2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Context-based Dataset for Analysis of Videos of Autistic Children

Sk Rahatul Jannat[1], Heather Agazzi[2] and Shaun Canavan[1]
[1] Department of Computer Science and Engineering, University of South Florida, Tampa, USA
[2] Department of Pediatrics, University of South Florida, Tampa, USA

***Abstract*— Autism affects as many as 1 in 44 youth, with many higher-functioning children not diagnosed until school-age or later. Currently, diagnosing autism is a lengthy process often delivered in varying settings (i.e., context) by a multi-disciplinary team, where the result can include subjective bias. Automatic approaches that can help professionals with diagnosis can result in earlier and quicker diagnosis. To help facilitate development of automatic approaches, we present a new, context-based dataset for analysis of videos of autistic children. The data was collected from 14 children, using the gold-standard RITA-T evaluation. Along with the dataset we also provide a baseline, context-based approach for classification of these videos. The baseline shows encouraging results that context matters for classification, and we also detail findings about which features (face, body, or gaze) are most encouraging within those contexts.**

I. Introduction

Autism affects as many as 1 in 44 youth [1], with many higher-functioning children not diagnosed until school-age or later [51]. Significant impairment in social-communication, adaptive, and school functioning is common, and compared to other types of pediatric psychopathology, autism is particularly severe and longstanding [5]. *While the importance of early diagnosis and intervention (e.g., 12-48 months) is well established [10], [14], [51], the average age of diagnosis of autism is currently between 4 and 5 years [51]*, greatly delaying access to intervention during a critical developmental period [10] associated with the most significant treatment-related gains in cognitive, language and adaptive skills [43]. Research has documented the importance of studying these behaviors, as well as the need to understand the development of typically developing children to find any deviations [36].

There is a wealth of information in the medical literature on diagnosing autism. For example, Frazier et al. [16] found a reliable pattern of gaze abnormalities for autistic individuals, suggesting a problem with selecting socially relevant versus irrelevant information. Helminen et al. [19] suggested autistic children lack the perceptual detection advantage of direct gaze and fail to respond to gaze with enhanced physiological orienting. Weiss et al. [45] looked at facial action units and found the autistic subjects had less differentiated facial responses, showing that reduced facial expressivity is characteristic of autistic subjects. Bishop et al. [4] showed that adults that were previously diagnosed with a language condition as children would have been given a diagnosis of autism with contemporary approaches to diagnosis, showing language is an important modality for diagnosis of autism. Park et al. [34] conducted a review on our understanding of autism, detailing that along with facial expression and gaze, body, head pose, and gestures are associated with repetitive behaviors such as body rocking and hand flapping.

Recent works for classifying autistic subjects often focus on brain scan data, and computer vision techniques (e.g., expression and gaze). For example, Kong et al. [25] looked at the connectivity between regions of interest (ROI) among brain scans. Bi et al. [3] used Random Support Vector Machine Clustering on resting-state functional MRI data. Similarly, Niu et al. [33] looked at the same data as Bi et al. They developed a multichannel deep attention neural network which showed better performance on the data. Seminal work by Rehg et al. [35] collected video, audio, physiological recordings, scoring datasheets, and parent questionnaires to analyze the social behavior of children. They presented baseline analyses of decoding social behavior of children, which showed they could reliably predict child social data with multiple modalities (although each modality was investigated individually). The research also resulted in a new multimodal dyadic behavior dataset consisting of adult-child social interactions. Rudovic et al. [37] showed that the intensity of engagement could be predicted in autistic children. This was done using facial expression and deep learning. Drimalla et al. [12] proposed a multimodal approach to detecting autism and showed that the combination of audio, video, and gaze features could increase the accuracy of detection. Jiang et al. [22] classified autism using facial expression and eye gaze. Samad et al. [41] analyzed facial action units to show that autistic subjects frowned more and had low correlations in temporal activations compared to their typically developing peers. Li et al. [28] used action units, facial expression, arousal, and valence to classify autism automatically using a convolutional neural network. Along with these works, there are interesting works on analyzing family home movies of autistic infants. Saint-Georges et al. [39], conducted a literature review of this. They detailed 18 studies showing the signs that differentiate autistic children from those with developmental delays. This includes, but is not limited to, less looking at others, lower eye contact quality, and less positive facial expressions. This also motivates the current study to use gaze and face modalities.

There has also been recent significant work in video understanding, which we propose to do here for autistic children. Wang et al. [44], analyze videos of human subjects to extract real-time 3D pose estimation. They propose to predict both real and virtual bones simultaneously. Luo et al. [30], investigated referring video object segmentation. They proposed Semantic-assisted Object Cluster which combines video and text data for cross-modal alignment. Along with video understanding, data is also an important aspect of analyzing video of autistic children. Unfortunately, the datasets

979-8-3503-9494-8/24/$31.00

used are often not publicly available [37], [12] due to the sensitive nature of the data. Along with this, the tasks used for data collection are often not directly related to autism (e.g., general tasks meant to elicit emotion [22]). While there are some publicly available datasets, they also have limitations such as general tasks not specifically related to autism [50], or the publicly available dataset from Pandey et al. where the control class was picked from another dataset [26] with different tasks. Considering this, our proposed dataset has the following extensions to state of the art. (1) The RITA-T autism screening test [9] was used for data collection; (2) Both autistic and control subjects took part in the same RITA-T screening; and (3) The videos are context-based (each separate RITA-T task). To summarize, the contributions of this work are 3-fold:

1) A new, publicly available, context-based dataset of children with low and high risk for autism is proposed.
2) A context-based baseline approach to classifying videos of children with risk of autism is proposed.
3) An analysis of how different modalities impact classification is given.

