Automatic Action Unit Detection in Infants Using Convolutional Neural Network

Zakia Hammal[1], Wen-Sheng Chu[1], Jeffrey F. Cohn[1,2], Carrie Heike[3], and Matthew L. Speltz[4]
[1]Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
[2]Department of Psychology, University of Pittsburgh, Pittsburgh, USA
[3]Seattle Children's Hospital, Seattle, USA
[4]University of Washington School of Medicine, Seattle, USA

***Abstract*—Action unit detection in infants relative to adults presents unique challenges. Jaw contour is less distinct, facial texture is reduced, and rapid and unusual facial movements are common. To detect facial action units in spontaneous behavior of infants, we propose a multi-label Convolutional Neural Network (CNN). Eighty-six infants were recorded during tasks intended to elicit enjoyment and frustration. Using an extension of FACS for infants (Baby FACS), over 230,000 frames were manually coded for ground truth. To control for chance agreement, inter-observer agreement between Baby-FACS coders was quantified using free-margin kappa. Kappa coefficients ranged from 0.79 to 0.93, which represents high agreement. The multi-label CNN achieved comparable agreement with manual coding. Kappa ranged from 0.69 to 0.93. Importantly, the CNN-based AU detection revealed the same change in findings with respect to infant expressiveness between tasks. While further research is needed, these findings suggest that automatic AU detection in infants is a viable alternative to manual coding of infant facial expression.**

1. Introduction

Before the onset of speech, facial expression, vocalization, and body movement are the infant's means to communicate emotion and communicative intent and co-regulate social interaction. Adults are able to read these communication channels with varying ability. Objective measurement for research and clinical use is elusive.

Manual objective measures, such as the Baby Facial Action Coding System (Baby FACS) [28] and AFFEX/MAX [15], [16] enable frame-by-frame manual annotation of infants's facial expression. Baby FACS is an extension of FACS [11]. Like FACS, Baby FACS is a sign-based approach that detects nearly all possible anatomic movements of the face, which are referred to as action units. Individually or in combination, action units (AUs) can describe all facial expressions.

A major challenge for manual FACS and Baby FACS is the extensive time involved in training expert coders and frame-by-frame annotation (or coding) from video. FACS is labor intensive. Training to criterion on the certification test for FACS can take months, and coding a single minute of video may require an hour or more. Real-time coding for research or clinical use is not possible. Given these considerations, there has been great interest in developing approaches for the automatic recognition of FACS AUs.

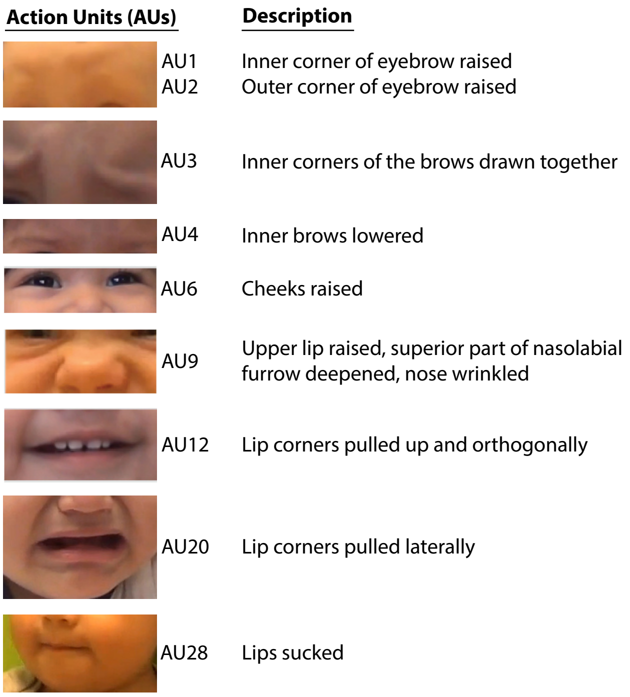
Action Units (AUs)
Description
AU1
Inner corner of eyebrow raised
AU2
Outer corner of eyebrow raised
AU3
Inner corners of the brows drawn together
AU4
Inner brows lowered
AU6
Cheeks raised
AU9
Upper lip raised, superior part of nasolabial furrow deepened, nose wrinkled
AU12
Lip corners pulled up and orthogonally
AU20
Lip corners pulled laterally
AU28
Lips sucked

