ARL-IL CNN for Automatic Facial Expression Recognition of Infants under 24 Months of Age

Simeng Yan¹, Wenming Zheng¹, Chuangao Tang¹, Yuan Zong¹, Nana Qiu² and Xiaoyan Ke²

¹Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China
²Affiliated Brain Hospital of Nanjing Medical University, Nanjing, China

*wenming_zheng@seu.edu.cn

Abstract. Automatic facial expression coding of infants plays an important role in infants-related applications, including computer-aided ASD diagnosis, automatic intervention for ASD children and diagnosis of ADHD, etc. However, most of existing facial expression researches focused on adult facial expression analysis, the infant facial expression recognition has been less investigated. Due to an age gap between the facial expression datasets of adults and infants, a facial expression recognition model trained on adult datasets usually shows poor generalization to infants datasets. A labeled infant facial expression dataset can mitigate this problem, and hence we first collect a facial expression dataset of 30 infants under 24 months of age by recording videos of infants’ facial expression during a face-to-face mother-infant interaction. Due to infants spontaneous facial behaviors, the dataset covers multiple challenges, such as large head-poses, occlusion, facial expression intensities, etc. To develop an automatic facial expression coding system, we propose a framework consisted of adaptive region learning and island loss, i.e., ARL-IL, to self-adaptively discover facial regions with higher discriminability between different emotion classes. The framework was verified on our collected dataset, and attained a classification accuracy of 86.86%, which has shown better performance than conventional method based on hand-crafted features and some basic CNN architectures. To interpret the effectiveness of ARL-IL, we also visualize the learned features and find that the proposed framework focuses on facial regions with more emotion information compared with other hand-crafted features or learned features from basic CNN architectures. The experimental results show that our proposed framework has robustness to the large head-poses and occlusion.

1. Introduction
Facial expressions of infants are important social signals in daily life. In clinical practice, infants’ facial expression is regarded as one of important evaluation indicators for early screening of some diseases, such as ASD[1]. An automatic facial expression analysis system can monitor healthy and at-risk infants, and report their emotional state. Thus, it alleviates human labor and has potential in large-scale medical applications.

In recent years, automatic recognition of adult facial expression has been mature and successfully applied in many fields, but few studies focus on infant facial expression recognition. Compared with
adults’ facial expressions from facial appearance, infants have smoother skin, fatter cheeks, shorter jaws, larger eyes, lighter eyebrows and hair, shallower wrinkles, etc[2]. In addition, infants have habits of sucking thumb, covering eyes or mouth, large and fast head movements, which easily caused occlusions. Because of the age gap between infant and adult, models trained on adults’ facial expression datasets usually show poor performance on infant datasets. Lack of large amount of infants’ facial expression data hinders the research progress on facial expressions recognition of infant. Challenges caused by large head pose and occlusion make it difficult to align facial key points or regions, thereby affecting the performance of facial expression recognition of infant. If we can catch the facial expression-related region and extract features self-adaptively from different perspectives, the problem could be solved. Therefore, this paper has three main contributions:

- An IFE-3C dataset was established for research. It consists of 9,133 spontaneous facial expression images from 30 infants, and covers multiple challenging variables, such as large head pose, occlusion, natural illumination, etc.
- An ARL-IL framework based on adaptive region learning and island loss was proposed. It is expected to automatically focus on local regions most relevant to facial expression, while reducing bias caused by irrelevant variables.
- The proposed method obtained a classification accuracy of 86.86% on the IFC-3C dataset. Visualization of feature maps show that ARL-IL model is more effective in learning facial expression-related regions’ feature under different head poses and occlusion.

2. Related work

Most existing facial expression datasets, including BU-3DFE[3], Multi-PIE[4], CK+[5] and etc., recruited adult subjects. Few studies focused on infant, for example, the TIF[6], CIF[7] datasets are both under variable control with small data scale. The samples in CIT[7], RCLA&NBH_Smile[8] and TIF[6] datasets are shown in figure 1, respectively.