II. Data and Collection

A. Participants and Recruitment

Participants included 14 toddlers ages 20-36 months (M = 27 months) who participated in a pilot study of the impact of multimodal data on autism risk diagnosis. Participants were recruited from two pediatric clinics in one medical building at a large university medical center in Tampa, FL. The first clinic was the local Part C early intervention program for children birth-3 years with documented developmental delays (DD). Children in this clinic were either participating in an initial eligibility evaluation or a comprehensive evaluation for autism. Children presenting for autism evaluation had already been found eligible for the Part C program at a prior appointment due to documented developmental delays. The second clinic was a primary care clinic to which children presented for sick or well-child visits. Study participants and their caregivers presenting to one of the two previously described clinics were approached by the attending healthcare provider at the conclusion of their scheduled visit as long as they met all inclusion criteria: 1) toddlers ages 20-36 months; 2) child must be able to sit up and move arms in response to the presentation of toys; 3) Child must have adequate vision/hearing so their responses to examiner social bids/toys can be assessed; and 4) Caregiver must be 18 years or older. The provider showed the caregiver a study flyer and briefly explained the study goals and asked if the caregiver would like to hear more information about the study. Data were not collected on the number of families who declined further study information. If the caregiver was interested in further details, the provider texted a study co-PI (psychologist) to come and get the family. The psychologist accompanied the family to a private evaluation room and explained the study. The psychologist was not blinded to the child's appointment type and thus, was aware if the child had presented for an autism evaluation. However, the psychologist did not have the evaluation results prior to the screening. Of the 14 caregivers who wanted more information, 100% consented to the study.

Over seven screening dates, we consented and enrolled 6 children with DD + autism from the Part C program, 6 children with DD only (5 from Part C and 1 from primary care), and 2 who were typically developing from the primary care clinic. Among them, 57% were Hispanic/Latino, 29% identified as non-white, and 28% were female suggesting oversampling among diverse individuals and females who have autism+DD. While the sample size was small, pilot studies serve to identify and address issues that could occur with respect to future study conceptualization, study design, data collection, data management, and data analysis [31]. In addition, the difference in effects we find in pilot studies can be used to plan future power and sample size of a larger trial.

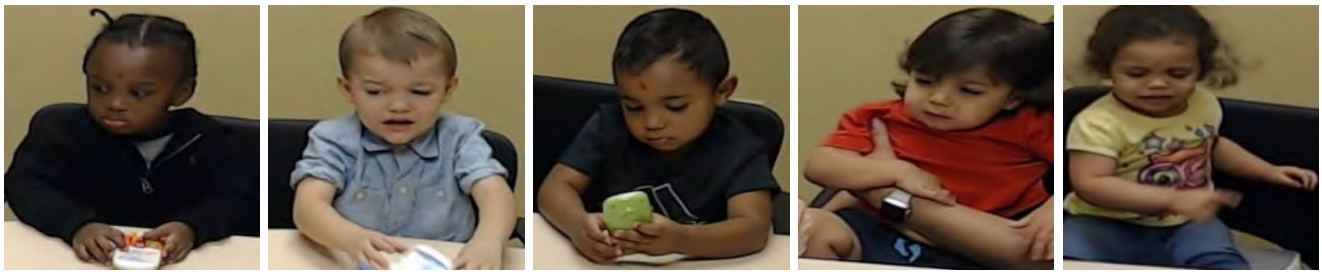
B. Collection Procedures

The Institutional Review Board at the University of South Florida approved this study and written informed consent was obtained for all participants. The child participant completed the Rapid Interactive Screening Test for Autism-Toddlers (RITA-T) with the study psychologist while the caregiver watched and completed a demographic form. The caregiver sat beside the child to ensure the child was comfortable and to redirect the child to their seat as needed. To address toddler movement, we structured the room as follows: 1) used a small room to minimize space for movement; 2) limited distractions in the room (e.g., walls were bare, devices on silent); 3) positioned the camera out of child's reach; 3) provided a sturdy chair for child with child booster seat; 4) positioned child chair directly next to caregiver; 5) once child was seated, they were scooted close to the table; 6) evaluation began immediately upon placing the child in the chair to minimize lag time; 7) if the child demonstrated need to leave the chair, the child was provided a brief 15 second break and then physically and verbally redirected to chair with caregiver assistance; 8) child received constant verbal and tangible reinforcement through play for on-task behavior (e.g., sitting in chair, looking at stimulus).

We were able to collect data accounting for child movement, straying from the chair to roam the room, or a child turning toward caregiver for feedback. Child movement did not affect our ability to collect multimodal data and in our experience, children are easily redirected to the seat after a brief break.

The examiner prompted children through the tasks and recorded their responses on a record form. The assessments averaged 6-10 minutes. Upon completion of the assessment, caregivers were offered a choice of toys (value under $20) and given a copy of the consent form. Caregiver questions were answered and then they were instructed that the assessment was completed and thanked for their time. Finally, the psychologist reviewed each child's medical record and completed the fidelity checklist as described below.

There are 14 videos that were collected using a Logitech HD Pro C920 webcam placed from a high angle to capture a range of child behaviors (e.g., facial expression, eye contact, gesture) that occurred as part of the RITA-T assessment and form the basis for the multimodal dataset (See Fig. 1).



(a) Toy blocking. (b) Object tease. (c) Blocked vision. (d) Magic ball. (e) Color constancy.

(f) Object vs face. (g) Rapid joint attention. (h) Sad face, still face. (i) Recognition.

Fig. 1: Samples subjects from dataset for each task (context). Note: these images also show samples used for the proposed body movement experiments. As can be seen in these images, the subjects are varied in their expressions, gaze, and body movement. This variation results in a challenging, real-world dataset.

TABLE I: Constructs assessed, behaviors observed, and rationale for sample RITA-T tasks A-D. Facial expressions, eye gaze, and body movement are behaviors observed in these tasks. Tasks E-I (not shown here) elicit different behaviors in expression, gaze, and body as well.