Figure 1: Coded baby FACS Action Units

Automatic recognition of AUs in infants, as illustrated in Figure 1, is challenging for several reasons. Infant faces have different proportions than those of adults (*e.g.*, larger eyes and smaller jaw relative to rest of the face), fatty cheek pads are prominent, their skin is smoother and less textured, their brows fainter, and their jaw contour less distinct. They have facial actions not present or rare in adults (*e.g.*, brow knitting and certain lip movements), wrinkling is less apparent or absent, and rapid and frequent occlusions are common. Although human observers accommodate these changes, these and other sources of variation represent considerable challenges for a computer vision system. Messinger and colleagues [26], [27], [41], have had some success using person-specific active appearance models (AAM) in small numbers of action units and infants. Person-independent, generic approaches to AU detection in larger samples of infants and for a broader set of action units are needed.

Automatic detection of AUs in infants would address the needs of researchers and clinicians for automated and objective measures of infant emotion and communicative intent. For instance, automatic detection of AUs could be used to identify infants at risk for insecure attachment or developmental disorders [8]. Objective individual assessment of infant expressiveness could be used to target children with cranial nerve abnormalities for specialized interventions and to assess pre- to post-surgery changes in facial movements.

Most approaches to automatic recognition of action units in adults can be divided into two main categories: static and dynamic approaches (for a complete review, please see [9], [31], [43]). Static approaches extract facial shape and/or appearance features (*e.g.*, SIFT, HOG, LBP) at the frame-by-frame level and train off-the-shelf classifiers for the recognition of AUs at the frame level. Representative approaches include neural networks [33], Bayesian networks [35], SVMs with paradigms that are either conventional supervised learning [5], [24] or transductive learning [7], [42], boosting based approaches [1], and more recently the end-to-end convolutional neural networks [14], [45]. Dynamic approaches consider temporal information by recognizing AUs at the segment level (*i.e.*, predefined consecutive frames) or video level. Dynamic approaches detect the spatiotemporal changes in the extracted shape and/or appearance features (*e.g.*, LBP-TOP, LPQ-TOP) for the recognition of AUs. Representative approaches include temporal rule-based models [29], [36], segment-based SVMs [10], [32], hidden Markov model [23], [37], dynamic Bayesian networks [22], [34], [39], conditional random field [4], [38], [40], and bidirectional long short-term memory [17].

To address the need for objective measurement of infant facial actions, we propose a Convolutional Neural Network (CNN) based approach. CNN has become one of the most powerful machine learning methods in large-scale object detection, image classification [21], [30], and more recently AU detection [14], [17]. Other approaches to AU detection first engineer hand-crafted features and then independently train classifiers. In contrast, CNN-based networks synergistically learn representations and classifiers [21]. This integration of feature and classifier learning is a great advantage. Learned features reduce person-specific biases that hand-crafted features introduce [6], [17], and their integration with training leads to improved performance relative to standard approaches. In a recent study conducted on two large spontaneous datasets (BP4D [44] and GFT [12]), Chu and colleagues [6] found that a CNN-based approach for AU detection outperforms ones that use hand-crafter features (*e.g.*, SIFT).

The current contribution extends previous research by combining a generic person-independent tracking method with a multi-label CNN [20] for automatic detection of 9 AUs in spontaneous video of infants. The CNNs are trained end-to-end, and allow for predictions of multiple AUs at the same time. The AUs chosen are ones from across the face that are critical to expressions of positive and negative emotion. To the best of our knowledge, this is the first time the multi-label CNN has been used for detecting AUs in infants.

To elicit a range of spontaneous positive and negative facial expressions, we used two age-appropriate emotion induction tasks. We then trained a multi-label CNN for frame-level AU detection. In addition to AU detection, we used the CNN to measure facial expressiveness and tested the hypothesis that automatic and manual coding of expressiveness would yield the same changes between emotion tasks. We thus evaluate both intersystem reliability for AU detection and intersystem validity for discriminating between tasks intended to elicit positive and negative emotion.

