KnowPain: Automated System For Detecting Pain In Neonates From Videos

Rajkumar Theagarajan, Bir Bhanu, Danilyn Angeles* and Federico Pala

Center for Research in Intelligent Systems, University of California, Riverside, CA 92521, USA
*Neonatal Intensive Care Unit (NICU), Loma Linda University, Loma Linda, CA 92354, USA

Email: rtthea001@ucr.edu, bhanu@cris.ucr.edu, dangeles@llu.edu, fedpala@ucr.edu

Abstract—Premature neonates are subjected to clinically required but painful procedures throughout their hospitalization. Since neonates are non-verbal, pain scoring tools are used to measure their pain responses. Although a number of pain instruments have been developed to assist health professionals, these tools are subjective and may underestimate the pain response of neonates. This could lead to the pain being misread resulting in mis-diagnosis and under/over treatment. In this paper, a deep learning based approach is used to detect pain in videos of premature neonates during painful clinical procedures. A Conditional Generative Adversarial Network (CGAN) is used to continuously learn the representation and classify painful facial expressions in neonates from real and synthetic data. A Long Short-Term Memory (LSTM) is used for modeling the temporal changes in facial expression to further improve the classification. Furthermore, the proposed approach is able to implicitly learn the intensity of pain as a probability score directly from the facial expressions without any manual annotation. Experimental results show that this approach achieves an accuracy of 95.34% on the iCOPE Classification Of Pain Expressions (video) dataset, 88.27% on the Loma Linda Infant Pain Expressions (video) dataset and 94.12% on the Infant Classification Of Pain Expressions (video) dataset outperforming state-of-the-art approaches.

Keywords—Neonatal pain assessment; Facial expressions; Background suppression; Convolutional Neural Networks

I. INTRODUCTION

The automatic detection of pain is a subject of high interest in the health domain where health professionals attempt to understand a patient’s medical condition and develop suitable treatment plans. One of the most challenging problems in neonatology is the continuous quantitative and objective assessment of pain in premature neonates. This is because neonates cannot verbally express their pain. Therefore, a continuous monitoring by a care provider is required. However, it is very difficult to have care providers continuously monitor neonates since this requires highly skilled professionals, who are not readily available. Therefore, it is vital that automatic tools be developed that accurately detect a neonate’s pain from a host of distress signals as captured by monitoring videos cameras.

Facial expressions are the most distinguishing features as pain indicators in neonates. In fact, most of the manual pain assessment tools such as COMFORT [1], CRIES [2] rely on facial expressions of a neonate. Even though the facial characteristics of neonates’ expressions of pain have been studied extensively, the primary problem is that these tools rely on the observations of health professionals, who can be skewed by bias.

To address this problem we developed a fully automated system to detect pain in video recordings of neonates taken during various clinical settings. Our approach uses a Conditional Generative Adversarial Network (CGAN) that learns the representation and classifies different facial expressions of neonates from real and synthetic images. To date, there has been a shortage of dataset for classifying facial expressions of neonates and there have been only two publicly available datasets [3] and [8]. In order to compensate for the lack of datasets, the CGAN is trained to generate synthetic images of different facial expressions of neonates.

Facial expressions are very dynamic and not instantaneous, making it very important to model the changes in facial expressions over time, especially in very critical tasks such as predicting the pain in neonates. To achieve this, we trained a Long Short-Term Memory (LSTM) that takes as input a series of feature vectors (extracted from the discriminator of the CGAN) over successive frames and predicts if the neonate is in pain. Furthermore, we show that our approach is able to implicitly estimate the intensity of pain of a neonate as a probability score directly from the videos, without manual annotation for the intensity of pain.

In this paper, we are concerned only if the baby is experiencing pain due to clinical procedures and not due to other factors such as hunger, diaper change, etc.

II. RELATED WORKS AND OUR CONTRIBUTIONS

Pain indicators in neonates can be divided into two categories: physiological and behavioral indicators. Physiological indicators utilize signals such as heart rate, blood pressure, vagal tone and palmar sweating to identify pain in neonates. Lindh et al. [4] show that during the process of heel lancing the heart rate of neonates significantly increased over the mean heart rate, thus, indicating pain. Ranger et al. [5] used a multidimensional approach for detecting pain in neonates using a NIRs-based tissue oxygenation monitor to measure the changes in the cerebral tissue oxygenation. They also used the oxygen saturation level, heart rate and
blood pressure to estimate the intensity of pain felt by a neonate during this procedure.

