy) and cooing.

Pai et al. [106] presented a spectral method to classify infants' cry as a whimper or vigorous. The database of this work was collected from 27 infants, average age is 36 gestational weeks, hospitalized in the NICU at a local hospital in Tampa, Florida. The audio data were recorded during acute painful procedure (i.e., heel lancing and immunization). Two types of pain cry were recorded, whimper cry (14 samples) and vigorous cry (20 samples). The ground truth labels for the recorded samples were given by trained nurses using NIPS pain scale (see Table 1.1). To obtain the power spectrum for each sample, Welch's method was applied in 20-milliseconds windows. After getting the spectrum, Linear Predictive Coefficients (LPC) along with other statistics (e.g., mean and standard deviation) were extracted from each sample and used to train kNN. The average accuracy of the classifier, evaluated using 10-fold cross validation, was 76.47%.

2.3.1.2.3 Cepstral Domain Analysis

The Cepstral Domain of a signal is generated by taking the Inverse Fourier transform (IFT) of the logarithm of the signal's spectrum. Mel Frequency Cepstral Coefficients (MFCC) is a common Cepstral Domain method that is used to extract a useful and representative set of features (coefficients) from a sound signal and discard noise and non-useful features.

One of the first studies to analyze infant cry using MFCC was introduced in [107]. The proposed method was applied to a database that consists of 230 cry episodes collected from 16 healthy neonates (2 to 6 months old). The crying episodes were recorded during three different stimuli: immunization (pain), jack-in-the-box (fear), and head restraint (anger). The cry signals of fear and

anger were combined together to represent the no-pain cry. Prior to the feature extraction stage, all episodes were filtered to 8000 Hz using low-pass filter, sampled at 16 kHz, and segmented into 256-sample frames (16 ms) with 50% overlap. For each segment, 10 MFCCs were extracted and fed into a neural network as input. The testing protocol was 10-fold cross validation. The highest correct classification rates for pain and no pain classes were 92.0% and 75.7% respectively.

Barajas-Montiel et al. [108] presented MFCC-based method to classify infant cry as pain cry, hunger cry, and no-pain-no-hunger cry (sleepy and discomfort) using Fuzzy Support Vector Machine (FSVM). FSVM is an extension of SVM that reduces the effect of outliers by assigning a fuzzy value or weight for each training point rather than assigning equal points as in SVM. The database utilized in this work consists of 1627 cry samples collected and labeled by medical doctors. A total of 209 samples were recorded during pain, 759 samples were recorded during hunger, and 659 samples were recorded during other states such as sleepiness and discomfort. In the preprocessing stage, each cry sample was filtered, normalized, and divided into segments of one second. Every one second segment was further divided into frames of 50-milliseconds and then 16 MFCC coefficients were extracted from each frame. This procedure generated, for each sample, a high-dimensional vector; PCA was used to reduce the vector dimensionality. Then, the reduced feature vector was used to train FSVM, which achieved 97.83% accuracy.

Yousra and Sharrifah [109] introduced a Cepstral Domain method to classify infant cry as pain or no-pain (hunger and anger). A set of 150 pain samples and 30 no-pain samples were recorded for infants ranging from newborns up to 12 months old. The pain samples were recorded during routine immunization procedures in the NICU at a local hospital. The no-pain samples were recorded at infants' homes. Of the 180 recorded samples, 881 samples were obtained by creating one second segments. These samples were then used to extract two sets of features, namely Mel Frequency Cepstral Coefficients (twelve MFCC coefficients) and Linear Predication Cepstral Coefficients (sixteen LPCC coefficients). The extracted features were fed to a neural network trained with the scaled conjugate gradient algorithm; 700 samples were used for training the network and 181 samples were used for testing. The proposed method achieved 68.5% and 76.2% accuracies for LPCC and MFCC, respectively. These results suggest that MFCC features outperform LPCC features in detecting infant pain cry.

Similarly, Vempada et al. [99] investigated the use of MFCC (i.e., 13 MFCCs, 13 delta MFCCs and 13 delta-delta MFCCs) along with other Time Domain features for classifying infant cry as pain, hunger, or wet-diaper. Each recorded sample was segmented into 20 milliseconds frames. Then, MFCC was applied to each frame to extract useful features for classification. Part of the extracted features was used to build the SVM model and part was used to evaluate its performance. The average accuracies for pain, hunger, and wet-diaper are 30.56%, 66.67%, and 86.11% respectively. Referring to the results of Vempada et al. [99] under Time Domain Analysis, it can be seen that the wet-diaper cry has good accuracy using both Time and Cepstral features. However, pain cry is poorly recognized using MFCC features and hunger cry is poorly recognized using Time features. To improve the overall performance, feature fusion and score fusion of Time and Cepstral Domains were performed. The feature fusion achieved 77.78%, 61.11%, and 83.33% accuracies for pain, hunger and wet-diaper. The average accuracies using score fusion for pain, hunger, and wet-diaper are 80.56%, 75%, and 86.11% respectively. These findings show that the fusion of different domains can be a good practice for analyzing infant cry.

Before we conclude, we would like to note that none of the works presented above have their database or code publicly available for research use, according to the authors who were contacted through email and our online search in public repositories.

2.3.1.3 Body Movement Analysis

Neonates tend to move their head, extend their arms/legs, and splay their fingers when they experience pain. Following the same structure, this section should provide a summary of the existing methods that analyze body movement for the purpose of assessing neonatal pain. However, we are not aware of any method, except the method presented in this dissertation, that investigated neonatal pain assessment from body movement.

2.3.2 Physiological Pain Responses

Pain analysis based on physiological responses can be defined as the process of extracting pain-relevant features from the body's physiological responses. Examples of the most common physiological responses include changes in vital signs and cerebral hemodynamic activity.

2.3.2.1 Vital Sign Analysis

Vital signs readings represent the changes in the body's basic functions such as changes in the heart rate. Caregivers monitor these signs at frequent intervals to check the body condition and understand underlying medical problems. The four main vital signs that are frequently checked by health professionals are Heart Rate (HR), Respiratory Rate (RR), Blood Oxygen Saturation (SpO2), and Blood Pressure (BP).

The adhesive electrodes and sensors are the most common technology for monitoring vital signs in the NICU. These sensors are placed on the infant's skin to record vital signs signals. Then, the recorded signals are transferred via a translating component to a format that can be displayed on the monitor. To further analyze vital signs data, most monitors provide a wireless data stream to an electronic medical record or allow exporting these signs as a time-stamped Excel file.

Different studies utilized vital signs data to study the association between these signs and pain. For example, Lindh et al. [110] described a method to study the association between heart data and neonatal acute pain by analyzing the Heart Rate Variability (HRV) in the Frequency Domain. Vital signs monitor was used to collect heart data from 25 infants (postnatal age of 72-96 hours) in four different events: 1) baseline; 2) sham heel prick (i.e., warming the foot and lancing it with intact lancet); 3) sharp heel prick; 4) and squeezing the heel for blood sampling. The recorded data were inspected for error detection and the artifact were removed by applying interpolation. Then, Statistical and Spectral analyses were carried out on the exported heart data to compute the Heart Rate mean (HR_{mean}), the Power in Low Frequency (P_{LF}) the Power in High Frequency (P_{HF}), and the Total Heart Rate Variability (P_{tot}). The computed values were used to perform Multivariate Statistics to illustrate the correlation between these variables and each of the four events. The results showed significant increases in HR_{mean}, P_{tot}, and P_{LF} between baseline and sharp prick. The results also showed that squeezing the heel for blood sampling during the heel lancing causes a significant increase in HR_{mean} and decrease in P_{tot} and P_{HF} as compared with baseline and sharp prick.

Faye et al. [111] presented a method to analyze the Heart Rate Variability (HRV) for 28 infants (age $>$ 34 gestational weeks) with chronic pain. EDIN pain scale [112] was used to score the pain and separate infants into two groups: (1) Low EDIN with EDIN pain score $<$ 5, and (2) High EDIN

with EDIN pain score ≥ 5. To study the association between chronic pain and cardiovascular data, Linear Regression Analysis was performed using the mean of Heart Rate (HR_{mean}), Respiratory Rate (RR_{mean}), Blood Oxygen Saturation ($SpO2_{mean}$), and High Frequency Variability Index (HFVI). The results showed that HRV changed (i.e., significant decrease) between the two groups; and no significant changes in RR and SpO2 were found between the two groups. The results also showed that HFVI (< 0.9 threshold) was able to assess pain with a sensitivity of 90%, a specificity of 75%, and 0.81 Area Under the ROC curve.

Although measuring vital signs using the readily available adhesive electrodes/sensors is the current standard for collecting these data, this standard is expensive, causes stress, and can damage the infants' delicate skin. Therefore, it has been suggested to use contactless and non-invasive methods for monitoring infants' vital signs. Examples of video-based vital signs detection methods are presented in [113, 114, 115, 116].

In summary, we presented above the current efforts for assessing neonatal pain using vital signs. Although studies have found a correlation between changes in vital signs and pain, vital sign changes are sensitive to other states (e.g., hunger and fear) and underlying illness [117]. Therefore, it has been suggested [117] to use vital signs in conjunction with behavioral indicators, which are considered more pain-specific, for pain assessment.

2.3.2.2 Cerebral Hemodynamics Analysis

Studies [118, 119, 120, 121, 75, 122, 123, 124] have shown that there is an association between changes in cerebral oxygenation and pain. The most popular methods to measure the cerebral oxygenation changes are Functional Magnetic Resonance Imaging (fMRI) and Near Infrared Spectroscopy (NIRS). fMRI is a safe method for measuring the brain hemodynamic activity. It produces an activation map that shows which parts of the brain get activated during an emotional event such as pain. NIRS is similar to fMRI but it is less invasive and more suitable for bedside monitoring. It measures, using small probes attached to the head, subtle changes in the concentration of oxygenated hemoglobin [HbO_2] and de-oxygenated hemoglobin [HbH].

Bartocci et al. [119] introduced a NIRS-based method to measure the brain hemodynamic activity for 40 infants, with age >= 26 gestational weeks, during three periods: 1) baseline (P_0); 2) tactile stimulus for cleaning (P_1); and 3) venipuncture painful stimulus (P_2). All the data were

collected in the NICU at local Hospitals in Sweden and Italy using a double-channel Near Infrared Spectroscopy Device (NIRO 300). This device is widely used in neonatal research to measure functional activation of the cortex. Each infant was recorded in the baseline period (P_0) when s/he was in a quiet, awake, and stable condition. The tactile stimulus period (P_1) was recorded after the disinfecting of the infant's skin with an alcohol-soaked cotton at room temperature. The painful period (P_2) was recorded for at least 60 seconds following the insertion of the needle. For all the 40 infants, NIRS data (i.e., HbH, HbO_2, and $HB_{total} = HbH + HbO_2$) along with vital signs data (i.e., HR and SaO_2) were collected during the three periods. The collected data were sampled and exported to a computer for further analysis. Next, $[HbO_2]_{dif}$, $[HbH]_{dif}$, and $[HB_{total}]_{dif}$ were computed by subtracting their values in P_0 from their values in P_1 and P_2 periods. Also, the average values of these measurements were computed and used to perform Student's t-test, ANOVA, and Newman-Keuls post hoc statistical tests. The results showed a significant increase in HR and decrease in SaO_2 between P_0 and P_2 periods. For the NIRS measurements, a significant increase was found in the HbO_2 concentrations in both hemispheres between P_0 and P_2 periods; HbO_2 increase was more pronounced in male than female infants.

Another NIRS-based method was presented in [122] to measure the brain hemodynamic activity for 18 infants in the NICU at the University College London Hospital, London. The infants' age ranges from 25 to 45 postmenstrual weeks. Vital signs readings along with NIRS data (i.e., HbH, HbO_2, and $HB_{total} = HbH + HbO_2$) were recorded, using NIRO 300 device, during baseline and heel lancing periods. The data collection of baseline was performed 20 seconds prestimulus. After the insertion of the lancet, the infant's foot was not squeezed for a period of 30 seconds to ensure that the evoked response occurred because of the initial stimulus and not the squeezing. The collected data were sampled and the maximum changes from the baseline were calculated for each measure. The result of the statistical analysis (t-test) indicated that the painful stimulus produced a clear cortical response that is measured as an increase in HB_{total} in the contralateral somatosensory cortex. This cortical response was more pronounced in awake infants than in sleeping infants. Moreover, the results showed that the response in the contralateral somatosensory cortex for awake infants increases with age. Extensions of this work are presented in [123] to study the relation between NIRS data and behavioral indicators of pain and in [125] to investigate the impact of age and frequency of painful procedures on the brain neuronal responses.

For postoperative pain, Ranger et al. [120] presented a NIRS-based method to assess infants' postoperative pain based on analysis of hemodynamic activity in brain regions. NIRS data (i.e., HbO_2 and HbH) for 40 infants ($<$ 12 months) were recorded, using NIRO 300 device, during the following periods: 1) chest-drain removal procedure following cardiac surgery (T_2); 2) removal of the dress (T_1); 3) and baseline (T_0). To verify associations between NIRS data and pain stimulus, Univariate Linear Regression was performed on the extracted measures. The results showed a significant increase in HbH during pain (i.e., the difference of HbH measurement between the baseline (T_0) and pain (T_2) was significant).

Before we conclude, we would like to draw the reader's attention to the difference of cortical response between the postoperative and procedural pain. Procedural pain produced changes measured as an increase in HbO_2 [119] or HB_{total} [122] while the postoperative pain caused an increase in HbH [120]. Also, we want to note that none of the databases described under this section are publicly available according to the authors, contacted via email, and our online search in public repositories.

2.3.3 Fusion of Pain Responses

The methods discussed so far utilize a single pain response or modality for assessment. Because pain is expressed through multiple responses, existing pediatric pain scales are multimodal incorporating both behavioral and physiological responses for assessment. Multimodality allows for a reliable assessment of pain in case of missing data due to occlusion, noise, gestational age (e.g., weak facial muscles in premature neonates), physical exertion or exhaustion, and sedation.

Pal et al. [104] described a multimodal emotion detection method that predicts the emotional state of neonates based on analysis of facial expression and crying. Facial features were extracted from the infant's facial expression and the fundamental frequency along with the first three formants were extracted from the crying signals. The extracted features for each modality were then used to build a single classifier and a decision-level fusion method was applied to combine the decision labels for both classifiers. Specifically, facial expression and crying modalities were combined by finding the conditional probability matrices and using the index for the maximum value of a belief vector, which is derived from the probability matrices, as the final fused decision. The overall accuracy for predicting neonates' emotions using a decision-level fusion was 75.2%.

Decision-level fusion methods are easy to implement because it depends on combining different classifiers' labels. However, this level of fusion can result in loss of information since it assumes that the modalities are independent (i.e., the correlation between modalities is ignored). Feature-level fusion can mitigate this issue by combining all the modalities together in a rich and high-dimensional feature vector. However, the high-dimensionality of the feature vector along with the scaling and missing data can raise several issues in practice. These issues can be handled using methods such as standardization for scaling, PCA for reduction, and interpolation for missing data. To the best of our knowledge, there is currently no work, except the work presented in this dissertation, that combines different pain responses at the feature level for the purpose of assessing neonatal pain.

2.4 Pain Recognition Methods

We divide the automatic pain recognition into two main classification tasks: pain detection and pain intensity estimation. We present next a description and a discussion of limitations for each task.

2.4.1 Pain Detection

Pain detection aims to identify the presence or absence of pain emotion. It is a typical classification problem in which discrete classes are considered the output of a classifier. For example, a classifier that is trained with pain-relevant features can be used to classify the emotional state of an infant as pain or no-pain.

SVM classifier is commonly used for pain detection (e.g., [126, 127, 49, 40, 90, 88, 83, 95, 60, 59]). Other classifiers that are used for pain detection are Neural Network [39, 107] and k-means [104]. Such classifiers achieved varying levels of performance in detecting the pain label.

Pain detection provides the pain label without the intensity or the level of the detected pain. For pain assessment application, detecting the pain without its intensity may not be enough for the following three reasons. First, providing the pain label without its level does not reflect the severity of pain. Second, it does not reflect the individual differences in response to a painful stimulus; an infant's level of pain might be different than that of another infant. Third, producing the label without its intensity does not provide information about the pain dynamic and how it changes

over time; an infant might experience different pain intensities at different time intervals. Due to these reasons, we believe estimating the intensity of pain is important and can lead to better understanding and intervention.

2.4.2 Pain Intensity Estimation

Estimating the intensity of the detected pain provides better pain assessment and might lead to better pain management. Several pain recognition methods were proposed for pain intensity estimation.

For example, Gholami et al. [84] presented a method to estimate pain intensity using RVM. Unlike SVM, RVM classifier outputs the probabilities of the class memberships or labels. The uncertainty for each class membership was used to estimate infants' pain intensity. For validation, the automated intensity estimation was compared with the intensity estimation provided by expert and non-expert observers using kappa coefficient. This coefficient ranges from 0 to 1 where larger values indicating better reliability. The agreement between RVM and human observers was 0.48 for experts and 0.52 for non-experts.

Hammal et al. [41] described a method to estimate pain intensities for 25 subjects with an orthopedic injury. Four SVM classifiers were built separately to automatically assess four levels of pain. To measure the reliability of judgments between the automatic estimation and the manual estimation, Intra-class Correlation Coefficient (ICC) was used. ICC has a range from 0 to 1 with values close to 1 indicates high similarity. The results showed moderate (0.55 ICC) to high (0.85 ICC) consistency between the manual and automated pain intensity assessment.

Similarly, Gruss et al. [61] introduced a method to estimate four levels of pain using SVM. Facial expression and biopotentials signals were recorded under four levels of pain (T_1 to T_4) as described in Section 2.5.1 (BioVid Heat Pain Database). Then, the recorded signals were analyzed to extract different mathematical features. These features were used to build the SVM classifiers, which were trained with 75% of the data and tested on 25% of data. The proposed method achieved 76.00% (sensitivity) and 82.59% (specificity) for baseline vs T_1, 80.00% (sensitivity) and 82.59% (specificity) for baseline vs T_2, 84.71% (sensitivity) and 85.18% (specificity) for baseline vs T_3, and 92.24%, (sensitivity) and 89.65% (specificity) for baseline vs T_4.

2.5 Existing Pain Databases

The quality, complexity, and capacity are three important factors that should be considered when collecting databases for pain assessment. Low-quality databases with a vague notion of suffering and inadequate annotations can lead to inaccurate results. Also, the complexity of the database, in terms of its modalities/dimensions, is critical to develop reliable multimodal pain assessment system that can still assess pain during the failure of recording a specific pain indicator. Finally, databases with a relatively small number of subjects are not sufficient to draw solid conclusions. Therefore, collecting high quality, multimodal, and large data is necessary for developing robust pain assessment systems.

Most of the existing pain databases are not publicly available for research use because of IRB (Institutional Review Board) regulations and restrictions to protect the privacy of subjects. This section provides brief descriptions of the publicly available pain databases for adults and neonates.

2.5.1 Adult Pain Databases

UNBC-McMaster Shoulder Pain Expression Archive [43] is one of the first databases to address the need for adequately annotated and publicly available databases of pain expression. The database consists of videos collected from 129 subjects (63 males and 66 females) during a series of movements to test their affected and unaffected shoulder. All videos were manually coded using FACS (48398 FACS coded frames). The database has self-report and observer ratings for each video sequence.

The BioVid Heat Pain Database is an advanced multimodal database introduced by Walter et al. [54]. This database contains video and biopotentials signals (i.e., Skin Conductance Level [SCL], Electrocardiogram [ECG], Electromyogram [EMG], and Electroencephalography [EEG]) for 90 subjects with age distributions of 18 to 35 (group 1), 36 to 50 (group 2), and 51 to 65 (group 3). Each group has a total of 30 subjects (15 male and 15 female). All subjects underwent experimentally induced heat stimulus with four intensities or pain levels (T_1 to T_4). To adjust the level of the stimulation, a subject-specific pain threshold and a pain tolerance were determined. Every pain level was stimulated 20 times (i.e., a total of 80 stimulation). In each stimulus, the maximum temperature the subject could take was held for four seconds and there was a pause duration of 8-12 seconds between the stimuli. This procedure was repeated twice, once when the subject's face

was recorded and once when the biopotentials sensors were attached. The subject's face and head pose were recorded using three cameras (AVT Pike F145C cameras) and a Kinect. The biopotentials data were recorded using a Nexus-32 amplifier. More discussion about the experiment setup, sensors' channels, and the synchronization procedure of this database can be found in [54].