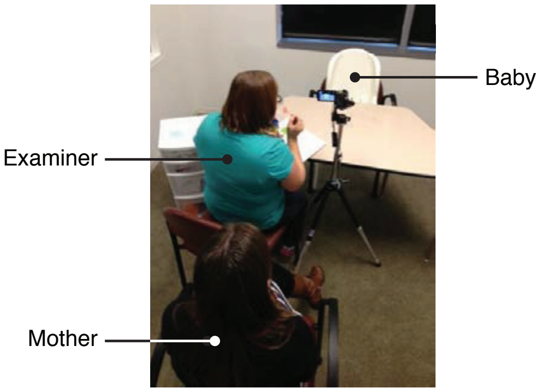

Figure 2: Configuration of observational procedure

2. Methods

2.1. Participants

Participants were 86 ethnically diverse 13-month-old infants (M = 13.06, SD = 0.62) recruited as a part of a multi-site study involving children's hospitals in Seattle, Chicago, Los Angeles, North Carolina, and Philadelphia. Two infants were African-American, 5 Asian-American, 28 Hispanic-American, 35 European-American, 1 Indian-American, 9 Multiracial, and 6 Unknown. Thirty seven were girls. Forty-nine infants were mildly affected with craniofacial microsomia (CFM). CFM is a congenital condition associated with varying degrees of facial asymmetry. Comparisons between CFM and unaffected infants will be a focus of future research. Participant recruitment and ascertainment of the full sample are not yet completed. All parents gave informed consent to the recruitment procedures.

2.2. Observational Procedures

We used two observational tasks, a positive and a negative task, where each consisting of multiple trials to elicit a range of infants' facial expressions. In the positive emotion task, an experimenter engaged the infant by blowing soap bubbles toward them and using her voice to build suspense and elicit positive engagement (*i.e.*, surprise, amusement, or interest). In the negative emotion task, the experimenter presented a toy car to the infant to generate interest and then gently took back the car and covered it with a clear plastic bin to elicit negative affect (*i.e.*, frustration, anger, or distress) [13]. Both observational tasks were repeated three times and were terminated if the infant became too upset or the mother became uncomfortable with the procedure. For both positive and negative emotion tasks, the experimenter sat across a table from the infant with the mother seated to the experimenter's side (Figure 2). Both tasks were recorded using a Sony DXC190 compact camera at 60 frames per second. Infants' faces were orientated approximately 15° from frontal, with considerable head movement.

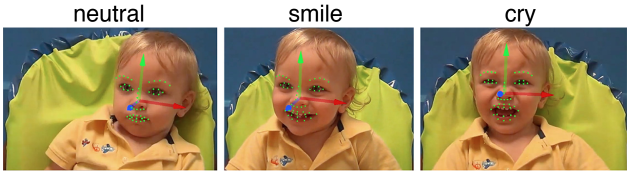

Figure 3: Examples of tracking results with head orientation pitch (green), yaw (blue), roll (red), and the 49 fiducial points during neutral expression (no AU present), smile (AU 6+12), and, cry (AU 4+6+20).

2.3. Manual AU Coding

Facial action units were coded manually using the Facial Action Coding System for Infants and Young Children (Baby FACS [28]). As noted above, Baby FACS is an extension of FACS [11] for use with infants. It includes an additional action unit in the brow region, as well as adaptations guided by variation in facial morphology and dynamics between infants and adults. Consistent with FACS, action units (AUs) correspond to discrete, minimally distinguishable actions of the facial muscles. For the current study, we sampled nine action units from the upper, middle, and lower face. The actions chosen are central to the communication of positive and negative affect (see Figure 1) [3], [25], [27], [28]. Smiles are indexed by AU 12 (lip corner puller) and cry faces by AU 20 (lip stretcher). AU 6 (cheek raiser) differentiates felt smiles from social smiles and is an intensifier of positive and negative affect. AU 1+2 (brow raiser) is key component of surprise. AU 3 and AU 4 figure in interest, concentration, and also negative affect. AU 9 (nose wrinkler) signals disgust and distress. AU 28 (lip suck) was selected as one of several candidate lip movements that are common in infants.

