A Review of Automated Pain Assessment in Infants: Features, Classification Tasks, and Databases

Ghada Zamzmi, *Member, IEEE,* Ruicong Zhi, *Member, IEEE,* Rangachar Kasturi, *Fellow, IEEE,* Dmitry Goldgof, *Fellow, IEEE,* Terri Ashmeade, *MD,* and Yu Sun, *Member, IEEE*

***Abstract*—Bedside caregivers assess infants' pain at constant intervals by observing specific behavioral and physiological signs of pain. This standard has two main limitations. The first limitation is the intermittent assessment of pain, which might lead to missing pain when the infants are left unattended. Second, it is inconsistent since it depends on the observer's subjective judgment and differs between observers. The intermittent and inconsistent assessment can induce poor treatment and, therefore, cause serious long-term consequences. To mitigate these limitations, the current standard can be augmented by an automated system that monitors infants continuously and provides quantitative and consistent assessment of pain. Several automated methods have been introduced to assess infants' pain automatically based on analysis of behavioral or physiological pain indicators. This paper comprehensively reviews the automated approaches (i.e., approaches to feature extraction) for analyzing infants' pain and the current efforts in automatic pain recognition. In addition, it reviews the databases available to the research community and discusses the current limitations of the automated pain assessment.**

***Index Terms*—Neonatal pain assessment, automated pain recognition, pain databases, facial expression, crying sound, physiological indicators.**

I. Introduction

PHYSICAL pain can be defined as the negatively-valenced experience associated with actual or potential tissue damage [1]. Pain in neonates can be categorized into two main types: acute procedural pain and acute prolonged pain [2]. Acute procedural pain is often caused by a short painful stimulus (e.g., immunization) and it ends as soon as the cause of pain (i.e., stimulus) is removed. The acute prolonged pain (a.k.a., postoperative) is triggered by a clear stimulus (e.g., surgical procedure) and has a clearly defined beginning and expected end point; the intensity of this type of pain decreases as a function of time since the painful stimulus occurred. Infants may experience different types of pain simultaneously.

The accurate assessment of pain is vital because it helps caregivers understand the severity of the patient's situation and develop appropriate treatments. The most well-known pain assessment method is the patient's self-evaluation. Another common pain assessment method is the Visual Analog Scale (VAS) that has symbols or numbers to denote different levels of pain. Although these methods are the gold standards for clinical assessment, they are not applicable for infants.

The current standard for assessing pain in this vulnerable population depends on the caregivers' observation of specific behavioral (e.g., facial expression) and physiological (e.g., vital signs) pain indicators. Table I summarizes the most common pediatric pain scales for different types of pain. As stated in [4], most of the existing pain scales are designed for procedural pain and a few are designed for prolonged pain. The interested reader is referred to [2], [3], [4] for more information about the validity and shortcomings of different neonatal pain scales.

Assessing infants' pain manually using the common pediatric scales has three limitations. First, caregivers assess pain at different time intervals and are not able to provide continuous assessment of pain. Continuous monitoring is important because infants might experience pain when they are left unattended. This is especially true for postoperative pain since it requires continuous intensive care and prompts pain detection and intervention. Second, caregivers' assessment of pain is highly biased and is affected by several idiosyncratic factors, such as the observer's cognitive bias, identity [9], [10], background and culture [9], [11], [12], and gender [13], that may lead to inconsistent assessment and treatment of pain. Third, the current practice for assessing infants' pain is time-consuming and requires a large number of trained and professional labors, which makes it infeasible in low-income countries where the medical professionals and resources are scarce.

The intermittent and inconsistent assessment of pain might lead to misdiagnosis and over/under treatment. Different pediatric studies [14], [15], [16], [17] reported that the inadequate pain treatment is associated with an increase in the avoidance behaviors and social hypervigilance and it can cause long-term changes in the brain structure (e.g., cause alterations in the cerebral white matter and subcortical grey matter). These alterations can lead to a variety of behavioral, developmental, and learning disabilities [17]. Consequently, developing automated systems that provide continuous and more consistent pain assessment is important.

