Attentional Generative Multimodal Network for Neonatal Postoperative Pain Estimation

Md Sirajus Salekin[1], Ghada Zamzmi[1], Dmitry Goldgof[1], Peter R. Mouton[2], Kanwaljeet J. S. Anand[3], Terri Ashmeade[1], Stephanie Prescott[1], Yangxin Huang[1], and Yu Sun[1(✉)]

[1] University of South Florida, Tampa, FL, USA
yusun@usf.edu
[2] SRC Biosciences, Tampa, FL, USA
[3] Stanford University, Stanford, CA, USA

Abstract. Artificial Intelligence (AI)-based methods allow for automatic assessment of pain intensity based on continuous monitoring and processing of subtle changes in sensory signals, including facial expression, body movements, and crying frequency. Currently, there is a large and growing need for expanding current AI-based approaches to the assessment of postoperative pain in the neonatal intensive care unit (NICU). In contrast to acute procedural pain in the clinic, the NICU has neonates emerging from postoperative sedation, usually intubated, and with variable energy reserves for manifesting forceful pain responses. Here, we present a novel multi-modal approach designed, developed, and validated for assessment of neonatal postoperative pain in the challenging NICU setting. Our approach includes a robust network capable of efficient reconstruction of missing modalities (e.g., obscured facial expression due to intubation) using an unsupervised spatio-temporal feature learning with a generative model for learning the joint features. Our approach generates the final pain score along with the intensity using an attentional cross-modal feature fusion. Using experimental dataset from postoperative neonates in the NICU, our pain assessment approach achieves superior performance (AUC 0.906, accuracy 0.820) as compared to the state-of-the-art approaches.

Keywords: Generative model · Multimodal learning · Neonatal pain · NICU · Postoperative pain

1 Introduction

It is known that newborns subjected to emergency surgical procedures experience variable levels of pain during the postoperative period. Studies [13] in human and animal neonates reported that postoperative pain leads to long-lasting and likely permanent harm to the normal development of the highly vulnerable nervous system of neonates. Thus, a major challenge for the scientific community

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13433, pp. 749–759, 2022.
https://doi.org/10.1007/978-3-031-16437-8_72

is to effectively assess, prevent, and mitigate where possible the impact of post-operative pain on neonates. The combination of artificial intelligence (AI)-based methods with continuous monitoring and capture of subtle visual signals have enhanced the capability for continuous assessment of pain behaviors in neonates [14,22,26]. Although these methods achieved promising results, the shortcomings of early methods have limited its applicability for widespread use in real-world clinical practice [13].

Prior approaches, except [14], focused on assessing neonatal acute procedural pain, i.e., short-term distress following brief medical procedure (e.g., immunization) that are routinely experienced by healthy newborns in the presence of caregivers. Given the relatively benign and transient impact of these painful experiences, there is a growing need for expanding pain assessments to help mitigate the long-term and potentially more harmful consequences of postoperative pain in the NICU [13]. Second, prior works were designed for clinical scenarios with full access to visual and audio signals with minimum occlusion and background noise; thus, these pain assessments would be expected to perform poorly or completely fail for intubated neonates, variable light conditions, and ambient sound. A third limitation is that prior works handled failure of signal detection (missing modalities) by ignoring the absent modality and making a final decision based on existing data. The resulting loss of relevant pain information and modality bias could lead to errors since all current manual scales for assessing postoperative pain rely on all modalities to generate the final pain scores.

Our technical contributions are as follows. First, we developed a deep feature extractor followed by an RNN (Recurrent Neural Network) autoencoder network to extract and learn spatio-temporal features from both visual and auditory modalities. Second, we designed a novel generative model that combines all the modalities while learning to reconstruct any missing modalities. Third, instead of using early or late fusion techniques, we used a transformer-based attentional model that learns cross-modal features and generates the final pain label along with its intensity. From an application standpoint, this work presents the first multimodal spatio-temporal approach for neonatal postoperative pain intensity estimation that is designed, developed, and evaluated using a dataset collected in a real-world NICU setting.