![Figure 1. Samples from Infants Datasets of CIT, RCLA&nbh_Smile and TIF.](image)

In the field of automatic infant facial expressions analysis, there are some datasets labeled for specific research. Hammal et al.[9] recorded 230,000 facial images of 13-month-old infants during positive and negative tasks, and proposed a convolutional network based on 9 AUs, which obtained an accuracy of 80.00% and a F1-score of 43.0% on their test set. Tang et al.[8] proposed a RCLA&NBH_Smile dataset with 77,039 infants’ facial images labeled by “smile” and “non-smile”, and a two-way convolutional network based on facial regions around AU6 and AU12, which achieved an accuracy of 87.16% and a F1-score of 62.54%. Sun et al. [10] collected 16,837 facial images of 24 infants under 13 months, which was divided into comfortable and uncomfortable categories. A fine-tuned DenseNet trained twice and they achieved an accuracy of 91%.

Head poses and facial occlusions are the main obstacles to facial expression recognition in real scenes. Some methods have been proposed to deal with the challenge caused by head poses or facial occlusions individually. For generative model methods, such as learning by GAN[11], filling occlusion pixels by DBNs[12] are proposed to generate new representations without occlusions. For multi-channel training methods, a multi-channel network with pose-aware was proposed for multi-view facial expression recognition [13], a two-channel network with or without occluded data [14] and a multi-channel network based on facial sub-regions and global region[15] are proposed to reduce the effect of facial occlusion. For weight adjustment method, occluded feature is weakened by facial key points[16], and weights of different facial landmarks are adjusted to adapt the head pose changes[17].
3. Dataset

3.1. Experimental Setup
The experiment was approved by the Medical Ethics Committee of Affiliated Brain Hospital of Nanjing Medical University (2017-KY089-01). All subject guardians agreed to participate in the study and signed informed consent. All infant subjects were required to complete a face-to-face experiment (3 minutes) with their mother in a behavioral observation room at Nanjing Brain Hospital. That is, under the fixed instruction, the mother interacted with the infant seated in a safety seat for 2 minutes, then stopped interacting and kept a still face for 1 minute. Three-way cameras (TC-980S) captured side-view facial behaviors of the mother and infant, frontal view facial expressions of the infant and mother frontal view facial expressions of the mother, respectively. The snapshot of the experimental scene is shown in Figure 2. The obtained video frame resolution is 1920×1080 pixels and the video stream is captured at a frame rate of 60 fps.

![Figure 2. A snapshot of the face-to-face experiment.](image)

3.2. Data Pre-Processing
The frontal-view video data of the infants were used for subsequent analysis. The pre-processing steps are as follows:
- Down-sampling: The video frames are down-sampled at a rate of 2 fps. Such a strategy is a trade-off between time cost for infant emotion labelling and infants emotion information loss.
- Human coding: Three psychologists and two non-psychologists were recruited to label infant facial expressions. A max-voting strategy was employed for multiple decisions. Positive facial expressions, including smile, surprise, laugh, interest, etc. are marked as 1, neutral facial expressions, including calmness and innocence are marked as 0 and negative facial expressions, including anxiety, crying, etc. are marked as -1.
- Image normalization: We employed the MTCNN[18] framework for face detection, then normalized the detected facial images to size of 128×128 pixels.

3.3. Dataset information
To evaluate the variation of head poses of all subjects in the dataset, we use the off-the-shelf Face++ API [19] to predict head pose. Based on the prediction results, we find that the yaw angle ranges between -72° and 72°, the pitch angle ranges between -38° and 38° and the roll angle ranges between -50° and 53°. The distributions of 3 types of head poses are shown in Figure 3, respectively.

![Figure 3. The distribution of head poses.](image)

Sample images with different head poses and facial occlusions caused by large head pose, hand, collar, table corners, etc.

![Figure 4. Samples from the IFE-3C dataset.](image)
The IFE-3C dataset contains 30 infants during 6-24 months of age, including 15 boys and 15 girls. Totally, 9,133 facial images with different head poses and facial occlusions are collected in this dataset. Some samples from IFE-3C are shown in Figure 4 and the information of the dataset is detailed in Table 1, respectively.

| Table 1. Information of IFE-3C dataset. |
|-----------------|--------|--------|--------|
| IFE-3C Dataset  | Sum    | Ratio  | Total  |
| Subjects        | Boy    | 15     | 50%    | 30     |
|                 | Girl   | 15     | 50%    |         |
| Images          | Positive | 3216 | 35.21% |         |
|                 | Neutral | 3132 | 34.29% | 9133    |
|                 | Negative | 2785 | 30.49% |         |