2.5.2 Neonatal Pain Databases

iCOPE/COPE, collected by Brahnam et al. [76], is the first pain expression database that is designed specifically for the automatic assessment of neonatal pain. The database consists of 204 static images captured, using Nikon D100 digital camera, for 26 healthy infants (50% female). The infants' age ranges from 18 hours to 3 days old. Before the data collection session, all infants were fed and they were swaddled to get an unobstructed image of the face. The images for each infant were taken during four stimuli: 1) the puncture of a heel lance; 2) friction on the external lateral surface of the heel; 3) transport from one crib to another; and 4) an air stimulus to provoke an eye squeeze. The main limitation of this database is the 2D static images that do not show the expression's dynamic and how it evolves over time. Currently, Dr. Brahnam and her team are working on collecting a new and more challenging video database (COPE 2). This database is not yet available for research use. Another limitation of this database is the single modality (i.e., facial expression). As discussed earlier, incorporating different pain indicators is important to ensure proper and reliable assessment of pain.

The YouTube videos database is another publicly available database for neonatal pain [77]. The database consists of YouTube videos recorded, by parents or a guardian, for neonates receiving immunizations; the infants' age ranges from less than a month to 12 months old. The recorded videos show the infant's face, body, and have sounds. Along with the raw videos, other data such as the infant's gender, number of injections, and the gender of the caregiver were collected. All videos were scored by experts using FLACC [128] (Face, Legs, Activity, Cry, Consolability) pain scale. The main limitation of this database is the low-quality of the recorded videos which leads to the exclusion of many videos from annotations.

As far as we are aware, COPE and YouTube databases are the only neonatal databases that are available per request for research in pain detection. This suggests that there is an essential need

for collecting high-quality, multimodal, and relatively large pain databases to advance the research of automatic pain assessment in neonates.

2.6 Limitations of Automatic Pain Assessment

There are several limitations that should be addressed to advance the state of automatic pain assessment. These limitations can be summarized as follows:

1. There are very few accessible databases for research in neonatal pain. At the time of writing this dissertation, we are only aware of two databases, COPE and YouTube videos, that are available per request for research in neonatal pain assessment. To advance the automated assessment of neonatal pain, researchers need to have access to advanced and multimodal databases that are collected and annotated by experts in the field.

2. Existing methods for automatic pain assessment focus on adults. We think this focus is attributed, in addition to the database-accessibility issue, to the common belief that the algorithms designed for adults should have similar performance when applied to infants. Contrary to this belief, we think the methods designed for assessing adults' pain will not have similar performance and might completely fail for two reasons. First, the facial morphology and dynamics vary between infants and adults. Furthermore, infants' facial expressions include additional movements and units that are not present in the Facial Action Coding System (FACS). As such, Neonatal FACS was introduced and designed specifically for infants. Second, we think the preprocessing stage (e.g., face and body movement tracking) is more challenging in infants because they are uncooperative subjects recorded in an unconstrained environment (i.e., NICU).

3. Most of the existing approaches assess pain based on analysis of a single pain response or modality (e.g., facial expression). Studies have shown that pain causes behavioral and physiological changes and suggested considering multiple modalities for better pain assessment. In addition, it has been reported that some infants have

limited ability to behaviorally express pain due to developmental stage, movement disorder, or physical exertion. Therefore, it is important to develop multimodal approaches that can better handle the missing data.

4. Existing methods for assessing pain do not take the contextual and clinical data (e.g., medication type and dose, age and gender) into account when analyzing pain. Studies found an association between infants' clinical data and their reaction to pain experience. For example, it has been shown [24, 25, 26, 28] that infants of different age groups respond to pain differently. Hence, incorporating clinical and contextual information with other pain responses is necessary to refine the assessment process and obtain a context-sensitive pain assessment method.

5. Existing methods for assessing pain focus on the procedural pain rather than the postoperative pain. Continuous monitoring is more needed for the postoperative pain since it requires prompt pain detection and immediate intervention.

This dissertation addresses some of the above-mentioned limitations and proposes an automatic and multimodal system for assessing neonatal pain. The proposed system was tested on a real-world database collected at the Tampa General Hospital (TGH). The method of data collection is described in the next chapter.

CHAPTER 3

NEONATAL PAIN DATABASE

In this chapter, we present our comprehensive Neonatal Pain Assessment Database (NPAD). The chapter starts by discussing some facts about the hospital where we collected the data (Section 3.1). A description of the data collection process is provided in Section 3.2. Finally, the data collection of Near-infrared Spectroscopy (NIRS) data is presented in Section 3.3.

3.1 Tampa General Hospital

The Neonatal Intensive Care Unit (NICU) at Tampa General Hospital (TGH) is one of the biggest NICUs in the state of Florida. The care team in the NICU consists of neonatologists from the USF Health Morsani College of Medicine, staff pediatricians at TGH, and neonatal nurse specialists. The NICU admits, on average, around 755 neonates (approx. 50% female) from all major racial and ethnic categories each year. Table 3.1 shows the racial and ethnic distribution of the neonates hospitalized in the NICU at TGH between 2014 and 2016. The neonates born in the NICU at TGH represent all neonatal gestational ages and weight categories.

Table 3.1: Neonates' racial/ethnic distribution.

Admission Date	**Caucasian**	**Black**	**Hispanic**	**Asian**	**Others**	**Total**
1/1 – 12/31, 2014	417	204	182	15	14	832
1/1 – 12/31, 2015	511	300	282	24	23	1141
1/1 – 5/25, 2016	78	37	41	5	0	161
Total	1006	541	505	44	37	2133

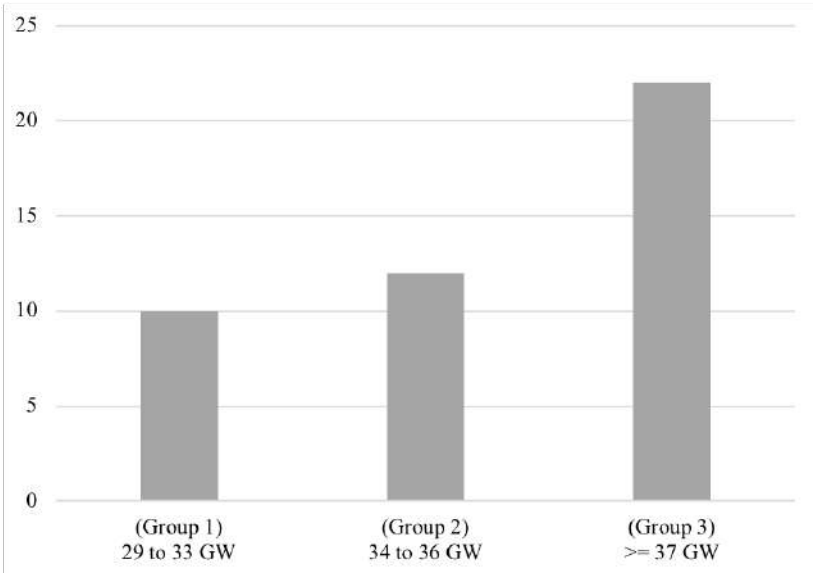


Figure 3.1: Histogram distribution for neonatal gestational age.

3.2 Data Collection

3.2.1 Participants' Demographics

Data was collected for a total of 44 Neonates (50% female) in the NICU at TGH. The age of the participating neonates ranged from 30 to 40 GW, with a mean age of 35.97 (SD = $\pm$ 2.87). Neonates who have significant facial abnormalities were excluded. Figures 3.1 to 3.3 present the distributions of age at birth, weight, and race among the participating neonates. It is important to note that this study is an IRB-approved study that requires an informed consent from the parents before the study's enrollment. The consent form of this study is presented in Figure 3.4.

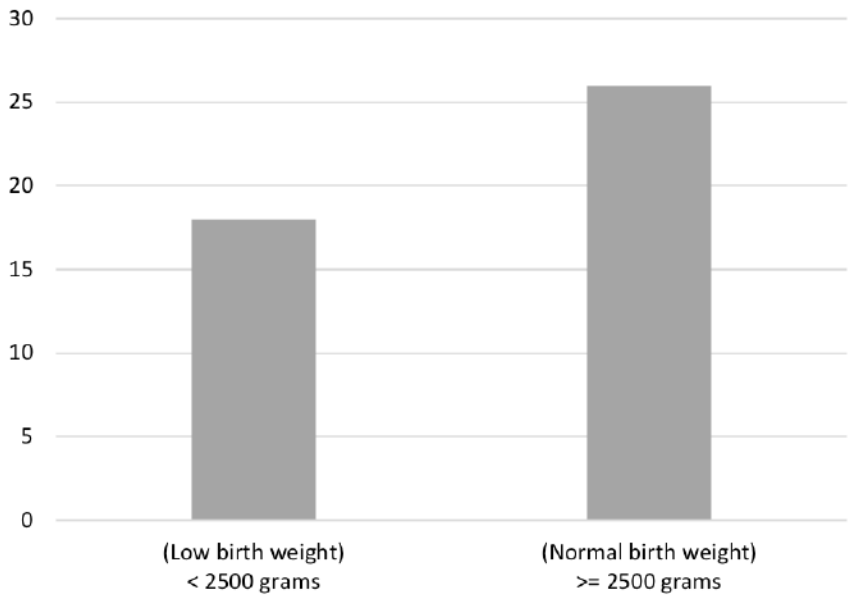


Figure 3.2: Histogram distribution for weight in grams.

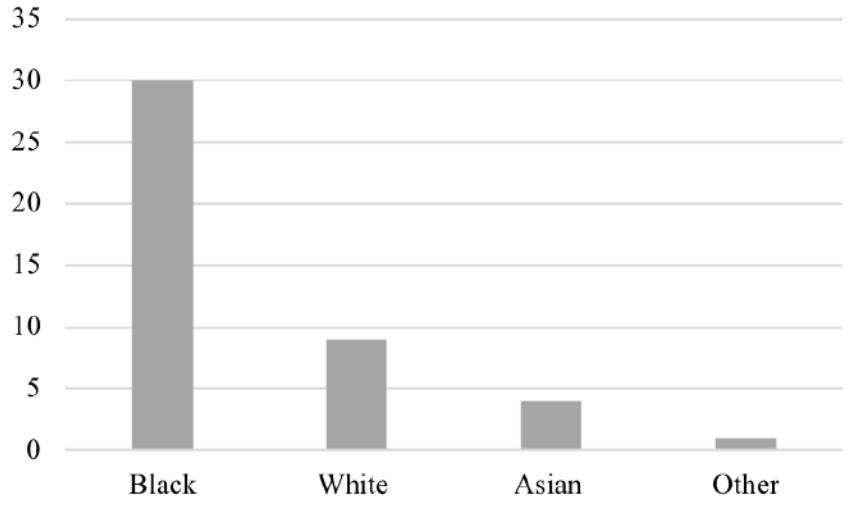
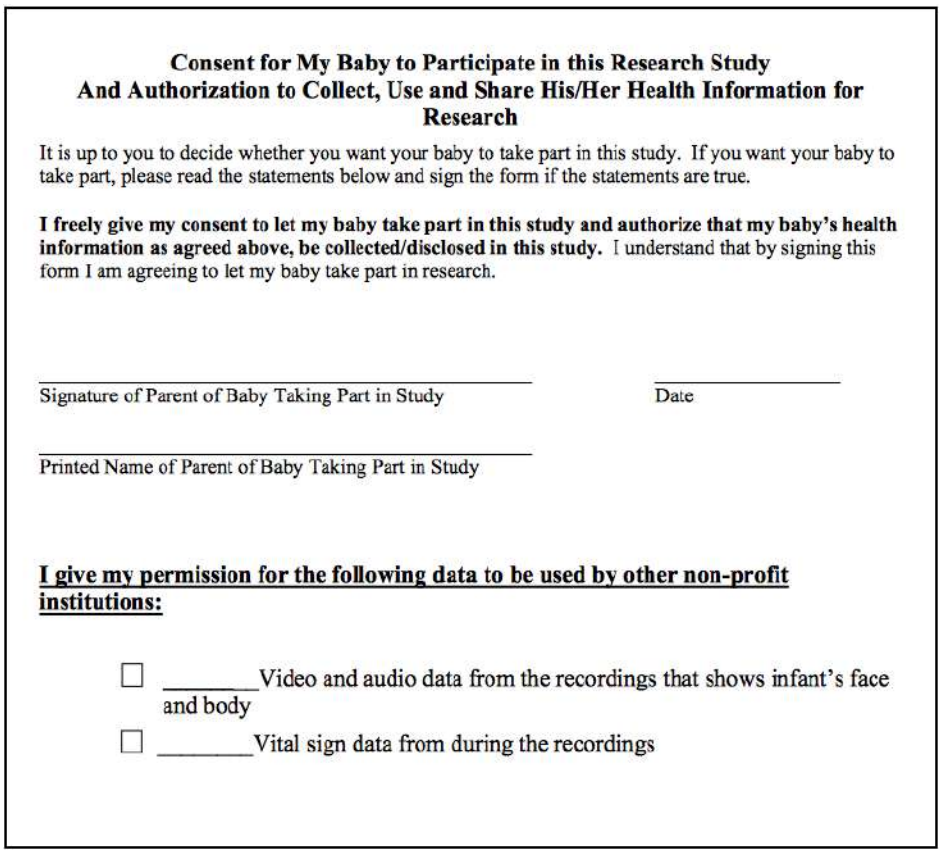


Figure 3.3: Histogram distribution for race.



Consent for My Baby to Participate in this Research Study
And Authorization to Collect, Use and Share His/Her Health Information for Research

It is up to you to decide whether you want your baby to take part in this study. If you want your baby to take part, please read the statements below and sign the form if the statements are true.

I freely give my consent to let my baby take part in this study and authorize that my baby's health information as agreed above, be collected/disclosed in this study. I understand that by signing this form I am agreeing to let my baby take part in research.

_____ _____
Signature of Parent of Baby Taking Part in Study Date

_____
Printed Name of Parent of Baby Taking Part in Study

I give my permission for the following data to be used by other non-profit institutions:

☐ _____Video and audio data from the recordings that shows infant's face and body

☐ _____Vital sign data from during the recordings

Figure 3.4: Consent form of this study.

3.2.2 Equipment and Setup

We collected video, audio, and physiological data from neonates while in a baseline state, and during painful procedures using the following equipment:

1. GoPro Hero camera was used to record video and audio signals. The camera was triggered remotely using the GoPro application installed on a smart device. The recorded data included the neonate's face, head, and body, as well as the sounds of neonates and background noise (e.g., sounds of equipment and nurses). The camera was installed on a stand that faces the neonate's incubator as illustrated in Figure 3.5.

2. Philips MP-70 monitor was used to collect vital signs such as heart rate (HR), respiratory rate (RR), oxygen saturation levels (SpO_2), and blood pressure (BP).

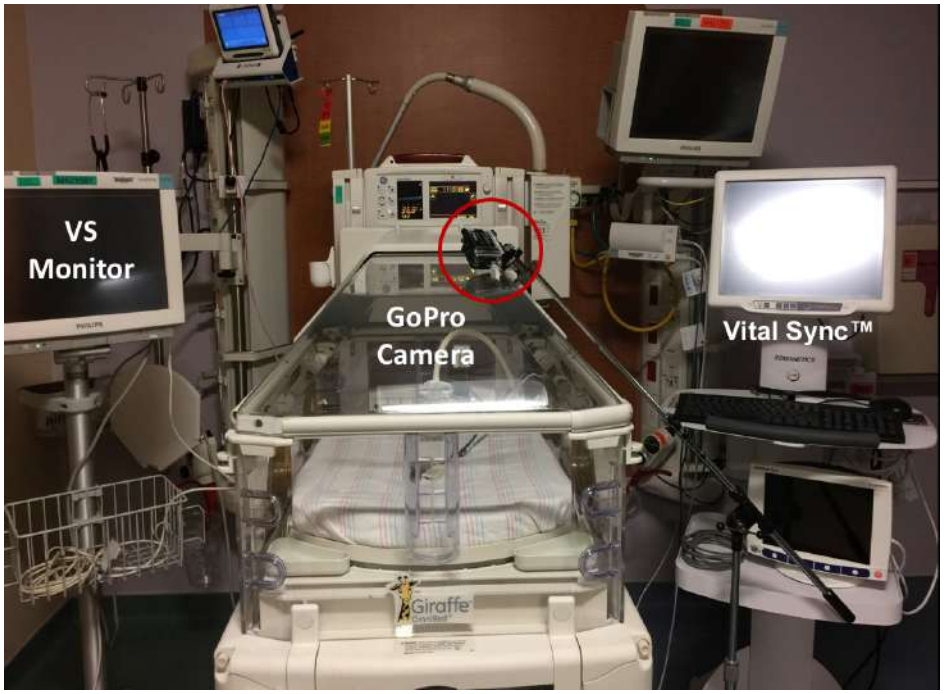


Figure 3.5: Setup of data collection.

All neonates hospitalized in the NICU at TGH have their vital signs continuously monitored as standard care using this monitor.

3. Vital Sync^{TM} system to integrate the neonatal data from bedside devices (Philips MP-70) into a single time-synched unit. By connecting Philips MP-70 monitor to the Vital Sync^{TM}, we were able to mark the start, end, and ground truth labeling events. Then, we exported, for each event, the collected time-stamped vital signs data as an excel file. Figure 3.5 shows the setup of the study's equipment.

To ensure data synchronization, we marked the start and end points of data collection by simultaneously inserting a timestamped-event to the Vital Sync^{TM} monitor and using a clapperboard with the video/audio stream. We also marked, using the same method, the time of assessing pain by the bedside caregivers. Before we conclude, it is important to point out that all the data was collected during routine clinical procedures and carried out in the normal clinical environment that was only modified by the addition of the cameras. This makes our database highly representative of the real-world condition.

3.2.3 Contextual and Medical Data

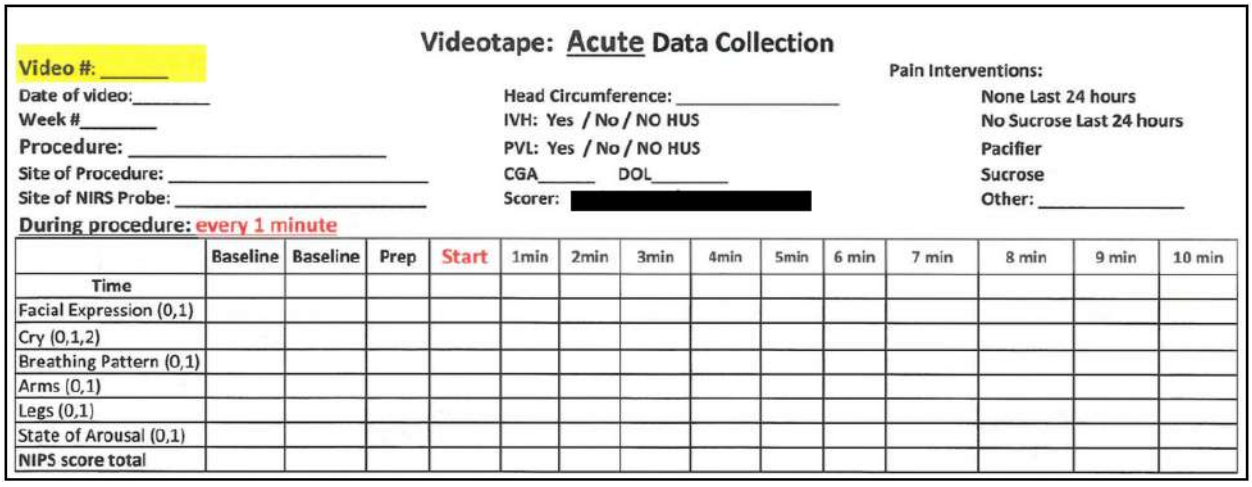
We collected, along with the video, audio, and physiological data, several contextual and medical data. Examples of these data include: a) pain characteristics such as procedural and postoperative pain; b) gestational age (GA) and day of life age; c) clinical data such as the medication type and dose, weight, race/ethnicity, and gender; and d) non-pharmacological interventions such as the mother's presence, sweeties, swaddling, holding/rubbing, or pacifier use. All these data were documented in the Electronic Medical Records (EMR) as well as the study data collection forms.