Task and Construct Assessed	**Behaviors Observed**	**Rational for Task**
(A) Toy blocking: social awareness and awareness of human agency	Eye contact and latency with examiner; Eye contact only with hand holding toy or give up on task	TD child looks to face/eyes to understand blocking; Autistic child will look at hand, give up, or look at face after prolonged period
(B) Object tease: social awareness and joint attention	Eye contact with examiner; Eye contact with caregiver; Eye contact with both examiner and caregiver; Eye contact only with hand holding toy or give up on task	TD child looks to face/eyes when teased; Autistic child will look at hand, or give up
(C) Blocked Vision: joint attention, awareness of human agency	Eye contact with examiner; Latency of eye contact with examiner	TD child looks to face/eyes of person blocking toy; Autistic child will continue to look at mirror and not look at person blocking
(D) Magic Ball: cognition, joint attention	Surprised reaction: facial expression, vocalization, gesture; Seeking object: eye contact, gesture, vocalization; Joint attention to caregiver or examiner	TD child searches for ball, amused by disappearance and look to caregiver/examiner; Autistic child keeps looking at empty cup, look for ball, give up, or not look at examiner

C. Collection Measures

Demographics. Caregiver and child demographic data were collected with a program specific demographic form.

RITA-T. The RITA-T [9] is a 9-item semi-structured play-based screening measure that looks at constructs that are impaired in autistic children including: joint attention, visual problem solving, human agency, social awareness, communication, and self-awareness. More specifically, the RITA-T has the following 9 tasks (context): (a) Blocked exploration of a toy; (b) Object tease; (c) Blocked vision; (d) Magic ball; (e) Color constancy; (f) Object vs. face; (g) Rapid joint attention (JA); (h) Sad face, still face; and (i) Recognition. Each play-based press looks at the child's integration of one or two of the previously mentioned constructs and three items look at developmental cognition. As can be seen in Table I, the tasks elicit responses from the face, gaze, and body/gesture that we evaluate in the baseline method (Section IV). Each press is coded and scored on a Yes or No scale (Yes = 0, No = 1) and some items have a latency scale (scores range from 0-2). In all circumstances, lower scores are better, and a total score is yielded by the sum of the 9 individual scores. Scores below 12 are not associated with autism; scores of 12-15 warrant further evaluation; and scores of 16+ are concerning for autism. The RITA- T has high sensitivity (.97), specificity (.71), positive predictive value (.95), and negative predictive value (.79) and does not correlate statistically with age or sex. Scores on the RITA-T correlate positively with the ADOS (r = 0.79), as well as to DSM-5 checklist items when completed by clinicians blinded to RITA-T test results (r = 0.76). The RITA-T also has discriminatory properties that result in different mean scores for children with autism as compared to children with DD and no autism. The RITA-T is fast, inexpensive,l and has excellent psychometric properties for children 18-84 months [8], [24], [27], making it a strong

TABLE II: Task description and scoring for RITA-T tasks A-D. The examiner adds up the total score from each task with lower scores preferred (less risk of autism). As can be seen here, the tasks elicit responses from gaze, gesture, and expression. We take each of these feature types into account during our baseline analysis (Sections III-IV). Selected tasks are shown as samples on how scoring is performed. For example, in task A, if the child looked at the examiner's eyes, but took $> 10s$ to do it, and abandoned the task, they would have a score of 3 (out of 4 max) for that task. The other tasks, including tasks E-I (not shown here), are scored similarly. Note: table recreated from official RITA-T [9] scoring sheet.

Task and Description	**Score**
A. Blocking of Phone: Done three times - take best score	
1. Looks at examiner's eyes	Y(0) N(1)
2. Latency to look at examiner's eyes	0-5s(0); 6-10s(1); >10s(2)
3. Abandons task	Y(1) N(0)
B. Phone tease: Done three times	
1. Looks at examiner's eyes	Y(0) N(1)
2. Looks at parent's eyes	Y(0) N(1)
3. Looks at both	Y(0) N(1)
C. **Blocked Vision: Done one time**	
1. Looks at examiner's eyes	Y(0) N(1)
2. Latency to look at examiner's eyes	0-5s(0); 6-10s(1); >10s(2)
D. Magic ball: Done three times	
1. Reaction surprise expression	Y(0) N(1)
2. Seeking object	Y(0) N(1)
3. Joint attention to parent or examiner	Y(0) N(1)

choice for early childhood autism research. See Table II for details, from the RITA-T scoring sheet, for tasks A-D.

Fidelity Measures. A fidelity checklist was used to ensure uniform data collection across participants. It included the following items: 1) developmental assessment scores; 2) RITA-T score; 3) medical record review to ascertain a) diagnosis of autism or not; b) scores on autism assessments (ADOS-2, CARS2); c) any history of DD. All participants presenting to USF BAES had items 1, 3a, 3c. Participants from Part C early intervention clinic for autism evaluations had items 1, 2, 3a, 3b, 3c. Participants from the primary care clinic had item 2 and we could not ascertain a priori if they had items 1, 3a, 3b, 3c.

Raters and Rater Training. A licensed psychologist conducted the play-based assessment. The psychologist has extensive training/experience in administering autism screenings and diagnostic tools and is certified in the RITA-T. The psychologist was formally trained in the measure and had to demonstrate competency in their ratings to be certified. More specifically, they had to submit their ratings for child cases, and they were compared against expert ratings. They had to achieve inter-rater reliability to pass, which they did.

III. Context-based Analysis of Videos of Autistic Children (Baseline)

A. Context and Pre-processing

In this paper, we propose a context-based baseline approach to classifying videos of autistic children. Here, context refers to reciprocal child behavior during administration of the 9 tasks on the RITA-T screener. For all 14 children in our study, each video contains the child taking part in all 9 of the tasks (as detailed in Section II-C). During the RITA-T, the child is given a score for each task [9]. The final score is based on the cumulative score for all tasks and provides a level of risk for autism. It is important to note, however, not all children will have a high score for all tasks (e.g., the child may have higher risk for autism overall, but for a particular task they scored similar to a typically developing child - low risk). Considering this, instead of performing classification on the entire video, we split the videos into 9 separate tasks (contexts). Classification is done separately for each individual context. More specifically, we set the training, validation and testing sets for individual contexts. According to the RITA-T scoring algorithm, higher scores are associated with higher risk for autism. By using the approach of a context-based train-test set, our proposed network is able to effectively learn better for specific contexts. We are motivated to do this by previous work from Rudovic et al. [37]. They were analyzing engagement intensity of autistic children from face images. Here, they found that having one model focused on each culture performed better for this task. Here, we extend this idea where each model is explicitly trained and tested on one context (compared to one culture in Rudovic et al.).

