A biologically-inspired computer vision pipeline for neonatal oral motor assessment

Danila Germanese[1][0000−0002−7814−5280], Serena Bardelli[2], Armando Cuttano[2], Giulio Del Corso[1], Fabrizia Festante[3], Andrea Guzzetta[3], M.Antonietta Pascali[1], Lucia Rocchitelli[3], and Sara Colantonio[1]

[1] Inst. of Inf. Science and Techn. (ISTI), National Research Council, Pisa, Italy
danila.germanese@isti.cnr.it
[2] Centro di Form. e Sim. Neon. (NINA), Az. Ospedaliero Univ. Pisana, Pisa, Italy
[3] Dept. Dev. Neur.,IRCCS Stella Maris F., Dept. Cl. and Exp. Med.,Univ. of Pisa, Pisa, Italy

Abstract. Automatic recognition of neonatal oral motor patterns may provide useful information about newborns' health. However, their distinctive facial morphology, together with their high movement variability, poses significant challenges for computer vision analysis. Here, we propose a biologically inspired computer vision pipeline for the automatic classification of neonatal mouth poses from video data. The approach includes structured preprocessing, representation learning, and supervised classification. First, mouth regions are localised and normalised, then encoded into a 512-dimensional feature vector using a Gabor+Sobel-based autoencoder. Finally, a multilayer perceptron classifier discriminates each feature vector between mouth closed, mouth open, and tongue protrusion. The autoencoder achieved high reconstruction quality (SSIM = 0.97 on unseen data), and the classifier reached a balanced accuracy of 0.83. These findings suggest that a biologically inspired feature extraction can effectively capture key morphological characteristics of neonatal faces, enabling robust automated behavioral analysis.

Keywords: Computer vision · Neonatal behavior analysis · Representation learning.

1 Introduction

Human facial appearance provides key indicators of health and neurodevelopmental status, particularly in early life, when atypical conditions may manifest through altered facial expressions or behaviours [1].

This has motivated the adoption of contactless computer vision (CV) approaches for extracting objective visual biomarkers to support clinicians and caregivers in evaluating newborns' overall well-being [2]. Current solutions include facial expression analysis [3], motor impairment evaluation [4], behavioral observation for autism spectrum disorder [5], and monitoring in neonatal intensive care units [6]. However, newborns represent a particularly challenging

population due to rapid developmental changes, distinct facial morphology, and high variability in pose and spontaneous movements [7]. As a result, most existing studies focus on older infants [8], while research on early newborns, especially preterm infants, remains limited. A significant open research question concerns neonatal imitation, i.e., the ability to reproduce observed gestures [9], which may provide insights into early social cognition and neurodevelopment [10]. Its analysis requires accurate facial action units (FACs) detection [11], but methods effective in adults [12], often fail in newborns due to marked differences in facial structure and motion dynamics [13].

In this paper, we addressed these challenges by analysing video-derived images from 10 newborns (8 preterm, 2 full-term, $\leq$4 weeks post-term equivalent age). We implemented a biologically-inspired CV pipeline, including a Gabor+Sobel-based autoencoder and a multilayer perceptron, for the automatic classification of mouth states (open/closed/ tongue protrusion) as a preliminary step towards computational assessment of neonatal imitation and, more broadly, towards unobtrusive monitoring of early neurodevelopment.

2 The Dataset

2.1 Data acquisition

For each subject, video acquisition included three 3-minute conditions (tongue protrusion, mouth opening, and control static stimulus). Infants were placed in a cushioned bassinet and recorded using a Canon Legria HFG70 4K camera. Preterm infants (32–36 weeks) and full-term infants ($\geq$37 weeks) were recruited from the Neonatal Unit of the University Hospital of Pisa if clinically stable, without significant brain abnormalities and in the absence of perinatal complications.

The study was approved by the Tuscany Region Paediatric Ethics Committee (123/2020, 12/2021) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all families prior to participation.

2.2 Data Pre-Processing

Videos were analyzed at the frame level (video frame rate = 25 fps). Inspired by our previous work [14], a dedicated pre-processing pipeline was designed. It included (see fig.1):

- *Face detection and extraction*, using Google MediaPipe face detection module [15]. Frames in which no face was detected were discarded.
- *Mouth localization and extraction*, using MediaPipe Face Landmarker module [15]. Also, a rotated bounding box aligned with the mouth corners was used to normalize infants' head pose variability and standardize the region of interest. Frames with unreliable landmarks were removed.