4. Methodology
Convolutional neural networks have achieved great success in the field of computer vision, pattern recognition. Here, we build our framework based on CNN architecture. The convolutional layers and pooling operation extract spatial local information and integrate channel information layer by layer, and gradually obtain the task-related deep feature maps with semantic information. A deep hierarchical CNN architecture has a pyramid shape, in reverse view, each point of deep feature map can be traced back to a local facial region of the inputted facial images, which carries emotion-related features for the local regions. Thus, we propose an adaptive region learning (ARL) method with using attention strategy on the higher layers in the network. An attention operation similar to [20] is introduced for our baseline CNN architecture.

For the ARL module, the front-end CNN output feature is \( F \in \mathbb{R}^{C \times H \times W} \). \( F_c \) and \( F_s \) represent the refined channel attention feature map and the refined spatial attention feature map, respectively. \( F_{arl} \) is for the adaptive region learning feature and can be calculated as follows:

\[
F_c = F \otimes \left( \sigma(MLP(F_c^{\max})) + MLP(F_s^{\max}) \right),
\]

\[
F_{arl} = F_c \otimes \sigma \left( \text{conv} \left( \left[ F_s^{\max}, F_s^{\max} \right] \right) \right) + \alpha F,
\]

where \( \otimes \) denotes element-wise multiplication, \( \sigma \) denotes sigmoid function, MLP (multi-layer perception) is two layers fully connect operation with relu function, \( F_c^{\max} \) and \( F_s^{\max} \) are based on spatial global max pooling and average pooling, respectively, the conv represents a convolution operation with kernel size of 3×3, \( F_s^{\max} \) and \( F_s^{\max} \) are based on the channel global max pooling and average pooling, respectively, \( \oplus \) denotes element-wise addition, and \( \alpha \) is used to adjust the weight of \( F \).

More irrelevant variables of facial expression, such as identity, poses, occlusion, illumination and etc., cause a larger intra-class variance in the same facial expression category. It seriously disrupts the classification performance of facial expression recognition systems. Here, we introduce the island loss[21] to learn representations with less intra-class variations while keeping larger inter-class distances. In our proposed framework, the island loss and the cross-entropy loss are both taken into consideration. The overall loss \( L_{nl} \) is as shown as follows:
\[ L_{\text{IL}} = L_{\text{softmax}} + \lambda L_{\text{island}}, \quad (3) \]

\[ L_{\text{island}} = \frac{1}{2} \sum_{i=1}^{m} \| x_i - c_{y_i} \|^2 + \lambda_1 \sum_{c_i \in \mathcal{C}} \sum_{c_j \in \mathcal{C}, j \neq i} \left( \frac{c_k \cdot c_j}{\| c_k \|_2 \| c_j \|_2} + 1 \right), \quad (4) \]

where \( \lambda \) and \( \lambda_1 \) are used to balance the two items. The former item in Eq.4 is center loss, which is used to calculate L2 distance in each class. The \( c_{y_i} \in \mathcal{R}^d \) denotes the center of the class to which the sample \( y_i \) belongs. The \( x_i \) denotes the \( i^{th} \) sample in the mini-batch. The latter one in Eq.4 represents a total of pairwise distances between class centers, which calculates cosine distance for two class centers. The \( c_k \) and \( c_j \) denote the \( k^{th} \) and \( j^{th} \) center of classes, respectively.

The integrated ARL-IL framework learns facial expression information from local facial regions adaptively under the island loss constriction and the overall framework is shown in Figure 5.

![Figure 5. The ARL-IL framework.](image)

5. Experiment and analysis

5.1. Implementation details

The system is implemented by the Keras toolbox. The \( \alpha \) in ARL method was set as 0.1, the \( \lambda \) and \( \lambda_1 \) in loss function were set as 0.005 and 0.5, respectively. Kernel parameters of convolution layers are initialized with random uniform method and bias initial values are set to 0. The SGD is used for optimization and the momentum is set to 0.95. Learning rate is initialized to 0.001 and decay rate is 1e-6 per 10 epochs. The input images are resized to 128×128 pixels. Each batch consists of 64 samples during the training stage. The strategies, including random horizontal flip, random rotation, random cropping and random width and height shift, are used for data augmentation.