3.2.4 Pain Stimuli and Ground Truth

Data was collected during procedural and postoperative painful stimuli by bedside caregivers in the presence of a research assistant and the principal investigator.

A total of 35 neonates were recorded during procedural pain. The stimuli that trigger procedural pain are routine heel lancing and immunization. The recording for procedural pain consists of eight time periods: baseline period (T_0), procedure preparation period (T_1), the painful procedure period (T_2), and five post-painful-procedure periods (T_3 to T_7). The pain assessment for each of these periods was documented by bedside caregivers using the Neonatal Infant Pain Scale (NIPS). This pain scale consists of facial expression, cry, arms and legs movement, vital signs, and state of arousal. The label for each pain response is 0 or 1 except for cry, which can be labeled as 0, 1, or 2. Adding the labels of NIPS's components generates a total pain score, which is used to obtain, through thresholding, three emotional states or labels: no pain state for a score of 0-2, moderate pain state for a score of 3-4, and severe pain state for a score greater than 4. These states provide the ground truth labels that are used to train the machine learning classifiers. Figure 3.6 presents the scoring sheet of the procedural or acute painful stimulus.

As for the postoperative pain, 9 neonates were recorded for 15 minutes prior to a major surgery (e.g., Gastrostomy tube and Omphalocele repair) to get their baseline state and for three hours after the surgery. The pain score was documented by bedside caregivers, during the baseline and every 15 minutes during the postoperative state, using Neonatal Pain, Agitation and Sedation Scale (N-PASS). This pain scale consists of facial expression, crying irritability, behavior state, extremities tone, and vital signs. The label for each pain response ranges from -2 to 2. Adding the



Videotape: Acute Data Collection

Video #: _____
Date of video: _____
Week # _____
Procedure: _____
Site of Procedure: _____
Site of NIRS Probe: _____
During procedure: every 1 minute

Head Circumference: _____
IVH: Yes / No / NO HUS
PVL: Yes / No / NO HUS
CGA_____ DOL_____
Scorer: [redacted]

Pain Interventions:
None Last 24 hours
No Sucrose Last 24 hours
Pacifier
Sucrose
Other: _____

	Baseline	Baseline	Prep	Start	1min	2min	3min	4min	5min	6 min	7 min	8 min	9 min	10 min
Time														
Facial Expression (0,1)														
Cry (0,1,2)														
Breathing Pattern (0,1)														
Arms (0,1)														
Legs (0,1)														
State of Arousal (0,1)														
NIPS score total														

Figure 3.6: Example of the scoring sheet for a procedural painful procedure.

labels of N-PASS components generates a total pain score that ranges from -10 to 10. Performing a thresholding on the generated score provides four emotional states: deep sedation (a score of -10 to -5), light sedation (a score of -5 to -2), normal state (a score of -2 to 3), and pain state (a score larger than 3). These states provide the ground truth labels that are used to train the machine learning classifiers. Figure 3.7 presents the scoring sheet of the postoperative painful stimulus.

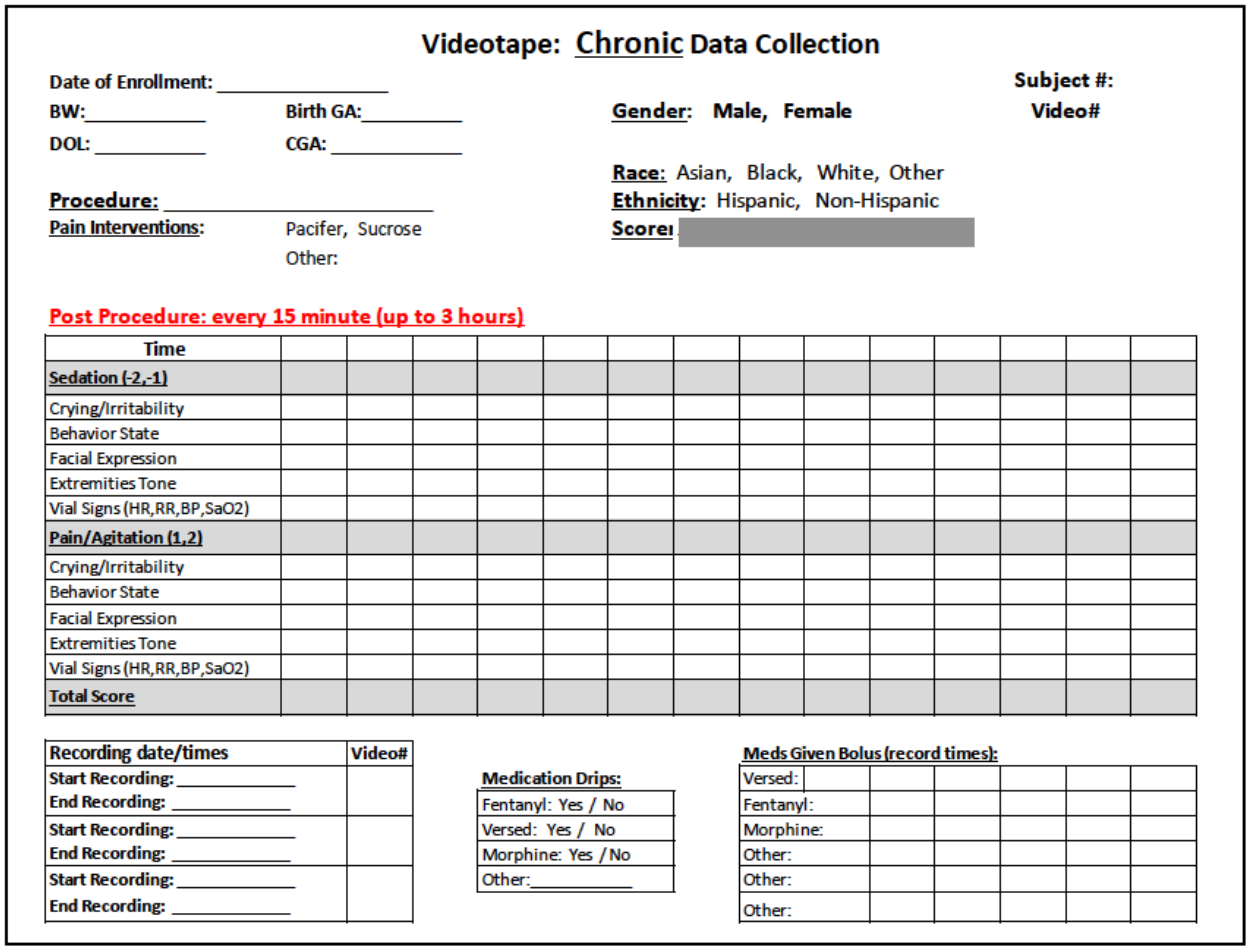
Before we conclude, we would like to point out the following:

1. The nurses who documented the NIPS and NPASS scores underwent a standardized training program to ensure proper utilization of the tools.

2. Each epoch was scored independently by two trained nurses so that inter-observer reliability can be established using kappa coefficient to measure the inter-observer agreement. We include all the cases of agreement and exclude the cases of disagreement (below 5%) from further analysis.

3. No procedures were done for the study purposes. All the procedures recorded have been ordered as clinically indicated procedures.

4. Portions of this database can be made available for research use upon request. The interested researcher should contact the research team in the Department of



Videotape: Chronic Data Collection

Date of Enrollment: _____ Subject #:

BW: _____ Birth GA: _____ Gender: Male, Female Video#

DOL: _____ CGA: _____

Race: Asian, Black, White, Other

Procedure: _____ Ethnicity: Hispanic, Non-Hispanic

Pain Interventions: Pacifer, Sucrose Score: _____

Other:

Post Procedure: every 15 minute (up to 3 hours)

Time														
Sedation (-2,-1)														
Crying/Irritability														
Behavior State														
Facial Expression														
Extremities Tone														
Vial Signs (HR,RR,BP,SaO2)														
Pain/Agitation (1,2)														
Crying/Irritability														
Behavior State														
Facial Expression														
Extremities Tone														
Vial Signs (HR,RR,BP,SaO2)														
Total Score														

Recording date/times	Video#
Start Recording: _____ End Recording: _____	
Start Recording: _____ End Recording: _____	
Start Recording: _____ End Recording: _____	

Medication Drips:
Fentanyl: Yes / No
Versed: Yes / No
Morphine: Yes / No
Other: _____

Meds Given Bolus (record times):

Versed:						
Fentanyl:						
Morphine:						
Other:						
Other:						
Other:						

Figure 3.7: Example of the scoring sheet for a postoperative painful procedure.

Computer Science and Engineering at the University of South Florida for further details.

3.2.5 Samples of the Database

Figure 3.8 shows examples from our NPAD database. The images were randomly selected and face-masked to ensure confidentiality. As can be seen from the figure, the clinical environment has different levels of illumination. Also, it can be seen that it is common to miss a specific pain indicator or response. For example, the infant's face is blocked in some images due to sleeping position (1^{st} row, 4^{th} column), pose (2^{nd} row, 3^{rd} column), tapes (3^{rd} row, 1^{st} and 4^{th} columns), or Oxygen mask (4^{th} row, 3^{rd} and 4^{th} columns). Also, swaddling the infants would block their body

as illustrated by the image in the 3^{rd} row and 4^{th} column. These images demonstrate the need for using multiple pain responses (multimodal) when assessing neonatal pain.

Figure 3.8: Image samples from our NPAD database.

Table 3.2: Infants' demographics for near infrared spectroscopy data.

Gender	7 Females		7 Males	
	Min	Max	Mean	Std Dev
Birth Weight (g)	1090	2360	1664	367
Birth GA (weeks)	27	34	31.1	1.9
Head Circumference (cm)	26	34.5	30	2
Ethnicity	Hispanic		Non-Hispanic	
	1		13	

3.3 Collection of Near Infrared Spectroscopy Data

We also collected the cerebral oxygenation data and used them as an objective indicator of pain for verification. We present below our method of collecting NIRS data for 14 neonates. Eligible infants included are those admitted to the NICU at TGH with a birth gestational age of less than 37 weeks. Table 3.2 presents the demographics for these 14 neonates.

The INVOS^{TM} Near Infrared Spectroscopy (NIRS) device was used to measure the cerebral oxygenation readings. This device uses near infrared light to determine the deoxyhemoglobin (HbH) and the oxy-hemoglobin (HbO_2) concentrations. HbH and HbO_2 are then added together to determine the total hemoglobin concentration (HbT). To obtain a regional oximetry value (rSO_2), HbO_2 is divided by HbT. The NIRS device's sample rate was 30 seconds. The Vital Sync^{TM} was used to timestamp and export the NIRS data as an Excel file. The Vital Sync^{TM} was also used to log events, such as the start and end of the painful procedure. Prior to the placing of the NIRS probes, the team verified the location of the painful procedure with the bedside nurse. The probe was then placed on the contralateral side of the forehead. Once the INVOS^{TM} had established its auto-baseline, an event report was made in the Vital Sync^{TM} and collection of the pre-procedure data was started. Pre-procedure data were collected for approximately ten minutes, at which point, the bedside nurse would start the painful procedure (i.e., heal lancing) and an additional event report was made on the Vital Sync^{TM}. Post-procedure data were then collected for 10 additional minutes after the painful procedure (i.e., heal lancing) had ended.

In this chapter, we presented our method to collect the real-world NPAD database. In the next chapter, we will discuss our novel algorithms for analyzing this database to create an automatic and multimodal pain assessment.

CHAPTER 4

AUTOMATIC NEONATAL PAIN ASSESSMENT

In this chapter, we present the algorithms of the individual components of our automatic pain assessment system. These components are: algorithms for analysis of facial expression (Section 4.2), algorithm for analysis of body movement (Section 4.3), algorithm for analysis of crying sounds (Section 4.4), and algorithm for analysis of vital signs (Section 4.5). We then present our methods for combining these components to create a multimodal pain assessment system (Section 4.6). We note that the technical background of the algorithms presented in this chapter can be found in appendices B and C.

4.1 Note to Reader

Portions of this chapter were published in the International Conference on Pattern Recognition (IEEE) [127] and the Scandinavian Conference on Image Analysis (Springer) [129]. Permissions from the publishers are included in Appendix A.

4.2 Facial Expression Analysis

Automatic recognition of neonatal pain using facial expression consists of three main stages: 1) preprocessing and face tracking, 2) facial features extraction, and 3) pain recognition. Each of these stages is presented in the next subsections.

4.2.1 Preprocessing and Facial Landmark Detection

The first step of preprocessing involved dividing the recorded videos into short segments of five, ten, fifteen, and twenty seconds. Then, histogram equalization was performed on low-light videos to enhance their contrast. Next, the neonate's face was tracked in each frame as described below.

Figure 4.1: ZFace tracker; 49 points (green), mesh points (blue), and head orientations.

To track the infant's face and detect the facial landmarks, we applied ZFace [130] (see Section B.5.3), which is a person-independent tracker, in each video to obtain 49 facial landmark points and the face's boundary points. The tracker outputs the coordinates of a mesh of points, 6 degrees of freedom of rigid head movement, and a failure message to indicate the failure frames. After obtaining the facial landmarks in each frame, we used them for registration and facial region cropping. Figure 4.1 shows the 49 points (green) as well as the mesh of 512 points (blue); the arrows indicate the head's orientations.

4.2.2 Facial Features Extraction

In this section, we describe three handcrafted algorithms (Section 4.2.2.1) and two deep learning based algorithms (Section 4.2.2.2) for extracting pain-relevant features from neonates' faces. The main difference between the deep methods and the handcrafted methods is that the deep methods extract features that are learned, at multiple levels of abstraction, directly from the training data. In contrast, the handcrafted methods are designed beforehand by human experts to extract specific features.

4.2.2.1 Handcrafted Methods

Three handcrafted methods, namely optical strain, geometric distances, and local binary pattern (LBP), were used to extract pain-relevant features from neonates' faces. These methods are presented below.

4.2.2.1.1 Optical Strain

Our optical strain method measures the non-rigid facial tissue deformations (i.e., strain magnitude). There are two ways to estimate the strain magnitude [131]: (i) integrate the strain definition into the optical flow equations, or (ii) derive strain directly from the flow vectors. Since the second method allows post-processing the flow vectors before calculating the strain, it is used to estimate the optical strain. The equations to compute the flow vectors and the strain magnitude can be found in [131].

The algorithm for extracting facial features using the optical strain consists of the following steps [127, 126]. First, the detected face region is divided into four regions (I,II,II, and IV). Next, the optical flow is calculated between consecutive frames of a video for each region of the face as well as the overall face region. Then, the optical strain is estimated over the flow fields to generate the strain components of the strain tensor. After generating the strain components, the strain magnitude is calculated, as discussed in [131], for each region of the face along with the overall face region and normalized. Each region generates a sequence corresponding to the amount of strain observed over time. Finally, a peak detector method is applied to the strain plots obtained for each region from I to IV to detect the points of maximum strain, which correspond to facial expressions.

To form the feature vector for classification, we computed several descriptive statistics (e.g., mean and 25th percentile) for the detected maximum strain values and concatenated them into $5 \times S$ dimensional vector, where 5 represents the facial regions (I to IV and overall face region) and S represents the number of statistics.

4.2.2.1.2 Geometric Distances

The Neonatal Facial Coding System [78] (NFCS) is an extension of Facial Action Coding System (FACS) [79] designed specifically for neonatal pain. Examples of NFCS pain-relevant facial movements include eye squeeze, bulging brow, deepening of the nasolabial furrow, horizontal mouth stretch, vertical mouth stretch, and pursed lips. Using the points detected by ZFace, we computed eleven Euclidean distances between the detected landmark points, to represent different NFCS facial movements, as follows:

1. Eye squeeze: Defined as a reduction of the distances between the upper and lower eyelids of the left eye (d1) and right eye (d2) or a reduction of distances between the highest arch's points and upper eyelids of the left eye (d3) and right eye (d4).

2. Bulging brow: Defined as a reduction of the distance between the inner corners of the eyebrows (d5) or a reduction of the distance between the highest arch's points of left and right eyebrows (d6).

3. Nasolabial furrow: Defined as an increase in the distance between the nose's left end point and the mouth's left corner point (d7) or an increase in the distance between the nose's right end point and the mouth's right corner point (d8).

4. Vertical mouth stretch: Defined as an increase in the distance between the mouth's upper and lower points (d9).

5. Horizontal mouth stretch: Defined as an increase in the distance between the mouth's left and right corner points (d10).

6. Jaw drops: Defined as an increase in the distance between the nose's tip and the lowest point of the lower face boundary or chin's tip (d11).

To form a feature vector for classification, several statistics (e.g., standard deviation and 75th percentile) for each distance were calculated across frames and concatenated into 11 x S dimensional vector, where S represents the number of statistics.

4.2.2.1.3 Local Binary Pattern

We applied LBP-TOP appearance descriptor, which stands for Local Binary Patterns on Three Orthogonal Planes, to small patches of the neonate's face to extract pain-relevant features; descriptions of these descriptors are presented in Section B.4. We applied the descriptors to small patches or regions instead of the entire face because of two main reasons. First, applying the descriptors to small regions allows to better capture the changes of skin texture at different local regions such as eyebrow corner, nasolabial furrow, and mouth corner. Second, applying the descriptors to small regions significantly reduces the computational complexity. We extracted, using LBP-TOP, appearance features from 32×32 patches located around 31 facial landmark points. The length of

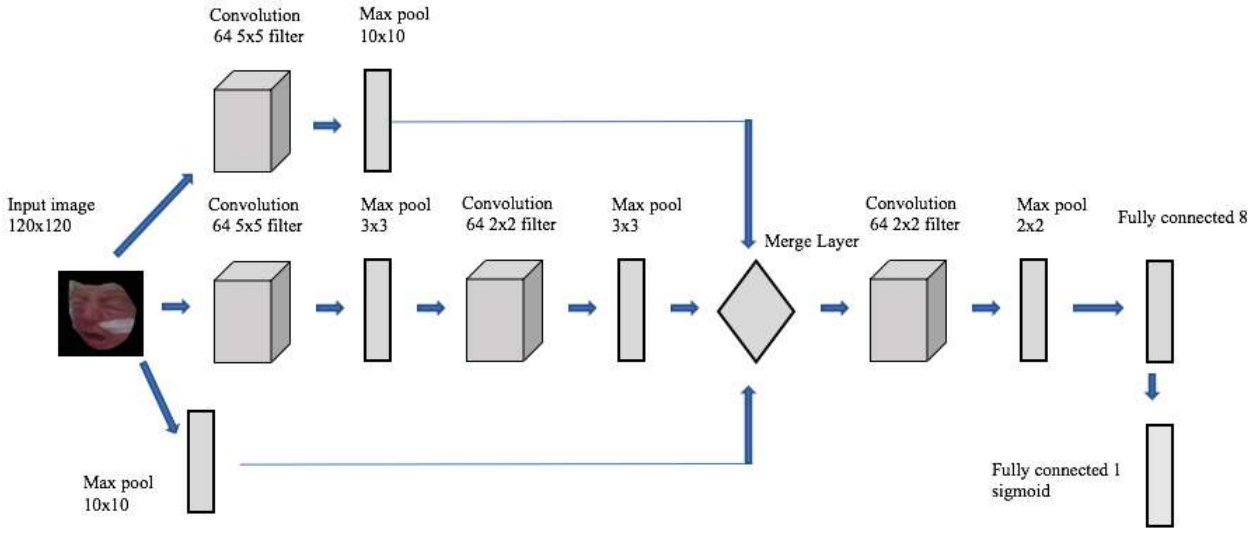


Figure 4.2: The architecture of N-CNN for pain expression recognition.

the extracted feature vector is 177. The dynamic appearance feature vector was then created for each video by concatenating all the local appearance descriptors over 31 facial landmark patches (177×31). Before we proceed, we would like to mention that the analysis of facial expression using LBP was conducted as a joint work with Dr. Ruicong Zhi. We refer the reader to [132] for a complete description of this work.