Chang and Li [6] used CNNs to classify the cry of neonates into hunger, pain and sleepiness with an accuracy of 78.5%. The voice signal is converted to a spectrogram using fast fourier transform and classified using a CNN. Barajas-Montiel and Reyes-Garcia [7] used Fuzzy Support Vector Machines to classify the cries of neonates into pain, hunger and no pain with 90% accuracy.

The first work for addressing neonate pain using behavioral indicators related to the facial expression was known as the COPE project [8]. The authors classified facial expression images of neonates as “Pain” and “No pain” using a SVM with an accuracy of 88%. A drawback of this work was the use of static images to classify facial expressions, ignoring the dynamic changes in facial expressions which in turn fails to identify if the pain is an acute or chronic pain.

Zamzmi et al. [9] used a multi-modal approach consisting of facial expression, body movement and physiological signals for detecting pain in neonates. The authors evaluated their approach on a dataset collected at NICU in Tampa General Hospital. The authors used an optical flow based strain estimation method to detect changes on the face as indicators of pain. If the strain is beyond a computed threshold, then the neonate is considered to be in pain. This approach works best only under constrained situations where majority of the face is visible to the camera. Additionally, the threshold for detecting the pain using both the facial expression and body movement is computed on a dataset collected by the authors and hence may not yield correct results when evaluated on a different dataset collected in a different environment with different illumination conditions.

Celona et al. [10] fused features extracted from CNNs along with hand-crafted features to classify facial expressions of neonates into pain and no pain. The authors used a feature set consisting of Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG) and features extracted from pre-trained VGG-face and MBP-CNN. The authors achieved an accuracy of 83.78% on the COPE dataset [8] by fusing the features from VGG-face and MBP-CNN.

Our work is significantly different from the work done by the authors of [8] - [10] and in light of the state-of-the-art, the contributions of this paper are as follows:

- Used Conditional Generative Adversarial Networks to classify and generate synthetic facial expression images of neonates.
- Trained the networks on one dataset and tested on two different datasets to evaluate the generalization of our approach.
- Demonstrated that our approach is able to implicitly estimate the intensity of pain of a neonate as a probability score computed directly from a video.

III. TECHNICAL APPROACH

This section describes the overall framework of our approach. Fig. 1 shows the framework of our approach.

![Diagram](https://via.placeholder.com/150)

Figure 1. Framework of our approach.

First, the face of the neonate is localized using a CNN and filtered using an Isotropic filter. Isotropic filter is used to suppress unwanted background information. Next, the input is passed into a Conditional Generative Adversarial Network, which consists of two CNNs (Generator and Discriminator) trained against each other. Fig. 2 shows the architecture of our Conditional Generative Adversarial Network.

The Generator is a CNN trained to take as input a noise vector Z and generate a realistic image as shown in Fig. 2. The discriminator is a CNN that takes as input either a real or synthetic image generated by the generator. The task of the discriminator is to predict if the given image is real or synthetic and also predict if the image is a painful facial expression or not. After training the discriminator, we extract the learned features from the discriminator and pass it as input to the LSTM. The LSTM takes a sequence of these features and predicts over time if the neonate is in pain. It should be noted that the LSTM is trained only after training the discriminator. The following sub-sections explain in detail the pre-processing of the data and the training of the individual networks.

**A. Pre-processing Videos**

In our approach the pre-processing of the video is done in two steps:

- Detecting and cropping the face of a neonate
- Isotropic filtering of the facial expression

1) Detecting and Cropping the Face of a Neonate: In our approach the face of the neonate is detected using the YOLO V2 CNN architecture [11]. YOLO V2 divides the input frame into a 11x11 grid and predicts B bounding boxes and the associated confidence score for each box in each grid. Formally, the confidence is defined as Pr(object) * IOU, where Pr(object) is the probability of an object present and IOU is the Intersection Over Union between the predicted bounding box and the ground-truth bounding box. YOLO V2 was initially trained on the Pascal VOC dataset [12] and fine-tuned to detect the faces of neonates using 154 manually annotated videos from the iCOPE dataset [3].
2) Isotropic Filtering: After detecting the face of a neonate, we perform isotropic filtering using the INFace toolbox [13]. Isotropic smoothing (IS) works in two steps. First, a set of 40 Gabor wavelet filters are used to extract discriminative features, while in the second step the kernel partial-least squares discrimination technique is used to reduce the dimensionality of the Gabor features and enhance the discriminatory power.