Three Baby FACS expert coders manually coded action units on a frame-by-frame basis for both tasks (see Figure 1). For each frame, action units were coded on a 2-level dimension (0 = absent, 1 = present) by one of three coders. Coders continuously coded the first 45 seconds of the positive emotion task and three 15-sec segments from the negative emotion task. The latter were the first 15 seconds following each of the three toy removal actions (15s × 3 repetitions = 45s total). Reliability of manual AU coding is described in Section 3.2.

2.4. Automatic AU Coding

The proposed automatic AU coding system involved three steps: 1) face tracking, 2) face registration to control for variation due to rigid head movement, and 3) detection of AU occurrence. Below we describe each step in turn.

2.4.1. Automatic Face Tracking and Registration. ZFace [18], a fully person-independent, generic approach, was used to track the registered face image. For each video frame, the tracker output the 3D coordinates of 49 fiducial points and 6 degrees of freedom of rigid head movement or a failure message when a frame could not be tracked (see Figure 3). To remove non-rigid head variation, tracked faces were registered to a reference face using similarity transform resulting in 200×200 face images. Faces were normalized for processing by the CNN Section 2.3 (see Figure 1).

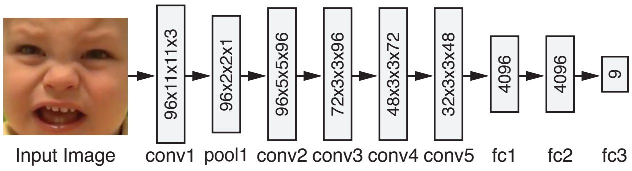

Figure 4: The architecture of the proposed 8-layer multi-label CNN used for automatic AU coding

2.4.2. Convolutional Neural Networks (CNNs) for AU Detection. An 8-layer multi-label CNN network was trained to learn a generalizable spatial representation for multiple AUs (see Figure 4). Each normalized face frame was labeled +1 if an AU is present, −1 if an AU is absent, and 0 otherwise (*e.g.*, missing tracking). The 8-layer network was modified from AlexNet [20], and composed of 5 convolutional layers, a max-pooling layer, and 2 fully-connected layers. The final fully-connected layer provided the classification output. Given a ground truth label $y \in \{-1,0,+1\}^L$ ($-1/+1$ indicates absence/presence, and 0 a missing label) and a prediction $\hat{y} \in \mathbb{R}^L$ for L AU labels, the multi-label CNN aims to minimize the multi-label cross entropy loss:

$$L_E(y,\hat{y}) = \frac{-1}{L}\sum_{l=1}^{L}[y_l > 0]\log(\hat{y}_l) + [y_l < 0]\log(1-\hat{y}_l),$$

where $[x]$ is an indicator function returning 1 if the statement x is true, and 0 otherwise. The outcome of the fc2 layer is L_2 normalized as the final representation, resulting in a 4096-D vector (see Figure 4). Due to dropout and ReLu, fc2 feature contains about 35% zeros out of 4096 values, resulting in a significantly sparse vector. The output of the multi-label CNN network (the activations of the final layer) denotes the confidence scores for the presence/absence of each AU.

The multi-label CNN network was trained with mini-batches of 512 samples, a momentum of 0.9 and weight decay of 0.005. The network was initialized with learning rate of $1e^{-3}$, which was further reduced every 5 epochs. The implementation was carried out using the Caffe toolbox [19] with modifications to support multi-label cross-entropy loss. All experiments were performed using one NVidia Tesla K40c GPU.

3. Results

We first present the results of automatic face tracking, manual FACS coding, and automatic AU coding. We then evaluate the hypothesis that manual and automatic coding yield the same pattern of findings with respect to differences in facial expressiveness between tasks intended to elicit positive and negative affect.