- *Mouth images resizing* to 64 × 64 pixels, in order to reduce computational cost.
- *Grayscale conversion* of mouth images, to reduce input dimensionality and the number of trainable parameters.
- *Histogram matching* using a reference image was applied to reduce illumination variability while preserving texture details, such as lip contours and subtle mouth movements.

2.3 Data Labeling

The images were independently assigned by three experts to one of the three categories: MC/MO/TP. Images without consensus among the three annotators were removed.

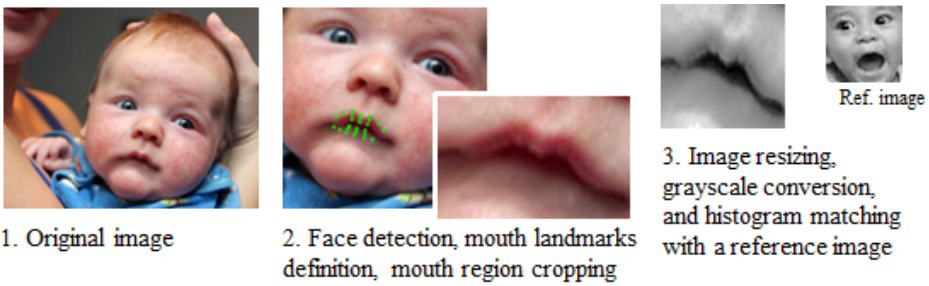

Fig. 1. Image pre-processing steps: from the video frame to ROI definition. (Image was taken from [13])

3 Methods

The proposed method had two objectives: (i) reducing high-dimensional image data into a compact numerical representation, through a Gabor+Sobel-based autoencoder, and (ii) performing mouth-state recognition through a four-layer multi-layer perceptron (MLP) classifier trained to distinguish between the three classes: MO, MC, and TP.

3.1 Representation learning: the autoencoder

A convolutional autoencoder was designed to reconstruct facial images while enforcing the extraction of structured features via biologically inspired filtering. At the encoder input stage, a bank of Gabor convolutional filters was introduced to mimic the human primary visual cortex [16], thereby providing the network with a structured prior before learning task-specific representations. Gabor filters

are selectively responsive to oriented spatial frequencies [17] and are particularly well suited for capturing fine facial structures: wrinkles, eyelashes, eyebrows, skin folds, and shading patterns. In addition, Sobel filters were incorporated to compute spatial image derivatives: the horizontal operator enhances vertical edges, while the vertical operator highlights horizontal edges, enabling the extraction of local luminance variations in facial regions [18]. By integrating these filters, the autoencoder is guided by perceptually meaningful features, rather than fully unconstrained representations, thus more closely approximating the early stages of human visual processing and reducing the degrees of freedom of the learned representation.

Hence, from an architectural point of view, the network applies three convolutional branches to each grayscale input image: a bank of 32 Gabor filters (9×9) and 2 Sobel operators (horizontal and vertical, 3×3). The Gabor filters are initialized following the standard formulation and fine-tuned during training, whereas the Sobel filters remain non-trainable. The resulting feature maps form a tensor of size $34 \times 64 \times 64$. The encoder stage consists of three convolutional blocks. The first two blocks include convolution, batch normalisation, ReLU activation, and 2x2 max pooling, increasing channels from 34→64→128 while reducing spatial resolution, thus producing a tensor of $128 \times 16 \times 16$. The third block maps 128→256 channels without further downsampling, producing a tensor of $256 \times 16 \times 16$. A 1×1 convolutional bottleneck expands the tensor to $512 \times 16 \times 16$, followed by batch normalisation, ReLU activation, and dropout ($p = 0.3$). Finally, a global adaptive average pooling applied to the bottleneck tensor produces a 512-dimensional latent vector per frame. The decoder mirrors the encoder to reconstruct grayscale images of size $1 \times 64 \times 64$. The network was trained using an adaptive loss that combines pixel-wise fidelity (Mean Squared Error, MSE) and Structural Similarity Index Measure, SSIM [19], on 6806 infant facial images (including those from InfAnFace public dataset [13] and those extracted from video frames of two preterm newborns from the Neonatal Unit of the University Hospital of Pisa, see subsection 2.1). Optimization used Adam (lr= 10^{-5}), a batch size of 16, and early stopping (patience = 15 epochs), with a 5-fold cross-validation scheme.