It has been shown that when analyzing facial expressions not all the information on the face is important. Theagarajan et al. [14] showed that information such as skin color do not give any additional information about the facial expression and can be considered as background and suppressed. In our experiments we observed that by performing isotropic filtering as compared to using the original RGB images, we get slightly improved accuracy, recall and precision.

B. Classifying and Generating Facial Expressions of Neonates

Generative Adversarial Networks consists of two CNNs trained against each other [15]. The generator \( G \) takes a random noise vector, \( z \), and generates an image, \( X_{gen} = G(z) \). The discriminator \( D \) takes either a real or a generated image, and outputs a probability distribution \( P(S|X) = D(X) \) over the two image sources. \( D \) is trained to maximize the log-likelihood of assigning the correct source while \( G \) tries to minimize it, thus learning to generate realistic images. The optimization function \( V \) is given by:

\[
\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))]
\]  

However, in an unconditioned generative model, there is no control on the facial expression of the neonate being generated. By conditioning the model on additional information we can generate facial expressions for a given particular class [16]. In our case the condition is the class “Pain” and “No pain”.

In our approach, the input condition \( L \) to the generator is a 1x1 vector which is concatenated to the noise vector \( Z \). \( L = 0 \) conditions the generator to generate an image belonging to the class “No pain” and \( L = 1 \) conditions the generator to generate an image belonging to the class “Pain”. The discriminator is designed such that it has two convolutional layer branches as shown in Fig. 2. In our approach the discriminator has two tasks which are:

- To classify the input image as either a real image or a synthetic image.
- To classify the input image as either “Pain” or “No pain”.

C. Long Short-Term Memory for Sequence Prediction

Long Short Term Memory (LSTM) networks have been shown to be effective in analyzing the dynamic changes in facial expressions [21]. In our approach, the input to the LSTM is a \( M \times N \) feature vector, where \( M \) is the length of the feature vector extracted from the discriminator and \( N \) is the length of the image sequence we want to model. In our approach the dimension of the feature vector is \( M = 2048 \times 1 \) and we chose \( N = 16 \) consecutive frames. The input sequences to the LSTM are created such that, each frame is the last of a sequence once, e.g., if the first sequence is \( S_0 = \{V_0, V_1, ..., V_{N-1}, V_N \} \), then the second sequence is \( S_1 = \{V_1, V_2, ..., V_N, V_{N+1} \} \). Each sequence \( S \) is labeled with one label \( t \), which is the label of the last frame in the sequence. In our case, \( t \) is a binary vector indicating “Pain” or “No pain”. In order to train the LSTM we need to be able to extract robust features from the discriminator which is achieved only after training the CGAN.
IV. EXPERIMENTAL RESULTS

A. Datasets
We evaluated our approach on three datasets namely:
- Infant Classification Of Pain Expressions [8]
- iCOPE Classification Of Pain Expressions [3]
- Loma Linda Infant Pain Expressions

Infant Classification Of Pain Expressions [8] contains 204 static facial expression images of neonates. Among the 204 images, 67 are rest, 18 are cry, 23 are air puff, 36 are friction, and 60 are pain. In our approach, all images are divided into two categories: “No pain” and “Pain”. The set of “No pain” consists of 144 images combining rest, cry, air puff and friction. The set of “Pain” images being a collection of the remaining 60 images.

iCOPE Classification Of Pain Expressions [3] consists of 234 HD videos of neonates taken during clinical trials. Each video has a duration of approximately 20 seconds, resulting in 57,078 “No pain” and 73,964 “Pain” sequences. The ground-truth for these videos were obtained on a frame by frame basis by multiple clinical experts with 25+ years experience in the NICU. Each frame of the video was labeled as either “No pain” or “Pain”.

Loma Linda Infant Pain Expressions dataset was obtained from the NICU facility at Loma Linda University and consists of 18 videos, collected after obtaining IRB approval and consent from the parents. These videos were recorded using a cellphone camera with occlusions due to the clinical procedure. The purpose of collecting videos in this setup was to evaluate the robustness of an approach in an unconstrained low resolution video setting. The duration of the videos range between 8 to 150 seconds, resulting in 9,682 “No pain” and 13,369 “Pain” sequences.

B. Data Partition

We split the 234 videos in the dataset [3] into 154 videos for training, 20 videos for validation and 60 videos for testing. Table I shows the summary of the data partition. The data partition was made such that no video belonging to the same neonate was present in both the training and validation/testing datasets.

