A Two-stage Multi-modal Affect Analysis Framework for Children with Autism Spectrum Disorder

Jicheng Li, Anjana Bhat, Roghayeh Barmaki

University of Delaware
{lijichen, abhat, rlb}@udel.edu

Abstract

Autism spectrum disorder (ASD) is a developmental disorder that influences communication and social behavior of a person in a way that those in the spectrum have difficulty in perceiving other people's facial expressions, as well as presenting and communicating emotions and affect via their own faces and bodies. Some efforts have been made to predict and improve children with ASD's affect states in play therapy, a common method to improve children's social skills via play and games. However, many previous works only used pre-trained models on benchmark emotion datasets and failed to consider the distinction in emotion between typically developing children and children with autism. In this paper, we present an open-source two-stage multi-modal approach leveraging acoustic and visual cues to predict three main affect states of children with ASD's affect states (positive, negative, and neutral) in real-world play therapy scenarios, and achieved an overall accuracy of 72.40%. This work presents a novel way to combine human expertise and machine intelligence for ASD affect recognition by proposing a two-stage schema.

1 Introduction

Autism is the fastest-growing developmental disorder in the United States: approximately 1 in 54 children is on the autism spectrum (Baio et al. 2018). Individuals with ASD are characterized by having significant social communication impairments such as inefficient use of social gaze, gestures, and verbal communication (National Institute of Health 2018). Thus, individuals in the spectrum have difficulty perceiving and presenting communication cues such as emotion. Previous research has shown that play therapy can improve children's social and emotional skills and perceive their internal emotional world better (Chethik 2003). The video recordings of play therapy interventions can provide a rich source to analyze children's emotion or affect states during treatment sessions.

In this paper, we present a two-stage affect prediction method using video data (see Figure 1). We use a subset of ASD-affect dataset from (Kaur and Bhat 2019) which includes more than four different therapeutic games for chil-
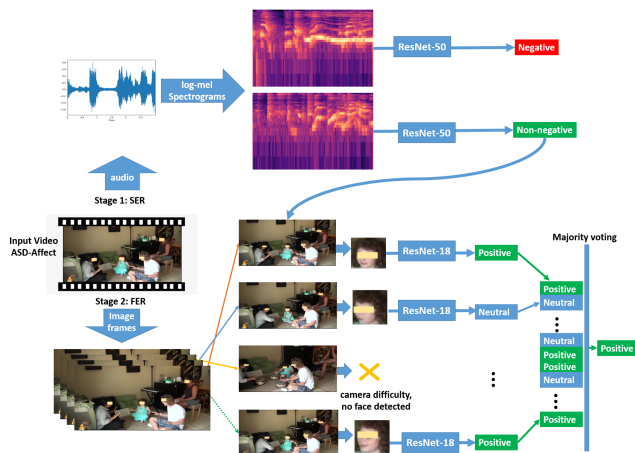
Figure 1: The workflow of our proposed affect prediction framework atop ASD-Affect dataset. Distinct emotional habits of ASD children inspire the two-stage schema, as they tend to scream and shout more in negative affect states and smile more in positive ones. Given a video input, we first classify negative data via speech emotion recognition (SER). Non-negatives will be passed to the second stage, where the model will decide whether it's positive or neutral based on facial expressions.

dren to play. Sample settings of ASD-affect are shown in Figure 2.

Emotion recognition is the process of identifying human emotion by multiple cues, including facial or spoken expressions, physiological and biological signals. Facilitated by machine learning techniques, computer vision, speech and signal processing, we can automate the process of emotion recognition. Researchers have shown that messages pertaining to feelings, affects and attitudes of interpersonal communication significantly reside in facial expressions and speech (Mehrabian et al. 1971; Dhall et al. 2012). Inspired by that, in this paper, we define the problem as automating emotion recognition for children with ASD using multi-modal inputs, especially from visual and audio signals.

Figure 2: Sample data of ASD-affect. ASD-affect is featured by rich diversities in subjects and scenarios. Children with ASD were conducting various activities under therapists' instructions. These activities include singing, dancing, drumming, yoga, and interaction with robots. The dataset consists of more than 20 hours of therapy recordings.