3.1. Reliability of Face Tracking Results

As described in Section 2.4.1, the tracker output the tracking results or a failure output (*i.e.*, missing) for each video frame. For the positive emotion task, 6.55% of

TABLE 1: Inter-observer agreement (reliability) for manual coding. Metrics include positive agreement (PA, which is equivalent to F1-score), negative agreement (NA), kappa, and accuracy (ACC).

AU	PA (F1)	NA	Kappa	ACC
1	0.55	0.83	0.60	0.80
2	0.50	0.87	0.69	0.85
3	0.45	0.86	0.70	0.85
4	0.44	0.94	0.86	0.93
6	0.65	0.89	0.80	0.90
9	0.40	0.97	0.90	0.95
12	0.61	0.94	0.85	0.93
20	0.42	0.91	0.75	0.88
28	0.53	0.95	0.86	0.93
Average	0.51	0.91	0.78	0.89

TABLE 2: Reliability of automatic AU coding on the subject-independent test set. Metrics include positive agreement (PA, which is equivalent to F1-score), negative agreement (NA), kappa, and accuracy (ACC).

AU (Base rate)	PA (F1)	NA	Kappa	ACC
1 (27.2%)	0.48	0.94	0.77	0.78
2 (22.1%)	0.33	0.94	0.77	0.73
3 (23.0%)	0.50	0.91	0.69	0.78
4 (11.7%)	0.19	0.96	0.84	0.74
6 (30.9%)	0.76	0.91	0.74	0.92
9 (7.0%)	0.26	0.98	0.93	0.77
12 (20.2%)	0.64	0.93	0.77	0.92
20 (18.4%)	0.48	0.92	0.72	0.82
28 (7.7%)	0.25	0.95	0.83	0.72
Average	0.43	0.94	0.78	0.80

frames were missing and could not be tracked. The corresponding figure for the negative emotion task was 18.28% missing frames that could not be tracked. The percentage of well tracked frames was lower for the negative emotion task than for the positive emotion task ($t = 5.78$, $p \leq 0.01$, $df = 85$).

3.2. Reliability of Manual AU Coding

To assess inter-coder agreement, two or more of the Baby FACS coders (blind to case/control status) independently annotated on a frame-by-frame basis 15s of randomly selected videos from the positive emotion and negative emotion tasks for 68 infants (30 cases and 38 controls). To quantify agreement, four reliability metrics were used: Accuracy (ACC) measures the percentage of correct predictions over total instances; the free-margin Kappa coefficient (Kappa), estimates chance agreement by assuming that each category is likely to be chosen at random [2]; and positive and negative agreement (PA and NA, respectively). PA here is equivalent to F1, the harmonic mean between recall and precision. The reliability metrics measure the extent to which FACS coders make the same judgment on a frame-by-frame basis. The selected reliability metrics capture different properties about the results. Choice of one metric over another depends on a variety of factors, including the purpose of the task, preferences of individual investigators, and the nature of the data (*e.g.*, base rates distribution, see Table 1).

3.3. Reliability of Automatic AU Detection

To guard against over-fitting (*e.g.*, [20]), we used independent train, validation, and test splits to evaluate the performance of the proposed model. Training was performed on 61 randomly selected participants (about 70% of the entire dataset), validation on 7 randomly selected participants (about 10% of the entire dataset), and test on 18 randomly selected participants (about 20%). All experiments were conducted in a subject independent manner, *i.e.*, each subject will appear only once in either training, validation or test split.

To quantify agreement between manual and automatic coding of AUs, we used the same metrics as for inter-observer coding. That is, positive agreement (or the conventional F1 measure), negative agreement, free-margin Kappa, and raw accuracy. By using the same metrics for both inter-coder agreement and agreement between manual and automatic coding, we could compare the nature of errors for each.

The performance of the multi-label CNN network on the test set is presented in Table 2. The automatic AU detection results vary with the choice of metric and track those of manual coding, as shown in Table 1. Similar to human coders, average positive agreement (F1) was moderate and negative agreement and kappa were high. For some individual AUs, results were variable. Low base rates may have attenuated the PA (F1) metrics in some cases, such as AU 4, AU 9 and AU 28. The other metrics all reveal good to high agreement between manual and automatic coding.

