Developing a Deep Learning-Based Affect Recognition System for Young Children

Amir Hossein Farzaneh[1(✉)], Yanghee Kim[2], Mengxi Zhou[2], and Xiaojun Qi[1]

[1] Department of Computer Science, Utah State University, Logan, UT 84322, USA
farzaneh@aggiemail.usu.edu,
xiaojun.qi@usu.edu

[2] Department of Educational Technology, Research, and Assessment, Northern Illinois University, DeKalb, IL 60115, USA
ykim9@niu.edu,
z1841378@students.niu.edu

Abstract. Affective interaction in tutoring environments has been of great interest among several researchers in this community, which has spurred the development of various systems to capture learners' emotional states. Young children are one of the biggest learner groups in digital learning environments, but these studies have rarely targeted them. Our current study leverages computer vision and deep learning to analyze young childrens' learning-related affective states. We developed an effective recognition system to compute the probability for a child to present neutral or positive affective state. Our results showed that the prototype was able to achieve an average affective state prediction accuracy of 93.05%.

Keywords: Emotion Recognition · Deep Learning · Computer Vision · Young Children · Learner Affect.

1 Introduction

Advances in Artificial Intelligence (AI) over recent decades have led to a growing interest in the development of AI-based approaches to education, as well as broadening the use of AI applications in education. The AIED community acknowledges the important role of affect for learning and has examined the relationship between affective states and learning gains in various domains and with various groups of learners [13] [2], [5]. Representative on-going efforts of the community include the quantitative method (called BROMP) to observe student behaviors and affective states [12], the virtual character to collect students' reported emotional states [15], and human annotators to detect learner emotions [14]. Likewise, AI-enhanced emotion recognition has also been proliferating, helping to interpret emotional states of users for both educational and therapeutic purposes. In particular, children's interactions with digital devices are rich in emotions and involve a lot of non-verbal responses [3, 18]. Designing

and implementing a system that recognizes children's emotions using a non-verbal channel such as facial expression, is needed to substantiate and expedite the analysis of children's behaviors in a digital learning environment.

A majority of conventional recognition systems are built on smaller datasets and rely on compact hand-crafted features [7, 10, 16]. As a result, they fail to incorporate the variability in facial expressions among different demographics. With the emergence of larger Facial Expression Recognition (FER) databases, modern deep learning techniques [4, 8, 9, 20] have increasingly been implemented to operate directly on image pixels to automatically extract complex features from facial images to represent emotions at different layers and handle challenging factors for emotion recognition in the wild. However, these emotion recognition engines have been built from adult face databases that represent the fine-tuned dynamics of mature faces. The performance of such predictive models for children is therefore sub-optimal.

In this study, we have leveraged deep learning techniques to predict the emotions of young children.

2 Deep Learning-Based Facial Emotion Recognition

In the current on-going study, we aim to develop a Deep Convolutional Neural Network (DCNN) based emotion recognition prototype that automates effective extraction of sophisticated facial features and thereby more accurate classification of affective states. We continuously test this prototype with kindergarten-aged children as they interact in a natural classroom environment.

To train DCNN, we use an enhanced FER (FER+) dataset [1] that contains 35,887 face images annotated with eight emotions: *neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, *fear*, and *contempt*. FER+ images have been captured under diverse illuminations, head poses, and occlusions and have a broad spectrum of demographics including people of different ages and races. On the other hand, other FER datasets containing children's faces such as National Institute of Mental Health Child Emotional Faces Picture Set (NIMH-ChEFS) [6] offer few labeled images under lab-controlled conditions with limited head poses. Since there are various and unexpected bodily movements of children in a natural classroom setting, we decided to use FER+ to train DCNN.

For the testing dataset, we use three video sequences which include children interacting with a teaching assistant. In these trial tests, we consider two affective categories: positive and neutral. For annotation, two researchers and two graduate students discussed the annotation criteria first, individually annotated ten clips of face images, and discussed the individual results until they reached consensus. They repeated this process four times for each child.