For the proposed context-based approach, the videos are divided into 9 contexts according to RITA-T scores. The 9 contexts assess for symptoms of risk of autism by evaluating a child's ability to engage in a range of social-communication and cognitive skills which are categorized as joint attention (JA), social awareness (SA), human agency (HA), and cognition (C). Each context has a score which can range from 0-1 to 0-4. A score of 0 indicates low risk for autism on a particular task and higher scores are associated with higher risk for autism. For example, Task-A scoring criteria range from [0,4]. On the other hand, Task-G scoring criteria range from [0,1]. As previously mentioned, each task is scored independent of other tasks and the sum of all scores indicates a risk level for autism. Given this, we looked at the context

of each task individually when determining the level of risk for autism. To do this, our class label (risk for autism) is determined by thresholding the individual scores for each context. *W*hen the total range of possible scores, for an individual context, was [0,2], then 0 is considered low risk for autism and a score of 1 or 2 is considered higher risk for autism. When the range was [0,3] or [0,4], then 0 or 1 is considered low risk for autism and the rest (e.g., 2, 3, or 4) are considered higher risk for autism. This scoring threshold is based on the RITA-T test where lower score mean low risk and higher score mean high risk. This score threshold was validated through the licensed psychologist that was formally trained in this measure. Considering this, once we split the videos into the 9 separate contexts, the same child could have class labels of both low- and high-risk for autism across the different contexts. This allows us to individually model context and to validate our approach.

To feed the video data into the network, along with splitting the video data by context, we have also done further pre-processing. For our study, we have evaluated three different modalities within our experiments. The pre-processing to extract each modality is detailed here. 1) **Body:** in this portion of data we focused on subjects' body movements and gestures in response to the context. To do this, we manually annotated all frames with the bounding box that contains the subject. See Fig. 1 for examples of this; 2) **Face:** we used MTCNN [47] to crop out the face data of subjects; 3) **Gaze:** we used OpenFace [2] to extract gaze features from the videos, which provides information on subject's eye movements in response to task contexts.

B. Architectures

To model face and body modalities, we use a vision transformer, as they have shown encouraging results in gesture recognition [13] and expression analysis [48]. Gestures and expressions are two features of interest for us in analyzing these modalities. For gaze data, we use a time-series transformer as it's a 2D matrix consisting of time-sequence and gaze features. Transformers have also shown encouraging results for gaze recognition [42]. Along with these approaches, we also fuse all modalities (Fig. 2).

1) Vision Transformer: To analyze face and body data we passed our context-based videos to our network as a sequence of images. We ran the same model for two modalities separately. The vision transformer model consists of multiple transformer blocks. To start, we passed the sequenced images to the model by splitting images into patches of size 8, which are then fed into the patch encoder layer. This layer projects the data into a vector of size 64, and adds a learnable positional embedding to this vector. The output from this layer is then fed into the transformer block. This block consists of one normalization layer, one multi-head attention layer, another normalization layer, and a final multi-perceptron layer (MLP). There is a total of 8 of these transformer blocks. Inside the transformer block multi-head attention is used as a self-attention mechanism, which has been applied to the sequence of patches. The output is divided by the number of heads, which is four in our experiments. In doing this, the

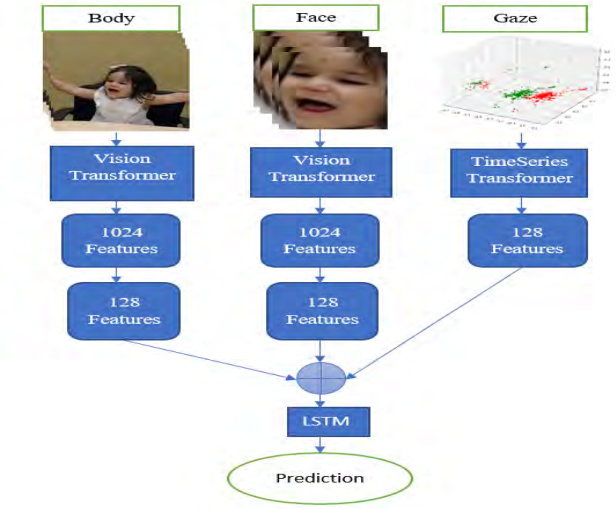


Fig. 2: Overview of multimodal architecture. Face, body, and gaze data are modeled separately, where 128 features are extracted from each modality. These features are then concatenated and used as input to an LSTM network.

model has multiple independent ways to understand the input. These heads get concatenated and transformed, so we get $MultiHead(Q, K, V) = Concat(head_a, \ldots, head_h)W^O$. Here $head_a = attention(QW_a^Q, kW_a^k, VW_a^V)$ and Q, K ,V are Queries, Keys and Values, respectively. In our experiments, Q = Number of Patches from an image + 1 class token, which is 1025 $(1024 + 1)$; K = number of channels of image (3); and V = shape of a patch which is 64 (8×8). The transformer blocks produce a tensor of size [batch size, number of patches, projection dim], which is then processed by a classifier head with softmax to produce the final class probabilities output. For our experiments, we use an image size of 256×256. The learning rate is 0.001, batch size is 8, number of epochs is 100, patch size is 7, number of transformer heads is 4, and number of transformer layers is 8. For this model, total number of parameters is $45,394,953$.