3.2 Classification: The Multi-Layer-Perceptron

Mouth images, each encoded into a compact 512-dimensional feature vector, were then used as input features for a supervised classifier. The classification task involved distinguishing between three classes: MC, MO, TP.

To this end, a fully connected MLP was designed. The network architecture was composed of four dense layers: (i) $512 \rightarrow 256$ neurons + Batch Normalization + ReLU + Dropout (p = 0.3), (ii) $256 \rightarrow 128$ neurons + Batch Normalization + ReLU, (iii) $128 \rightarrow 64$ neurons + ReLU, (iv) $64 \rightarrow 3$ output neurons (logits). Batch normalization was employed to stabilise the training process, while dropout regularization was introduced to reduce overfitting given the limited dataset size. A leave-one-patient-out cross-validation (LOPO-CV) protocol was used: at each iteration, data from one subject was reserved for testing while the

remaining subjects were used for training, preventing identity leakage and providing a realistic estimate of generalization to unseen individuals. Because the dataset exhibited class imbalance (see Fig. 2), the network was optimized using Focal Loss, implemented on top of the cross-entropy loss computed from logits (converted into probabilities via the softmax function), and defined as in eq. 1:

$$FL(p_t) = (1 - p_t)^{\gamma} \cdot CE(p_t) \tag{1}$$

where:

- p_t represents the predicted probability of the true class
- CE is the standard cross-entropy loss
- $\gamma = 2$ controls the focusing effect, giving higher weight to hard-to-classify or minority samples

The model was trained using the Adam optimizer (lr= 10^{-4}) on a dataset of 9,912 compact 512-dimensional feature vectors derived from six preterm and two full-term newborns from the Neonatal Unit of the University Hospital of Pisa.

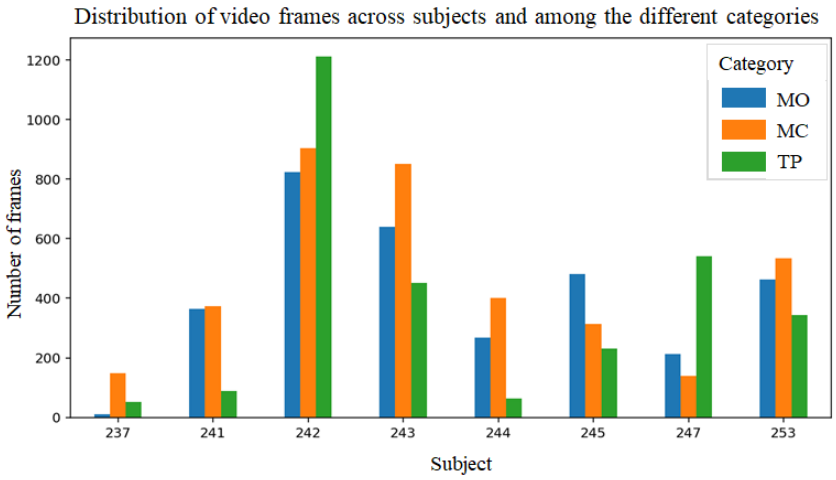

Fig. 2. The distribution of video frames per subject and across the different class labels.

4 Results

4.1 Gabor+Sobel convolutional autoencoder recontruction

The Gabor+Sobel–based convolutional autoencoder was evaluated on a held-out test set, achieving a mean SSIM of 0.97 (Fig. 3). This high structural agreement

between reconstructed and original images indicates that the model effectively captures the fine-grained textural patterns of infant faces, thus confirming the suitability of the learned latent representation for subsequent feature-based classification tasks.