To explore the relation of agreement between manual and automatic AU coding, we compared kappa coefficients for both (see Figure 5). With exception of AU 1, manual and automatic coding of AUs were highly similar and agreement between methods was high. AU 1 (inner-brow raise) and AU 2 (outer-brow raiser) in infants are especially challenging given the faint contrast of the brows in infants. Overall, the results suggest that the CNN for AU detection in infants performed comparably with that of manual AU coding by human observers.

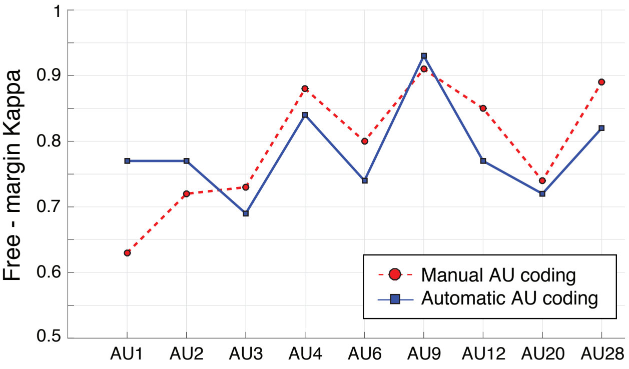

Figure 5: Comparison of agreement (*i.e.*, free-margin kappa) between manual and automatic AU coding

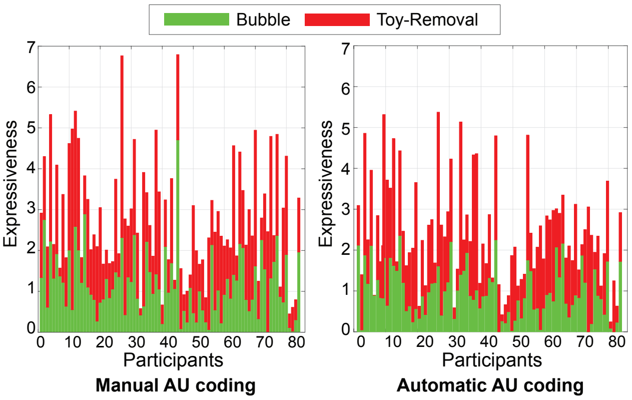

Figure 6: Comparison of expressiveness distribution between manual AU coding (left) and automatic AU coding (right) across the positive (Bubble) and negative (Toy-Removal) emotion tasks.

3.4. Validity Analyses

The analyses so far suggest moderate to high reliability between manual and automatic coding of action units. Here, we ask about their validity. Would manual and automatic coding reveal consistent differences between infant facial expressiveness in the positive emotion and negative emotion tasks? The dependent measure of interest was infant facial expressiveness. Facial expressiveness for manual coding was operationalized as the continuous sum of all manually observed AUs on a frame-by-frame basis. Similarly, for automatic coding, facial expressiveness was operationalized as the continuous sum of all automatically detected AUs on a frame-by-frame basis. The goal was to address two questions:

1) Does facial expressiveness differ between positive and negative emotion tasks?
2) Do manual and automated coding reveal the same differences in facial expressiveness between positive and negative emotion tasks?

To answer these questions, we began by automatically coding the entire dataset using the classifier from the train set. Table 3 reports the resulting agreement between manual and automatic coding of AUs. We then computed facial expressiveness scores for both.

Expressiveness Comparison: We first assessed whether expressiveness within each task differs between manual coding and automatic coding. Figure 6 shows the distribution of expressiveness measured by manual and automatic coding, respectively. Within each task, no significant effects of method were found ($F = 3.82$, $p > 0.05$ and $F = 1.04$, $p > 0.5$ for positive and negative emotion tasks, respectively). F is the F-statistic from an analysis of variance and p is the probability that the F-statistic occurred by chance alone, assuming that the null hypothesis is true.