2) Time-series Transformer: To analyze gaze data, we use a time-series transformer [49] to process a tensor of shape (batch size, sequence length, features), where sequence length is the number of time steps and features is each input of the time-series. To extract gaze features, we apply OpenFace to the body movement video data (manually annotated frames as detailed in Section III-A). The (x, y, z) coordinates are then extracted for left and right eyes. This is done over the entire video. In our time-series transformer we used four transformer encoder blocks, an average pooling layer, and a MLP classification head. Finally, we used a dense layer with softmax to classify the video from gaze data. For this model, the total number of parameters is $19,593,240$.

3) Multi-modal Fusion model: Along with single modalities (face, body, and gaze), we also aim to investigate how the fusion of these modalities can be used for our context-based approach. To do this, the face and body features are extracted from the vision transformer. From the transformer block, 1024 features are extracted and then fed to a dense

layer to extract a total of 128 features. This is done separately for both face and body data resulting in a total of 128 features for each. From the time-series transformer, we extract 128 features in total for the gaze data. After extracting all features from face, body, and gaze, we concatenated all features into a new vector of length 384 (128×3). This approach to feature fusion is based on the total number of gaze features being 128. Considering this, we also use 128 features for body and face features to ensure a consistent number of features across each modality. Once we have 128 features for each modality, they are concatenated into the larger feature vector. This multimodal approach to feature fusion has had success in other affect related works [17], [18]. This new vector is then used as input to a long short-term memory (LSTM) network. We are motivated to use an LSTM based on previous works that have shown success in using them with a multimodal approach to analyzing human data [38], [6], [32]. These works have shown success with similar modalities (e.g., face and body data), when temporal information is available. Along with these works Li et al. [29] also use an LSTM in their pipeline for searching videos, showing encouraging results for video-based understanding. The learning rate is set to 0.0003 and the Adam optimizer [23] is used. For this model, the total number of parameters is $14,690$. See Fig. 2 for an overview of the proposed baseline approach.

IV. Baseline Experiments and Results

A. Experimental Design

Given the preprocessed video frames (as detailed in Section III-A), we created an 80/20, subject-independent, split of the data for training and testing, respectively. For the train/test split we chose the participants according to their RITA-T score. For example, participants with RITA-T score 12-16 are at medium risk of autism, > 16 are at high risk and < 12 are at low risk. Considering this, we created the training/testing set using the combination of all risk levels. As we are evaluating low vs. high risk, we chose participants with < 12 as low risk and > 12 as high risk. An 80/20 split was chosen instead of leave-one-subject-out cross-validation due to combining all risk levels. More specifically, we ensure there was a low, medium, and high risk for autism, based on the RITA-T scores. This was to ensure testing on each scenario. For face and body data, they were used as input to the vision transformer as a sequence of images where prediction happened on each image. More specifically, for each subject we had a sequence of images as input and also the same number of outputs as prediction in terms of high- and low-risk for autism. For the final classification, we used majority voting across all input images. For example, given 179 images as input and 150 were predicted as high risk and 29 as low risk for autism, the final output would have a classification of high risk for autism. For some of the subjects we were unable to extract (crop) the facial data. Due to this, we had to discard some images for total occlusion of the face. This resulted in some subjects/tasks not being used due to a small number of images being pre-processed. For example, in some cases only $10 - 20$ images were cropped.

For the Gaze data we used the time-series transformer. In that network, we used the gaze data extracted from OpenFace as input. We used the (x, y, z) coordinates of the left and right eyes. Considering this, the input to the transformer is a time-sequence vector of size 6. Unlike the vision transformer (majority voting), there is one output prediction from the entire video (time-series).

For our multimodal fusion, we extract all of the features from the three networks. As there is a large difference in values between face and body, and gaze, we normalized the range for all modalities to $[0, 1]$. Given these normalized features, we then sum up the features as $m_i = b_i + f_i + g_i$ for $i = 1$ to N. Here, $N = 128$, b_i, f_i, and g_i are the i^{th} body, face, and gaze features, respectively. This results in the new features vector $v_f = [m_1, m_2, m_3, \ldots, m_N]$, which is used as input to an LSTM network for classification.

As mentioned for training, in Section III-A, the class labels for each context were determined based on a threshold for the individual scores. For testing, this same thresholding technique was used, however, if a particular context had a class label that was different from the ground truth diagnosis, that particular context was removed from the testing. For example, if a child had high risk for autism and the thresholding technique resulted in a class label of low-risk for autism, that context was removed for testing. While we are looking at context, we also want to be able to further verify which context correctly classifies the videos (i.e., a ground truth of high risk for autism results in a classification of high risk for autism). Considering this, for body and gaze data, we have 14 contexts for high risk for autism and 6 for low risk for autism, resulting in 20 total contexts for testing. These testing contexts specifically came from participants 5 and 10 for high risk for autism, and participant 9 for low risk for autism, in the proposed dataset. For face data there are a total of 19 videos as no cropped facial data was returned for one task.

B. Results

The results (accuracy and F1) for each context and modality (unimodal and multimodal) can be seen here in Table III. For each modality (body, gaze, or face), there is one context that results in the highest accuracy and F1 score. More specifically, for **body**, the context **object vs. face** (Task F) resulted in an accuracy of 100% and an F1 of 0.93. For **face**, the context **rapid joint attention** (Task G) resulted in an accuracy of 100% and an F1 of 0.72. Finally, for **gaze**, the context **toy blocking** (Task A) resulted in an accuracy of 100% and an F1 of 1.0. Across all contexts the accuracy ranges from $[33\%, 100\%]$, and the F1 ranges from $[0.02, 1.0]$. Due to the large range, for both accuracy and F1, these results support that context matters when classifying videos of autistic children. These results also support that within individual contexts, different modalities will have different performance. As also can be seen in Table III, in the last row, the averages across all contexts for each modality are similar. For example, the average accuracies for body, face, and gaze are 0.61, 0.59, and 0.57, respectively. Where it is different is in the actual contexts. As previously noted Task G resulted in an accuracy of 100% for the face modality, however, when

TABLE III: Average accuracy and F1-score, across testing data, for each context and corresponding modality. Each row represents one of the following 9 contexts: (A) Blocked exploration of a toy; (B) Object tease; (C) Blocked vision; (D) Magic ball; (E) Color constancy; (F) Object vs. face; (G) Rapid join attention (JA); (H) Sad face, still face; (I) Recognition. **Bold text** corresponds to best context for each individual modality. B/F/G corresponds to a multimodal approach (body, face, and gaze). Metrics are listed as **Accuracy** | **F1** in each table cell, from 0 to 1. Last row is the average accuracy and F1 across all contexts. Higher is better.