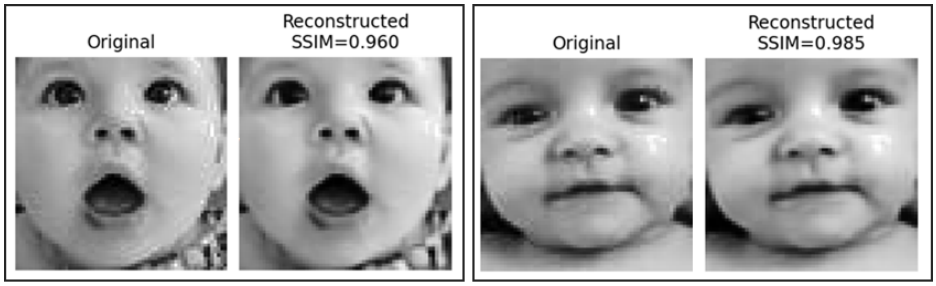

Fig. 3. Two images from [13] reconstructed by the convolutional autoencoder

4.2 MLP Classification results

Table 1 reports fold-wise performances, with balanced accuracy ranging from 0.76 to 0.91 and macro F1-score ranging from 0.72 to 0.91, indicating good generalisation across subjects, despite inter-individual variability in facial appearance and motion patterns. Also global results (Fig. 4) confirm accurate classification across classes. The confusion matrix demonstrates that the classifier accurately identifies the majority of samples in each class. However, misclassifications predominantly occur between physiologically similar categories, rather than uniformly across classes or at specific temporal phases (e.g., onset or offset of a gesture). In particular, MC and TP are occasionally predicted in place of MO, while MC is also partially confused with TP. This behaviour is likely attributable to residual illumination-related artefacts (not fully corrected during preprocessing) which can reduce the visual discriminability between these states (see Fig. 5).

Table 1. For each cross-validation fold of the mlp, the corresponding balanced accuracy (BA) and macro-averaged F1-score (F1 M) are reported.

	Folds							
Metrics	***F1***	***F2***	***F3***	***F4***	***F5***	***F6***	***F7***	***F8***
BA	0.80	0.90	0.80	0.91	0.76	0.77	0.82	0.83
F1 M	0.84	0.88	0.78	0.91	0.72	0.74	0.80	0.84

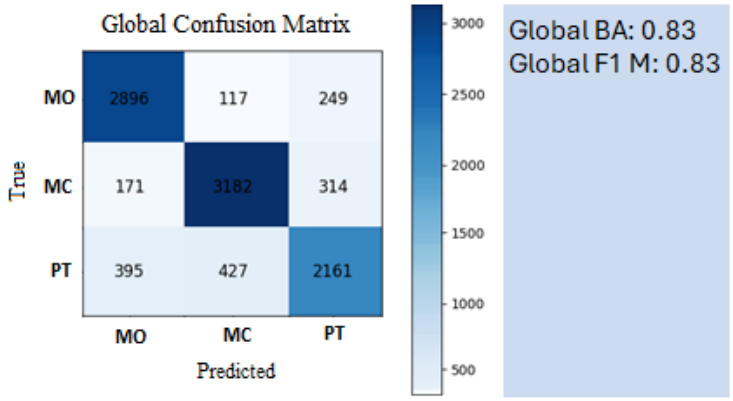

Fig. 4. The global balanced accuracy (BA) and macro-averaged F1-score (F1 M) together with the global confusion matrix

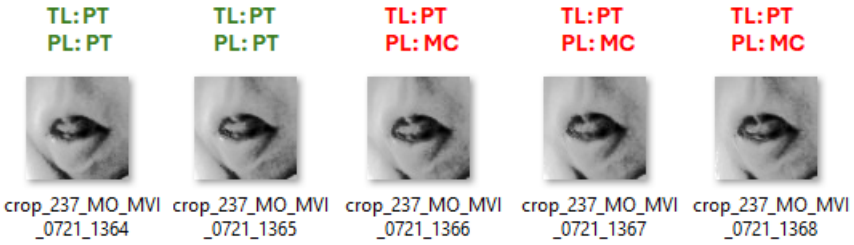

Fig. 5. Example of misclassification errors: MC confused with PT

We also compared the proposed multilayer perceptron (MLP) with two conventional machine learning classifiers: a support vector machine (SVM) with a radial basis function (RBF) kernel ($C = 10$, $\gamma = 0.001$) and a random forest (RF) classifier (n_estimators = 200, max_depth = None, min_samples_split = 10, min_samples_leaf = 2). The proposed MLP-based approach achieved superior performance, with a global balanced accuracy and F1-score of 0.83, compared to the SVM (0.79 and 0.79, respectively) and the RF (0.72 and 0.72, respectively). These findings suggest that the learned latent representation, combined with the MLP classifier, is more effective at capturing discriminative patterns of neonatal mouth states than classical machine learning approaches in a subject-independent setting.