We next compared differences in facial expressiveness between tasks. To do so, we used repeated measures analyses of variance (ANOVA), sex was entered as a between-subjects factor, and task condition as a within-subjects factor. Separate ANOVAs were used for manual and automatic AU coding. Student's paired t-tests were used for post-hoc analyses following significant ANOVAs.

TABLE 3: Agreement (Kappa) between automatic and manually coded AUs on combined training and test set.

	Kappa								
AU	1	2	3	4	6	9	12	20	28
Avg.	0.79	0.78	0.83	0.89	0.83	0.92	0.88	0.80	0.93

TABLE 4: Post-hoc paired t-test with $p < 0.01$ (following significant ANOVAs) for expressiveness. M (SD): Mean (standard deviation) of expressiveness, t: t-ratio, df: degrees of freedom.

	Positive	Negative	Paired t-test	
AU coding	M (SD)	M (SD)	t	df
Manual	1.19 (0.81)	1.62 (0.99)	-3.50	85
Automatic	0.98 (0.63)	1.46 (0.94)	-4.62	83

For both manual and automatic coding, facial expressiveness was greater during negative emotion task than during positive emotion task ($F = 8.45$, $p < 0.05$ and $F = 13.94$, $p < 0.01$, respectively). There was no difference in facial expressiveness between males and females ($F = 0.07$, $p > 0.1$ and $F = 2.01$, $p > 0.1$, for manual and automatic coding, respectively). Similarly, there was no task by sex interaction ($F = 0.82$, $p > 0.1$ and $F = 0.2$, $p > 0.1$, for manual and automatic coding, respectively).

Overall participants' facial expressiveness using automatic coding of AUs was higher during the negative emotion task compared with the positive emotion task with no difference between males and females. The study findings and resulting inferences are consistent between manual and automatic coding (see Table 4).

4. Conclusion

We have developed an end-to-end multi-label convolutional neural network (CNN) for automatic AU coding in infants, and evaluated the CNN model by detecting the presence from the absence of 9 reliably coded AUs. To our knowledge, this study possesses one of the largest amount of AUs ever coded in infants, and the largest amount of infants and of automatically coded video.

From our results, automatic coding of AUs showed moderate to strong reliability with manual coding. To assess validity of automated coding, we compared facial expressiveness in positive and negative emotion tasks. The same differences between positive and negative emotion tasks were found for both automatic and manual coding. For both, infants' facial expressiveness was higher during the negative emotion task than during the positive emotion task. The obtained results suggest that automatic measurement of facial expressiveness in infants is interchangeable with manual coding and could be a feasible option for research. Clinical applications are worth pursuing.

Automatic detection of AU and facial expressiveness in infants is in the early stages of research. The current contribution paves the way for more extensive investigations. One next direction would be to include the dynamics for AU recognition. The goal is to evaluate whether the obtained results could be improved by taking into account the dynamic changes in facial expressiveness.

Acknowledgement: Research reported in this paper was supported in part by the National Institute of Health under Award Numbers DE026513, DE022438 and MH096951. We also thank NVidia for providing a Tesla K40c GPU to support this research.

References

[1] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

[2] R. L. Brennan and D. J. Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(3):687–699, 1981.

[3] L. A. Camras, H. Oster, J. Campos, R. Campos, T. Ujiie, K. Miyake, L. Wang, and Z. Meng. Production of emotional facial expressions in european american, japanese, and chinese infants. *Developmental Psychology*, 34(4):616, 1998.

[4] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. Learning partially-observed hidden conditional random fields for facial expression recognition. In *Computer Vision and Pattern Recognition*, 2009.

[5] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan. In the pursuit of effective affective computing: The relationship between features and registration. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):1006–1016, 2012.

[6] W.-S. Chu, F. De la Torre, and J. F. Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *Automatic Face and Gesture Recognition*, 2017.

[7] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(3):529–545, 2017.

[8] J. F. Cohn, S. B. Campbell, and S. Ross. Infant response in the still-face paradigm at 6 months predicts avoidant and secure attachment at 12 months. *Development and Psychopathology*, 3(4):367–376, 1991.

[9] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568, 2016.