5.2. Baseline Models

Among the manual feature operators, the local binary pattern (LBP) can describe local texture features, and has significant advantage of invariance in greyscale and rotation. SVM could solves high-dimensional problems through feature mapping. Extracting LBP features on original picture and build three-classes classifier based on linear SVMs, which is referred as LBP-SVM.

Fine-tuning a pre-trained model is common for network transferring, and VGG network performs well on the facial expression datasets. Fixed the weights of the first 15 layers of the VGG-Face model and changed the final output number of the fully connected layer to 3. The methods using cross entropy loss and island loss are referred as FVGG-CEL and FVGG-IL, respectively.
The comparing experiments are conducted by removing sub-mechanisms for verifying effectiveness of each component. The baseline CNN networks using cross-entropy loss and island loss are referred as CNN-CEL and CNN-IL, respectively. The CNN network with adaptive region learning using cross-entropy loss and island loss are referred as ARL-CEL and ARL-IL, respectively.

5.3 Performance Evaluation.
We conducted five-folds subject-independent cross-validation experiments on IFE-3C dataset. Randomly group data for every 6 subjects. In each fold, a group of data were used as test data, and the rest were used as training data. The classification results for each baseline method and the proposed ARL-IL are shown in Table 2 and Figure 6, respectively.

| Index | Method       | Accuracy   |
|-------|--------------|------------|
| 1     | LBP-SVM      | 73.69%     |
| 2     | FVGG-CEL     | 79.19%     |
| 3     | FVGG-IL      | 79.31%     |
| 4     | CNN-CEL      | 84.47%     |
| 5     | CNN-IL       | 85.02%     |
| 6     | ARL-CEL      | 84.68%     |
| 7     | ARL-IL       | 86.86%     |

From Table 2, we can find that the ARL-IL framework achieves the best classification performance compared with the rest methods. The CNN-based methods outperform the conventional hand-crafted feature-based method, i.e., LBP-SVM. It is one of the verifications that the learned feature representation has better discriminative ability. From the view of different losses, our introduced island loss shows better performance than the cross-entropy loss.

![Figure 6. Confusion matrix for each model.](image-url)
As can be seen in the confusion matrices in Figure 6, the recall rates for negative facial expression samples among all models are lower compared with those for neutral and positive facial expressions. In addition, negative facial expressions are more likely to be confused with neutral. Such a result could be attribute to low label stabilities for negative facial expression among all labellers.

Based on the results in Table 2 and Figure 6, we can come to the conclusion that joint learning with region-based attention mechanism and island loss shows better classification accuracy. It is an effective framework for automatic facial expression recognition of infants.

5.4 Visualization-based Analysis
We employ the Grab-CAM[22] method to compare the region of interest (ROI) for the aforementioned deep learning-based models and the feature maps are visualized in Figure 7. The LBP-based features characterize the local texture in the facial images from pixel level, which has poor ability to focus on key facial regions for facial images. From the visualization of the fine-tuned models, namely FVGG-CEL/IL, we find that the facial regions with less importance for facial expression classification are emphasized by the network. Compared with the fine-tuned models, the CNN-based models show better abilities to focus on expressions-related facial regions of infant. Especially for the ARL-IL framework, it can learn features from more precise key facial regions, such as regions around mouth, eyes and eyebrows, with discarding irrelevant regions, including ears, cheeks, and hands on the face (occlusion). In addition, as can be seen in the last column of Figure 7, the ARL-IL framework shows robustness for facial occlusion and large head poses.

Figure 7. The visualization result for each model.
6. Conclusion and future work

In this paper, an IFE-3C infant facial expression dataset is collected for investigation. An ARL-IL framework is proposed for automatic facial expression coding system, which is based on self-adaptive region learning and island loss. Experimental results proved that it can effectively realize infant facial expression recognition in natural scenes.

In the future, we will collect more data and develop a more robust infant facial expression recognition system with integrating temporal information hidden in the video. In addition, we will also apply the system for mental health monitoring of infant in real life.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, in part by the National Natural Science Foundation of China under Grant 61921004, Grant 61902064, and Grant 81971282, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242018K3DN01 and Grant 2242019K40047. In addition, thanks to Xiaoyan Ke team of Affiliated Brain Hospital of Nanjing Medical University for their contributions to the data collection work.