However, there are several challenges in the process of automatic emotion recognition on ASD-affect dataset:

Insufficient public dataset: having adequate labeled training data that include as many variations of the populations and environments as possible is important for the design of a deep expression recognition system. However, due to privacy concerns, ASD dataset, especially from children is very scarce. ASD-related multi-modal dataset, which records children's behaviors in play therapy, is even more sparse.

Domain shifts: existing methods (Doyran et al. 2019) have directly applied pre-trained model of normal people for play therapy analysis, either explicitly or implicitly under the assumption that emotional traits such as facial expressions of typically developing people and ASD children are the same or similar. We argue that such simplifications are not always appropriate given that ASD children suffer from affect and communication disorder and cannot express their emotions appropriately, especially for children diagnosed at level three (Weitlauf et al. 2014).

Data noise: many existing benchmarks for emotion recognition (Lucey et al. 2010; Valstar and Pantic 2010; Burkhardt et al. 2005) are posed and collected in a controlled laboratory setting. In contrast, our dataset was collected in an in-the-wild manner featuring various backgrounds, people, activities, and durations. Therefore, ASD-affect has lots of noise, which requires substantial data cleaning and post-processing.

Sparse labeling: unlike other multi-modal benchmark datasets (Dhall et al. 2012; Nojavanasghari et al. 2016) where the duration of data samples are in the scale of seconds (usually less than 5 seconds), examples of ASD-affect dataset may last for minutes, equivalent to lacking ground truth or introducing excessive noise to the dataset. This is because benchmark datasets were intentionally collected and labeled for autonomous recognition by machine, but ASD-affect was initially compiled and annotated to serve human experts.

Despite all these challenges, we used transfer learning, fine-tuning, and data post-processing - listed in the following sections - to prepare ASD-affect for further analysis using speech and facial emotion recognition methods.

In this paper, we propose a two-stage framework to evaluate affect states of children in play therapy scenarios using multi-modal emotion clues. This method effectively combines prior knowledge from human experts with machine intelligence. To distinguish children between three different affect states - neutral, positive, and negative - in stage 1, the model predicts whether children are in a negative state based on negative symptoms (shouting and screaming) residing in speech. In stage 2, children's emotions in positive and neutral states are recognized by distinct facial expressions. The workflow of our framework is presented in Figure 1. Our approach enables physical therapists to better and more efficiently analyze the effectiveness of play therapy interventions since human professionals require a fair amount of training to better understand the behaviors and emotional states of ASD children. This method can be further applied to data annotation and label verification for other ASD datasets, as actions of ASD children resemble relatively well.

This paper is organized as follows. We first summarize related works in emotion recognition and play therapy analysis in Section 2. Section 3 describes the method we proposed, followed by experiment and results discussion presented in Section 4, 5 and 6. Lastly, section 7 outlines the conclusion and future steps for this research.

2 Related Work

Multi-modal Emotion Recognition

Emotions are convoluted psychological states composed of several components: personal experience, physiological, behavioral, and communicative reactions. There are two mainstream emotion representations: discrete model (Ekman 1994) and dimensional model. In this paper, we used discrete emotion models.

Emotions can be carried in various modalities of inputs. Mehrabian shows that 55% of messages pertaining to feelings and attitudes of interpersonal communication is in facial expressions (Mehrabian et al. 1971). Besides, Dhall suggests that audio modality can bring extra gain in emotion recognition accuracy (Dhall et al. 2012). Thus, multi-modal emotion recognition approaches usually outperform unimodal ones. Two main sub-sets of multi-modal emotion recognition models are facial expression recognition (FER) and speech emotion recognition (SER), which are also the main focus of this work.