4.2.2.2 Deep Learning Methods

We present two deep learning approaches to recognize neonatal pain. First, we assessed neonatal pain using our novel Neonatal Convolutional Neural Network (N-CNN). Second, we used pre-trained CNN architectures as fixed feature extractors followed by pain recognition using supervised machine learning classifiers. Both approaches are presented below.

4.2.2.2.1 Convolutional Neural Network

Before we present our Neonatal Convolutional Neural Network (N-CNN), we want to note that understanding this section requires basic knowledge of CNN concepts such as convolutions, pooling, activations, and augmentation. Appendix C of this dissertation provides an explanation for these concepts.

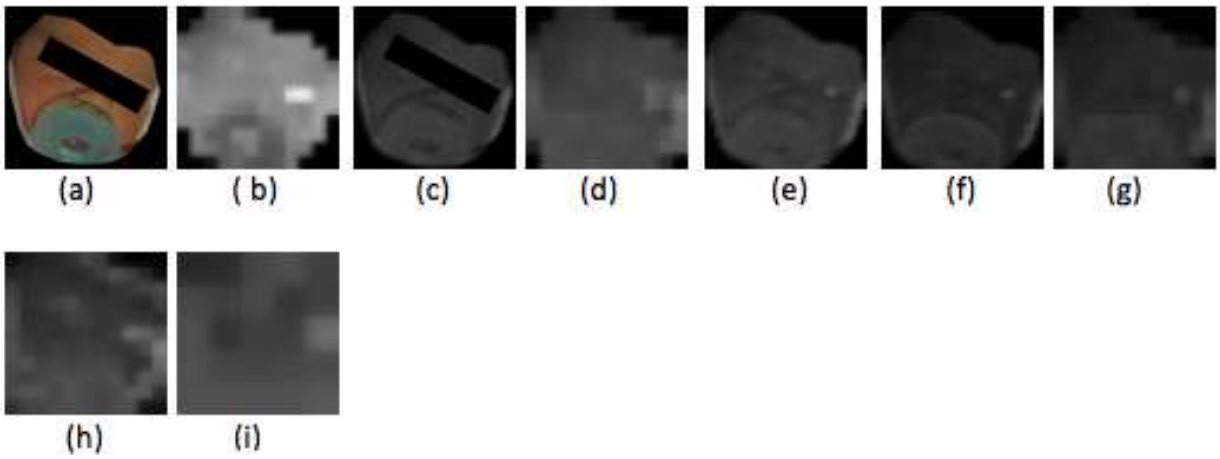


Figure 4.3: Visualizations of the output of different layers for no-pain input image.

Our Neonatal Convolutional Neural Network (N-CNN), inspired by [133, 134], is a cascaded CNN that has three main branches. The first branch consists of a pooling layer that performs max pooling operation using 10×10 filters. The second branch consists of two convolutions layers with 64 filters of size 5×5 followed by pooling layers with 3×3 filters (i.e., $conv1 \rightarrow pool1 \rightarrow conv2 \rightarrow pool2$). The last branch consists of two layers: a convolutional layer with 64 filters of size 5×5 and a pooling layer to perform a max pooling operation using 10×10 filters. Each of these branches performs a specific task and captures different set of features. For example, the first branch would downsample the image size and captures the most prominent features, whereas the third branch captures more generic convolutional features such as the image's texture and color blobs. The second branch extracts deeper features from the image since it has four layers. After feeding the image into these three branches, we merged the outputs of the three branches by concatenation. Our experiments showed that this cascaded CNN architecture achieves better classification performance than the regular CNN architecture. Figure 4.2 presents the architecture of our N-CNN. Figures 4.3 and 4.4 show the visualizations of the output of different layers during no-pain and pain states; (a) represents the input image, (b) is the output of max-pool 1, (c) is the output of convolutional 1, (d) is the output of max-pool 2, (e) is the output of max-pool 3, (f) is the output of convolutional 3, (g) is the output of max-pool 4, (h) is the output of convolutional 4, and (i) is the output of max-pool 5.

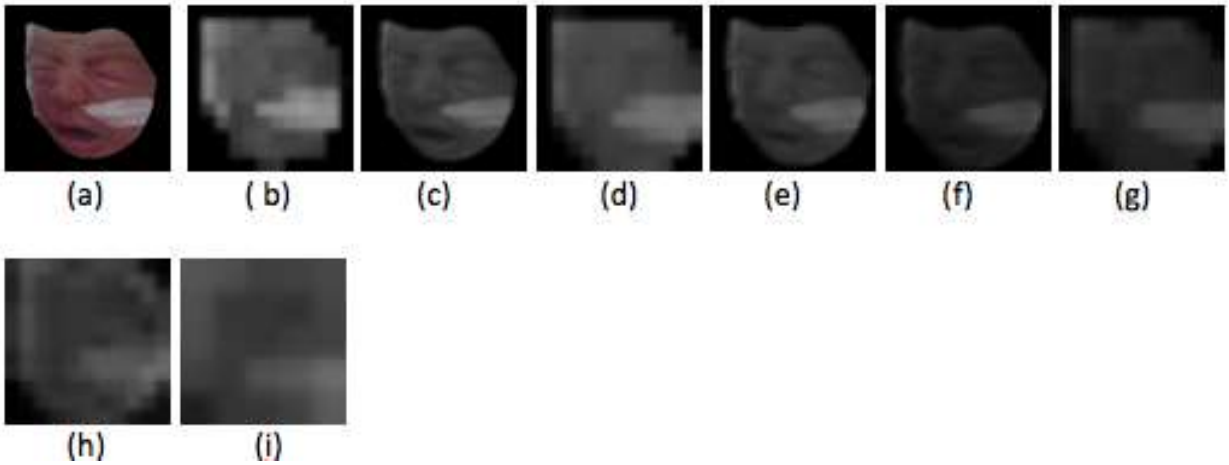


Figure 4.4: Visualizations of the output of different layers for pain input image.

We trained the N-CNN from scratch with random weights initialization and 7,2593 training parameters, and used RMSprop [135] as a gradient descent optimization algorithm along with a constant learning rate of 0.0001. For both training and validation, we used a batch size of 16. Also, we applied L2 regularizer [136] and dropout [137] before the final classification layer to prevent over-fitting. The parameters of the N-CNN are presented in Table 4.1. It is worth noting that we have designed three different CNN architectures and chosen N-CNN as the final architecture since it achieves the best performance.

Since training a CNN from scratch requires a large amount of data, we performed data augmentation in the training set as follows:

1. Each image was rotated by 15 degrees. This procedure generates a total of 24 images for each frame. That is, we randomly rotate each image by 15 degrees to obtain 24 images.

2. Each rotated image was flipped horizontally and vertically. This procedure generates a total of 72 (24 original + 24 horizontal-flip + 24 vertical-flip) augmented images for each frame.

The exact number of images for both training and testing is given in Chapter 5. All the images were re-sized to 120×120 using bi-cubic interpolation method.

Table 4.1: Parameters of N-CNN

Branch	Layer	Type	Input	#Filters	Filter Size
Left	Layer 1	Max Pool 1	120 x 120 x 3	–	10 x 10, st. 10, pd. 0
Central	Layer 2	Conv 1	120 x 120 x 3	64	5 x 5, st. 1, pd. 0
	Layer 3	Max Pool 2	Layer 2	–	3 x 3, st. 3, pd. 0
	Layer 4	Conv 2	Layer 3	64	2 x 2, st. 1, pd. 0
	Layer 5	Max Pool 3	Layer 4	–	3 x 3,st. 3, pd. 0
Right	Layer 6	Conv 3	120 x 120 x 3	64	5 x 5, st. 1, pd. 0
	Layer 7	Max Pool 4	Layer 6	–	10 x 10, st. 10, pd. 0
Merge Layer (Left *II* Central *II* Right)					
	Layer 8	Conv 4	Merge Layer	64	2 x 2, st. 1, pd. 0
	Layer 9	Max Pool 5	Layer 8	–	2 x 2, st. 2, pd. 0
	Layer 10	FC1	Layer 9	–	–
	Layer 11	FC2	Layer 10	–	–

4.2.2.2.2 Transfer Learning

Instead of training CNN end-to-end, using a pre-trained CNN is an attractive and more realistic alternative. Therefore, we decided to investigate the feasibility of classifying pain using a CNN that was trained for different classification task. Specifically, we used VGG (Visual Geometry Group) Face CNN [138], which was originally trained on a large face database (approx. 2.6M face images of 2,622 identities) for face recognition, to extract deep features from neonates' faces. The reason for choosing an architecture trained to recognize faces instead of emotions is that face recognition is well-studied and validated on large volume databases as compared to emotion classification. In addition, the features of face recognition and facial expression recognition are rather similar since both tasks involve analyzing human faces [139]. Table 4.3 presents the architecture of VGG-Face, where each layer (e.g., Conv 1-1) is followed by ReLU and each block (e.g., Conv 1-1 and Conv 1-2) is followed by pooling.

In addition to VGG-Face, we used three other CNN architectures that were originally trained on a relatively different database (ImageNet dataset; approx. 1.2M images and 1000 classes) for image classification. These CNNs architectures are VGG-F, VGG-M, and VGG-S [140]; F, M, and S are abbreviations for fast, medium, and slow, respectively. The main reason for using these three architectures is to investigate the difference between using CNNs trained on a relatively similar dataset (face dataset) and CNNs trained on a relatively different dataset (ImageNet dataset). Tables 4.4 to 4.6 provide the architectures for VGG-F, VGG-M, and VGG-S, respectively. Each

layer in these tables, except Full 8, is followed by ReLU; k x n x n indicates the number of filters and their size.

Before feeding the images to the pre-trained CNN architectures, we re-sized them to 224×224 to accommodate with their size requirement ($224 \times 224 \times 3$, RGB images). The CNN architectures are implemented in a MATLAB Toolbox called MatConvNet [141].

Table 4.2: VGG-Face architecture.

Conv 1-1	$64 \times 3 \times 3$, stride 1, padding 1
Conv 1-2	$64 \times 3 \times 3$, stride 1, padding 1
Conv 2-1	$128 \times 3 \times 3$, stride 1, padding 1
Conv 2-2	$128 \times 3 \times 3$, stride 1, padding 1
Conv 3-1	$256 \times 3 \times 3$, stride 1, padding 1
Conv 3-2	$256 \times 3 \times 3$, stride 1, padding 1
Conv 3-3	$256 \times 3 \times 3$, stride 1, padding 1
Conv 4-1	$512 \times 3 \times 3$, stride 1, padding 1
Conv 4-2	$512 \times 3 \times 3$, stride 1, padding 1
Conv 4-3	$512 \times 3 \times 3$, stride 1, padding 1
Conv 5-1	$512 \times 3 \times 3$, stride 1, padding 1
Conv 5-2	$512 \times 3 \times 3$, stride 1, padding 1
Conv 5-3	$512 \times 3 \times 3$, stride 1, padding 1
Full 6	4096 dropout
Full 7	4096 dropout
Full 8	2622

Table 4.3: VGG-F architecture.

Conv 1	$64 \times 11 \times 11$, stride 4, padding 0
Conv 2	$256 \times 5 \times 5$, stride 1, padding 2
Conv 3	$256 \times 3 \times 3$, stride 1, padding 1
Conv 4	$256 \times 3 \times 3$, stride 1, padding 1
Conv 5	$256 \times 3 \times 3$, stride 1, padding 1
Full 6	4096 dropout
Full 7	4096 dropout
Full 8	1000 softmax

Table 4.4: VGG-M architecture.

Conv 1	$96 \times 7 \times 7$, stride 2, padding 0
Conv 2	$256 \times 5 \times 5$, stride 2, padding 1
Conv 3	$512 \times 3 \times 3$, stride 1, padding 1
Conv 4	$512 \times 3 \times 3$, stride 1, padding 1
Conv 5	$512 \times 3 \times 3$, stride 1, padding 1
Full 6	4096 dropout
Full 7	4096 dropout
Full 8	1000 softmax

Table 4.5: VGG-S architecture.

Conv 1	$96 \times 7 \times 7$, stride 2, padding 0
Conv 2	$256 \times 5 \times 5$, stride 1, padding 1
Conv 3	$512 \times 3 \times 3$, stride 1, padding 1
Conv 4	$512 \times 3 \times 3$, stride 1, padding 1
Conv 5	$512 \times 3 \times 3$, stride 1, padding 1
Full 6	4096 dropout
Full 7	4096 dropout
Full 8	1000 softmax

4.2.3 Pain Recognition

We discussed above several handcrafted and deep learning methods for facial feature extraction. The next stage is about using these features for pain recognition or classification after selecting the most relevant features. We used two main feature selectors, namely Relief-f [142] and Symmetric Uncertainty [143], to select the most important features from the extracted facial features. Then, the selected features were used to train four supervised machine learning classifiers: Naive Bayes (NB), Nearest Neighbors (kNN), Support Vector Machines (SVMs), and Random Forests (RF). We chose these classifiers because they have been successfully used in automatic pain assessment applications (see Chapter 2). We used NB, kNN, SVMs, and RF with the handcrafted features and the deep features extracted using the pre-trained CNNs. As for the N-CNN, the classification was performed in the last layer of N-CNN.

4.3 Body Movement Analysis

The automatic recognition of neonatal pain from body movement consists of three main stages: 1) preprocessing and body tracking, 2) feature extraction, and 3) pain recognition. Each of these stages is presented in the next subsections.

Figure 4.5: First row: original and binary images; second row: filtered binary image and ROI.

4.3.1 Preprocessing and Body Tracking

The first step of preprocessing involved dividing the recorded videos into short segments of five, ten, fifteen, and twenty seconds. Then, a standard histogram equalization was performed on low-light videos to enhance their contrast. Next, the neonate's body was detected in each frame as described below.

To detect the neonate's body region, we implemented a color-based tracking method to detect the body region in each frame. The first step of our method involved creating a total of 6,052 patches, half of which were body patches and the other half were non-body patches, with size 128 x 128. After creating the patches, we converted them to YCbCr (Luminance; Chroma: Blue; Chroma: Red) color space and generated the Cb and Cr histograms of these patches. Next, we generated the normal distribution of these histograms. The normal distributions of Cb channel for body and non-body patches showed a relatively high overlapping while the normal distributions of Cr channel were relatively separated. Then, we detected the cut-off point (i.e., the cross point

of body and non-body normal distributions) of Cr histogram that gives the smallest error. The detected cut-off point was used as a threshold to convert the frame into binary images, which was pruned using morphological operations. Finally, we used the nose-tip point and considered the region below this point as the region of interest (ROI). We would like to mention that this method fails in cases when the neonate's body is occluded or when it is covered with a blanket that has a similar color to the background. In these cases, we manually detect the location of the body in the first frame and track it over all frames. Figure 4.5 shows our algorithm's result in detecting the body region of a neonate.

4.3.2 Body Feature Extraction

Our method to analyze body movement depends on the motion image, which is a simple and efficient method to estimate an infant's body movement in video sequences [127]. It identifies the change of each pixel value between consecutive frames. Each pixel in the motion image $M(x, y)$ has a value of 0 to represent no movement or 1 to represent movement. To analyze the infant's body movement, we computed the motion images between consecutive video frames. Then, we applied a filtering method to reduce noise and get the maximum visible movement.

Since caregivers focus on observing the amount of body movement when assessing neonatal pain, we used the amount of body motions in each video frame as the main feature for classification. This feature is computed as follows:

$$A_m = \frac{1}{N_x N_y} \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} M(x, y) \tag{4.1}$$

where N_x and N_y represent the image's height and width. To find the total amount of motion in each video sequence, we summed A_m as:

$$Total_{motion} = \sum_{k=1}^{F} A_m^k \tag{4.2}$$

where F is the total number of frames. The generated $Total_{motion}$ value is the main feature that is used for classification.

4.3.3 Pain Recognition

The $Total_{motion}$ feature extracted as described above is the main feature used for classification. We used a simple thresholding to classify a given instance. That is, a given instance is classified as pain instance if it exceeds the threshold, and no-pain instance otherwise. In addition to thresholding, we used this feature to train Naive Bayes, k Nearest Neighbors (kNN), Support Vector Machines (SVMs), and Random Forests (RF). We chose these classifiers because they have been successfully used in automatic pain assessment applications (see Chapter 2).

4.4 Crying Sound Analysis

The automatic recognition of neonatal pain using crying sound consists of three main stages: 1) preprocessing, 2) feature extraction, and 3) pain recognition. Each of these stages is presented in the next subsections.

4.4.1 Sound Signal Preprocessing

The first step of preprocessing involved dividing the recorded audio into short segments of five, ten, fifteen, and twenty seconds. The segmented audio signal was divided into several Hamming windows of 32-milliseconds that shift every 16-milliseconds and Hamming windows of 30-milliseconds that shift every 10-milliseconds to minimize the signal discontinuities. The first windowing scheme is used with the Linear Predictive Cepstral Coefficients (LPCC) and the second is used with the Mel Frequency Cepstral Coefficients (MFCC). No filtering or background noise removal operations were performed prior to feature extraction.

4.4.2 Sound Feature Extraction

In this section, we describe handcrafted and deep learning methods for extracting pain-relevant features from neonates' sounds.

4.4.2.1 Handcrafted Methods

To analyze neonatal sounds, we applied[1] two Cepstral Domain methods: Linear Predictive Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC). LPCC (20 coefficients) were computed from 32-milliseconds window with 16-milliseconds offset while MFCC (20 coefficients) were computed from 30-milliseconds window with 10-milliseconds offset. This generates LPCC and MFCC feature vectors with length:

$$(\frac{L}{shift} - (\frac{W}{shift} - 1)) \times 20 \tag{4.3}$$

where L represents the length of the audio, shift represents the window offset, and W represents the window size. The extracted feature vectors were then reduced using vector quantization method as follows:

1. We clustered the extracted features or coefficients of all audio instances using k-means algorithm.
2. We computed the centroid (a.k.a. codeword) for each cluster or group.
3. We generated a codebook matrix from the groups' centroids, the rows of this matrix represent the group ID and the columns represent the centroids. The generated codebook has a total of 32 groups.
4. We used the stored codebook to map a new instance to the group whose center is close to this instance features.

4.4.2.2 Deep Learning Methods

We investigated the use of CNN for neonatal sounds' analysis. In particular, we converted the audio signals into spectrogram images of size 120 $\times$ 120 and used these images as input for N-CNN (see Figure 4.2 and Table 4.1). The spectrogram is a 2-D visual representation of change for every frequency component of an audio signal with respect to time. In the spectrogram image, the frequency is shown on one axis and the time is shown on the other axis; the color indicates the amplitude of each frequency during a particular time. Figure 4.6 shows the spectrogram images of neonates' sounds during a pain and no-pain events.

[1]The algorithms for analyzing crying sound was mostly implemented in [106, 129].