<table>
<thead>
<tr>
<th>Data Partition</th>
<th>Number of videos</th>
<th>Number of Pain sequences</th>
<th>Number of No Pain Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>154</td>
<td>38,738</td>
<td>51,352</td>
</tr>
<tr>
<td>Validation</td>
<td>20</td>
<td>6,200</td>
<td>5,300</td>
</tr>
<tr>
<td>Testing</td>
<td>60</td>
<td>12,140</td>
<td>17,111</td>
</tr>
</tbody>
</table>

The 154 training videos were used to train the localization network (YOLO V2), the Conditional Adversarial Generative Networks and the Long Short-Term Memory. It should be noted that the Infant Classification Of Pain Expressions [8] and the Loma Linda Infant Pain Expressions were used exclusively only for evaluating our approach. Furthermore, all of the above three datasets were collected in different environments and conditions, thus training on one dataset and cross-evaluating on two other datasets helps to evaluate the robustness of an approach.

C. Network Hyper-parameters

We performed random hyper-parameter selection using only the validation dataset for the individual networks to obtain the best learning rate, momentum and weight decay. Table II shows the best hyper-parameters for each network. We chose random values for the learning rate, momentum and weight decay within a given range and step size and trained the networks for three epochs. The combination of hyper-parameters that gave us the highest classification accuracy after three epochs was chosen as the best hyper-parameter set for a network. We employ early stopping to prevent the network from over-fitting during training.

<table>
<thead>
<tr>
<th>Network</th>
<th>Learning rate</th>
<th>Momentum</th>
<th>Weight decay</th>
<th>Optimizer</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLO V2</td>
<td>5x10^-4</td>
<td>0.9</td>
<td>1x10^-4</td>
<td>SGD</td>
</tr>
<tr>
<td>Generator</td>
<td>1x10^-3</td>
<td>-</td>
<td>5x10^-4</td>
<td>Adam</td>
</tr>
<tr>
<td>Discriminator</td>
<td>1x10^-5</td>
<td>-</td>
<td>5x10^-4</td>
<td>Adam</td>
</tr>
<tr>
<td>LSTM</td>
<td>3x10^-3</td>
<td>-</td>
<td>1x10^-4</td>
<td>Adam</td>
</tr>
</tbody>
</table>

D. Experiments

In this section we evaluate our approach on the above mentioned three datasets and compare our results with the results obtained from the state-of-the-art approaches [8], [10], [18], [19] and [20]. We implemented the state-of-the-art approaches from scratch. All the image sequences were classified after the pre-processing described in Section III. Table III shows the comparison of our approach with state-of-the-art approaches on the testing dataset of iCOPE Classification Of Pain Expressions Dataset [3] shown in Table I, Loma Linda Infant Pain Expression Dataset and Infant Classification Of Pain Expressions Dataset [8]. Since the state-of-the-art approaches are suitable only for static images, they cannot be used directly for classifying video sequences. In order to maintain a fair comparison, we evaluate these approaches by extracting feature vectors for all images in a given sequence and pass them as input to train a Support Vector Machine (SVM) with linear kernel to predict “Pain” and “No pain”. The final classification is obtained by taking a majority vote over the input sequence.

E. Analysis of Results

Table IV shows the confusion matrices obtained from our approach on the iCOPE Classification Of Pain Expressions Dataset [3], Loma Linda Infant Pain Expression Dataset and Infant Classification Of Pain Expressions Dataset [8].
<table>
<thead>
<tr>
<th>Approach</th>
<th>Classification Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>iCOPE Classification Of Pain Expressions dataset [3]</td>
</tr>
<tr>
<td>Brahnam et al. [8]</td>
<td>71.56%</td>
</tr>
<tr>
<td>Celona et al. [10]</td>
<td>86.41%</td>
</tr>
<tr>
<td>LBP [18]</td>
<td>69.15%</td>
</tr>
<tr>
<td>HOG [19]</td>
<td>67.33%</td>
</tr>
<tr>
<td>VGG Face [20]</td>
<td>84.64%</td>
</tr>
<tr>
<td>Proposed</td>
<td>94.85%</td>
</tr>
</tbody>
</table>

Table IV

<table>
<thead>
<tr>
<th>Ground-truth →</th>
<th>No pain</th>
<th>Pain</th>
<th>Ground-truth →</th>
<th>No pain</th>
<th>Pain</th>
<th>Ground-truth →</th>
<th>No pain</th>
<th>Pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pain</td>
<td>16,232</td>
<td>879</td>
<td>No pain</td>
<td>8,336</td>
<td>1,346</td>
<td>No pain</td>
<td>133</td>
<td>11</td>
</tr>
<tr>
<td>Pain</td>
<td>625</td>
<td>11,515</td>
<td>Pain</td>
<td>1,872</td>
<td>11,497</td>
<td>Pain</td>
<td>1</td>
<td>59</td>
</tr>
</tbody>
</table>