Facial Expression Recognition FER systems can be divided into two main categories based on the feature representations: static and dynamic. In static-based methods, the feature representation is encoded with only spatial information from a single image frame. In contrast, dynamic-based approaches consider temporal relations among contiguous frames in the input facial expression sequence (Li and Deng 2018). Li proposed a bi-modality method (Li et al. 2019),

where convolutional networks (CNNs) were used to recognize static facial expressions while a bi-direction long short term Memory (Bi-LSTM) was employed to learn dynamic facial expression sequences extracted by CNNs. Liu also embodied facial landmarks in the FER system (Liu et al. 2018). However, these works were conducted on benchmark datasets (Dhall et al. 2012) where sequential relation of images is well-preserved so that sequential methods are able to function. Conversely, our ASD dataset was recorded in the natural or in-the-wild settings; so we could only use a static-based method to classify facial expressions in each frame, without considering temporal information.

Speech Emotion Recognition Speech is a rich, dense form of communication that can convey information effectively. There are two classical ways to extract emotional features from speech. First is to obtain low-level discriminator features of speech, such as Mel-frequency cepstral coefficients (Yeh, Lin, and Lee 2019; Yoon et al. 2020). Another way is to convert audio to spectrograms then use CNNs as feature extractors (Zhang et al. 2018; Zhao, Mao, and Chen 2019). In this paper, we use spectrograms as audio representations.

Play Therapy Analysis

Play therapy is an approach to psychotherapy where a child is engaging in play activities. Doyran and colleagues (Doyran et al. 2019) proposed a visual and text-based framework to track the affective state of a child during a play therapy session. However, audio modality was less explored in their work, and categorical representations of facial expressions needed more investigation. Bangerter investigated the spontaneous production of facial expressions of individuals with ASD as a response to entertaining videos (Bangerter et al. 2020). It turned out that individuals with ASD showed less evidence of facial action units relating to positive facial expression than typically developing children. Due to small face sizes and low resolution of ASD-affect dataset, using facial action units approach in the current work was not feasible, but we are looking into it in future.

3 Method

In this paper, we propose an open-source two-stage multimodal framework to predict children's affect states in play therapy leveraging visual and audio information[1]. First, we distinguished negative videos from non-negative ones (neutral and positive) using spectrograms generated from audio. Next, to differentiate between positive and neutral videos, we used static-based facial expression recognition methods. The workflow of this method is illustrated in Figure 1.

Two-stage Schema

Our data assessment on ASD-affect inspires the two-stage approach. Children who negatively and passively participated in play interventions tend to shout and scream more often, and such characteristic is manifest in speech. However,

---

[1]The source code is available to download at GitHub: https://github.com/Li-Jicheng/Autism-Affect-and-Emotion-Recognition.

there are no significant diversities in speech emotion between neutral and positive recordings. Instead, children are smiling when positively engaged in therapy, while their facial expressions remain neutral more often in neutral states. Therefore, we chose to leverage the variance in facial expressions to distinguish between positive and neutral data.

Stage 1: Negative vs Non-negative

Since distinct speech emotions exhibit different patterns in the energy spectrum, to capture emotion features from speech, we selected log-Mel spectrograms which have been effective in speech emotion recognition tasks in the past (Zhao, Mao, and Chen 2019; Zhang et al. 2018; Chen et al. 2018). A spectrogram is a visual representation of the spectrum of a signal's frequencies as it varies with time. It is a graph with two geometric dimensions: time and frequency. The amplitude of a particular frequency at a particular time is represented by the intensity or color of each pixel in the spectrogram. A Mel-spectrogram is a spectrogram where the frequencies are converted to the Mel scale - a perceptual scale of pitches assessed by listeners to be equal in distance from one another (Stevens, Volkmann, and Newman 1937). We used the logarithmic form of Mel-spectrogram to better reflect emotions, since humans perceive sound in a logarithmic scale (Venkataramanan and Rajamohan 2019).

Stage 2: Neutral vs Positive

As noted earlier, due to image resolution constraints, temporal information was not well preserved as adjacent frames were discarded frequently in the data cleaning stage, causing sequential models to fail to converge. Therefore, we needed to use static-based methods that solely depend on one frame to predict facial expression. We choose ResNet-18 (He et al. 2016) with a decreased input size to better fit the average face sizes detected in ASD-affect. We pre-trained the model using EmoReact (Nojavanasghari et al. 2016), a multi-modal emotion dataset of children and fine-tuned it on ASD-affect dataset.