	Body	**Face**	**Gaze**	**B/F/G**
A	0.66 | 0.62	0.5 | 0.2	**1.0** | **1.0**	0.66 | 0.67
B	0.66 | 0.42	0.5 | 0.53	0.66 | 0.53	0.66 | 0.67
C	0.66 | 0.48	0.66 | 0.59	0.33 | 0.33	0.33 | 0.33
D	0.33 | 0.2	0.8 | 0.9	0.66 | 0.53	0.33 | 0.33
E	0.66 | 0.65	0.66 | 0.5	0.66 | 0.67	0.66 | 0.67
F	**1.0** | **0.93**	0.33 | 0.46	0.33 | 0.2	0.66 | 0.67
G	0.33 | 0.43	**1.0** | **0.72**	0.66 | 0.67	0.66 | 0.67
H	0.66 | 0.46	0.33 | 0.02	0.33 | 0.2	0.66 | 0.67
I	0.5 | 0.54	0.5 | 0.03	0.5 | 0.33	0.5 | 0.33
Avg	0.61 | 0.53	0.59 | 0.44	0.57 | 0.55	0.57 | 0.56

body and gaze were used for this context it resulted in an accuracy of 33% and 66%, respectively. Similarly, Task F resulted in 100% accuracy for body, however, both face and gaze had an accuracy of 33%. Task A resulted in 100% accuracy for gaze, however, body and face had accuracies of 0.66% and 0.5%, respectively.

To give a broader picture regarding classification across all contexts, and testing subjects, for our proposed context-based approach, we detail the confusion matrices for face, body, and gaze in Tables IV, V, and VI, respectively. Here, the numbers in the confusion matrix correspond to all contexts across each of each of the subjects. For face and body, the models more often classified the context as high risk for autism. For example, in Table IV, it can be seen that high risk was the classification for 10 out of 19 instances, for face. This can be explained, in part, as high risk is the majority class and it has been shown that classifiers trained on imbalanced data often predict the majority class [15]. Showing the confusion matrices in this way (all contexts across all subjects in the testing sets) supports the need to evaluate and discuss individual contexts. With only the information given in Table IV, an overall accuracy of 55% can be calculated (10 out of 18 instances were correctly classified). On the other hand, if we look at individual contexts, a better picture of the classification appears. More specifically, as detailed previously, the context rapid joint attention (Task G) had an accuracy of 100%. A similar result is found with body data, where 11 out of 20 instances were classified as high-risk for autism. Conversely, the opposite is true with gaze as the modality. Here, low-risk for autism was more often the classification (Table VI). Here, 11 out of 20 instances were classified as low-risk for autism. Only one instance of low-risk for autism was incorrectly classified, the rest were correct. Intuitively, we would expect the models

TABLE IV: Face confusion matrix for low vs. high risk for autism. Numbers correspond to total of all contexts, for face, across all subjects in testing set.

	High-risk	**Low-risk**
High-risk	7	6
Low-risk	3	3

TABLE V: Body confusion matrix for low vs. high risk for autism. Numbers correspond to total of all contexts, for body, across all subjects in testing set.

	High-risk	**Low-risk**
High-risk	7	7
Low-risk	4	2

to be biased towards the majority class [11]. One potential reason for this could be our use of a time-series transformer compared to the vision transformer for face and body data. While this is an interesting finding, it is currently out of scope of the current paper and left for future work.

We also evaluated our proposed multimodal approach (Section III-B.3). These results can also be seen in Table III (last column, labelled B/F/G). While the multimodal approach still had differences across each context, the results here are less significant. For example, the range of the accuracies and F1 are smaller with $[0.33,\ 0.66]$ and $[0.33,\ 0.67]$, respectively. As can be seen in Table VII, the confusion matrix for the multimodal approach looks similar to the others (face, body, and gaze). The main difference being no one context was able to be used to achieve 100% accuracy. This type of result (i.e., unimodal approach outperforms multimodal approach) has been seen in other related applications. While it has been shown that a multimodal approach can outperform a unimodal approach [46], the opposite has also been found to be true. More specifically, a lower accuracy and F1-score are reported for the multimodal approach [40], as shown here.

These results make sense from an intuitive standpoint. For example, in the toy blocking context (Task A), gaze had the highest accuracy and F1. This would be an instance where the child should look at the examiner after they block the toy. More discussions on this, the other contexts and modalities, and the proposed approach (unimodal and multimodal) can be found in Section V.

TABLE VI: Gaze confusion matrix for low vs. high risk for autism. Numbers correspond to total of all contexts, for gaze, across all subjects in testing set.

	High-risk	**Low-risk**
High-risk	8	6
Low-risk	1	5

TABLE VII: Multimodal confusion matrix for low vs. high risk for autism. Numbers correspond to total of all contexts, for multimodal approach, across all subjects in testing set.

	High-risk	**Low-risk**
High-risk	7	7
Low-risk	1	5

V. Discussion

The proposed dataset addresses some challenges and limitations such as (1) data not being publicly available [37], [12] due to the sensitive nature of the data; (2) the tasks used for data collection are often not directly related to autism [22], [50]; and (3) the control class was picked from another dataset [26] with different tasks.