2 Related Works

Multimodal Learning: The common approach for multimodal learning involves training different modalities followed by an early or late fusion [4,7,19]. Since these approaches assume that both individual modalities and combined modalities have the same ground truth (GT) labels, it is not suitable for post-operative pain analysis, where modalities have individual GT labels that might differ from the final GT label. For example, the GT label for a specific modality (e.g., sound) could be no-pain while the final assessment of the neonate's state could be pain. To handle this issue, a recent work [14] uses a multimodal Bilinear CNN-LSTM network that was trained using individual modalities' labels and the

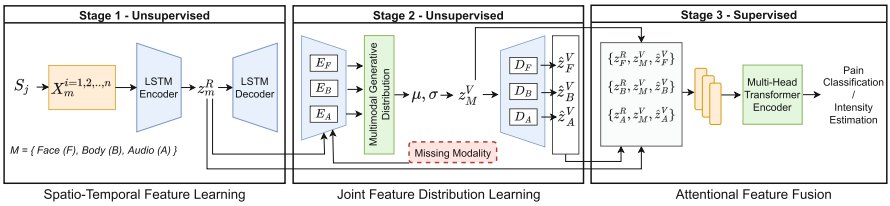

Fig. 1. Proposed approach for postoperative pain assessment (best viewed in color).

final label. The main limitation of this work is that the proposed network can not appropriately handle missing modalities; i.e., it makes the final assessment of pain based on the existing modalities, which can introduce an inherent bias towards available modalities.

Autoencoder (AE): AE [28] is a representational and unsupervised learning technique that learns a compressed feature embedding. In [17], an LSTM-based autoencoder is proposed to reconstruct and predict video sequences. To address the issue of non-regularized latent space in AE, Variational Autoencoder (VAE) [8] generates a probability distribution over the latent space. In multimodal learning, VAE has been extended to utilize multimodal joint learning for generative models such as CMMA [11], JMVAE [18], and MVAE [23].

3 Methodology

Figure 1 shows our novel approach in three stages: spatio-temporal feature learning (Stage 1), joint feature distribution learning (Stage 2), and attentional feature fusion (Stage 3). The following section presents the important notations and pre-processing steps then describes the details of each stage.

Definition and Notations: Let S be the number of visual samples in the video modality, and each sample consists of n number of frames i.e. $S_j = f_1, f_2, f_3, ..., f_n$ where $S_j \in S$. In the case of auditory modality, S_j is just one audio signal. Each pain episode contains three sensory signals $m \in M : face(F)$, $body(B)$, $audio(A)$, i.e. $M = \{F, B, A\}$. For any given sample, we have individual GT labels for F, B and A sensory signals. It is worth mentioning that unlike prior multimodal learning works [6,10,25] in postoperative pain assessment, the individual GT labels are provided based on the observation of the entire modality, not per frame. Finally, a final GT label is provided based on all sensory signals. This final label provides the assessment as pain or no-pain along with an intensity score. As any of these modalities or sensory signals can be missing in the real-world, we aim to detect the pain or no-pain class along with the pain intensity level. If any particular modality is missing, we reconstruct the modality and integrate it into the pain assessment contrary to the previous work [14], which entirely discards the missing modality.

Pre-processing and Augmentation: To prepare the multimodal dataset (Sect. 4) for the proposed approach, we performed the following. First, we extracted the visual (face and body) frames and audio signals from the raw data. To detect the facial region from the images, we used a YOLO-based face detector [12]. This detector was pre-trained using the WIDER face dataset [24] (≈393,703 labeled faces, 32,203 images). As for the body region, we used another YOLO-based [12] detector, which was pre-trained using the COCO object dataset [9] (≈1.5M object instances, 330K images). After detecting the face and body regions, we resized (224×224) all images to provide a consistent data flow in the multimodal network. In the case of the audio modality, we converted all the audio signals to $16K$ mono signals. Due to the partial occlusion of the neonate's face or body in some sequences, some frames were not detected which led to a different number of frames belonging to face and body modalities. To fix this issue and remove repetitive frames, we extracted the salient frames from these sequences with an equal time distribution. Inspired by [21], we divided each sequence into N equal segments. From each segment, we chose F-number of random frames. This has proven to be an efficient frame extraction method in several computer vision tasks [21]. In our experiments, we empirically chose the value of N and F as 10 and 1, respectively. Finally, we performed video augmentation by random rotation (±30) and horizontal flip. This augmentation was applied to all frames of a particular sequence dynamically during the training time.