1) iCOPE Classification Of Pain Expressions Dataset [3]: It is observed from Table III and Table IV, that our approach achieved a classification accuracy of 94.85%, precision of 94.86% and recall of 96.29%. We observed that majority of the misclassified sequences occurred during the transition phase (i.e., when a neonate’s facial expression changes from No pain to Pain and vice-versa). The reason for this is that during this phase when the facial expression is changing, the features extracted are a mix of both the classes “Pain” and “No Pain” leading to a boundary misclassification. Since the transition phase is a very short duration of time, any misclassification occurring during this interval can be safely ignored.

2) Loma Linda Infant Pain Expressions Dataset: It is observed from Table III and Table IV, that our approach achieved a classification accuracy of 86.10%, precision of 96.29% and recall of 81.66%. This is lower compared to the performance of our approach on dataset [3]. The reason for this is that the environment in which the both the datasets were captured are very different. The infants in dataset [3] were placed in an open lid cradle and had no devices attached to their face. On the other hand, the infants in the Loma Linda Infant Pain Expressions Dataset were placed in a closed lid cradle causing reflection of light and had a nasal breathing device attached on their face. Our approach achieved the highest classification accuracy, demonstrating that our approach generalizes better compared to the state-of-the-art.

3) Infant Classification Of Pain Expressions Dataset [8]: Our approach achieved a classification accuracy of 94.12%, precision of 92.36% and recall of 99.25%. Since this dataset consists of static images we cannot use the LSTM during this interval can be safely ignored. We observed that majority of the misclassified sequences occurred during the transition phase (i.e., when a neonate’s facial expression changes from No pain to Pain and vice-versa). The reason for this is that during this phase when the facial expression is changing, the features extracted are a mix of both the classes “Pain” and “No Pain” leading to a boundary misclassification. Since the transition phase is a very short duration of time, any misclassification occurring during this interval can be safely ignored.

F. Estimating the Intensity of Pain

In our approach we estimate the intensity of pain as a probability score directly from the Softmax output of the LSTM. In order to determine the accuracy of the estimated pain, we annotated 150 sequences into 4 levels of pain intensity namely: “No Pain”: 0.0 - 0.35, “Minimal Pain”: 0.36 - 0.5, “Moderate Pain”: 0.51 - 0.8 and “Severe Pain”: 0.81 - 1.0. The annotators were shown 5 sequences from each class as a reference. Table V shows the confusion matrix in estimating the intensity of pain for the 150 sequences. It is observed that our approach estimates intensity of pain with an accuracy of 94% on these 150 sequences.

G. Data Augmentation using Synthetic Images

In order to observe the effect of augmenting the dataset, we generated 10,000 images for both classes (pain/no pain) using the CGAN shown in Fig. 2. We used these 20,000 synthetic images in addition to the 154 training videos from Table I to train the networks. Table VI shows the performance of the system before and after augmenting the dataset.
On observing Table VI, we find a slight improvement in classification accuracy after generating 10,000 synthetic images per class, but beyond this limit, the classification accuracy did not significantly increase while the time required for training increased significantly.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy after augmentation</th>
<th>Accuracy before augmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>iCOPE Classification Of Pain Expressions [3]</td>
<td>95.34%</td>
<td>94.85%</td>
</tr>
<tr>
<td>Loma Linda Infant Pain Expressions</td>
<td>88.27%</td>
<td>86.04%</td>
</tr>
<tr>
<td>Infant Classification Of Pain Expressions [8]</td>
<td>94.12%</td>
<td>94.12%</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

We developed a system for automatically detecting and estimating the pain of neonates from videos. We performed exhaustive evaluation on three different datasets and our approach outperformed the state-of-the-art on all three datasets and is also robust to noisy data. Our approach performed well even on unseen low resolution videos collected in a different environment and in the presence of occlusions, demonstrating its generalization capabilities. We observed that augmenting our dataset with synthetic images helped slightly improve the performance. Furthermore, our approach implicitly learns the intensity of pain as a softmax probability score directly from the video without any annotation.

ACKNOWLEDGEMENT

This work was supported in part by Bourns Endowment funds and in part by NIH grant R01 NR011209.

REFERENCES