5 Discussion

This study proposes a pipeline for the classification of neonatal oral-motor patterns, integrating structured preprocessing, biologically inspired representation learning, and a downstream multiclass classifier. The core component is a con-

volutional autoencoder enhanced with Gabor and Sobel filters to emulate early human visual processing. This architecture encodes infant mouth images into a 512-dimensional latent representation, subsequently used to train an MLP classifier. By imposing an explicit inductive bias toward low-level visual features such as edges and orientations, the model improves feature stability and generalization, particularly in low-data regimes typical of neonatal studies. The framework was evaluated on a dataset acquired at the Neonatal Unit of the University Hospital of Pisa, comprising video recordings from 10 newborns (8 preterm and 2 full-term). Two infants were used for autoencoder training, while the remaining eight were reserved for subject-independent evaluation of the classifier.

Results show that the learned latent representation proved informative, enabling effective discrimination of oral-motor behaviors, including mouth opening, closing, and tongue protrusion. Performance remained consistently strong, although variability across cross-validation folds was observed, reflecting inter-subject differences rather than model instability. This variability is likely driven by multiple factors, including changes in illumination across recordings, intrinsic inter-infant variability in movement dynamics, and anatomical differences such as lip morphology and facial structure, all of which influence the learned representation despite not being directly related to the target classes.

Previous approaches in neonatal facial and behavioural analysis have largely focused on pain assessment [20], facial expression recognition [3], or general motor activity [1, 4, 8], typically relying on either handcrafted features or end-to-end deep learning architectures applied to facial images or video sequences [5]. In contrast, the present study addresses a more fine-grained task, focusing specifically on oral-motor behaviours, subtle but clinically relevant indicators of early neurodevelopment. Moreover, while most existing methods adopt purely data-driven learning strategies [20], without incorporating explicit prior knowledge, this study introduces a biologically inspired feature extraction stage based on Gabor and Sobel filtering, embedding prior knowledge of early visual processing into the model.

The main limitation of this study is the modest dataset size, which includes a relatively small number of subjects. Although a large number of frames is available, it is important to emphasize that these frames are not independent observations, as they are temporally correlated within the same recording sessions. Consequently, the effective diversity and variability of the training data are substantially lower than what the raw frame count alone would suggest, thereby further constraining the generalisability of the findings. However, we would like to emphasise that data collection in this context is particularly challenging. The study involves newborns—predominantly preterm infants—who represent a highly sensitive and clinically vulnerable population. Recruitment is therefore inherently limited by strict ethical requirements, clinical stability criteria, and the need to minimise stress and disturbance during recording sessions. Moreover, the acquisition protocol requires controlled conditions and active participation (i.e., exposure to structured stimuli), which further restricts the number of eligible subjects and usable recordings. For these reasons, assembling larger and more

heterogeneous datasets in this setting is non-trivial. To mitigate these issues, we adopted a strict leave-one-patient-out cross-validation protocol, explicitly designed to assess subject-independent performance and to reduce optimistic bias arising from intra-subject correlations. This evaluation strategy provides a more realistic estimate of generalisation to unseen individuals compared to random or frame-based splitting approaches.

Also, oral motor patterns such as mouth opening and tongue protrusion are inherently dynamic phenomena. Failure to consider temporal evolution may result in ambiguities between visually similar static configurations.

Therefore, future research should aim to include a larger, more heterogeneous population to promote subject-independent learning. Furthermore, incorporating temporal modelling could represent a substantial and potentially important enhancement to the proposed framework, as it would enable the explicit exploitation of motion dynamics over time, thereby improving the ability to distinguish between visually similar static mouth configurations that may correspond to different underlying gestures. Another promising development is the integration of geometric information, or kinematic descriptors. Combining learned visual embeddings with explicit motion descriptors may have the potential to enhance interpretability and improve classification accuracy.

Acknowledgements This work was supported by the Italian Ministry of Health - Grant RC L1 (and the 5x1000 voluntary contributions).