Spatio-Temporal Feature Learning (Stage 1): We train an LSTM-based AE to capture the spatial and temporal features from the video (F, B) and auditory (A) modalities. Initially, we extracted spatial features from each facial image using FaceNet-based model [16]. This model was pre-trained on the VGGFace2 dataset [2]. For the body region, we used a Resnet18-based model, which was trained on the popular ImageNet [5] dataset. For the auditory modality, we used Google's VGGish model, which was pre-trained with YouTube-8M[1] dataset. Finally, feature sequences of each modality were used to train the LSTM-based AE in an unsupervised manner, where the encoder learns a compressed spatio-temporal feature representation from the deep features. For a spatial feature vector X_m^i with d_m feature-length and n sequence length, this AE maps the sequence as follows:

$$E_R : X_m^{i=1,2,\ldots,n} \rightarrow z_m^R \quad and \quad D_R : z_m^R \rightarrow \hat{X}_m^{i=1,2,\ldots,n} \tag{1}$$

$$L_R = \frac{1}{n} \sum_{i=1}^{n} (X_m^i - \hat{X_m^i})^2 \tag{2}$$

where $m \in M$, E and D are the RNN encoder and decoder functions, z_m^R is the fixed size latent feature space of the RNN AE, and $\hat{X}$ are the reconstructed features. We used the mean square error (MSE) as the loss function (L_R) to learn the feature reconstruction.

[1] http://research.google.com/youtube8m/.

Joint Feature Distribution Learning (Stage 2): After training the LSTM AE, we extracted the latent feature z_m^R for each sensory signal. z_m^R is the feature vector for a particular video (F, B) or audio (A). To learn the joint probability distribution of these vectors, we used VAE [8,11,18,23]. A basic VAE consists of a generative (θ) model and inference (ϕ) model, and it is optimized through Evidence Lower Bound (ELBO). We initially generated a parameterized inference model to estimate the probability distribution $(\mu,\ \sigma)$ of the latent space for each sensory signal (F, B, A). We used a product of expert approximation (POE) [3] to generate a joint-posterior distribution. This POE acts as a common parameterized inference network to estimate the final probability distribution of the joint latent space. ELBO can be defined based on the combination of the likelihood and Kullback-Leibler (KL) divergence as follows:

$$ELBO(z_m^R) := \mathbb{E}_{q_\phi|z_m^R}[\lambda \log p_\theta(z_m^R|z^V)] - \beta KL[q_\phi(z^V|z_m^R), p(z^V)] \tag{3}$$

where z_m^R and z^V are the observation and the latent space, respectively; $p_\theta(z_m^R|z^V)$ and $q_\phi(z^V|z_m^R)$ are the generative model and inference network respectively; $p(z^V)$ is the prior; λ and β are the controlled parameters [1,23]. To incorporate the POE over multiple sensory signals, this equation can be extended as:

$$ELBO(z_M^R) := \mathbb{E}_{q_\phi|z_m^R}[\sum_{m \in M} \lambda_m \log p_\theta(z_m^R|z^V)] - \beta KL[q_\phi(z^V|z_m^R), p(z^V)] \tag{4}$$

Theoretically, training an ELBO consisting of N sensory signals requires 2^N combinations, which is computationally expensive. Therefore, we only optimized ELBO of the joint signals instead of individual signals. We passed $Null$ values for the ELBO of the individual signals, and defined the joint learning loss (L_V) from this multimodal AE as follows:

$$L_V = ELBO(z_M^R) + ELBO(z_F^R) + ELBO(z_B^R) + ELBO(z_A^R) \tag{5}$$

Based on the equation above, the multimodal AE can be trained under different missing data conditions. Specifically, if any signal is missing in the test case, the POE can still create the generative probability distribution, which is used to generate the common latent features (z_M^R) that acts as a common joint feature for all signals. Then, the multimodal AE can reconstruct the individual features $(\hat{z}_m^R)$ again from the common feature space (z_M^R). We used MSE as the loss function for the reconstruction.