Along with the dataset we also present a baseline study evaluating face, gaze, and body data along with the context to classify videos for low- and high-risk for autism. The results are encouraging and show some results we would expect to see. As detailed in Section IV-B, the toy blocking context (Task A) resulted in 100% accuracy from gaze. As noted in Table I, the behaviors observed are eye contact and latency with examiner; eye contact only with hand holding toy or give up on task. Considering this, the results could be explained by the child looking (or not looking) at the examiner when the toy was blocked. Although there are other tasks where eye contact is an important behavior, these tasks do not do as well as Task A. For example, Task C also has eye contact as an observed behavior, however, the accuracy of this task for gaze was 33%. These results suggest that the context in which the modality is seen has an impact on the ability to classify the videos. Similar results were observed for both face and body. For example, face data worked best (highest accuracy) with the rapid joint attention context (Task G). Here, the examiner looks at the child and points somewhere for the child to look. This action could have large variations in facial affect (e.g., child is surprised at what they are looking at versus child shows a neutral face). Body movement data worked best for the object vs. face context (Task F). Here, the child is shown two images (one with object and one with face). Often, instead of simply looking at the image they are most interested in, the child orients their entire body towards that image. The children who scored high-risk for autism, would often orient towards the image of the object and not the face. These results are promising and motivate us to conduct further research.

Overall, due to the public availability of the proposed dataset and the presented baseline approach, this work has the potential to help advance the field for both machine learning research into autism diagnosis and medical practitioners. These advancements include, but are not limited to, (1) giving practitioners a tool to help expand autism diagnosis to a larger cohort of subjects; (2) lowering the average age range of diagnosis for autism; and (3) giving researchers a publicly available, RITA-T and context-based dataset of videos of autistic children. The initial results on the dataset, from the baseline, are encouraging. They suggest a context-based approach can improve classification of risk level for autism among young children. Along with this, they also suggest that the use of RITA-T for context can elicit appropriate responses in face, gaze, and body movement that can be used to train machine learning classifiers.

A. Limitations and Future Work

Although these results are encouraging, there are some limitations to our work. First, the current size of the dataset is relatively small with 14 subjects. To address this, we will continue to collect a larger cohort of children which will result in a larger available dataset over time. Second, there is a large class imbalance (11 subjects with high risk, 3 with low). We will address this limitation by over-sampling in a well child pediatric clinic with the aim to bring balance among participant enrollment for children with high and low risk for autism. Third, only three of the subjects were evaluated in the test set. In our experimental design, we have an $80/20$ split between training and testing data. For future work, we will perform leave-one-subject-out cross-validation on the current dataset and the larger dataset that will be collected. Finally, the multimodal approach overall, did not perform as well as the unimodal approach. We will further evaluate the proposed approach with the larger dataset, leave-one-subject-out cross-validation, and we will evaluate different fusion approaches such as score-level fusion [20].

B. Ethics Statement

This work and all its future applications are intended to be used with consent from all parties that can include, but is not limited to, the child's caregiver, the examiner, and any medical practitioner that is involved in diagnosis of the child. A major ethical concern involved in working with human subjects is privacy. In this work, we collect video data from 14 children which also includes the examiner and the caregivers in the full videos. For the collected dataset, all participants consented to have their data recorded and used for future research purposes. This was approved by IRB ensuring no ethical oversight. There are also issues with class imbalance and the use of machine learning. This imbalance may introduce bias towards the majority class [7]. For this reason, we evaluate our models against F1-score and accuracy to give more insight into model performance. Jeni et al. [21] have shown that reporting multiple evaluation metrics can help when class imbalance is shown.

VI. Conclusion

In this paper, we proposed a new context-based dataset for analyzing videos of autistic children. The new baseline results, on the proposed dataset, show encouraging results on the usefulness of the dataset and that context matters when analyzing videos of autistic children. The collected dataset, used in this study, will be released to the larger scientific community for replication of this work, as well as furthering advancements in video-based analysis of autistic children. The dataset and all code will be available to request and download through the University of South Florida.

REFERENCES

[1] J. Baio. Prevalence of asd among children aged 8 years-autism and developmental disabilities monitoring network. *11 sites, United States*, 2014.

[2] T. Baltrušaitis et al. Openface: an open source facial behavior analysis toolkit. In *WACV*, pages 1–10. IEEE, 2016.

[3] X.-a. Bi et al. Classification of autism spectrum disorder using random support vector machine cluster. *Frontiers in genetics*, 9:18, 2018.

[4] D. V. Bishop et al. Autism and diagnostic substitution: evidence from a study of adults with a history of developmental language disorder. *Developmental Medicine & Child Neurology*, 50(5):341–345, 2008.

[5] L. Burke and K. P. Stoddart. Medical and health problems in adults with high-functioning autism and asperger syndrome. In *Adolescents and adults with autism spectrum disorders*, pages 239–267. Springer, 2014.

[6] L. Cai, J. Dong, and M. Wei. Multi-modal emotion recognition from speech and facial expression based on deep learning. In *2020 Chinese automation congress (CAC)*, pages 5726–5729. IEEE, 2020.

[7] J. Chakraborty, S. Majumder, and T. Menzies. Bias in machine learning software: Why? how? what to do? In *ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440, 2021.

[8] R. Choueiri et al. Improving early identification and access to diagnosis of autism spectrum disorder in toddlers in a culturally diverse community with the rapid interactive screening test for autism in toddlers. *Journal of Autism and Developmental Disorders*, 51(11):3937–3945, 2021.

[9] R. Choueiri and R.-T. Team. Manual of administration: The rapid interactive screening test for autism in toddlers (rita-t). In *University of Massachusetts*, 2019.

[10] G. Dawson et al. Randomized, controlled trial of an intervention for toddlers with autism: the early start denver model. *Pediatrics*, 125(1):e17–e23, 2010.

[11] Q. Dong et al. Imbalanced deep learning by minority class incremental rectification. *IEEE PAMI*, 41(6):1367–1381, 2018.

[12] H. Drimalla et al. Detecting autism by analyzing a simulated social interaction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 193–208. Springer, 2018.