References

[1] Cassel T D, Messinger D S, Ibanez L V, Haltigan J D, Acosta S I and Buchman A C 2007 J. Aut. Devel. Dis. pp 122-32
[2] Camras L A and Shutter J M 2010 Emotional facial expressions in infancy Emot. Rev. pp 120-9
[3] Yin L, Wei X, Sun Y, Wang J and Rosato M J 2006 A 3D facial expression database for facial behavior research In 7th Int. Conf. on Automatic Face and Gesture Recognition (FGR06) (Southampton) pp 211-216
[4] Gross R, Matthews I, Cohn J, Kanade T and Baker S 2010 Multi-pie Image and Vision Computing 28(5) 807-813
[5] Lucey P, Cohn J F, Kanade T, Saragih J, Ambadar Z and Matthews I 2010 Conf. on Computer Vision and Pattern Recognition-Workshops (San Francisco) pp 94-101
[6] Maack J K, Bohne A, Nordahl D, Lvisdatter L, Lindahl A, Øvervoll M, Wang C and Pfuhl G 2017 The tromso infant faces database (TIF): development, validation and application to assess parenting experience on clarity and intensity ratings Frontiers in Psychology 8 409
[7] Webb R, Ayers S and Endress A 2018 The city infant faces database: a validated set of infant facial expressions Behavior Research Methods 50(1) 151-159
[8] Tang C, Zheng W, Zong Y, Cu Z, Qiu N, Yan S and Ke X 2018 Automatic smile detection of infants in mother-infant interaction via CNN-based feature learning Proc. of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data (Seoul: ACM) pp 35-40
[9] Hammal Z, Chu W S, Cohn J F, Heike C and Speltz M L 2017 Automatic action unit detection in infants using convolutional neural network 2017 7th Int. Conf. on Affective Computing and Intelligent Interaction (San Antonio) pp 216-221.
[10] Sun Y, Shan C, Tan T, Tong T, Wang W, Pourtaherian A and Peter H N 2019 Detecting discomfort in infants through facial expressions Physiological Measurement 40(11) 115006
[11] Zhang F, Zhang T, Mao Q and Xu C 2018 Joint pose and expression modeling for facial expression recognition Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition pp 3359-68

[12] Ranzato M A, Mnih V, Susskind J M and Hinton G E 2013 Modeling natural images using gated MRFs IEEE Transactions on Pattern Analysis and Machine Intelligence 35(9) 2206-2222

[13] Liu Y, Zeng J, Shan S and Zheng Z 2018 Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition 2018 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition (Xi'an) pp 458-465

[14] Xu M, Cheng W, Zhao Q, Ma L and Xu F 2015 Facial expression recognition based on transfer learning from deep convolutional networks 2015 11th Int. Conf. on Natural Computation (Zhangjiajie) pp 702-708.

[15] Li Y, Zeng J, Shan S and Chen X 2018 Occlusion aware facial expression recognition using cnn with attention mechanism IEEE Transactions on Image Processing 28(5) 2439-2450

[16] Cheng Y, Jiang B and Jia K 2014 A deep structure for facial expression recognition under partial occlusion 2014 Tenth Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing (Kitakyushu) pp 211-214

[17] Zhang T, Zheng W, Cui Z, Zong Y, Yan J and Yan K 2016 A deep neural network-driven feature learning method for multi-view facial expression recognition IEEE Transactions on Multimedia 18(12) 2528-2536

[18] Zhang K, Zhang Z, Li Z and Qiao Y 2016 Joint face detection and alignment using multitask cascaded convolutional networks IEEE Signal Processing Letters 23(10) 1499-1503

[19] https://www.faceplusplus.com.cn/

[20] Woo S, Park J, Lee J Y and Kweon I S 2018 Cbam: convolutional block attention module Proc. of the European Conf. on Computer Vision (ECCV) pp 3-19

[21] Cai J, Meng Z, Khan A S, Li Z, O'Reilly J and Tong Y 2018 Island loss for learning discriminative features in facial expression recognition 2018 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition (Xi'an) pp 302-309

[22] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D 2017 Grad-cam: visual explanations from deep networks via gradient-based localization Proc. of the IEEE Int. Conf. on Computer Vision pp 618-626