Attentional Fusion (Stage 3): After generating the spatio-temporal latent space (z_m^R) and reconstructing missing modalities $(\hat{z}_m^R)$ from the joint probability latent space (z_M^R) in Stage 1 and Stage 2, we stacked the latent features of F, B, and A signals and applied an attentional fusion using the Transformer encoder [20] as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

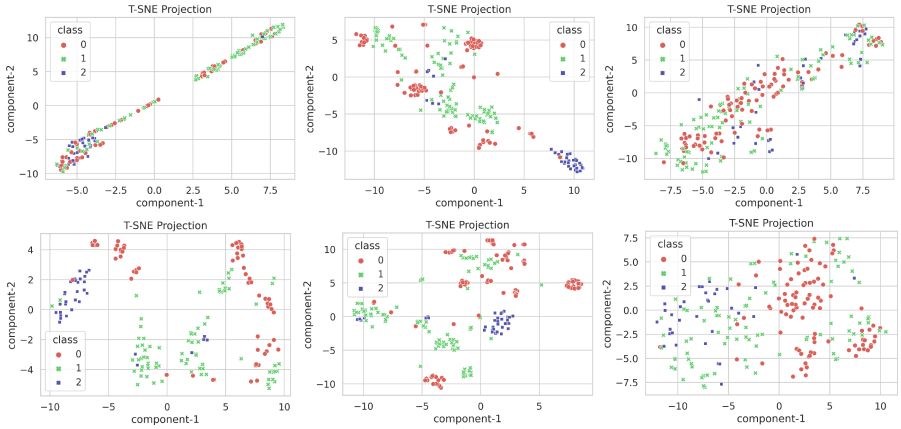

Fig. 2. t-SNE projection of spatio-temporal features using perplexity of 40. Each column represents face, body, and audio (left to right). Top and bottom rows are the baseline and proposed features, respectively (best viewed in color).

where, Q, K, V, and d_k are the query, key, value matrix, and the scaling factor, respectively. As shown in Fig. 1, attentive features were generated using the latent features from all modalities (F, B, A). The generated attentive features were then concatenated and used for pain assessment. Specifically, the pain assessment was produced as follows. The spatio-temporal feature (z_m^R) or reconstructed feature (z_m^V) was selected for each sensory signal. Then, the selected features were stacked followed by performing a multi-head attention to learn the cross-modal relation while focusing on the salient features. Finally, the attentive features were concatenated and used as a final feature vector. This vector was used for assessing pain and estimating its intensity.

4 Experimental Setup and Results

This section presents evaluation of the proposed approach (three stages) and the performance of both pain classification and intensity estimation. We used the accuracy, F-1 score, and AUC to report the performance of binary classification, and MSE and mean absolute error (MAE) to report the performance of intensity estimation. All the models were developed based on PyTorch environment using a GPU machine (Intel core i7-7700K@4.20 GHz, 32 GB RAM, and NVIDIA GV100 TITAN V 12 GB GPU).

Dataset: We used the USF-MNPAD-I [15] neonatal pain dataset, which is the only publicly available neonatal postoperative pain dataset for research use [13]. This dataset has 36 subjects recorded during acute procedural pain, and 9 subjects during postoperative pain. Each subject has videos (face and body) and audios (crying and background noises) recorded in the NICU of a local hospital. Each video and audio contain pain and no-pain segments that are labeled

Table 1. Performance of our approach and previous works when all signals are present.