In the past several years, there has been an increasing interest in the use of machine-learning methods for understanding human behavioral responses to pain based on analysis of facial expressions ([18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48],

The authors, except Ruicong Zhi, are with Computer Science and Engineering Department (Ghada Zamzmi, Rangachar Kasturi, Dmitry Goldgof, and Yu Sun) and Pediatrics Department (Terri Ashmeade), University of South Florida, Tampa, FL, 33620.
Ruicong Zhi is with the School of Computer and Communication Engineering, University of Science and Technology Beijing; Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, P.R.China (E-mail:{ghadh,r1k,goldgof,tashmead,yusun}@mail.usf.edu; zhirc@ustb.edu.cn)

TABLE I: Examples of Common Pediatric Pain Scales

Pain Scale	Pain Type	Age Range	Behavioral Measures	Physiological Measures	Psychometric Properties [1]
[5] Neonatal Infant Pain Scale (NIPS)	Procedural	28-38 gestations weeks	Facial expression, crying, arms/legs, movement, and arousal state	Breathing patterns	Inter-rater reliability: (r=0.92-0.97) Internal Consistency: (Cronbach's α= 0.87-0.95) Content validity Concurrent validity: (r=0.53-0.83)
[6] Neonatal Facial Coding System (NFCS)	Procedural	Preterm $\geq$ 25 gestations weeks to term infants	Brow bulge, Eye squeeze Nasolabial furrow, Open lips, Horizontal mouth, Vertical mouth, Lips pursed, Taut tongue, Chin quiver, Tongue protrusion	NA	Inter-rater and Intra-rater reliability $\geq$ 0.85 Internal Consistency: (Cronbach's α = 0.87-0.95) Content and face validity Construct validity
[7] Neonatal Pain, Agitation, and Sedation Scale (N-PASS)	Postoperative	23-40 gestations weeks	Facial expression, behavior movements, crying/irritability, and extremities tone	Heart rate, respiratory rate, blood pressure, and oxygen saturation	Inter-rater reliability: (r=0.85-0.95) Intra-rater reliability: (r=0.87) Internal consistency: (Cronbach's α = 0.84-0.89) Construct validity: (P $\leq$ .0001)
[8] Crying, Requires O2, Increased VS, Expression, and Sleepless (CRIES)	Postoperative	32 – 60 gestations weeks	Facial expression, crying, and, sleeping state	Requires increased oxygen and VS	Inter-rater reliability: (r=0.98) Construct and content validity

[49], [50], [51], [52], [53], [54], [55], [56]), crying sound ([42], [57], [58], [59], [60], [61], [62]), and body movement ([63], [64]). Also, studies have shown that automated systems can be used to detect emotions from physiological responses such as pupil dilation ([65], [66], [67], [68]), galvanic skin response (GSR) ([37], [35], [69], [70]), changes in heart rate ([70], [71], [72], [69]), and cerebral hemodynamic changes ([73], [74]). A short review of the current efforts in analyzing pain emotion automatically and a discussion of challenges is presented in [75].

In this review, we extensively and specifically explore the current efforts for assessing infants' pain automatically. The main contributions of this paper can be summarized as follows:

- We present a structured review of the current methods for extracting pain-relevant features from infants' data (Section II). We divided these methods into three main categories, behavioral-based, physiological-based, and multimodal-based. These categories were divided further, based on the utilized pain indicator, into facial expression, body movement, crying sound, vital signs, and cerebral hemodynamic. Then, each of these categories was divided further as illustrated in Figure 1.
- We propose to categorize the existing pain recognition works into pain detection and pain intensity estimation. 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). Description and a discussion of limitations for each classification task is presented in Section III.
- We review the pain databases that are available for research use (Sections IV), discuss the current limitations of automated pain assessment, and suggest directions for future research (Section V).