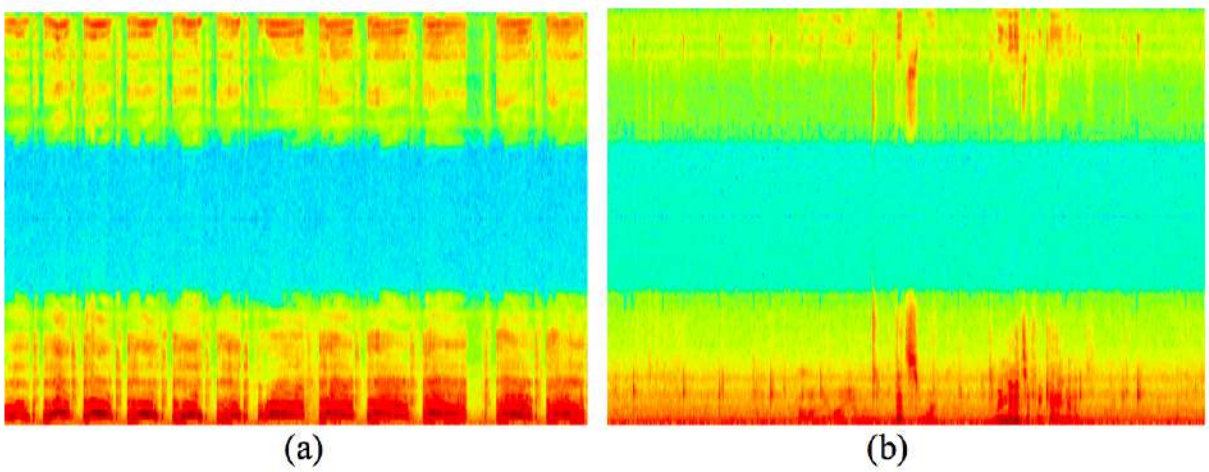


Figure 4.6: Spectrogram images; (a) pain and (b) no-pain.

Since training a CNN requires a large amount of data, we performed data augmentation in our audio database, which has 182 audio signals collected during pain and no-pain events. Each audio signal was augmented by adding three frequencies (f/2, f/3, 4f/3), six different levels of noise (0.01, 0.001, 0.002, 0.003, 0.004, and 0.005), and a combination of both frequency and noise level (f/2 with noise 0.01 or f/3 with noise 0.001). This augmentation process generates a total of 27 augmented audio signals for each audio signal.

4.4.3 Pain Recognition

The group ID feature, which was extracted using LPCC and MFCC (see Section 4.4.2.1), was used to classify the emotional state of a neonate as pain or no-pain. We used a simple thresholding to classify a given instance. That is, a given instance is classified as pain instance if it exceeds the threshold, and no-pain instance otherwise. In addition to thresholding, we used this feature to train Naive Bayes, Nearest Neighbors (kNN), Support Vector Machines (SVMs), and Random Forests (RF). We chose these classifiers because they have been successfully used in automatic pain assessment applications. The deep learning features, which were extracted from the spectrogram images, were used by N-CNN to assess neonatal pain.

4.5 Vital Signs Analysis

Vital sign readings such as Heart Rate (HR), Respiratory Rate (RR), and Oxygen Saturation levels (SpO2) were collected using Vital Sync^{TM} device. These data were then exported as a timestamped Excel file for further analysis. To remove the outliers from the exported vital sign (i.e., HR, RR, and SpO2) numbers, we applied median filter with different window sizes. Then, we calculated several descriptive statistics (e.g., mean, standard deviation, max) for vital sign readings across the pain and no pain event (i.e., 3 x statistics dimensional vector for each event).

After generating the feature vector, we applied Relief-f [142] and Symmetric Uncertainty [143] to select the most important features followed by classification using Naive Bayes, k Nearest Neighbors (kNN), Support Vector Machines (SVMs), and Random Forests (RF).

4.6 Fusion of Pain Responses

In this section, we describe decision-level and feature-level methods for combining different pain responses to generate a multimodal pain assessment. The fusion methods combine facial expression, crying, body movement, and vital signs readings. To the best of our knowledge, this is the first work that assesses neonatal pain using a combination of these pain responses.

4.6.1 Decision-level Fusion

The decision-level fusion represents a variety of methods designed to merge the decisions or outcomes of multiple classifiers into one single ensemble decision. To combine the outcomes of different pain responses, we applied a majority voting scheme. In the majority-voting scheme, each pain response contributes one vote (i.e., class label) and the majority label in the combination is chosen as the final decision or outcome. If the combination of different pain responses has a tie, we chose the class that has the highest confidence score as the final outcome.

4.6.2 Feature-level Fusion

Feature-level fusion is the process of combining multiple modalities (i.e., pain responses) in the early stage by concatenating the features of these responses into a single high-dimensional feature

vector. The concatenated vector is then used for classification. Feature-level fusion has three main issues: scaling, the high-dimensionality of the feature vector, and the missing data or feature.

The extracted features of each pain response were normalized in the range of [0,1] for scaling before concatenating them into a single feature vector. For feature reduction, two feature selectors, namely Relief-f and Symmetric Uncertainty, were applied. To handle the missing data, we trained a classifier for each case of the missing data (i.e., a single classifier for each case of missing feature). For example, given that $S = \{x_1, x_2, x_3, ..., x_f\}$, where x represents a feature, f represents the total number of features, and x_1 feature is missing, the classifier would be trained using all the features except x_1.

In this chapter, we presented several algorithms for analyzing neonatal pain. The results of applying these algorithms to our database (NPAD) as well as discussions of performance are presented in the next chapter.

CHAPTER 5

IMPLEMENTATION AND RESULTS

This chapter presents the experimental design and the results of testing our automatic pain assessment system on NPAD database. We conducted three sets of experiments. In the first set, we tested the performance of pain assessment using a single pain response at a time (Section 5.2: Unimodal). In Section 5.3, we present the results of assessing neonatal pain using a combination of different pain responses (multimodal). In the third set, we created different pain assessment models according to the neonates' age, gender, race, and weight (Section 5.5). Finally, we compared the results of our pain assessment methods with other state of the art methods (Section 5.6). The performance of pain assessment for all experiments is shown using the metrics of confusion matrix and the area under the Receiver Operating Characteristic (ROC curve).

5.1 Note to Reader

Portions of this chapter were published in the International Conference on Pattern Recognition (IEEE) [127] and the Scandinavian Conference on Image Analysis (Springer) [129]. Permissions from the publishers are included in Appendix A.

5.2 Unimodal Pain Assessment

We used a single pain response to classify the emotional state of 31 neonates into pain or no-pain (unimodel). It is important to note that we included the severe-pain and no-pain labels of NIPS pain scale in the analysis. Moderate pain label was excluded from further analysis because the number of epochs for this label in the current NPAD database is small.

5.2.1 Pain Assessment From Facial Expression

To assess neonatal pain using facial expression, we investigated several handcrafted and deep learning methods. The results of pain assessment for both the handcrafted and deep learning methods are presented next.

5.2.1.1 Handcrafted Methods

Before presenting the results of assessing neonatal pain, we want to mention that subject-level 10-fold cross validation was used for evaluation as follows:

1. We divided the 31 subjects into 10 folds (i.e., each fold has approximately 3 subjects).
2. The classifier is trained using 9 folds and tested on the 10th fold.
3. We repeated the process (step 2) 10 times and calculated the accuracy by averaging the accuracies of testing the classifier on the testing folds.

5.2.1.1.1 Optical Strain

The extracted facial strain feature vector (Chapter 4) was reduced using Relief-F [142] and Symmetric Uncertainty [143] to obtain the best 5, 10, and 15 features. Then, several supervised machine learning classifiers (e.g., SVM and Random Forest) were used to classify the recorded events to pain or no-pain events. Assessing neonatal pain using the facial strain features achieved 83.88% average accuracy and 0.75 AUC. The second column of Table 5.1 presents the confusion matrix of assessing neonatal pain using the facial strain features. From the matrix, we can see that the FPR (False Positive Rate) is lower than the FNR (False Negative Rate). In some applications, minimizing the FPR is more important than minimizing the FNR. In case of pain assessment, we believe it is equally important to minimize both FPR and FNR as several pediatric studies have reported serious outcomes of both over- and under-treatment.

5.2.1.1.2 Geometric Distances

The geometric feature vector extracted (11 distances $\times$ statistics) as described in Chapter 4 was reduced using Relief-F [142] and Symmetric Uncertainty [143]. The reduced feature vector was then used to train several machine learning classifiers such as SVM, Random Forest, kNN, and

Table 5.1: Confusion matrices of neonatal pain assessment from facial expression.

	Strain Features		**Geometric Features**	
	Pain	No Pain	Pain	No Pain
Pain	TPR = 64.1%	FNR = 35.9%	TPR = 79.5%	FNR = 20.5%
No Pain	FPR = 9.3%	TNR = 90.7%	FPR = 8.4%	TNR = 91.6%
	LBP-TOP Features		**Learned Features (N-CNN)**	
	Pain	No Pain	Pain	No Pain
Pain	TPR = 78.9%	FNR = 21.1%	**TPR = 82.9%**	FNR = 17.1%
No Pain	FPR = 7.7%	TNR = 92.3%	FPR = 6.5%	**TNR = 93.5%**

Naive Bayes. The output of the classifier is a binary label that indicates if the pain is present or not. Assessing neonatal pain using the geometric features achieved 88.13% average accuracy and 0.85 AUC. Table 5.1 presents the confusion matrix of using geometric features for pain assessment. From the matrix, we can see that the FNR and FPR of the geometric features are lower than the strain features. However, both FNR and FPR should be minimized further to mitigate over- and under- treatment.

5.2.1.1.3 Local Binary Pattern

The LBP-TOP feature vector extracted as described in Chapter 4 was reduced using a Supervised Locality Preserving Projections (SLPP) algorithm [144]. The reduced feature vector was then used to train several machine learning classifiers such as SVM, Random Forest, kNN, and Naive Bayes. The output of the classifier is a binary label that indicates if the pain is present or not. Assessing neonatal pain using LBP-TOP features achieved 87.66% average accuracy. Table 5.1 presents the confusion matrices of using LBP-TOP features for pain assessment. The matrix shows that the FPR of LBP-TOP features is lower than the strain and geometric features. However, both rates should be minimized further to mitigate over- and under- treatment. We would like to note that the analysis of facial expression using LBP was conducted in collaboration with a visiting professor. A complete description of this collaboration can be found in [132].

5.2.1.2 Deep Learning Methods

We divided the experiments of this section into two main folds: pain assessment using CNN and pain assessment using transfer learning.

To create the training and testing sets for our deep learning methods, we extracted the key frames, thereby removing many similar frames, from each video sequence after cropping the face using ZFace tracker. The total number of key frames obtained from all the videos of 31 subjects was 3026 frames. These frames were randomly divided into equal training set (1513 frames) and testing set (remaining 1513 frames).

Before presenting the results, we want to mention that the trained CNN was evaluated and the performance was computed as follows:

1. We randomly split our database (3026 images) into a training set (1513 images) and a testing set (1513 images) three times to obtain three training sets ($TR_1 - TR_3$) and testing sets ($TS_1 - TS_3$).

2. We used the training set to train the CNN followed by testing on its corresponding testing set (e.g., TR_1 for training and TS_1 for testing).

3. We averaged the accuracies and the AUC values of the three testing sets.

5.2.1.2.1 Neonatal Convolutional Network

Before training our proposed N-CNN, we randomly divided the training set (1513 frames) into final-training (70%), validation (20%) and testing (10%) sets. Then, we performed image augmentation to increase the size of the training set as follows:

1. Each frame was rotated randomly by 15 degrees. This procedure was repeated 24 times to generate 24 augmented images.

2. Each rotated image was flipped horizontally and vertically. This procedure generated a total of 72 (24 original + 24 horizontal-flip + 24 vertical-flip) augmented images for each frame.

It is important to note that we have not performed any data augmentation on the separated testing set of 1513 frames. Also, the images of both the training and testing sets were re-sized to 120 $\times$ 120 using a bi-cubic interpolation method. We used Keras [145] and Tensorflow [146] for training and testing our proposed Neonatal-CNN (N-CNN).

The assessment of pain using our N-CNN achieved 91.5% average accuracy and 0.93 AUC. Table 5.1 presents the confusion matrix. As shown in the matrix, both FNR and FPR of the N-CNN are lower than the handcrafted features. The difference of AUC between N-CNN and the three handcrafted methods (i.e., strain, geometric, and LBP-TOP) is statistically significant (P=0.05). This result indicates that the learned features might be better in assessing neonatal pain than the handcrafted features.

5.2.1.2.2 Transfer Learning

We present here the results of classifying the emotional states of neonates into pain or no-pain using several pre-trained CNN models. Using a pre-trained CNN offers an attractive and more practical alternative to training the CNN from scratch. We refer the reader to Appendix C for further discussion about the advantages of using transfer learning in medical applications. To extract deep features from the neonates' facial images, we used the following pre-trained CNNs as fixed feature extractors:

1. VGG-F, VGG-M, and VGG-S CNN architectures, which were originally trained on ImageNet dataset (approximately 1.2M images and 1000 classes) for image classification. The architectures of these CNNs are presented in tables 4.4 to 4.6.

2. VGG-Face CNN, which was trained on a large face dataset (approximately 2.6M face images of 2622 identities) for face recognition. We hypothesize that this architecture should achieve higher pain classification results than VGG-F, VGG-M, and VGG-S since it is trained originally on a dataset relatively similar to our infant's faces dataset. Table 4.3 presents the architecture of VGG-Face.

Before feeding the images to the four pre-trained CNNs, we re-sized them to 224 X 224 to accommodate the size requirement of these CNNs (244 x 224 x 3, RGB images). All CNN architectures were implemented in a MATLAB Toolbox called MatConvNet [141].

We used the four pre-trained CNNs as fixed feature extractor to extract deep features from all the images in the training (1513 images) and testing (1513 images) sets. We then performed feature selection to reduce the high dimensionality of the extracted deep feature vectors. We applied two feature selection methods, namely Relief-F [142] and Symmetric Uncertainty [143]. For

Table 5.2: Pain classification performance using deep features of higher layer.

CNNs	VGG-F		VGG-M		VGG-S		VGG-Face	
	PostR	PreR	PostR	PreR	PostR	PreR	PostR	PreR
Dimensions	4096	4096	4096	4096	4096	4096	4096	4096
Selector	RF(5)	SU(5)	SU(10)	SU(10)	RF(10)	SU(5)	SU(15)	SU(5)
Classifier	SVMs	NB	RFT	NB	NB	RFT	kNN	kNN
Accuracy	83.86	89.29	83.13	83.86	**90.41**	87.10	**90.34**	89.55
AUC	0.74	0.74	0.74	0.75	**0.74**	0.72	**0.84**	0.86

Table 5.3: Pain classification performance using deep features of lower layer.

CNNs	VGG-F		VGG-M		VGG-S		VGG-Face	
	PostR	PreR	PostR	PreR	PostR	PreR	PostR	PreR
Dimensions	43264	43264	86528	86528	147968	147968	100352	100352
Selector	SU(10)	SU(15)	SU(10)	SU(15)	RF(15)	SU(10)	SU(15)	SU(10)
Classifier	NB	NB	NB	RFT	NB	RFT	NB	kNN
Accuracy	**87.13**	84.72	86.32	83.06	86.31	84.13	**88.23**	82.47
AUC	**0.71**	0.76	0.75	0.66	0.71	0.70	**0.80**	0.70

classification, we experimented with Naive Bayes (NF), k Nearest Neighbors (kNN), Support Vector Machines (SVMs), and Random Forests (RF) classifiers. We chose these classifiers because they have shown good classification performance in transfer learning applications. We experimented with the feature selection methods and the classifiers as implemented in Weka (version 3.7.13).

We used the deep features extracted from the last fully connected layer of VGG-Face and VGG-F, M, and S (Full 7 in tables 4.3, 4.4, 4.5, and 4.6). These features are more relevant to the utilized database. We also used deep features extracted from the last convolutional layer of VGG-Face and VGG-F, M, and S (Conv 5 in tables 4.3, 4.4, 4.5, and 4.6). These features are more generic (e.g., edge detector and colors) and independent of the utilized database. The main motivation for extracting features from both higher (Full 7) and lower (Conv 5) layers is to investigate which layers would give better pain classification results; what is the best layer to transfer? Tables 5.2 and 5.3 show the performance of pain classification using learned features extracted from the higher (Table 5.2) and lower (Table 5.3) layers of VGG-F, M, S and VGG-Face. PostR and PreR are abbreviations that indicate features were extracted after and before applying Rectified Linear Unit (ReLU) function. NB, RFT, RF, and SU represent Naive Byes, Random Forest Trees, RelieF, and

Symmetric Uncertainty, respectively; (#) indicates the number of features that were selected by RF or SU.

As shown in Table 5.2, VGG-S achieved the highest pain classification performance (90.4% accuracy and 0.74 AUC) as compared to VGG-F and VGG-M. However, the difference of AUC between VGG-S, VGG-F, and VGG-M is not statistically significant (P=0.05). VGG-Face achieved the highest pain classification performance and the AUC difference between VGG-Face (0.84) and VGG-S (0.74) is statistically significant (P=0.05). Table 5.3 shows the performance of pain classification using the lower layers of the pre-trained CNNs. The best pain classification accuracy among VGG-F, VGG-M, and VGG-S was obtained using VGG-F (87.13%). Although VGG-F has the highest accuracy, the AUC difference between VGG-F, and VGG-M and S is not statistically significant (P=0.05). VGG-Face achieved the highest pain classification performance and the AUC difference between VGG-Face (0.80) and VGG-F (0.71) is statistically significant (P=0.05). This result is consistent with our hypothesis that VGG-Face would achieve better pain classification since it was previously trained on a relatively similar database (face database).

As we mentioned earlier, the reason for extracting features from higher and lower layers is to investigate which layers would yield better pain classification results. Therefore, we compared the best result obtained from the higher layer of VGG-Face with the best result obtained from the lower layer. Although the accuracy of the former is approximately 2.1% higher than the latter, the AUC difference between them is not statistically significant at the P=0.05 level.

In summary, this section (Section 5.2.1) presents the results of assessing neonatal pain using facial features extracted by handcrafted and deep learning methods. The handcrafted methods that were used to extract facial features include optical strain, LBP variations, and geometric distances. The facial strain achieved the lowest accuracy (83.88%) as compared to LBP-TOP (87.66%) and geometric distances (88.13%). We speculate that the strain features achieved lower performance because the strain, derived from the optical flow, is sensitive to other motions. The texture features and the geometric distances, which were computed according to the NFCS, are considered more pain-specific than the strain features. The proposed N-CNN achieved the highest pain classification performance. These results suggest that the learned features are probably more discriminative than the handcrafted features in pain assessment.

Table 5.4: Confusion matrix of neonatal pain assessment from body movement.

	Body Motion Feature	
	Pain	No Pain
Pain	TPR = 61.9%	FNR = 38.1%
No Pain	FPR = 5.6%	TNR = 94.4%

Table 5.5: Confusion matrix of neonatal pain assessment from crying sound.

	Handcrafted Features		**Learned Features (N-CNN)**	
	Pain	No Pain	Pain	No Pain
Pain	TPR = 57.14%	FNR = 42.86%	**TPR = 75%**	FNR = 25%
No Pain	FPR = 6.87%	TNR = 93.13%	FPR = 3.7%	**TNR = 96.3%**

5.2.2 Pain Assessment From Body Movement

To assess neonatal pain from body movement, we used the Motion Image as described in Section 4.3. Assessing neonatal pain based on analysis of body movement achieved 84.41% average accuracy and 0.77 AUC. The average accuracy was computed as follows: 1) we divided the 31 subjects into 10 folds, 2) we trained the classifier using 9 folds and tested on the 10th fold, and 3) we repeated the process 10 times and calculated the average accuracy over the ten testing folds. Table 5.4 presents the confusion matrix of assessing neonatal pain from body movement.

5.2.3 Pain Assessment From Crying Sound

To assess neonatal pain from sounds, we used both handcrafted and learned features. We present next the results of pain assessment using both types of features.