Approach	Accuracy	Precision	Recall	F1-score	TPR	FPR	AUC
CNN-LSTM [14]	0.7895	0.7913	0.7895	0.7863	0.8761	0.3243	0.8791
EmbraceNet [4]	0.7921	0.7919	0.7921	0.7920	0.8182	0.2405	0.8790
Proposed	0.8202	0.8230	0.8202	0.8207	0.8080	0.1646	0.9055

with two manual pain scales: NIPS scale for procedural pain and N-PASS scale for postoperative pain. We used the procedural part of the dataset to learn the spatio-temporal features. The postoperative part was used to learn the joint feature distribution and reconstruct the missing modalities.

Network Architectures and Training: In Stage 1, we used state-of-the-art models (Sect. 3) to extract spatio-temporal feature vectors with 512-d, 512-d, and 128-d length from F, B, A signals, respectively. For temporal learning, we used an individual LSTM AE with 2 layers, taking the respective spatial feature vector of input sequences to produce a spatio-temporal 128-d latent space. As mentioned above (Sect. 3), the video has a sequence length of $\approx$10 s. In Stage 2, we used MLP encoder-decoder following $128 \rightarrow 128 \rightarrow 64$ and $64 \rightarrow 128 \rightarrow 128 \rightarrow 128$ encoder and decoder layers for each sensory signal. In Stage 3, a transformer encoder layer with 2 multi-heads had been used to initially perform the scale-dot-product attention. After that, all the features were concatenated $(128 + 128 + 128 = 384)$. Next, an MLP layer following $384 \rightarrow 256 \rightarrow 128 \rightarrow Y$ was used. In case of binary classification, a sigmoid function was used for pain and no-pain classes. As for estimation, $Y = 1$ is just a linear point for pain intensity estimation. A total of 218 postoperative videos (50% pain) were included in our experiments. Following previous approaches [14,27], we performed a leave-one-subject-out (LOSO) evaluation. For the spatio-temporal training, we used the procedural dataset to learn the spatio-temporal features until convergence. For RNN autoencoder, we used Adam optimizer with 0.001 learning rate and 16 batch size. In the joint learning and attentional feature learning, we followed LOSO and used Adam optimizer with 0.0001 learning rate and batch size of 8.

Visualization of Spatio-temporal Features: Spatio-temporal features were computed using FaceNet (face) [16], ResNet18 (body), and VGGish (sound). To evaluate the quality of the extracted features, we generated the t-SNE projections for all modalities as shown in Fig. 2. Note that all modalities are trained on the procedural pain set (unsupervised) and tested on the postoperative set. From the figure, we can observe that the feature points are scattered in the first row, which shows the baselines for face, body, and sound. The baseline for face and body signals are the raw pixels obtained from the video modality while the baseline for the sound is the mel frequency cepstral coefficients (MFCCs) calculated from the auditory modality. On the contrary, the second row shows the feature points, which are generated by stage 1, grouped into clusters indicating a good differentiation capability of the extracted features.

Table 2. Performance of the proposed approach and [14] when dropping each modality.

Approach	Modalities	Reconstruction?	Accuracy	F1-score	TPR	FPR	AUC
CNN-LSTM [14]	Drop_{Face}	No	0.7719	0.7522	0.9897	0.5135	0.8763
	Drop_{Body}	No	0.6901	0.6703	0.8866	0.5676	0.8396
	Drop_{Sound}	No	0.7076	0.6630	1.0000	0.6757	0.8353
Proposed	Drop_{Face}	Yes	0.7921	0.7928	0.7576	0.1646	0.9022
	Drop_{Body}	Yes	0.8258	0.8257	0.8485	0.2025	0.9086
	Drop_{Sound}	Yes	0.6854	0.6374	0.9899	0.6962	0.8028

Table 3. Ablation study of the attentional feature fusion.