Before we proceed, we would like to note that this review does not discuss preprocessing operations (e.g., image or signal enhancement, noise reduction, region of interest detection, facial landmark detection, etc.) since they are beyond the paper's scope. We refer the interested reader to [76], [77] for a review of image enhancement methods, to [78] for signal denoising methods, to [79], [80] for region of interest detection, and to [81], [82] for facial landmark detection. In addition, we note that understanding this paper requires basic knowledge of machine-learning concepts such as feature (i.e., a measurable property of an object), feature vector (i.e., n-dimensional vector of numerical features), classifier's accuracy, and other performance evaluation techniques. A simple, yet comprehensive, explanation of these concepts can be found in [83], [84].

Organization: Section II presents the current methods that analyze pain automatically to extract pain relevant features. Section III discusses the current state-of-art for pain recog-

[1]Properties to define instruments' reliability (i.e., consistency) and validity (i.e., accuracy).

nition. Section IV provides summary of pain databases that are available for research use. We list several challenges and discuss future directions of pain assessment in Section V. Finally, we conclude in Section VI.

II. Pain Analysis

The automated analysis of infants' pain is an emerging topic in artificial intelligence due to the increasing demands for continuous and consistent monitoring of pain in clinical environments and homecare. Numerous methods have been introduced to automatically detect infants' pain based on analysis of behavioral or physiological pain indicators or a combination of both. We grouped these methods into three main categories, namely behavioral measures based pain analysis, physiological measures based pain analysis, and multimodal based pain analysis, and divided these categories further as illustrated in Figure 1.

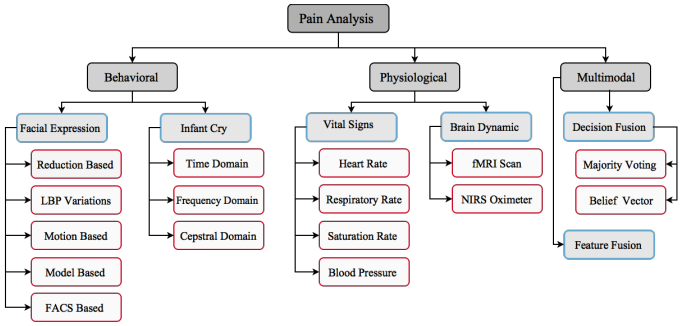

Fig. 1: Tree Diagram of Infants' Pain Analysis Methods.

A. Behavioral Measures Based Pain Analysis

Behavioral measures based pain analysis can be defined as the task of automatically extracting pain-relevant features from behavioral pain indicators such as facial expression and crying sound. In this section, we discuss the existing methods that analyze facial expression, crying, or body movement to extract useful features for classification.

1) Facial Expression: Facial expression is one of the most common and specific indicators 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 infants 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 [6].

Automatic recognition of pain expression consists of three main stages: (1) face detection and registration; (2) feature extraction; and (3) expression recognition (see Section III). Face detection is a mature area of research and, therefore, will not be discussed further. Several methods have been proposed to extract pain-relevant features from images. We broadly divided these methods based on their underlying algorithms into five groups: Feature Reduction Based methods, Local Binary Pattern Variation based methods, Motion-based methods, Model-based methods, and Facial Action Coding System [FACS] (see Figure 1). The first and second categories focus on analyzing static images and they both fall under texture-based methods. The last three categories focus on the temporal analysis of facial expression in videos. For each category, we discuss the underlying algorithms and the exiting works that utilize them. Table II presents a summary of the works we discussed in this section.

*a) **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 [85].

Another well-known method for feature selection is SFFS [86]. Sequential feature selection 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 or removing a single feature until there is no improvement in the classifier performance. SFFS is an extension of Sequential Forward Selection (SFS), which 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 (i.e., allow to dynamically increase and decrease the features until the best subset is reached).