4 Experiment

Data Processing

ASD-affect Dataset Bhat and colleagues proposed that use of embodied, multisystem interventions can enhance various social communication, perceptuo-motor, and cognitive-behavioral impairments of children with ASD (Kaur and Bhat 2019). They have studied the effects of various embodied creative interventions, including the themes of robotic, musical, physical activity, yoga, and dance therapy interventions for children with ASD. The video recordings of such interventions, known as the ASD-affect dataset, have provided a rich source for analyzing children's affect states in play therapy. In this paper, we used a subset of ASD-affect from six children. Sample data of ASD-affect are shown in Figure 2.

Data Reconstruction Originally, there were eight different types of labels in the ASD-affect: neutral, interested, positive, positive and talking, odd positive, runs away, camera difficulties, and negative. For our work, we reconstructed

the dataset, and excluded - runs away and camera difficulties and odd positive labels - or merged some labels - interested, positive, positive and talking were all considered as positive labels. After this reconstruction step, we had a total of 471 clips from six children in three classes of positive (68 clips), neutral (384 clips), and negative (19 clips). Clips lengths were varied. See Figure 3 for reconstructed data distribution.
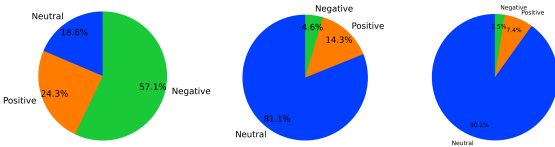

Figure 3: A comparison of ASD-affect and benchmark dataset AFEW (Dhall et al. 2012). Left: sample distribution of reconstructed AFEW. We relabeled fear, disgust, angry and sad as negative. Middle: data count distribution of ASD-affect. Right: data duration percentage of ASD-affect.

Log-Mel Spectrograms We first extracted audio tracks from video recordings. Audio files were stored in Waveform Audio File Format to retain high fidelity. We then applied noise reduction to audio files and removed silent utterances. Afterwards, we split each audio file into equal-length segments of 3 seconds, and zero-padding was applied to the utterances whose duration is less than 3 seconds. We set this sequence length since the average audio duration in selected benchmark datasets for SER was 3 seconds (Burkhardt et al. 2005; Livingstone and Russo 2018). After that, log-Mel spectrograms were generated from each audio segment using *librosa* toolkit (McFee et al. 2015). We set the Fast Fourier Transform (FFT) window length and hop length to 2,048 and 512, respectively. 64 Mel bands were used in the spectrogram generation. A total of 9,968 log-Mel spectrograms were generated, including 134 negative samples and 9,834 non-negative samples. We then used downsampling on non-negative spectrograms to even out data, and reduce data imbalance. In both the training and testing phase, all log-Mels were normalized by the global mean and standard deviation of the training set. All spectrograms were resized to 224×224 to match the network's input size.

Facial Images We first extracted image frames from raw video clips at a specific sampling rate. Considering that duration of neutral clips were typically longer than positive ones in ASD-affect, we set the sampling rate to 3 frames per second (FPS) for positive video clips and 1 FPS for the neutral to stratify data proportionally. Then we used MTCNN (Zhang et al. 2016) to detect human faces in each frame. We selected 1,756 template faces of children (about 2% of the total detected faces) to create a facial expression database for ASD-affect, consisting of 1,159 neutral and 706 positive faces. Each selected face was manually labeled as either neutral or positive based on facial expressions. The children face dataset served as training, validation, and test set for the FER model used in the second stage via 5-fold cross-validation. We used random crop, rotation, shifting, illumination adjust-

ment, and normalization techniques for data augmentation and noise reduction. Before training, all facial images were resized to 48×48 offline, then random cropped to 44×44 on-the-fly during training. In testing, faces were directly resized to 44×44.