5.2.3.1 Handcrafted Methods

To assess neonatal pain using crying sound, we computed LPCC and MFCC as described in Section 4.4.2.1. For classification, we experimented with Naive Bayes (NF), k Nearest Neighbors (kNN), Support Vector Machines (SVMs), and Random Forests (RF) classifiers. The classifiers were evaluated as follows: 1) we divided the 31 subjects into 10 folds, where each fold has approximately 3 subjects; 2) we trained the classifier using 9 folds and tested on the 10th fold; and 3) we repeated the process 10 times and calculated the average accuracy over the ten testing folds. Assessing

neonatal pain using handcrafted features extracted from sounds achieved 82.35% average accuracy and 0.69 AUC. The confusion matrix is presented in Table 5.5.

5.2.3.2 Deep Learning Methods

The audio signals were augmented as described in Section 4.4.2.2. Then, all the signals were converted to spectrogram images, which were used to build the N-CNN. We divided the entire dataset of sound signals into training and testing sets three times to obtain three training sets ($TR_1 - TR_3$) and testing sets ($TS_1 - TS_3$). The final performance was then reported by averaging the performances of the three folds. Assessing neonatal pain using the learned features achieved 93.95% average accuracy and 0.83 AUC. The confusion matrix is presented in Table 5.5. As the matrix shows, the FPR and FNR of the learned features are significantly lower than the FPR and FNR of the handcrafted features. Also, the difference of AUC between the handcrafted and learned features is statistically significant (P=0.05). This result indicates that the learned features might be better in assessing neonatal pain than the handcrafted features.

Before we proceed, we would like to note that we extracted the previously mentioned features from unfiltered audio signals that contain several sources of noise: sounds of nurses, parents, other infants, and the equipment. We believe using advanced methods to separate the neonate's sound from the environmental noise would improve the pain assessment performance.

5.2.4 Pain Assessment From Vital Signs

To assess neonatal pain using vital signs, we computed several descriptive statistics from the exported vital signs. The length of the vital signs feature vector for each video is 3 x S, where 3 represents HR, RR, and SpO2, and S represents the number of computed statistics. Then, we applied Relief-F and Symmetric Uncertainty (see Appendix D) followed by several machine learning classifiers (e.g., SVM and kNN). These classifiers were evaluated using subject-level 10-fold cross-validation. Assessing neonatal pain based on analysis of vital signs achieved 81.73% average accuracy and 0.72 AUC. The accuracy was computed by averaging the accuracies of the ten folds. Table 5.6 presents the confusion matrix of assessing neonatal pain using vital signs.

Table 5.6: Confusion matrix of neonatal pain assessment from vital signs.

	Vital Signs Statistics	
	Pain	No Pain
Pain	TPR = 56.76%	FNR = 43.24%
No Pain	FPR = 6.84%	TNR = 93.16%

Table 5.7: Confusion matrices for decision-level and feature-level fusion.

	Decision-Level Fusion		**Feature-Level Fusion**	
	Pain	No Pain	Pain	No Pain
Pain	TPR = 84.5%	FNR = 15.5%	TPR = 83.9%	FNR = 16.1%
No Pain	FPR = 1%	TNR = 99%	FPR = 6.9%	TNR = 93.1%

5.3 Multimodal Pain Assessment

We combined different pain indicators using decision-level fusion and feature-level fusion to classify the emotional state of 31 neonates into pain or no-pain. The multimodal pain assessment is mandatory because it allows to assess pain during circumstances when not all pain responses are available due to occlusion (e.g., stomach-down sleep and swaddling), clinical condition (e.g., Bell's palsy), level of activity (e.g., physical exertion), and sedation.

The accuracy of pain assessment using decision-level fusion (95.1%) is higher than the accuracy using feature-level fusion (92.3%), whereas the AUC of feature-level fusion (0.89) is higher than the decision-level fusion (0.85). The AUC difference between decision and feature fusion is not statistically significant (P=0.05). Table 5.7 presents the confusion matrices for both decision-level and feature-level fusion. As can be seen from the table, both the FNR and FPR of decision-level is lower than the FNR and FPR of feature-level fusion. Since minimizing both FPR and FNR is important to prevent over- and under-treatment, we can conclude that the decision-level fusion is better than the feature-level fusion.

5.4 Summary and Discussion

The previous sections (Section 5.2 and Section 5.3) present the results of unimodal and multimodal pain assessment. The performance of the unimodal pain assessment using handcrafted methods is presented in Table 5.8.

Table 5.8: Performance of the unimodal pain assessment.

	Facial Expression			**Body Motion**	**Crying Sounds**	**Vital Signs**
Handcrafted	St.	Geo.	LT.	Motion Image	LPCC/MFCC	Statistics
Accuracy	83.88%	88.13%	87.66%	84.41%	82.35%	81.73%
AUC	0.75	0.85	0.85	0.77	0.69	0.72

Table 5.9: Pain assessment from facial expression and crying sounds using N-CNN.

	Facial Expression	**Crying Sounds**
Accuracy	91.5%	93.95%
AUC	0.93	0.83

For facial expression pain response, the highest assessment accuracy (88.13%) was obtained using the geometric facial features (Geo. in Table 5.8). The second highest accuracy (87.66%) was obtained using LBP-TOP features (LT. in Table 5.8). The lowest accuracy (83.88%) was obtained using the facial strain features (St. in Table 5.8). We are of the opinion that the strain features achieved lower performance because the strain, derived from the optical flow is sensitive to other motions. As the table shows, facial expression achieved the highest pain assessment performance as compared to body movement (84.41%), crying sounds (82.35%), and vital signs (81.73%). These results are consistent with previous findings [147] that facial expression is the most pain specific response and vital signs are less pain specific since they can be associated with other conditions such as noise, hunger, age, or underlying disease. The highest pain assessment performance was achieved using the multimodal approach (decision-level: 95.1% and feature-level: 92.3%). The multimodal approach was able to accurately assess pain in case of missing data (e.g., occlusion of face), which is known to be common in the clinical environment. Figure 5.1 presents a summary of the unimodal and multimodal pain assessment using handcrafted methods.

Table 5.9 presents the performance of pain assessment from facial expressions and sounds using N-CNN. The performance of both facial expression and sound using the learned features is higher than the performance using handcrafted features. The AUC difference between facial expression assessment using handcrafted features and facial expression assessment using learned features is statistically significant (P=0.05). Similarly, the AUC difference between sound assessment using handcrafted features and sound assessment using learned features is statistically significant (P=0.05). These results suggest that learned features can yield better pain assessment perfor-

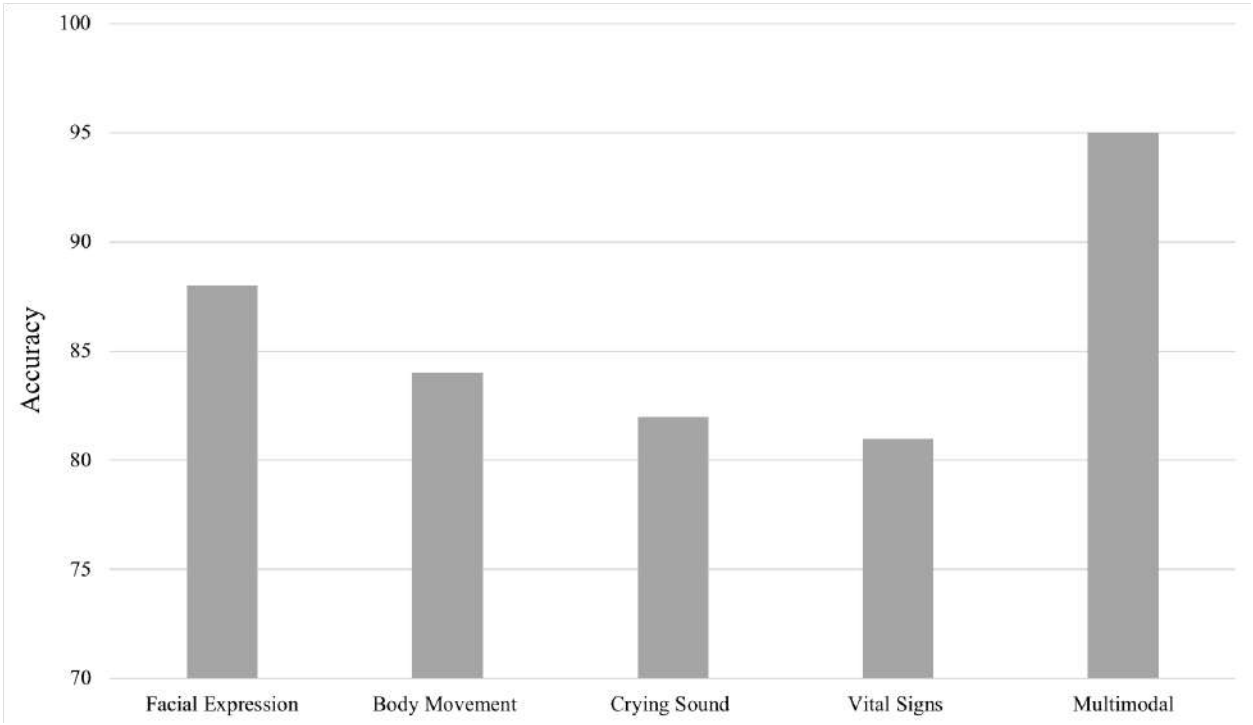


Figure 5.1: Performance of automatic pain assessment: unimodal and multimodal.

mance. We are currently developing a Multimodal Neonatal Neural Network (MN-NN) that assess the neonates' pain and condition using facial expression, crying sound, body movement, and vital signs data. We are investigating a feature-level and decision-level fusions to combine multiple pain modalities. In the first this scheme, the multimodal network would learn the features of each modality separately, and then combine these features in a shared representation layer followed by the classification layer. This scheme allows the network to learn each modality separately and learn the correlations/dependencies between these modalities before making a final decision. The second scheme involves developing a single network for each modality and then combine the outcomes of these networks using weighted voting or Bayesian approach.

5.5 Model Specific Pain Assessment

We present here the result of assessing pain for a specific group of neonates. We divided the neonates into groups according to their gender (female and male), gestational age (pre-term and full term), birth-weight (low weight and normal weight), and race (Hispanic and Non-Hispanic),

Table 5.10: Distribution of neonates across different groups.

Gender	Female	15
	Male	16
Age	pre-term (<37 GW)	14
	full-term (37 to 42 GW)	17
Ethnicity	Hispanic	9
	Not Hispanic	22
birth weight	Low (<2,500 gm)	10
	Normal (>= 2,500 gm)	21

built a model for each group, and reported each model's performance using the average accuracy and the AUC. The average accuracy was obtained by averaging the accuracies of subjects (i.e., leave-one-subject-out cross validation); Table 5.10 shows the number of subjects in each group.

5.5.1 Gestational Age Model

The pre-term model achieved a higher pain assessment accuracy (approx. 76%) than the full-term model (approx. 72%). However, the AUC difference between the pre-term model and the full-term model is not significant ($p < 0.05$). We think the lower accuracy of full-term model might be attributed to the wider range of expressions this group has, as compared to the pre-term group whose dominant expression is the pain expression.

In case of body movement, the pre-term model has a higher pain assessment accuracy (approx. 83%) than the full-term model (approx. 73%). The AUC for the pre-term model is also higher; however, the difference of AUC between the two models is not significant ($p < 0.05$). We think this result might be attributed to the high frequency of body movement of the older neonates in our dataset as compared to the pre-term neonates. This result suggests that there might be an association between the neonates' movement and their gestational age. However, we believe the number of subjects in each group is small to draw a conclusion; further investigation, on a larger dataset, is required to validate these results.

As for crying, the difference in the accuracy of pain assessment between the pre-term and full-term models is approximately 2%, and the difference of AUC between them is not significant ($p < 0.05$). The accuracy of assessing pain based on vital signs for pre-term neonates (approx. 80%) is relatively higher than full-term neonates (approx. 70%). However, the difference of AUC between

the two models is not significant ($p < 0.05$). We think the inter-individual differences might have an impact on the accuracy of vital signs between the pre-term and full-term groups. To validate these results and draw a solid conclusion, further investigation on a larger dataset is needed.

The feature fusion's performance of the pre-term model (approx. 84%) is higher than the full-term model (approx. 74%). Similarly, the accuracy of the pre-term model (approx. 94%) in the case of decision fusion is higher than the full-term model (approx. 89%). Although the accuracy of pain assessment differs between the pre-term and full-term models, the difference of AUC between them is not significant ($p < 0.05$) for both feature fusion and decision fusion.

Previous pediatric studies [26] reported a strong association between neonates' age and their response to painful stimulus. The results presented above suggest that the gestational age of neonates might have an impact on their pain response. However, further investigation, on a larger dataset, is required to validate these results because the number of subjects in each group is small to draw conclusions.

5.5.2 Gender Model

The male model has a higher pain assessment accuracy (approx. 76%) than female model (approx. 73%) in case of facial expression. However, the AUC difference between the male model and the female model is not significant ($p < 0.05$). In case of body movement, the male model has almost the same accuracy as the female model (approx. 78%), and the AUC values of both models are very similar. Likewise, the pain assessment accuracy of crying for the male model (approx. 74%) is very similar to the female model (approx. 75%), and the difference of AUC between female and male is not significant ($p < 0.05$). As for the vital signs, the accuracy of the female model (approx. 80%) is higher than the male model (approx. 78%), but the difference of AUC between female and male is not significant ($p < 0.05$).

The feature fusion's performance of the female model (approx. 85%) is higher than the male model (approx. 80%). Similarly, the accuracy of the female model (approx. 92%) in the case of decision fusion is higher than the male model (approx. 89%). Although the accuracy of pain assessment differs between the female and male models, the difference of AUC between them is not significant ($p < 0.05$) for both feature fusion and decision fusion.

5.5.3 Weight Model

In case of facial expression, the accuracies of pain assessment for low and normal weight are similar (low: approx. 74% and normal: approx. 73%), and the difference of AUC between low and normal models is not significant ($p < 0.05$). Also, the accuracies of pain assessment for low and normal weight using crying analysis and vital signs changes are similar (low: approx. 74% and normal: approx. 72%), and the difference of AUC between them is not significant ($p < 0.05$). As for the body movement, the normal birth weight neonates have lower accuracy than the low birth weight neonates (normal: approx. 77% and low: approx. 84%). Note that many of the neonates in the low birth weight group are pre-term neonates.

The feature fusion's performance of the low birth weight model is similar to the normal birth weight model (low: approx. 76% and normal: approx. 78%). On the other hand, the decision fusion's performance of the low birth weight model is higher than the normal birth weight model (low: approx. 93% and normal: approx. 89%). The difference of AUC between the low and normal models is not significant ($p < 0.05$) for both feature fusion and decision fusion.

The results presented above suggest that there is no association between neonates' birth weight and their response to pain. To further examine the influence of birth weight on neonates' pain responses, more investigation, on a larger dataset, is needed.

5.5.4 Race Model

The performance of pain assessment for Hispanic group (approx. 76%) is similar to Non-Hispanic group (approx. 77%) in the case of facial expression. The accuracy of pain assessment of the Hispanic model (approx. 81%) is higher than Non-Hispanic model (approx. 78%) for body movement, but lower for crying sound (Hispanic: 65% and Non-Hispanic: 72%). However, the difference of AUC between Hispanic and Non-Hispanic models is not significant ($p < 0.05$) for both body movement and crying. As for the vital signs, the difference of accuracy between the Hispanic and Non-Hispanic models is around 2%, and the difference of AUC is not significant ($p < 0.05$). The pain assessment performance of both the feature fusion and decision fusion is lower for the Hispanic group, but the difference of AUC between the Hispanic and Non-Hispanic groups is not significant ($p < 0.05$) for both feature fusion and decision fusion.

5.6 Comparison With the State of the Art

There are two ways to compare our work with the state of the arts in neonatal pain assessment. The first way is to re-implement the state of the art methods and apply them to our dataset. We have re-implemented Fotiadou et al. [95] method and applied it to our dataset. However, the obtained performance of our re-implementation was different than the performance reported in [95]. We think our choice of specific parameters and thresholds, due to the limited technical details provided in [95], affected our re-implementation and led to a different result. Therefore, we decided to compare our performance of pain assessment from facial expression with the performance of [95] as reported in the paper.

Fotiadou et al. [95] reported 0.98 AUC value in detecting pain expression of eight neonates (15 videos). This performance was obtained by varying the decision thresholds of the SVM classifier, which was evaluated using leave-one-subject-out cross-validation. Our algorithm achieved 0.85 AUC using handcrafted features with a subject level 10-fold cross-validation and 0.93 AUC using N-CNN; the number of subjects is 31 ($>$ 200 videos).

In case of crying sound, we compared our performance with Vempada et al. [99]. Vempada et al. [99] reported 80.56% average accuracy in detecting the pain cry from 120 cry corpus. This performance was obtained using SVM classifier, which was evaluated on a testing set. Our algorithm achieved 82.35% average accuracy in detecting pain cry of 31 neonates ($>$ 200 corpus). This accuracy was obtained using SVM classifier, which was evaluated using a subject level 10-fold cross validation. As previously mentioned, our work is the first to assess neonatal pain based on analysis of visual, vocal, and physiological signals. Therefore, we have not provided any comparison for pain assessment using body movement or multimodal.

The second way of comparison with the state of the art is to apply our pain assessment approach to existing publicly available neonatal databases and report the results. We followed this way and applied our method of assessing neonatal pain to COPE database. The database consists of static images of neonates' faces taken during four different stimuli:

1. Pain stimulus during the heel lancing (60 images).
2. Rest/cry stimulus during the transportation of an infant from one crib to another (63 rest images and 18 cry images).

Table 5.11: Confusion matrix of applying N-CNN to COPE database.

	Pain	**No Pain**
Pain	TPR = 79%	FNR = 21%
No Pain	FPR = 11%	TNR = 89%

3. Air stimulus to the nose (23 images).

4. Friction stimulus, which involves receiving friction on the external lateral surface of the heel with cotton soaked in alcohol (36 images).

Similar to [84], we divided COPE images into two sets: no-pain set (144 images) and pain set (60 images). The pain set contains images of neonates during acute painful stimulus (heel-lancing) while the no-pain set contains images of neonates during the other three stimuli. Then, we applied our N-CNN to COPE neonatal database (i.e., N-CNN trained in our database and tested on COPE database).

Applying our proposed N-CNN to COPE database achieved 84.5% average accuracy. Table 5.11 presents the confusion matrix. We think the lower accuracy of applying our N-CNN to COPE, in comparison with the accuracy of applying N-CNN to our database, is attributed to the difference between these two databases. Our neonatal database consists of two main stimuli: pain (heel lancing) and no-pain (normal or rest states). However, as mentioned above, COPE database was divided into pain set (heel lancing) and no-pain set (rest/cry, air, and friction). Figure 5.2 shows examples of the images that were incorrectly classified by our N-CNN.

5.7 Pain Monitoring in Real Life Clinical Setting

The results presented above are generated using segments of only pain or no-pain events. In real-life clinical setting, the pain assessment system should be applied continuously over segments that can contain both pain and no-pain events; i.e., pain and no-pain events fall into the same segment.

To validate our approach and gain insight of real-life pain assessment performance, we asked the nurses to manually selects segments that have both pain and no-pain events from videos of 8 infants. We called the selected segments "overlap segments". We then applied the trained model to full

Figure 5.2: Images of COPE database that were mislabeled by N-CNN.

video sequences. The model starts by classifying the first fixed window (w_1) followed by classifying the following window (w_n) followed by classifying the next window (w_{n+1}) until it reaches the last window. Each of these windows, which has a length of 5 seconds, has a label and confidence score. If the confidence score of a specific window is not high (i.e., below a pre-determined threshold), we used information from neighbor windows to decide the class of the current window. After applying the trained model over the entire video using the fixed window, we computed the performance in detecting overlap and non-overlap segments as follows:

- Non-overlap segments: Non-overlap segments are the segments that contain only pain or no-pain events. The number of non-overlap segments for 8 infants is 64 (8 neonates $\times$ 8 segments). A non-overlap segment was classified correctly if and only if most (>= 80%,

empirically determined) of the windows within this segment classified as belonging to one class (pain or no-pain).