One of the first studies in machine recognition of pain is presented by Brahnam et al. in [48], [49]. A feature reduction based approach was proposed [49] and applied on the Classification of Pain Expressions (COPE) dataset. This dataset consists of 204 color images captured for 26 Caucasian infants, half girls, using Nikon D100 digital camera. The infants' age ranges from eighteen hours to three 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 100×120 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, specifically

TABLE II: Summary of Automated Methods for Analyzing Infants' Pain Expression

Ref. & Year	Database	Category	Extraction Method	Classification	Results
[49]: 2006	COPE Database **Subjects:** 26, half girls **Race:** Caucasian **Age range:** 18 hours to 3 days **Stimuli:** Pain stimulus and 3 other stimuli: air, friction, and rest/cry **Data:** 204 static images	Feature Reduction Based	Column stacking image's intensities and dimensionality reduction (PCA)	PCA/LDA with L1 and SVM: Pain/nopain, pain/rest, pain/cry pain/air-puff, and pain/friction **Testing Protocol:** 10-fold cross-validation	SVM avg. accuracy: Pain/nopain (88%) Pain/rest (95%) Pain/cry (80%) Pain/air-puff (83%) Pain/friction (93%)
[50]: 2007	COPE Database	Feature Reduction Based	Column stacking image's intensities and dimensionality reduction	NNSOA, PCA/LDA and SVM: Pain (60 images) vs no pain (144 images) **Testing Protocol:** Leave-one-subject-out cross-validation	Average accuracy: NNSOA (90.20%) SVM (82.35%) PCA w/ L1 (80.35%) LDA w/L1 (76.96%)
[51]: 2010	COPE Database	Feature Reduction Based	Column stacking image's intensities	RVM: Pain/nopain Pain Intensity Estimation **Testing Protocol:** Leave-one-image-out cross-validation	Weighted Kappa Coeff.: 0.47 (expert/RVM) 0.46 (non-expert/RVM)
[52]: 2010	COPE Database	LBP Variation Based	LBP, LTP, ELTP, and ELBP descriptors	SFFS feature selection and SVMs **Testing Protocol:** Leave-one-out cross-validation (SFFS) Train/Test Split	Highest (0.93) area under the curve of ROC (ELTP)
[53]: 2015	**Subjects:** 10, half girls **Race:** Caucasian **Age range:** 32 to 41 gestations weeks **Stimuli:** Pain stimulus (i.e., heel lancing) and normal state **Data:** 10 videos and NIPS scores	Motion Based (Optical Flow)	Strain magnitude estimated from flow vectors	KNN and SVM: Pain vs no pain expression **Testing Protocol:** 10-fold cross-validation	Highest overall accuracy: KNN, (96%)
[54]: 2014	**Subjects:** 10 **Race:** N/A **Age range:** N/A **Stimuli:** Heel puncture, diaper change, hunger, and resting **Data:** 15 videos, ranges from few seconds to minutes	Model Based (AAM)	AAM-based features SPTS, SAPP, and CAPP	SVM: Discomfort vs comfort **Testing Protocol:** Leave-one-subject-out cross-validation	AUC of ROC: 0.98
[55]: 2015	**Subjects:** 50 children, 35% boys **Race:** 35 Hispanic, 9 non-Hispanic white, 5 Asian, and 1 Native American **Age range:** 5 to 18 years **Stimuli:** appendectomy (ongoing) and pressing surgical site (transient) **Data:** videos, self-report, and by proxy rating by a nurse and parent	FACS Based (Optical Flow)	Strain magnitude estimated from flow vectors	KNN and SVM: Pain expression vs no pain expression **Testing Protocol:** 10-fold cross-validation	Highest overall accuracy: KNN, (96%)