Detected faces may belong to children, other persons in the scene or due to noise. To localize children's faces properly, we leveraged the children's face dataset to create a face embedding database, where each face was encoded as a 128-dimensional vector. Whenever a new face is encountered, we can compare its embedding with the embedding database we have established to find matches. A 'match' is defined as the cosine similarity between the new face embedding and a known face embedding is less than a given confidence threshold. Only matched faces were used for predictions, and unmatched faces were excluded.

Speech Emotion Recognition

Since the whole dataset is imbalanced, where negative video clips are much less than non-negative ones, we applied weighted sampling to enhance negative samples' occurrence while working with spectrograms. We chose a batch size of 32, and the network was trained for 25 epochs. We used Adam (Kingma and Ba 2014) optimizer, and the learning rate was set to 0.001.

Facial Expression Recognition

The training set was the selected ASD children's faces, as mentioned above. We set the batch size to 64 for training while the total training epochs was 25, and chose Adam (Kingma and Ba 2014) optimizer with an initial learning rate of 0.001. The learning rate was decreased by a factor of 0.1 every 20 epochs. Unlike the training phase, in testing, input images were captured every five frames from videos on-the-fly. Note that inputs in testing were not face crops but image frames, indicating that a face detector has to be applied to capture human faces from frames. MTCNN was then applied to test images to capture human faces. Detected faces were compared with the established children's face database. Once children's faces were matched and located, a trained model predicted children's facial expressions, and such predictions were considered valid votes. Frames were discarded if no target children's faces were detected, including no faces or only faces from others (e.g., therapists or parents). At the end of each video, for all the valid votes, if the portion of positive predictions exceeds a certain threshold, the whole video is predicted as positive, otherwise neutral. In this experiment, we set the threshold to 0.5, equivalent to majority voting. The workflow of the test phase is explained in Algorithm 1.

5 Results

We used 5-fold cross-validation to report findings from our participant videos (recordings from two children were merged together due to small number of video clips, totalling five batches of data from six children. Since we had

Algorithm 1 Stage_2_testing(input video v, face detector $face\_det$, children face embedding $child\_embeds$, classifier $model$, threshold T)

1: $vote\_list = []$ ▷ Initialize a list to store valid votes
2: Load video v, capture one image as input every 5 frames. All captured images are stored in $frame\_list$
3: **for** img in $frame\_list$ **do**
4:     Detected face list $fl = face\_det(img)$
5:     **for** each face f in fl **do**
6:         **if** $f \in child\_embeds$ **then** ▷ Locate children
7:             $pred = model(f)$
8:             $vote\_list.append(pred)$
9:         **end if**
10:     **end for**
11: **end for**
12: $pos\_ratio = \frac{count(p_i == Positive) \text{ in } vote\_list}{vote\_list.length}$
13: **if** $pos\_ratio >= T$ **then** ▷ Majority voting
14:     return Positive
15: **else**
16:     return Neutral
17: **end if**

---

imbalanced classes, in addition to accuracy, we reported recall, F1 score, G-mean value, and ROC-AUC score for more in-depth analysis.

Stage 1: Negative vs. Non-Negative

We achieved an accuracy of 94.48% and an F1 score of 0.97. The recall of negative and non-negative labels is 68.42% and 95.57%. Besides, the G-mean value and ROC-AUC score are 0.92 and 0.93, respectively. The confusion matrix of stage 1 is shown in Figure 4. The classification results from recordings of each participant is shown in Figure 5.
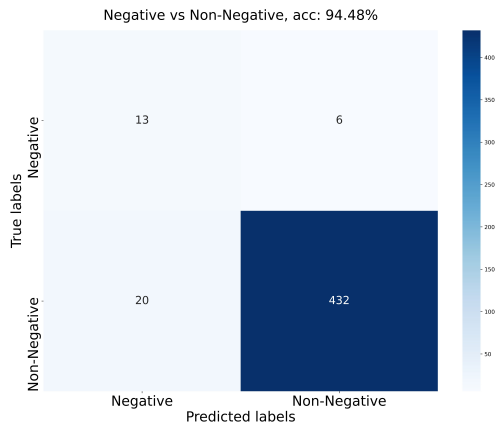