- Overlap segments: An overlap segment contains both pain and no-pain events. The number of overlap segments selected by nurses from 8 infants is 19. Contrary to a non-overlap segment, an overlap segment was classified correctly if the windows within this segment classified as belonging to both pain and no-pain class (e.g., 60% pain and 40% no-pain).

Assessing pain using this approach achieved 90.36% overall accuracy (59/64 for non-overlap and 16/19 for overlap). This result demonstrates the feasibility of using our system for pain monitoring in real-life. The proposed system would continuously monitor neonates and send a notification to caregivers when pain event is detected.

This chapter presented the results of applying the automatic pain assessment methods to our real-world NPAD database and provided comparison with the state of the art. The reported results are encouraging and demonstrate the feasibility of automatic neonatal pain assessment. The next chapter summarizes our contributions, discusses current limitations, and lists several potential future directions.

CHAPTER 6

CONCLUSIONS

This dissertation has proposed an automatic and multimodal pain assessment system that mitigates the shortcomings of the current practice. This chapter summarizes the dissertations' main contributions (Section 6.1), presents several possible directions for future work (Section 6.2), and makes closing remarks (Section 6.3).

6.1 Dissertation's Summary

This dissertation advances the state of the art in automatic pain assessment in the following ways:

1. Comprehensive Review of Pain Assessment Methods: This dissertation is the first to present a comprehensive review of automatic pain assessment methods, identify their limitations, and propose several possible directions to advance research. It also presented a review of existing pain databases and their limitations. This review is important because it enables researchers to understand the current status of automatic pain assessment.

2. Multidimensional Pain Database for Neonates: We are the first to collect a well-annotated and comprehensive neonatal database (NPAD) that consists of video, audio, and physiological data. Our approach for collecting the database can be easily integrated into the clinical environment. Also, this database is representative of real-world conditions since it was collected without any modifications to the clinical environment. This database addresses the need for pain databases to advance the research in automatic pain assessment.

3. Novel Handcrafted Features for Pain Assessment: We present several handcrafted descriptors to extract pain-relevant features from videos (strain, geometric, LBP-TOP, and motion image) and audios (LPCC and MFCC) of neonates. We are aware of the efficiency of using FACS-based handcrafted descriptor for pain expression recognition. However, we did not implement FACS-based method because it would require manually labeling the AUs in each frame by FACS-experts. The process of manually labeling AUs is very time-consuming (3 hours to code one minute [62]). In the future, we plan to provide AUs labels for our NPAD database and develop a FACS-based method for neonatal pain recognition.

4. Novel Learned Features for Pain Assessment: This dissertation introduces deep learning methods for assessing neonatal pain. The results have shown that deep learning methods performed better than handcrafted methods. However, the handcrafted methods help to understand the significance of various features and their ability to recognize the level of pain.

5. Fully Automated and Multimodal Pain Assessment System: This dissertation presents a multimodal neonatal pain assessment system that can be easily adopted and integrated into clinical environments since it uses inexpensive and non-invasive devices for pain monitoring. To the best of our knowledge, this dissertation is the first to introduce an automatic multimodal system for neonatal pain assessment.

6.2 Future Directions

While this dissertation demonstrates the feasibility of automatic neonatal pain assessment, it can be extended and improved in several ways:

1. Neonatal Face Tracking and Alignment: Accurately tracking the neonates' face and facial landmarks is still an open problem. In this dissertation, we have used several well-known face trackers that have shown excellent performance for adult and child populations. Unfortunately, the vast majority of these trackers performed poorly when applied to neonates. A noticeable extension would be to design and implement,

as part of our proposed system, a neonatal face tracker that can detect the face under severe occlusions, extreme poses, and different illumination conditions. This extension would lead to a more complete and stand-alone neonatal pain assessment system.

2. Multimodal Neonatal Neural Network: As discussed in Chapter 5, CNN achieved better pain classification performance, as compared to the handcrafted methods. Therefore, an immediate extension of this dissertation would be to design and build a Multimodal Neonatal Neural Network (MN-NN) for the automatic assessment of neonatal pain. This multimodal NN would integrate crying sounds, body movements, and facial expressions to the physiological readings and use them together to predict the neonate's state.

3. Contactless Detection of Vital Signs: The sensors of vital signs are expensive, cause stress, and can damage the infants' delicate skin. A possible extension would be to develop or use a video-based vital signs detection method for collecting vital signs values. This extension provides a cheaper and user-friendly alternative to the use of electrodes/sensors making our proposed pain assessment system more suitable for home care monitoring.

4. Assessment of Other Emotions and Modalities: Although this dissertation focuses on pain, we believe automatically recognizing other emotions such as hunger and discomfort is important because it can reduce the caregivers' time-commitment. Also, studies [121, 120, 119, 123] have reported an association between changes in cerebral oxygenation and pain. Therefore, we think incorporating the brain's hemodynamic activity and using it as an objective measure of pain is another possible future direction. In addition to the physiological data, it would be interesting to explore the association between neonatal pain and eye movements or pupil dilation because studies [148, 149, 150] reported strong ties between pupil-size and emotions.

5. Collection of a Larger Database: Collecting data from several hundreds of neonates is necessary to advance this work and draw a solid conclusion. In the future, we plan

to collect data for 600 infants divided into four gestational age groups: 1) 23 to 28 weeks, 2) 29 to 33 weeks, 3) 34 to 36 weeks, and 4) 37 weeks or older. The collected data would include neonates from different racial groups while experiencing both procedural and postoperative pain.

6. Real Time Monitoring of Neonatal Pain: While this dissertation proved the feasibility of neonatal pain assessment, various types of optimization should be done to make it suitable for real-time monitoring. The code should be optimized so it requires minimum size and achieves maximum speed (executes efficiently in real-time).

6.3 Closing Remarks

Pain exposure in early life is a serious issue that causes several short- and long-term outcomes. Using analgesic medications such as Morphine and Fentanyl is one way to mitigate this issue. However, recent studies [16, 17, 18, 19] found that the excessive use of analgesic medications is associated with impaired cerebellar growth in the neonatal period and poorer neurodevelopmental outcomes in the early childhood period. These findings suggest that both the failure to recognize and treat neonatal pain as well as the administration of certain analgesic medications in the absence of pain may lead to serious outcomes and cause permanent alterations in brain structure and function.

The current practice for assessing neonatal pain is inconsistent because it depends highly on the observer bias. Additionally, it is discontinuous and requires a large number of well-trained nurses to ensure the proper utilization of the pain scale. The discontinuous nature of the current practice as well as the inter-observer variations may result in delayed intervention and inconsistent treatment of pain. Since pain assessment is the cornerstone of pain management, developing automatic and continuous scales that generate immediate and more consistent pain assessment is crucial.

This dissertation investigates a challenging and important area of research and proves its feasibility. By continuing to explore this area and develop a highly accurate automatic assessment of pain, we hope to improve the effectiveness of pain intervention while mitigating the short- and long-term outcomes of pain exposure in early life.

LIST OF REFERENCES

[1] M. McCaffery, *Nursing practice theories related to cognition, bodily pain, and man-environment interactions*. University of California Print. Office, 1968.

[2] R. Grunau, "Self-regulation and behavior in preterm children: effects of early pain," *Progress in pain research and management*, vol. 26, pp. 23–56, 2003.

[3] R. E. Grunau, J. Weinberg, and M. F. Whitfield, "Neonatal procedural pain and preterm infant cortisol response to novelty at 8 months," *Pediatrics*, vol. 114, no. 1, pp. e77–e84, 2004.

[4] A. T. Bhutta and K. Anand, "Vulnerability of the developing brain: neuronal mechanisms," *Clinics in perinatology*, vol. 29, no. 3, pp. 357–372, 2002.

[5] K. Anand and F. M. Scalzo, "Can adverse neonatal experiences alter brain development and subsequent behavior?," *Neonatology*, vol. 77, no. 2, pp. 69–82, 2000.

[6] R. E. Grunau, L. Holsti, and J. W. Peters, "Long-term consequences of pain in human neonates," in *Seminars in Fetal and Neonatal Medicine*, vol. 11, pp. 268–275, Elsevier, 2006.

[7] M. DiLorenzo, R. Pillai Riddell, and L. Holsti, "Beyond acute pain: Understanding chronic pain in infancy," *Children*, vol. 3, no. 4, p. 26, 2016.

[8] J. Vinall, S. P. Miller, V. Chau, S. Brummelte, A. R. Synnes, and R. E. Grunau, "Neonatal pain in relation to postnatal growth in infants born very preterm," *Pain*, vol. 153, no. 7, pp. 1374–1381, 2012.

[9] A. A. of Pediatrics, Fetus, N. Committee, *et al.*, "Prevention and management of pain in the neonate: an update," *Pediatrics*, vol. 118, no. 5, pp. 2231–2241, 2006.

[10] S. Brummelte, R. E. Grunau, V. Chau, K. J. Poskitt, R. Brant, J. Vinall, A. Gover, A. R. Synnes, and S. P. Miller, "Procedural pain and brain development in premature newborns," *Annals of neurology*, vol. 71, no. 3, pp. 385–396, 2012.

[11] R. E. Grunau, M. T. Tu, M. F. Whitfield, T. F. Oberlander, J. Weinberg, W. Yu, P. Thiessen, G. Gosse, and D. Scheifele, "Cortisol, behavior, and heart rate reactivity to immunization pain at 4 months corrected age in infants born very preterm," *The Clinical journal of pain*, vol. 26, no. 8, p. 698, 2010.

[12] R. E. Grunau, M. F. Whitfield, J. Petrie-Thomas, A. R. Synnes, I. L. Cepeda, A. Keidar, M. Rogers, M. MacKay, P. Hubber-Richard, and D. Johannesen, "Neonatal pain, parenting stress and interaction, in relation to cognitive and motor development at 8 and 18 months in preterm infants," *Pain*, vol. 143, no. 1-2, pp. 138–146, 2009.

[13] S. M. Walker, "Translational studies identify long-term impact of prior neonatal pain experience," *Pain*, vol. 158, pp. S29–S42, 2017.

[14] A. Marchant, "'neonates do not feel pain': a critical review of the evidence," *Bioscience Horizons: The International Journal of Student Research*, vol. 7, 2014.

[15] B. Stevens, C. Johnston, P. Petryshen, and A. Taddio, "Premature infant pain profile: development and initial validation," *The Clinical journal of pain*, vol. 12, no. 1, pp. 13–22, 1996.

[16] J. G. Zwicker, S. P. Miller, R. E. Grunau, V. Chau, R. Brant, C. Studholme, M. Liu, A. Synnes, K. J. Poskitt, M. L. Stiver, *et al.*, "Smaller cerebellar growth and poorer neurodevelopmental outcomes in very preterm infants exposed to neonatal morphine," *The Journal of pediatrics*, vol. 172, pp. 81–87, 2016.

[17] S. Hu, W. S. Sheng, J. R. Lokensgard, and P. K. Peterson, "Morphine induces apoptosis of human microglia and neurons," *Neuropharmacology*, vol. 42, no. 6, pp. 829–836, 2002.

[18] D. Bajic, K. G. Commons, and S. G. Soriano, "Morphine-enhanced apoptosis in selective brain regions of neonatal rats," *International Journal of Developmental Neuroscience*, vol. 31, no. 4, pp. 258–266, 2013.

[19] A. T. Bhutta, C. Rovnaghi, P. M. Simpson, J. M. Gossett, F. M. Scalzo, and K. Anand, "Interactions of inflammatory pain and morphine in infant rats: long-term behavioral effects," *Physiology & behavior*, vol. 73, no. 1-2, pp. 51–58, 2001.

[20] A. S. Butler, R. E. Behrman, *et al.*, *Preterm birth: causes, consequences, and prevention.* National Academies Press, 2007.

[21] K. J. Anand, B. J. Stevens, and P. J. McGrath, *Pain in neonates and infants.* Elsevier Health Sciences, 2007.

[22] S. W. Derbyshire, "Gender, pain, and the brain," *Pain Clinical Updates*, vol. 16, no. 3, pp. 1–4, 2008.

[23] C. Miller and S. E. Newton, "Pain perception and expression: the influence of gender, personal self-efficacy, and lifespan socialization," *Pain Management Nursing*, vol. 7, no. 4, pp. 148–152, 2006.

[24] S. Gibbins, B. Stevens, P. J. McGrath, J. Yamada, J. Beyene, L. Breau, C. Camfield, A. Finley, L. Franck, C. Johnston, *et al.*, "Comparison of pain responses in infants of different gestational ages," *Neonatology*, vol. 93, no. 1, pp. 10–18, 2008.

[25] G. G. Page, "Are there long-term consequences of pain in newborn or very young infants?," *The Journal of perinatal education*, vol. 13, no. 3, pp. 10–17, 2004.

[26] B. O. Valeri and M. B. M. Linhares, "Pain in preterm infants: Effects of sex, gestational age, and neonatal illness severity.," *Psychology & Neuroscience*, vol. 5, no. 1, p. 11, 2012.

[27] N. Witt, S. Coynor, C. Edwards, and H. Bradshaw, "A guide to pain assessment and management in the neonate," *Current emergency and hospital medicine reports*, vol. 4, no. 1, pp. 1–10, 2016.

[28] L. Holsti, R. E. Grunau, M. F. Whifield, T. F. Oberlander, and V. Lindh, "Behavioral responses to pain are heightened after clustered care in preterm infants born between 30 and 32 weeks gestational age," *The Clinical journal of pain*, vol. 22, no. 9, p. 757, 2006.

[29] J. Lawrence, D. Alcock, P. McGrath, J. Kay, S. B. MacMurray, and C. Dulberg, "The development of a tool to assess neonatal pain.," *Neonatal network: NN*, vol. 12, no. 6, pp. 59–66, 1993.

[30] R. V. Grunau and K. D. Craig, "Pain expression in neonates: facial action and cry," *Pain*, vol. 28, no. 3, pp. 395 – 410, 1987.

[31] P. Hummel, M. Puchalski, S. Creech, and M. Weiss, "Clinical reliability and validity of the n-pass: neonatal pain, agitation and sedation scale with prolonged pain," *Journal of perinatology*, vol. 28, no. 1, p. 55, 2008.

[32] S. W. Krechel and J. BILDNER, "Cries: a new neonatal postoperative pain measurement score. initial testing of validity and reliability," *Pediatric Anesthesia*, vol. 5, no. 1, pp. 53–61, 1995.

[33] P. Hummel and M. van Dijk, "Pain assessment: current status and challenges," in *Seminars in Fetal and Neonatal Medicine*, vol. 11, pp. 237–245, Elsevier, 2006.

[34] S. H. Simons, M. van Dijk, K. S. Anand, D. Roofthooft, R. A. van Lingen, and D. Tibboel, "Do we still hurt newborn babies?: A prospective study of procedural pain and analgesia in neonates," *Archives of pediatrics & adolescent medicine*, vol. 157, no. 11, pp. 1058–1064, 2003.

[35] S. Dekel, B. Gedaliahu, A. Gori, A. Vasarri, M. C. Sorella, G. Di Nino, and R. M. Melotti, "Medical evidence influence on inpatients and nurses pain ratings agreement," *Pain Research and Management*, vol. 2016, 2016.

[36] R. R. Pillai Riddell and K. D. Craig, "Judgments of infant pain: The impact of caregiver identity and infant age," *Journal of pediatric psychology*, vol. 32, no. 5, pp. 501–511, 2006.

[37] R. R. Pillai Riddell, M. A. Badali, and K. D. Craig, "Parental judgments of infant pain: Importance of perceived cognitive abilities, behavioural cues and contextual cues," *Pain Research and Management*, vol. 9, no. 2, pp. 73–80, 2004.

[38] G. Zamzmi, R. Kasturi, D. Goldgof, R. Zhi, T. Ashmeade, and Y. Sun, "A review of automated pain assessment in infants: Features, classification tasks, and databases," *IEEE Reviews in Biomedical Engineering*, 2017.

[39] M. Monwar, S. Rezaei, and K. Prkachin, "Eigenimage based pain expression recognition," *IAENG International Journal of Applied Mathematics*, vol. 36, no. 2, pp. 1–6, 2007.

[40] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face – pain expression recognition using active appearance models," *Image and Vision Computing*, vol. 27, no. 12, pp. 1788 – 1796, 2009. Visual and multimodal analysis of human spontaneous behaviour:.

[41] Z. Hammal and M. Kunz, "Pain monitoring: A dynamic and context-sensitive system," *Pattern Recognition*, vol. 45, no. 4, pp. 1265 – 1280, 2012.

[42] Z. Hammal and J. F. Cohn, "Towards multimodal pain assessment for research and clinical use," in *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges*, RFMIR '14, (New York, NY, USA), pp. 13–17, ACM, 2014.

[43] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *Face and Gesture 2011*, pp. 57–64, March 2011.

[44] P. Lucey, J. Howlett, J. Cohn, S. Lucey, S. Sridharan, and Z. Ambadar, "Improving pain recognition through better utilisation of temporal information," in *International conference on auditory-visual speech processing*, vol. 2008, p. 167, NIH Public Access, 2008.

[45] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin, "Automatically detecting pain in video through facial action units," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, pp. 664–674, June 2011.

[46] Z. Wei and X. Li-min, "Pain expression recognition based on slpp and mksvm," *International Journal of Engineering and Manufacturing (IJEM)*, vol. 1, no. 3, p. 69, 2011.

[47] Z. Chen, R. Ansari, and D. Wilkie, "Automated detection of pain from facial expressions: a rule-based approach using aam," in *Medical Imaging 2012: Image Processing*, vol. 8314, p. 83143O, International Society for Optics and Photonics, 2012.

[48] K. Sikka, A. Dhall, and M. Bartlett, "Weakly supervised pain localization using multiple instance learning," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, April 2013.

[49] K. Sikka, "Facial expression analysis for estimating pain in clinical settings," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, (New York, NY, USA), pp. 349–353, ACM, 2014.

[50] S. Zhu, "Pain expression recognition based on plsa model," *The Scientific World Journal*, vol. 2014, 2014.

[51] R. Niese, A. Al-Hamadi, A. Panning, D. G. Brammen, U. Ebmeyer, and B. Michaelis, "Towards pain recognition in post-operative phases using 3d-based features from video and support vector machines.," *International Journal of Digital Content Technology and its Applications-IJDCTA*, vol. 3, no. 4, pp. 21–33, 2009.

[52] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain recognition from video and biomedical signals," in *2014 22nd International Conference on Pattern Recognition*, pp. 4582–4587, Aug 2014.

[53] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm, "Multimodal data fusion for person-independent, continuous estimation of pain intensity," in *Engineering Applications of Neural Networks*, pp. 275–285, Springer, 2015.

[54] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, S. Crawcour, P. Werner, A. Al-Hamadi, and A. O. Andrade, "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system," in *2013 IEEE International Conference on Cybernetics (CYBCO)*, pp. 128–131, June 2013.

[55] S. Walter, S. Gruss, H. Traue, P. Werner, A. Al-Hamadi, M. Kächele, F. Schwenker, A. Andrade, and G. Moreira, "Data fusion for automated pain recognition," in *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pp. 261–264, May 2015.

[56] M. Velana, S. Gruss, G. Layher, P. Thiam, Y. Zhang, D. Schork, V. Kessler, S. Meudt, H. Neumann, J. Kim, *et al.*, "The senseemotion database: A multimodal database for the development and systematic validation of an automatic pain-and emotion-recognition system," in *IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction*, pp. 127–139, Springer, 2016.

[57] M. Kächele, M. Amirian, P. Thiam, P. Werner, S. Walter, G. Palm, and F. Schwenker, "Adaptive confidence learning for the personalization of pain intensity estimation systems," *Evolving Systems*, vol. 8, no. 1, pp. 71–83, 2017.

[58] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges," in *Proceedings of the British Machine Vision Conference*, pp. 119–1, 2013.