[10] X. Ding, W.-S. Chu, F. De la Torre, J. F. Cohn, and Q. Wang. Cascade of tasks for facial expression analysis. *Image and Vision Computing*, 51:36–48, 2016.

[11] P. Ekman, W. V. Friesen, and J. Hager. Facial action coding system: the manual. *Network Information Research Corp*, 2002.

[12] J. M. Girard, W.-S. Chu, L. A. Jeni, and J. F. Cohn. Sayette group formation task (GFT) spontaneous facial expression database. In *Automatic Face and Gesture Recognition*, 2017.

[13] H. H. Goldsmith and M. K. Rothbart. The laboratory temperament assessment battery. *Locomotor Version*, 3, 1999.

[14] E. P. Ijjina and C. K. Mohan. Facial expression recognition using kinect depth sensor and convolutional neural networks. In *International Conference on Machine Learning and Apps*, 2014.

[15] C. E. Izard. The maximally discriminative facial movement cody system, (rev. ed.). *Instructional Resources Center, University of Delaware, Newark, Delaware*, 1983.

[16] C. E. Izard, L. M. Dougherty, and E. A. Hembree. *A system for identifying affect expressions by holistic judgments (AFFEX)*. 1983.

[17] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[18] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3d face alignment from 2d videos in real-time. In *Automatic Face and Gesture Recognition*, 2015.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International conference on Multimedia*, 2014.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[21] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[22] X. Li and Q. Ji. Active affective state detection and user assistance with dynamic bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 35(1):93–105, 2005.

[23] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, 31(3):131–146, 2000.

[24] S. Lucey, A. B. Ashraf, and J. F. Cohn. Investigating spontaneous facial action recognition through aam representations of the face. In *Face recognition*. 2007.

[25] R. Matias and J. F. Cohn. Are max-specified infant facial expressions during face-to-face interaction consistent with differential emotions theory? *Developmental Psychology*, 29(3):524, 1993.

[26] W. I. Mattson, J. F. Cohn, M. H. Mahoor, D. N. Gangi, and D. S. Messinger. Darwin's duchenne: Eye constriction during infant joy and distress. *PloS one*, 8(11):e80161, 2013.

[27] D. S. Messinger, W. I. Mattson, M. H. Mahoor, and J. F. Cohn. The eyes have it: Making positive expressions more positive and negative expressions more negative. *Emotion*, 12(3):430, 2012.

[28] H. Oster. Baby facs: Facial action coding system for infants and young children. new york university; 2000. *Unpublished monograph and coding manual*.

[29] M. Pantic and I. Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(2):433–449, 2006.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[31] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, 2015.

[32] T. Simon, M. H. Nguyen, F. De La Torre, and J. F. Cohn. Action unit detection with segment-based svms. In *Computer Vision and Pattern Recognition*, 2010.

[33] Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[34] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):258–273, 2010.

[35] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10), 2007.

[36] F. Tsalakanidou and S. Malassiotis. Real-time 2d+ 3d facial action and expression recognition. *Pattern Recognition*, 43(5):1763–1775, 2010.

[37] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1):28–43, 2012.

[38] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic. Variable-state latent conditional random fields for facial expression recognition and action unit detection. In *Automatic Face and Gesture Recognition*, 2015.

[39] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *IEEE International Conference on Computer Vision*, 2013.

[40] S. Yang, O. Rudovic, V. Pavlovic, and M. Pantic. Personalized modeling of facial action unit intensity. In *International Symposium on Visual Computing*, pages 269–281, 2014.

[41] N. Zaker, M. H. Mahoor, D. S. Messinger, and J. F. Cohn. Jointly detecting infants' multiple facial action units expressed during spontaneous face-to-face communication. In *International Conference on Image Processing*, 2014.

[42] J. Zeng, W.-S. Chu, F. D. la Torre, J. F. Cohn, and Z. Xiong. Confidence preserving machine for facial action unit detection. *IEEE Transactions on Image Processing*, 25(10):4753–4767, 2016.

[43] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[44] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *Automatic Face and Gesture Recognition*, 2013.

[45] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Computer Vision and Pattern Recognition*, 2016.