Figure 4: Confusion matrix for prediction of negative and non-negative videos extracted from audio modality and SER.
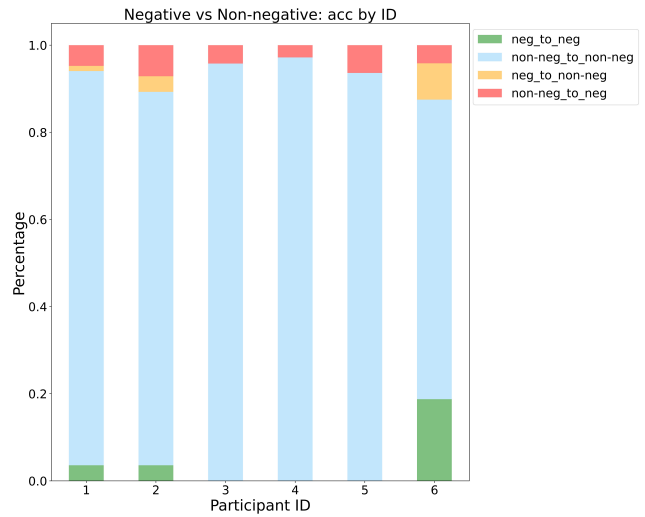

Figure 5: Classification results per participant in stage 1, negative vs. non-negative. Y-axis shows the percentages of predicted labels for video clips from every participant (data is normalized based on total video clips from every participant, because video counts vary significantly). The legend format is: true label to predicted label (e.g. $neg\_to\_non-neg$ means a negative video is misclassified as a non-negative one).

Stage 2: Neutral vs. Positive

We reached an overall accuracy of 75.93%, where recall for neutral and positive class was 78.29% and 63.24%, respectively. Confusion matrix is shown in Figure 6. Also, F-1 score is calculated as 0.79. The results from each participant's videos are shown in Figure 7.

Overall Accuracy

According to the confusion matrix for all three classes (Figure 8), we have correctly classified 285 neutral videos, 43 positive videos, and 13 negative ones. Adding them up, we had 341 out of 471 precisely classified clips, leading to an overall accuracy of 72.40% and an F1 score of 0.75.

6   Discussion

Overall, our method achieves an acceptable performance in both stages. However, there is a noticeable accuracy gap between stage 1 and stage 2 and between dominant and non-dominant classes of each stage. If we consider our problem a two binary classification problems –stage one with negative vs non-negative samples, and stage two of non-negative samples classified to positive and neutral samples– for non-dominant labels in both stages, which are negatives in stage 1, and positives in stage 2, the recall rates are very comparable: 68.42% and 63.24% respectively. On the other hand, for both of dominant classes in stages 1 and 2, the recall for non-negatives is significantly higher than neutral labels (95.57% vs. 78.29%). This may because speech emotion
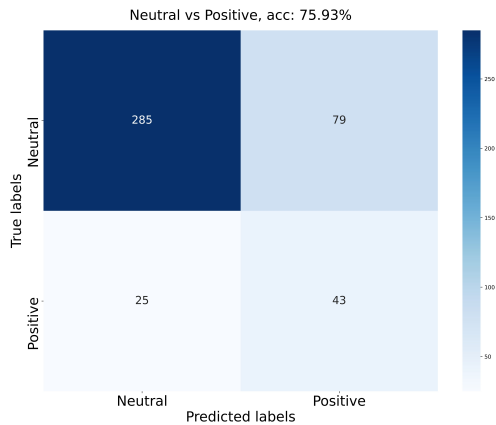
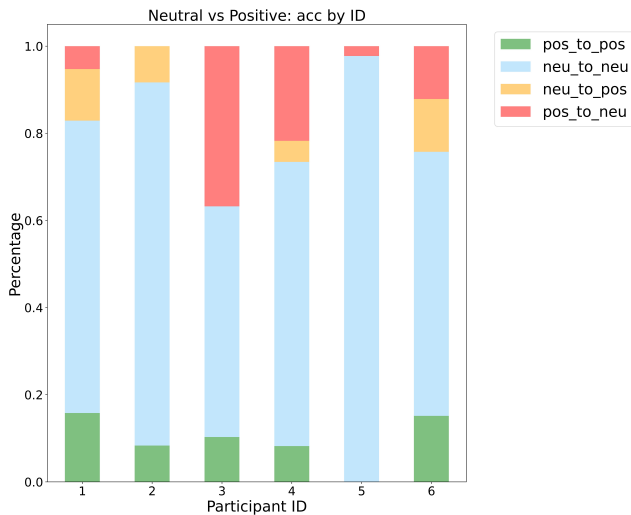Figure 6: Confusion matrix for prediction of positive and neutral videos extracted from FER.