References

1. C. Einspieler, A. F. Bos, M. E. Libertus, and P. B. Marschik. The general movement assessment helps us to identify preterm infants at risk for cognitive dysfunction. Frontiers in psychology, 7:178796, 2016.
2. D. Germanese, S. Colantonio, M. Del Coco, P. Carcagni, and M. Leo. Computer vision tasks for ambient intelligence in children's health. Information, 14(10), 2023.
3. A. Maroulis. Baby facereader AU classification for infant facial expression configurations. Measuring Behavior 2018, 2018.
4. C. Chambers, N. Seethapathi, R. Saluja, H. Loeb, S.R. Pierce, D.K. Bogen, et al. Computer Vision to Automatically Assess Infant Neuromotor Risk. IEEE Trans. Syst. Rehabil. Eng. 2020.
5. A. Ali, F.F. Negin, F.F. Bremond, and S. Thummler. Video-based behavior understanding of children for objective diagnosis of autism. In VISAPP 2022-17th Int. Conf. on Comp. Vis. Theory and Appl., 2022.
6. L. Sacks, E. Kunkoski, M. Noone. Digital Health Technologies in Pediatric Trials. Ther. Innov. Regul. Sci., 56, 929–933, 2022.
7. I. WR Bushnell. Mother's face recognition in newborn infants: Learning and memory. Infant and Child Development: An International Journal of Research and Practice, 10(1-2):67–74, 2001.
8. M. Leo, G.M. Bernava, P. Carcagni, C. Distante. Video-Based Automatic Baby Motion Analysis for Early Neurological Disorder Diagnosis: State of the Art and Future Directions. Sensors, 22, 866, 2022.

9. A. N. Meltzoff and M. K. Moore. Newborn infants imitate adult facial gestures. Child development, pages 702–709, 1983.
10. J. Davis, J. Redshaw, T. Suddendorf, M. Nielsen, S.K.Costantini, J. Oostenbroek, and V. Slaughter. Does neonatal imitation exist? Insights from a meta-analysis of 336 effect sizes. Persp. on Psych. Sci., 16(6),2021.
11. H. Oster. Baby facs: Facial action coding system for infants and young children. Unpublished monograph and coding manual. NY Univ., 2006.
12. S.C. Leong, Y.Ming Tang, C.Hin Lai, and C.K.M. Lee. Facial expression and body gesture emotion recognition:A systematic review on the use of visual data in affective computing. Comp.Sci.Rev.,48:100545,2023.
13. M. Wan, S. Zhu, L. Luan, G. Prateek, X. Huang, R. Schwartz-Mette, et al. Infanface: Bridging the infant–adult domain gap in facial landmark estimation in the wild. In IEEE 26th Int.C.on Pat.Rec.(ICPR),p.4486–4492,2022.
14. G. Del Corso, D. Germanese, M.A. Pascali, S. Bardelli, A. Cuttano, F. Festante, et al. Facial Landmark Identification and Data Preparation Can Significantly Improve the Extraction of Newborns' Facial Features. In IEEE 18th Int.C.on Aut.Face and Gest. Recogn.,Istanbul,Turkiye,2024.
15. C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, et al. Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172, 2019.
16. B. Ameur, S. Masmoudi, A.G. Derbel, A.B. Hamida. Fusing Gabor and LBP feature sets for KNN and SRC-based face recognition. In 2nd Int.C. Adv.Techn.for Sign.and Image Proc.(ATSIP),pp.453-458,IEEE,2016.
17. S. Xie, S. Shan, X. Chen, X. Meng, W. Gao. Learned local Gabor patterns for face representation and recognition, Sign.Proc., 89, 12, 2009.
18. S. Liu, X. Tang and D. Wang. Facial Expression Recognition Based on Sobel Operator and Improved CNN-SVM. IEEE 3rd Int. C. Inf. Comm. and Sig. Proc. (ICICSP), Shanghai, pp. 236-240, 2020.
19. Z. Wang, A.C. Bovik, H.R. Sheikh and E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity. in IEEE Trans. on Image Proc., vol.13, 4, pp. 600-612, 2004.
20. G. Zamzmi, R. Kasturi, D. Goldgof, R. Zhi, T. Ashmeade, Y. Sun. A Review of Automated Pain Assessment in Infants: Features, Classification Tasks, and Databases. IEEE Rev Biomed Eng, 11:77-96, 2018
5Oct 2017