2.1 Training

We use a VGG-like standard deep architecture [1] to train an inference model on the FER+ set. This VGG model achieves close to state-of-the-art performance

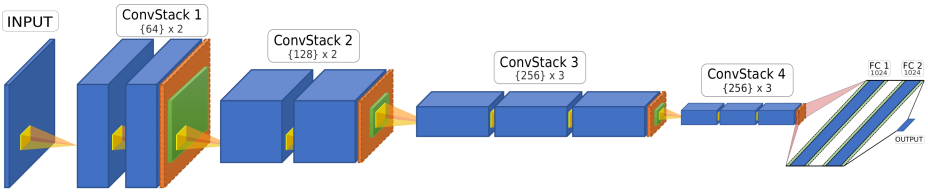

Fig. 1. VGG-12 architecture: blue, yellow, orange, and green are activation, convolution, max-pooling, and drop-out layers, respectively. Each number in the bracket reflects the depth of the corresponding convolutional layer.

while offering a simple architecture [17]. Fig. 1 presents the architecture of the VGG-like CNN, where 10 layers are convolutional layers in 4 stacks and 2 layers are linear classification layers. The final layer produces the probability of *positive* emotion and the probability of being *neutral* for each candidate.

Before the training process, we re-label *happiness* and *neutral* face images of the FER+ dataset as *positive* and *neutral* to match with the expected labels of children in a classroom setting. We then scale the original face images to the size 48×48 and feed the scaled and labeled 8,733 *positive* and 7,284 *neutral* face images to the VGG-12 network.

Training is carried out by optimizing the cross-entropy loss using the back-propagation algorithm. To provide better generalization on the test data, we perform on-the-fly data augmentation during training by applying random affine transformations [19] and random horizontal flipping on input images to generate significantly more perturbed training images.

2.2 Testing

To test the performance of our emotion recognition system, we first track each child's face in three video sequences using a CNN-based Multi-Domain Convolutional Neural Network (MDNet) tracker [11]. We then crop the face region, apply histogram equalization to increase its contrast, and pass the processed face through the emotion recognition system. The predicted emotion for each child in

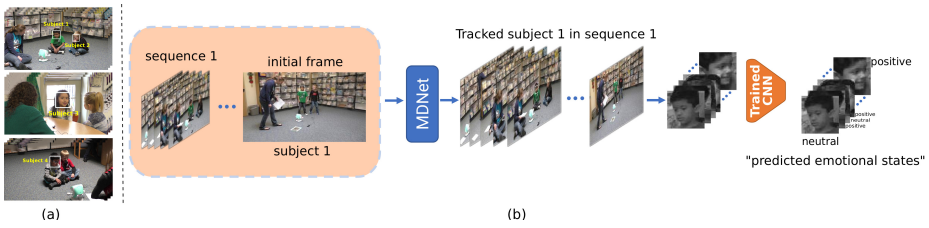

Fig. 2. Illustration of the proposed system. (a) Tracking results on three sample video frames of the test dataset. (b) Proposed emotion recognition system for subject 1 in sequence 1.

all frames is saved for evaluation. Fig. 2a presents the MDNet-tracked face image bounding boxes in yellow in three sample frames. Fig. 2b shows the proposed end-to-end emotion recognition system for subject 1 in test video sequence 1.

3 Evaluation

Table 1 summarizes the prediction accuracy results for four children in each test video sequence. We calculated the accuracy of the proposed emotion recognition system as the ratio of *the number of correct predictions* and *the total number of predictions* and evaluated prediction at the rate of one frame per second.

Due to the random parameter configuration for DCNNs during training, we evaluated our system in five trials and reported an average prediction accuracy for each child. Our proposed model achieved an average accuracy of 93.05% with above 90% of accuracy for three children and 83.46% for the fourth child. The lower rate for the fourth kid was mainly due to poor lighting in the room and more non-frontal faces.