[59] S. Walter, S. Gruss, K. Limbrecht-Ecklundt, H. C. Traue, P. Werner, A. Al-Hamadi, N. Diniz, G. M. d. Silva, and A. O. Andrade, "Automatic pain quantification using autonomic parameters," *Psychology & Neuroscience*, vol. 7, pp. 363 – 380, 12 2014.

[60] P. Werner, A. Al-Hamadi, and R. Niese, "Comparative learning applied to intensity rating of facial expressions of pain," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 05, p. 1451008, 2014.

[61] S. Gruss, R. Treister, P. Werner, H. C. Traue, S. Crawcour, A. Andrade, and S. Walter, "Pain intensity recognition rates via biopotential feature patterns with support vector machines," *PloS one*, vol. 10, no. 10, p. e0140330, 2015.

[62] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Faces of pain: Automated measurement of spontaneous all facial expressions of genuine and posed pain," in *Proceedings of the 9th International Conference on Multimodal Interfaces*, ICMI '07, (New York, NY, USA), pp. 15–21, ACM, 2007.

[63] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1797 – 1803, 2009. Visual and multimodal analysis of human spontaneous behaviour:.

[64] M. S. Bartlett, G. C. Littlewort, M. G. Frank, and K. Lee, "Automatic decoding of facial movements reveals deceptive pain expressions," *Current Biology*, vol. 24, no. 7, pp. 738 – 743, 2014.

[65] C. Florea, L. Florea, and C. Vertan, "Learning pain from emotion: Transferred hot data representation for pain intensity estimation.," in *ECCV Workshops (3)*, pp. 778–790, 2014.

[66] S. Kaltwang, O. Rudovic, and M. Pantic, *Continuous Pain Intensity Estimation from Facial Expressions*, pp. 368–377. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[67] S. D. Roy, M. K. Bhowmik, P. Saha, and A. K. Ghosh, "An approach for automatic pain detection through facial expression," *Procedia Computer Science*, vol. 84, pp. 99 – 106, 2016. Proceeding of the Seventh International Conference on Intelligent Human Computer Interaction (IHCI 2015).

[68] D. L. Martinez, O. Rudovic, D. Doughty, J. A. Subramony, and R. Picard, "Automatic detection of nociceptive stimuli and pain intensity from facial expressions," *The Journal of Pain*, vol. 18, no. 4, p. S59, 2017.

[69] N. Rathee and D. Ganotra, "Multiview distance metric learning on facial feature descriptors for automatic pain intensity detection," *Computer Vision and Image Understanding*, vol. 147, pp. 77 – 86, 2016. Spontaneous Facial Behaviour Analysis.

[70] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Pain detection through shape and appearance features," in *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, July 2013.

[71] N. Rathee and D. Ganotra, "A novel approach for pain intensity detection based on facial feature deformations," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 247 – 254, 2015.

[72] D. Liu, F. Peng, A. Shea, and R. Picard, "Deepfacelift: Interpretable personalized models for automatic estimation of self-reported pain," *CoRR*, vol. abs/1708.04670, 2017.

[73] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE transactions on cybernetics*, 2017.

[74] X. Zhang, L. Yin, and J. F. Cohn, "Three dimensional binary edge feature representation for pain expression analysis," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1, pp. 1–7, IEEE, 2015.

[75] A. Marquand, M. Howard, M. Brammer, C. Chu, S. Coen, and J. Mourão-Miranda, "Quantitative prediction of subjective pain intensity from whole-brain fmri data using gaussian processes," *NeuroImage*, vol. 49, no. 3, pp. 2178 – 2189, 2010.

[76] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack, "Svm classification of neonatal facial images of pain," in *International Workshop on Fuzzy Logic and Applications*, pp. 121–128, Springer, 2005.

[77] D. Harrison, M. Sampson, J. Reszel, K. Abdulla, N. Barrowman, J. Cumber, A. Fuller, C. Li, S. Nicholls, and C. M. Pound, "Too many crying babies: a systematic review of pain management practices during immunizations on youtube," *BMC Pediatrics*, vol. 14, p. 134, May 2014.

[78] H. Oster, "Baby facs: Facial action coding system for infants and young children," *Unpublished monograph and coding manual. New York University*, 2006.

[79] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[80] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[81] D. Ververidis and C. Kotropoulos, "Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition," *signal processing*, vol. 88, no. 12, pp. 2956–2970, 2008.

[82] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack, "Machine recognition and representation of neonatal facial displays of acute pain," *Artificial Intelligence in Medicine*, vol. 36, no. 3, pp. 211 – 222, 2006.

[83] S. Brahnam, L. Nanni, and R. Sexton, *Introduction to Neonatal Facial Pain Detection Using Common and Advanced Face Classification Techniques*, pp. 225–253. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[84] B. Gholami, W. M. Haddad, and A. R. Tannenbaum, "Relevance vector machine learning for neonate pain intensity assessment using digital imaging," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 1457–1466, June 2010.

[85] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[86] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[87] S. Liao and A. C. Chung, "Face recognition by using elongated local binary patterns with average maximum distance gradient magnitude," in *Asian conference on computer vision*, pp. 672–679, Springer, 2007.

[88] L. Nanni, S. Brahnam, and A. Lumini, "A local approach based on a local binary patterns variant texture descriptor for classifying pain states," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7888–7894, 2010.

[89] M. N. Mansor and M. N. Rejab, "A computational model of the infant pain impressions with gaussian and nearest mean classifier," in *Control System, Computing and Engineering (ICCSCE), 2013 IEEE International Conference on*, pp. 249–253, IEEE, 2013.

[90] L. Celona and L. Manoni, "Neonatal facial pain assessment combining hand-crafted and deep features," in *International Conference on Image Analysis and Processing*, pp. 197–204, Springer, 2017.

[91] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[92] G. J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," in *European conference on computer vision*, pp. 581–595, Springer, 1998.

[93] Y. Cheon and D. Kim, "Natural facial expression recognition using differential-aam and manifold learning," *Pattern Recognition*, vol. 42, no. 7, pp. 1340–1350, 2009.

[94] R. Beichel, H. Bischof, F. Leberl, and M. Sonka, "Robust active appearance models and their application to medical image analysis," *IEEE transactions on medical imaging*, vol. 24, no. 9, pp. 1151–1169, 2005.

[95] E. Fotiadou, S. Zinger, W. T. a Ten, and S. B. Oetomo, "Video-based facial discomfort analysis for infants," in *Visual Information Processing and Communication*, vol. 9029, pp. 9029 – 9029 – 14, International Society for Optics and Photonics, 2014.

[96] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *Face and Gesture 2011*, pp. 298–305, March 2011.

[97] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pp. 149–149, IEEE, 2006.

[98] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully automatic facial action recognition in spontaneous behavior," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pp. 223–230, IEEE, 2006.

[99] R. R. Vempada, B. S. A. Kumar, and K. S. Rao, "Characterization of infant cries using spectral and prosodic features," in *2012 National Conference on Communications (NCC)*, pp. 1–5, Feb 2012.

[100] G. Várallyay, "Future prospects of the application of the infant cry in the medicine," *Periodica Polytechnica Electrical Engineering (Archives)*, vol. 50, no. 1-2, pp. 47–62, 2006.

[101] C. Z. Boukydis and B. M. Lester, *Infant crying: Theoretical and research perspectives*. Springer Science & Business Media, 2012.

[102] D. Lederman, "Estimation of infants' cry fundamental frequency using a modified SIFT algorithm," *CoRR*, vol. abs/1009.2796, 2010.

[103] G. J. Varallyay, Z. Benyo, A. Illenyi, Z. Farkas, and L. Kovacs, "Acoustic analysis of the infant cry: classical and new methods," in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, pp. 313–316, Sept 2004.

[104] P. Pal, A. N. Iyer, and R. E. Yantorno, "Emotion detection from infant facial expressions and cries," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 2, pp. II–II, May 2006.

[105] B. F. Fuller and Y. Horii, "Spectral energy distribution in four types of infant vocalizations," *Journal of Communication Disorders*, vol. 21, no. 3, pp. 251 – 261, 1988.

[106] C.-Y. Pai, *Automatic Pain Assessment from Infants' Crying Sounds*. PhD thesis, University of South Florida, 2016.

[107] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens, "Identification of pain from infant cry vocalizations using artificial neural networks (anns)," in *Proc.SPIE*, vol. 2492, pp. 2492 – 2492 – 10, 1995.

[108] S. E. Barajas-Montiel and C. A. Reyes-Garcia, *Fuzzy Support Vector Machines for Automatic Infant Cry Recognition*, pp. 876–881. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[109] Y. Abdulaziz and S. M. S. Ahmad, "Infant cry recognition system: A comparison of system performance based on mel frequency and linear prediction cepstral coefficients," in *2010 International Conference on Information Retrieval Knowledge Management (CAMP)*, pp. 260–263, March 2010.

[110] V. Lindh, U. Wiklund, and S. Håkansson, "Heel lancing in term new-born infants: an evaluation of pain by frequency domain analysis of heart rate variability," *Pain*, vol. 80, no. 1, pp. 143 – 148, 1999.

[111] P. M. Faye, J. De Jonckheere, R. Logier, E. Kuissi, M. Jeanne, T. Rakza, and L. Storme, "Newborn infant pain assessment using heart rate variability analysis," *The Clinical journal of pain*, vol. 26, no. 9, pp. 777–782, 2010.

[112] T. Debillon, V. Zupan, N. Ravault, J. Magny, and M. Dehan, "Development and initial validation of the edin scale, a new tool for assessing prolonged pain in preterm infants," *Archives of Disease in Childhood-Fetal and Neonatal Edition*, vol. 85, no. 1, pp. F36–F41, 2001.

[113] L. Scalise, N. Bernacchia, I. Ercoli, and P. Marchionni, "Heart rate measurement in neonatal patients using a webcamera," in *2012 IEEE International Symposium on Medical Measurements and Applications Proceedings*, pp. 1–4, May 2012.

[114] L. A. Aarts, V. Jeanne, J. P. Cleary, C. Lieber, J. S. Nelson, S. B. Oetomo, and W. Verkruysse, "Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit — a pilot study," *Early Human Development*, vol. 89, no. 12, pp. 943 – 948, 2013.

[115] J. H. Klaessens, M. van den Born, A. van der Veen, J. Sikkens-van de Kraats, F. A. van den Dungen, and R. M. Verdaasdonk, "Development of a baby friendly non-contact method for measuring vital signs: first results of clinical measurements in an open incubator at a neonatal intensive care unit," in *Proc. SPIE 8935, Advanced Biomed. and Clin. Diagnostic Syst. XII*, pp. 89351p–1, 2014.

[116] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Opt. Express*, vol. 18, pp. 10762–10774, May 2010.

[117] C. V. Bellieni, "Pain assessment in human fetus and infants," *The AAPS journal*, vol. 14, no. 3, pp. 456–461, 2012.

[118] M. Ranger, C. C. Johnston, C. Limperopoulos, J. E. Rennick, and A. J. du Plessis, "Cerebral near-infrared spectroscopy as a measure of nociceptive evoked activity in critically iii infants," *Pain Research and Management*, vol. 16, no. 5, pp. 331–336, 2011.

[119] M. Bartocci, L. L. Bergqvist, H. Lagercrantz, and K. Anand, "Pain activates cortical areas in the preterm newborn brain," *PAIN*, vol. 122, no. 1, pp. 109 – 117, 2006.

[120] M. Ranger, C. C. Johnston, J. E. Rennick, C. Limperopoulos, T. Heldt, and A. J. Du Plessis, "A multidimensional approach to pain assessment in critically ill infants during a painful procedure," *The Clinical journal of pain*, vol. 29, no. 7, p. 613, 2013.

[121] M. Ranger and C. Gélinas, "Innovating in pain assessment of the critically ill: Exploring cerebral near-infrared spectroscopy as a bedside approach," *Pain Management Nursing*, vol. 15, no. 2, pp. 519 – 529, 2014.

[122] R. Slater, A. Cantarella, S. Gallella, A. Worley, S. Boyd, J. Meek, and M. Fitzgerald, "Cortical pain responses in human infants," *Journal of Neuroscience*, vol. 26, no. 14, pp. 3662–3666, 2006.

[123] R. Slater, A. Cantarella, L. Franck, J. Meek, and M. Fitzgerald, "How well do clinical pain assessment tools reflect pain in infants?," *PLoS medicine*, vol. 5, no. 6, p. e129, 2008.

[124] J. E. Brown, N. Chatterjee, J. Younger, and S. Mackey, "Towards a physiology-based measure of pain: patterns of human brain activity distinguish painful from non-painful thermal stimulation," *PloS one*, vol. 6, no. 9, p. e24124, 2011.

[125] R. Slater, L. Fabrizi, A. Worley, J. Meek, S. Boyd, and M. Fitzgerald, "Premature infants display increased noxious-evoked neuronal activity in the brain compared to healthy age-matched term-born infants," *NeuroImage*, vol. 52, no. 2, pp. 583 – 589, 2010.

[126] G. Zamzami, G. Ruiz, D. Goldgof, R. Kasturi, Y. Sun, and T. Ashmeade, "Pain assessment in infants: Towards spotting pain expression based on infants' facial strain," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 5, pp. 1–5, IEEE, 2015.

[127] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun, "An approach for automated multimodal analysis of infants' pain," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 4148–4153, IEEE, 2016.

[128] S. K. Jaskowski, "The flacc: A behavioral scale for scoring postoperative pain in young children," *AACN Nursing Scan In Critical Care*, vol. 8, no. 1, p. 16, 1998.

[129] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, Y. Sun, and T. Ashmeade, "Automated pain assessment in neonates," in *Scandinavian Conference on Image Analysis*, pp. 350–361, Springer, 2017.

[130] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d videos in real-time," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1, pp. 1–8, IEEE, 2015.

[131] M. Shreve, *Automatic macro-and micro-facial expression spotting and applications*. University of South Florida, 2013.

[132] R. Zhi, G. Zamzmi, D. Goldgof, and Y. Sun, "Infants' pain recognition based on facial expression: Dynamic hybrid descriptions," *The Institute of Electronics, Information and Communication Engineers (IEICE)*, 2018.

[133] R. Paul, S. H. Hawkins, M. B. Schabath, R. J. Gillies, L. O. Hall, and D. B. Goldgof, "Predicting malignant nodules by fusing deep features with classical radiomics features," *Journal of Medical Imaging*, 2018, (Accepted).

[134] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5325–5334, 2015.

[135] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[136] A. Y. Ng, "Feature selection, l 1 vs. l 2 regularization, and rotational invariance," in *Proceedings of the twenty-first international conference on Machine learning*, p. 78, ACM, 2004.

[137] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[138] O. M. Parkhi, A. Vedaldi, A. Zisserman, *et al.*, "Deep face recognition.," in *BMVC*, vol. 1, p. 6, 2015.

[139] A. Rassadin, A. Gruzdev, and A. Savchenko, "Group-level emotion recognition using transfer learning from face identification," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 544–548, 2017.

[140] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[141] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 689–692, ACM, 2015.

[142] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the ninth international workshop on Machine learning*, pp. 249–256, 1992.

[143] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[144] W. K. Wong and H. Zhao, "Supervised optimal locality preserving projection," *Pattern Recognition*, vol. 45, no. 1, pp. 186–197, 2012.

[145] N. Ketkar, "Introduction to keras," in *Deep Learning with Python*, pp. 97–111, Springer, 2017.

[146] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[147] R. V. Grunau and K. D. Craig, "Pain expression in neonates: facial action and cry," *Pain*, vol. 28, no. 3, pp. 395–410, 1987.

[148] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 185 – 198, 2003. Applications of Affective Computing in Human-Computer Interaction.

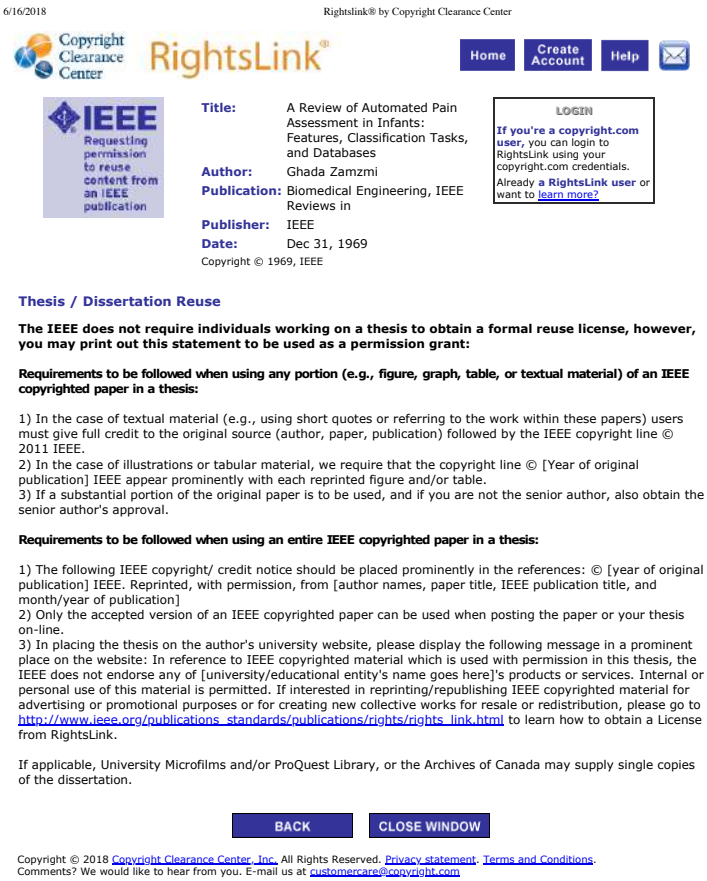
[149] A. Lanatà, A. Armato, G. Valenza, and E. P. Scilingo, "Eye tracking and pupil size variation as response to affective stimuli: A preliminary study," in *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pp. 78–84, 2011.

[150] D. Al-Omar, A. Al-Wabil, and M. Fawzi, "Using pupil size variation during visual emotional stimulation in measuring affective states of non communicative individuals," in *Universal Access in Human-Computer Interaction. User and Context Diversity: 7th International Conference, UAHCI 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part II*, (Berlin, Heidelberg), pp. 253–258, Springer Berlin Heidelberg, 2013.

APPENDIX A

COPYRIGHT PERMISSIONS

The permission below is for the use of material in Chapter 2.



6/16/2018 Rightslink® by Copyright Clearance Center

Copyright Clearance Center RightsLink®

Home | Create Account | Help

IEEE Requesting permission to reuse content from an IEEE publication

Title: A Review of Automated Pain Assessment in Infants: Features, Classification Tasks, and Databases
Author: Ghada Zamzmi
Publication: Biomedical Engineering, IEEE Reviews in
Publisher: IEEE
Date: Dec 31, 1969
Copyright © 1969, IEEE

LOGIN
If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials.
Already **a RightsLink user** or want to learn more?

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK | CLOSE WINDOW

Copyright © 2018 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

https://s100.copyright.com/AppDispatchServlet#formTop 1/1

The permissions below are for the use of materials in Chapter 4 and Chapter 5.

RightsLink®

Home | Account Info | Help

Title: Automated Pain Assessment in Neonates
Author: Ghada Zamzmi, Chih-Yun Pai, Dmitry Goldgof et al
Publication: Springer eBook
Publisher: Springer Nature
Date: Jan 1, 2017

Logged in as:
Ghada Zamzmi
University of South Florida
LOGOUT

Order Completed

Thank you for your order.

This Agreement between University of South Florida -- Ghada Zamzmi ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

printable details

License Number	4370790619777
License date	Jun 16, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Automated Pain Assessment in Neonates
Licensed Content Author	Ghada Zamzmi, Chih-Yun Pai, Dmitry Goldgof et al
Licensed Content Date	Jan 1, 2017
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	yes
Title	Dissertation: A Multimodal Neonatal Pain Assessment Using Computer Vision
Instructor name	Rangachar Kasturi and Dmitry Goldgof
Institution name	University of South Florida
Expected presentation date	Aug 2018
Requestor Location	University of South Florida 8516 Island Breeze Lane