Figure 7: Classification results per participant in stage 2, neutral vs. positive. Y-axis shows the percentages of predicted labels for video clips from every participant (data is normalized based on total video clips from every participant, since video counts vary significantly). The legend format is: true label to predicted label (e.g. neu_to_pos means a neutral video is misclassified as a positive one). Blue and green represent correct predictions for positive and neutral video clips, while red and orange stand for miss-classified ones. The more considerable sum of green and blue areas are, the better accuracy our model achieves. This figure illustrates that our model generalizes well among all individuals in ASD-affect in distinguishing neutral and positive videos.
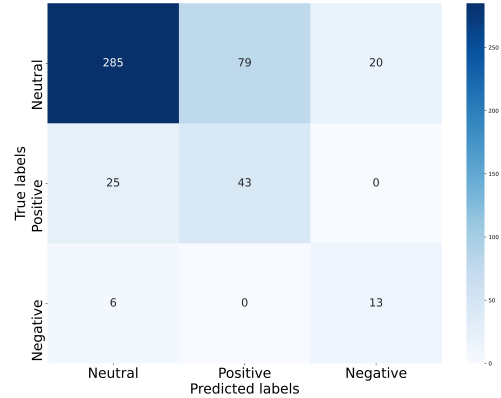

Figure 8: Confusion matrix for three classes.

features, such as shouting and screaming, are more distinct and recognizable and describe negative videos better. Moreover, the distinction of positive from neutral labels in stage 2 was very tough even for subject matters experts due to data noise and low video resolution. As such, SER performed relatively better than FER for our ASD-Affect dataset.

7 Conclusion

This paper proposed a novel framework for automatic emotion recognition of children with ASD using multi-modal information (facial and speech emotion), providing a baseline model to affect states analysis in play therapy. This work also has implications on automated affect annotation for play therapy video recordings. Besides, the framework leverages human expertise to a great extent by proposing a two-stage schema, a novel way to combine human knowledge and machine intelligence in ASD-related research.

We anticipate expanding this project in the future in multiple directions. We aim to offer a semi-automated annotation framework to assist subject-matter experts swiftly annotating the recordings from children with autism. We discussed some challenges of the pre-recorded videos in our dataset, especially the low-resolution issue. To overcome the problem, we plan to collect more audio-visual data with higher resolution to deploy other FER techniques, including sequential and action-unit based approaches mentioned in the paper. Furthermore, deficits in mutual gaze, and shared gaze is also known as a strong predictor of autism among children (Zhao et al. 2017), which we are interested in investigating in the future, as a next line of our previous research (Guo and Barmaki 2020) on automatic detection of mutual gaze among adults.

8 Acknowledgments

findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

Baio, J.; Wiggins, L.; Christensen, D. L.; Maenner, M. J.; Daniels, J.; Warren, Z.; Kurzius-Spencer, M.; Zahorodny, W.; Rosenberg, C. R.; White, T.; et al. 2018. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries* 67(6): 1.