Table 1. Emotional state testing accuracy for 4 children in 3 test video sequences

	Trials					
Target Subject	1	2	3	4	5	Average Accuracy
Subject 1|Sequence 1	89.01 %	96.70 %	87.91 %	96.70 %	94.51 %	92.97 %
Subject 2|Sequence 1	98.04 %	96.08 %	96.08 %	100.00 %	98.04 %	97.65 %
Subject 3|Sequence 2	96.96 %	98.48 %	98.18 %	98.78 %	98.18 %	**98.12 %**
Subject 4|Sequence 3	82.69 %	84.62 %	82.69 %	82.69 %	84.62 %	83.46 %
Average Accuracy	91.67 %	93.97 %	91.21 %	**94.54 %**	93.84 %	

4 Discussion and Future Work

In this study, we developed a system that recognizes young children's affective states (e.g., positive and neutral). The system achieves an average prediction accuracy of 93.05% in the five running trials with four children.

Some challenges at this stage are in line with previous research in the AIED community. These include the need for psychological and theoretical frameworks to more clearly define the categories of children's learning-related emotions. Some emotions in the highly recognized emotion database like FER+ are not specifically related to learning behaviors [2].

Lastly, the team acknowledges the need for the detection and analysis of dynamic affect (i.e., transition and reciprocity between affective states) beyond static affect to be able to fully understand learning behaviors in natural settings. This will be achieved effectively when complemented by other behavioral data that include speech, voice, and bodily movements, which leads us to continuous computational exploration to coordinate multi-modal datasets and interpret multiple sources of information meaningfully.

References

1. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 279–283 (2016)
2. Bosch, N., DMello, S.: The affective experience of novice computer programmers. International Journal of Artificial Intelligence in Education **27**(1), 181–206 (2017)
3. Breazeal, C.L.: Designing sociable robots. MIT press (2004)
4. Caramihale, T., Popescu, D., Ichim, L.: Emotion classification using a tensorflow generative adversarial network implementation. Symmetry **10**(9), 414 (2018)
5. DMello, S., Graesser, A.: Dynamics of affective states during complex learning. Learning and Instruction **22**(2), 145 – 157 (2012)
6. Egger, H.L., Pine, D.S., Nelson, E., Leibenluft, E., Ernst, M., Towbin, K.E., Angold, A.: The NIMH child emotional faces picture set (NIMH-ChEFS): a new set of children's facial emotion stimuli. International journal of methods in psychiatric research **20**(3), 145–156 (2011)
7. Kalsum, T., Anwar, S.M., Majid, M., Khan, B., Ali, S.M.: Emotion recognition from facial expressions using hybrid feature descriptors. IET Image Processing **12**(6), 1004–1012 (2018)
8. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1805–1812 (2014)
9. Matsugu, M., Mori, K., Mitari, Y., Kaneda, Y.: Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Networks **16**(5), 555 – 559 (2003)
10. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing **4**(2), 151–160 (2013)
11. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4293–4302 (2016)
12. Ocumpaugh, J., Baker, R., Rodrigo, M.: Monitoring protocol (BROMP) 2.0 technical & training manual. NY, NY: Teachers College (2015)
13. Ocumpaugh, J., Baker, R.S.J.d., Gaudino, S., Labrum, M.J., Dezendorf, T.: Field observations of engagement in reasoning mind. In: Artificial Intelligence in Education. pp. 624–627. Berlin, Heidelberg (2013)
14. Okur, E., Aslan, S., Alyuz, N., Arslan Esme, A., Baker, R.S.: Role of socio-cultural differences in labeling students' affective states. In: Artificial Intelligence in Education. pp. 367–380 (2018)
15. Ranjbartabar, H., Richards, D., Makhija, A., Jacobson, M.J.: Students' responses to a humanlike approach to elicit emotion in an educational virtual world. In: Artificial Intelligence in Education. pp. 291–295 (2018)
16. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing **27**(6), 803 – 816 (2009)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Turkle, S.: Alone together: Why we expect more from technology and less from each other. Hachette UK (2017)

19. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. pp. 435–442 (2015)
20. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A.M.: Facial expression recognition via learning deep sparse autoencoders. Neurocomputing **273**, 643 – 649 (2018)