Approach	Accuracy	Precision	Recall	F1-Score	TPR	FPR	AUC
ST + JF	0.5229	0.7559	0.5229	0.3824	0.9999	0.9541	0.5757
ST + JF + AF	0.7890	0.7899	0.7890	0.7888	0.7615	0.1835	0.8870

* ST = Spatio-Temporal, JF = Joint Features, AF = Attentional Fusion

Pain Assessment w/o and w/ Missing Modalities: We compared our proposed classifier with CNN-LSTM approach [14] and another multimodal approach named EmbraceNet [4]. In this experiment, we performed pain assessment in a subset of USF-MNPAD-I [15] that has all the sensory signals present (F, B, A). From Table 1, we can see that the proposed approach outperformed [14] and achieved 0.820 accuracy and 0.906 AUC. Although our approach achieved a lower TPR as compared to [14], it improved the FPR (0.165) by almost 50%. Similarly, our approach significantly outperformed EmbraceNet [4] ($p < 0.01$).

To evaluate the performance of our approach and the novel reconstruction method, we completely dropped (100%) each sensory signal, reconstructed the features of the dropped signal, combined them with the features of other signals, and reported the performance of multimodal pain classification. We also reported the pain assessment performance using CNN-LSTM as it is the most recent work in the literature that uses USF-MNPAD-I [15] dataset. Recall that this approach [14] discarded missing modalities when making a final assessment. We note that missing a sensory signal is common in clinical practices due to several factors including sensor failure, swaddling, or intbaution, among others. Our model can classify any case with missing modalities as it can reconstruct these modalities and integrate them into the assessment. From Table 2, we observe that reconstructing the features of face and body using our approach improved the performance as compared to CNN-LSTM. The lower performance of sound suggests that sound reconstruction has a higher impact on the final pain/no-pain decision, which is consistent with a similar trend observed in our previous work [14].

Multimodal Assessment with Attentional Feature Fusion: Unlike other approaches, we used an attentional fusion to examine the cross-modal influence on the decision. To evaluate this fusion approach, we performed an ablation

study, in which we reported the performance of pain classification with and without attentional fusion. In Table 3, we can observe that the proposed attentional fusion (ST + JF + AF) improved the pain classification performance by a large margin, demonstrating the effectiveness of this fusion approach.

Postoperative Pain Estimation: As the pain intensity in USF-MNPAD-I [15] dataset ranges from 0 to 7, we performed a regression-based training to generate the intensity score. We found an MSE of 3.95 and an MAE of 1.73, which are reasonable for this relatively small and challenging dataset. We further minimized the intensity range and found better results which are 0–4 (MSE 0.75, MAE 0.73) and 0–1 (MSE 0.13, MAE 0.27). We also found that the proposed approach is capable of understanding the no-pain/pain/no-pain transitions while estimating pain intensity with a success rate of 71.15%.

5 Conclusion

This work presents a novel approach for neonatal postoperative pain assessment. Our results demonstrated the efficacy of our novel approach in constructing missing signals, a common situation in NICU settings. Further, our results demonstrated the efficacy of our fusion method in enhancing multimodal pain assessment. These results are promising and suggest the superiority of our approach, which was evaluated on a challenging real-world dataset, as compared to similar works in the literature. In the future, we plan to further evaluate the proposed approach using a large-scale multi-site neonatal multimodal postoperative pain dataset as well as investigate the performance of the proposed approach when two or more modalities are missing.

Acknowledgement. This research is supported by National Institutes of Health (NIH), United States Grant (NIH R21NR018756). Although the second author (G.Z.) is currently affiliated with NIH, this work was conducted while being with the University of South Florida. The opinions expressed in this article are the G.Z.'s own and do not reflect the view of NIH, the Department of Health and Human Services, or the United States government.

References

1. Bowman, S., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pp. 10–21 (2016)
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE (2018)
3. Cao, Y., Fleet, D.J.: Generalized product of experts for automatic and principled fusion of Gaussian process predictions. arXiv preprint arXiv:1410.7827 (2014)
4. Choi, J.H., Lee, J.S.: EmbraceNet: a robust deep learning architecture for multimodal classification. Inf. Fusion **51**, 259–270 (2019)