Bangerter, A.; Chatterjee, M.; Manfredonia, J.; Manyakov, N. V.; Ness, S.; Boice, M. A.; Skalkin, A.; Goodwin, M. S.; Dawson, G.; Hendren, R.; et al. 2020. Automated recognition of spontaneous facial expression in individuals with autism spectrum disorder: parsing response variability. *Molecular autism* 11(1): 1–15.

Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W. F.; and Weiss, B. 2005. A database of German emotional speech. In *Ninth European Conference on Speech Communication and Technology*.

Chen, M.; He, X.; Yang, J.; and Zhang, H. 2018. 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition. *IEEE Signal Processing Letters* 25(10): 1440–1444.

Chethik, M. 2003. *Techniques of child therapy: Psychodynamic strategies*. Guilford Press.

Dhall, A.; Goecke, R.; Lucey, S.; and Gedeon, T. 2012. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia* 19(3): 34–41.

Doyran, M.; Türkmen, B.; Oktay, E. A.; Halfon, S.; and Salah, A. A. 2019. Video and Text-Based Affect Analysis of Children in Play Therapy. In *2019 International Conference on Multimodal Interaction*, ICMI '19, 26–34.

Ekman, P. 1994. Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. .

Guo, Z.; and Barmaki, R. 2020. Deep neural networks for collaborative learning analytics: Evaluating team collaborations using student gaze point prediction. *Australasian Journal of Educational Technology* 36(6): 53–71.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kaur, M.; and Bhat, A. 2019. Creative Yoga Intervention Improves Motor and Imitation Skills of Children With Autism Spectrum Disorder. *Physical Therapy* 99(11): 1520–1534.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Li, S.; and Deng, W. 2018. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348* .

Li, S.; Zheng, W.; Zong, Y.; Lu, C.; Tang, C.; Jiang, X.; Liu, J.; and Xia, W. 2019. Bi-Modality Fusion for Emotion Recognition in the Wild. In *2019 International Conference on Multimodal Interaction*, ICMI '19, 589–594.

Liu, C.; Tang, T.; Lv, K.; and Wang, M. 2018. Multi-Feature Based Emotion Recognition for Video Clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, 630–634.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13(5): e0196391.

Lucey, P.; Cohn, J. F.; Kanade, T.; Saragih, J.; Ambadar, Z.; and Matthews, I. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94–101.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.

Mehrabian, A.; et al. 1971. *Silent messages*, volume 8. Wadsworth Belmont, CA.

National Institute of Health. 2018. Autism Spectrum Disorder. https://www.nimh.nih.gov/health/publications/autism-spectrum-disorder/index.shtml.

Nojavanasghari, B.; Baltrušaitis, T.; Hughes, C. E.; and Morency, L.-P. 2016. EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, 137–144.

Stevens, S. S.; Volkmann, J.; and Newman, E. B. 1937. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America* 8(3): 185–190.

Valstar, M.; and Pantic, M. 2010. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 65. Paris, France.

Venkataramanan, K.; and Rajamohan, H. R. 2019. Emotion Recognition from Speech.

Weitlauf, A. S.; Gotham, K. O.; Vehorn, A. C.; and Warren, Z. E. 2014. Brief report: DSM-5 "levels of support:" a comment on discrepant conceptualizations of severity in ASD. *Journal of autism and developmental disorders* 44(2): 471–476.

Yeh, S.; Lin, Y.; and Lee, C. 2019. An Interaction-aware Attention Network for Speech Emotion Recognition in Spoken Dialogs. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6685–6689.

Yoon, S.; Dey, S.; Lee, H.; and Jung, K. 2020. Attentive modality hopping mechanism for speech emotion recogni-

tion. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3362–3366. IEEE.

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.

Zhang, S.; Zhang, S.; Huang, T.; and Gao, W. 2018. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Transactions on Multimedia* 20(6): 1576–1590.

Zhao, J.; Mao, X.; and Chen, L. 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control* 47: 312–323.

Zhao, S.; Uono, S.; Yoshimura, S.; Kubota, Y.; and Toichi, M. 2017. Atypical gaze cueing pattern in a complex environment in individuals with ASD. *Journal of autism and developmental disorders* 47(7): 1978–1986.