TABLE III: Summary of Automated Methods for Analyzing Infant Cry

Ref. & Year	Database	Category	Extraction Method	Classification	Results
[61]: 2012	**Subjects:** 120 infants **Race:** N/A **Age range:** 12-40 weeks **Stimuli:** N/A **Data:** 120 samples; 30 Pain, 60 hunger, and 30 wet-diaper	Time Domain Analysis	Short-time energy (STE) and pause duration	SVM **Testing protocol:** Splitting samples into train and test	Accuracy per class: Pain, 83.33% Hunger, 27.78% Wet-diaper, 61.11% Avg. accuracy: 57.41%
[42]: 2006	N/A	Frequency Domain Analysis	F0 Fundamental frequency and 3 first formants	K-means: Pain, hunger, fear, sadness, and anger	Classification accuracy: 91%
[62]: 1988	**Subjects:** 41 infants **Race:** Caucasian **Age range:** 2 to 6 months old **Stimuli:** Immunization for pain, feeding time for hunger, naptime for fussy, and fondling for cooing **Data:** 109 samples; 16 hunger, 23 cooing, 42 pain, and 28 fussy	Frequency Domain Analysis	Mean value of spectral energy	Statistical analysis (ANOVA)	Unique spectral characteristics of pain-induced cry
[57]: 2016	**Subjects:** 27 infants **Race:** Caucasian, Hispanic, African american, and Asian **Avg. age:** 36 gestation weeks **Stimuli:** Immunization and and heel lancing **Data:** 34 samples; NIPS score	Frequency Domain Analysis	LPC and statistics (e.g., mean and std)	kNN: Whimper cry Vigorous cry **Testing Protocol:** 10-fold cross validation	Average accuracy: 76.47%
[60]: 1995	**Subjects:** 16 **Race:** N/A **Age range:** 2 to 6 months old **Stimuli:** immunization (pain), jack-in-the-box (fear), and head restraint (anger). **Data:** 230 cry samples	Cepstral Domain Analysis	10 MFCC coeff.	Neural Network: Pain cry vs no-pain cry (fear & anger) **Testing Protocol:** 10-fold cross validation	Classification accuracy: Pain (92.0%) No-pain (75.7%)
[106]: 2006	**Subjects, race, age, and stimuli:** N/A **Data:** 1627 samples; 209 pain, 759 hunger, and 659 others (Data collected and labelled by doctors)	Cepstral Domain Analysis	16 MFCC coeff. Dimensionality reduction (PCA)	FSVM: Pain cry Hunger cry No-pain-no-hunger cry **Testing protocol:** 10-fold cross-validation	Average accuracy: 97.83%
[107]: 2010	**Subjects and race:** N/A **Age range:** newborns to 1 year **Stimuli:** Immunization (pain) and spontaneous emotions **Data:** 180 sample; 150 pain and 30 no-pain	Cepstral Domain Analysis	12 MFCC coeff. 16 LPCC coeff.	Neural Network: Pain/no-pain **Testing Protocol:** Splitting samples to train and test	Classification accuracy: MFCC: 76.2% LPCC: 68.5%
[61]: 2012	Database in 1^{st} row	Cepstral Domain Analysis	13 MFCC, Δ MFCC, and Δ Δ MFCC	SVM: Pain, hunger, and and wet-diaper **Testing Protocol:** Splitting samples to train and test	Accuracy per class: Pain(30.56%) Hunger (66.67%) wetdiaper (86.11%)

TABLE IV: Summary of Publications for Pain Analysis using Physiological Measures

Ref. & Year	Measures	Database	Extracted Data	Analysis Method	Results
[119]: 1999	Vital Signs	**Subjects:** 25 infants **Age range:** 72 - 96 hours **Stimuli:** baseline, sham heel prick, sharp heel prick, and heel squeezing	HR_{mean}, the power in low frequency (P_{LF}), and high-frequency (P_{HF}) and total heart rate variability (P_{tot})	Multivariate statistics	Increase in HR_{mean}, P_{tot}, and P_{LF}, between baseline and sharp prick
[71]: 2010	Vital Signs	**Subjects:** 28 infants **Age:** $>$ 34 gestational weeks **Stimuli:** baseline and a major surgery (postoperative)	Heart Rate Variability Index (HRVI)	Linear regression analysis	Sensitivity (90%) Specificity (75%) Area under ROC (0.81)
[120]: 2006	Cerebral Hemo-dynamics (NIRS)	**Subjects:** 40 infants, half male **Age:** $\geq$ 26 gestational weeks **Stimuli:** Baseline, tactile, and venipuncture pain stimulus	Difference of concentration of oxygenated $[HbO_2]$ and de-oxygenated $[HbH]$ and total ($HbH + HbO_2$) hemo-globin from baseline	Student t-test ANOVA NewmanKeuls post hoc test	$[HbO_2]$ increases in both hemispheres; more pronounced increase in male
[121]: 2006	Cerebral hemo-dynamics (NIRS)	**Subjects:** 18 infants **Age:** 25 - 45 postmentsural weeks **Stimuli:** Baseline and heel lancing	Vital signs data and mean of $[HbO_2]$, $[HbH]$, and $HB_{total} = HbH + HbO_2$	Statistical t-test	Significant increase in $[HB_{total}]$; more pronounced increase in awake infants
[122]: 2013	Cerebral hemo-dynamics (NIRS)	**Subjects:** 40 infants **Age:** $<$ 12 months **Stimuli:** Baseline (T_0), tactile (T_1), and painful (T_2) stimuli	$[HbH]_{mean}$, $[HbO_2]_{mean}$, and $[HR]_{mean}$	Univariate linear regression	ΔHbH differed significantly between T_0 and T_2

s/he was in a quiet, awake, and stable condition. The tactile stimulus period (P_1) was recorded after the disinfecting of an 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 time 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 the values in P_0 from their values in P_1 and P_2 periods. Also, the average of these measurements were computed and used to perform Student's t-test, ANOVA, and NewmanKeuls 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 [121] to measure the brain hemodynamic activity for eighteen infants in the NICU at 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 pre-stimulus. 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 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, it has been found that the response in the contralateral somatosensory cortex for awake infants increases with age.

Extensions of this work are presented in [131] to study the relation between NIRS data and behavioral indicators of pain and in [132] to investigate the impact of age and frequency of painful procedures on the brain neuronal responses.

For chronic pain, Ranger et al. [122] presented a NIRS-based method to assess infants' chronic pain based on analysis of hemodynamic activity in brain regions. NIRS data (i.e., HbO_2 and HbH) for forty 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 ΔHbH during pain (i.e., the difference of ΔHbH measurement between the baseline (T_0) and pain (T_2) was significant).

In a different population (i.e., adults), fMRI analysis was performed during baseline and different events of thermal

ical signals (i.e., Electrocardiogram [ECG], Galvanic Skin Response [GSR], and Electromyography [EMG]) collected for 90 subjects undergoing heat stimulus[2]. To extract pain-relevant features from videos, facial landmark points were detected in each frame and head pose was estimated. These points were used to compute several geometric features (distances) and gradient-based features. Then, the frames were grouped into time windows of 5.5 seconds to form a temporal descriptor or vector for classification. For the biomedical signals, all the signals were divided into windows of 5.5 seconds and filtered to remove noise using a Butterworth bandpass filter. To extract features for classification, different statistics (e.g., mean and standard deviation) were computed from the filtered biomedical signals. In the final step, the features extracted from both video and biomedical signals were fused together to form a single high-dimensional vector, which was used to train random forest model. The proposed method achieved up to 80.6% mean accuracy for a fusion of video and biomedical signals.

To summarize, we discuss above two levels, namely decision fusion and feature fusion, for combining different pain indicators. Decision-level fusion assumes that the modalities are independent and ignores the correlation between them. 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 (i.e., z-scores) for scaling, PCA for reduction, and interpolation for the missing data.

III. Pain Recognition

We divided 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.

A. 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., facial expression [30], [22], [24], [27], [34], [29], [30], [39], [49], [52], [53], [54], [50], [87], cry [61], [57], [106], and body movement [117], [118]). Other classifiers that are used for pain detection are Neural Network [18], [60], [87], k-nearest neighbors [57], [53], and k-means [42]. Such classifiers achieved varying levels of performance in detecting the pain label.

Pain detection provides the pain label, but does not provide the intensity or level of the detected pain. For pain assessment application, detecting the pain without its intensity may not be enough due to three main 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 painful stimuli. 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. For these reasons, we believe estimating the intensity of pain is important and can lead to better understanding and intervention.

[2]Description of this database (BioVid) is provided in Section IV.A

B. Pain Intensity Estimation

Estimating the intensity of the detected 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. [51] presented a method (see Table II, 3rd row) 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 experts and non-expert observers. The agreement between RVM and human observers, measured using kappa coefficient, was 0.48 for experts and 0.52 for non-experts.

Hammal et al. [19] 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) that ranges from -1 to 1 was used. 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. introduced a method [69] to estimate four levels of pain using SVM. Facial expression and biopotentials signals were recorded under four levels of pain (T1 to T4) as described in Section IV.A (BioVid Heat Pain Database). Then, the recorded signals were analyzed to extract complex mathematical features. These features were used to build SVM classifiers 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 T1, 80.00% (sensitivity) and 82.59% (specificity) for baseline vs T2, 84.71% (sensitivity) and 85.18% (specificity) for baseline vs T3, and 92.24%, (sensitivity) and 89.65% (specificity) for baseline vs T4.

IV. 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 term of its modalities/dimensions, is critical to develop reliable multimodal pain assessment system that can still assess pain in case of missing data. Finally, databases with relatively small number of subjects are not sufficient to evaluate the system

performance and draw conclusions. Therefore, collecting high-quality, multimodal, and large databases is necessary for developing robust pain assessment systems.

Most of the existing pain databases are not publicly available, due to legal/ethical reasons, for research use. This section provides brief descriptions of the publicly available pain databases for adults and infants.

A. Adult

UNBC-McMaster Shoulder Pain Expression Archive [23] is one of the first databases that addressed the need for adequately annotated and publicly available database 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). In addition to the videos, the database has self-report and observer ratings for each sequence.

Instead of recording a single moadilty/indicator, Walter et al. introduced [38] an advanced and multimodal database, known as the *BioVid Heat Pain Database*. 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 (T1 to T4). 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 can 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 [38].

B. Infant

COPE/iCOPE, collected by Brahnam et al. [48], is the first pain expression database that is designed specifically to assess infants' pain automatically. 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 photography 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 collaborators are working on collecting a new and 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 in [136], [135], incorporating different pain indicators is important to ensure proper and reliable assessment of pain.

Another publicly available neonatal pain database is described in [139]. The database consists of *YouTube videos* recorded, by parents or a guardian, for infants receiving immunization injections; 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 (Face, Legs, Activity, Cry, Consolability) [140] pain scale. The main limitation of this database is the low-quality of the recorded videos which leads to exclude many videos from annotations.

As far as we are aware, COPE and YouTube databases are the only available neonatal databases for research in pain detection. Therefore, collecting high-quality, multimodal, and relatively large pain databases is needed to advance the automated assessment of neonatal pain.

V. Limitation and Future Directions

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

- As discussed above, there are very few accessible databases for research in neonatal pain. At the time of writing this paper, 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.
- Existing methods for automatic pain assessment focus mainly 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 as reported in [6]. 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 tracking) is more challenging in infants because they are uncooperative subjects recorded in an unconstrained environment (i.e., NICU).
- Most of the existing approaches assess pain based on analysis of a single modality (e.g., facial expression). Studies [136], [135] have shown that pain causes be-