[13] A. D'Eusanio et al. A transformer-based network for dynamic hand gesture recognition. In *3DV*, pages 623–632. IEEE, 2020.

[14] J. H. Elder, C. M. Kreider, S. N. Brasher, and M. Ansell. Clinical impact of early diagnosis of autism on the prognosis and parent–child relationships. *Psychology research and behavior management*, 2017.

[15] C. Esposito et al. Ghost: adjusting the decision threshold to handle imbalanced data in machine learning. *Journal of Chemical Information and Modeling*, 61(6):2623–2640, 2021.

[16] T. W. Frazier et al. A meta-analysis of gaze differences to social and nonsocial information between individuals with and without autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(7):546–555, 2017.

[17] H. Gunes and M. Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE international conference on systems, man and cybernetics*, volume 4, pages 3437–3443. IEEE, 2005.

[18] H. Gunes and M. Piccardi. Fusing face and body gesture for machine recognition of emotions. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 306–311. IEEE, 2005.

[19] T. M. Helminen et al. Atypical physiological orienting to direct gaze in low-functioning children with autism spectrum disorder. *Autism Research*, 10(5):810–820, 2017.

[20] Y. Huang et al. Fusion of facial expressions and eeg for multimodal emotion recognition. *Computational intelligence and neuroscience*, 2017.

[21] L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 245–251. IEEE, 2013.

[22] M. Jiang et al. Classifying individuals with asd through facial emotion recognition and eye-tracking. In *EMBC*, pages 6063–6068. IEEE, 2019.

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

[24] X.-J. Kong et al. Validation of rapid interactive screening test for autism in toddlers using autism diagnostic observation schedule 2^{nd} edition in children at high-risk for asd. *Frontiers in psychiatry*, 12, 2021.

[25] Y. Kong, J. Gao, Y. Xu, Y. Pan, J. Wang, and J. Liu. Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing*, 324:63–68, 2019.

[26] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[27] J.-F. Lemay et al. Experience with the rapid interactive test for autism in toddlers in an autism spectrum disorder diagnostic clinic. *Journal of Developmental & Behavioral Pediatrics*, 41(2):95–103, 2020.

[28] B. Li et al. A facial affect analysis system for autism spectrum disorder. In *ICIP*, pages 4549–4553. IEEE, 2019.

[29] S. Li, X. Li, J. Lu, and J. Zhou. Structure-adaptive neighborhood preserving hashing for scalable video search. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2441–2454, 2021.

[30] Z. Luo, Y. Xiao, Y. Liu, S. Li, Y. Wang, Y. Tang, X. Li, and Y. Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *arXiv preprint arXiv:2305.17011*, 2023.

[31] C. G. Moore et al. Recommendations for planning pilot studies in clinical and translational research. *Clinical and translational science*, 4(5):332–337, 2011.

[32] W. Nie, Y. Yan, D. Song, and K. Wang. Multi-modal feature fusion based on multi-layers lstm for video emotion recognition. *Multimedia Tools and Applications*, 80:16205–16214, 2021.

[33] K. Niu et al. Multichannel deep attention neural networks for the classification of autism spectrum disorder using neuroimaging and personal characteristic data. *Complexity*, 2020, 2020.

[34] H. R. Park et al. A short review on the current understanding of autism spectrum disorders. *Experimental neurobiology*, 25(1):1, 2016.

[35] J. Rehg et al. Decoding children's social behavior. In *CVPR*, 2013.

[36] J. M. Rehg, A. Rozga, G. D. Abowd, and M. S. Goodwin. Behavioral imaging and autism. *IEEE Pervasive Computing*, 13(2):84–87, 2014.

[37] O. Rudovic et al. Culturenet: a deep learning approach for engagement intensity estimation from face images of children with autism. In *IROS*, pages 339–346. IEEE, 2018.

[38] O. Rudovic, M. Zhang, B. Schuller, and R. Picard. Multi-modal active learning from human data: A deep reinforcement learning approach. In *2019 International Conference on Multimodal Interaction*, pages 6–15, 2019.

[39] C. Saint-Georges, R. S. Cassel, D. Cohen, M. Chetouani, M.-C. Laznik, S. Maestro, and F. Muratori. What studies of family home movies can teach us about autistic infants: A literature review. *Research in Autism Spectrum Disorders*, 4(3):355–366, 2010.

[40] M. S. Salekin et al. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Computers in biology and medicine*, 129:104150, 2021.

[41] M. D. Samad et al. A pilot study to identify autism related traits in spontaneous facial actions using computer vision. *Research in Autism Spectrum Disorders*, 65:14–24, 2019.

[42] D. Tu et al. End-to-end human-gaze-target detection with transformers. In *CVPR*, pages 2192–2200. IEEE, 2022.

[43] G. Vivanti and C. Dissanayake. Outcome for children receiving the early start denver model before and after 48 months. *Journal of autism and developmental disorders*, 46(7):2441–2449, 2016.

[44] G. Wang, H. Zeng, Z. Wang, Z. Liu, and H. Wang. Motion projection consistency based 3d human pose estimation with virtual bones from monocular videos. *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[45] E. M. Weiss et al. Less differentiated facial responses to naturalistic films of another person's emotional expressions in adolescents and adults with high-functioning autism spectrum disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 89:341–346, 2019.

[46] G. Zamzmi et al. Automated pain assessment in neonates. In *SCIA*, pages 350–361. Springer, 2017.

[47] K. Zhang et al. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[48] Z. Zhao and Q. Liu. Former-dfer: Dynamic facial expression recognition transformer. In *ICMI*, pages 1553–1561, 2021.

[49] H. Zhou et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, pages 11106–11115, 2021.

[50] A. Zunino, P. Morerio, A. Cavallo, C. Ansuini, J. Podda, F. Battaglia, E. Veneselli, C. Becchio, and V. Murino. Video gesture analysis for autism spectrum disorder detection. In *2018 24th international conference on pattern recognition (ICPR)*, pages 3421–3426. IEEE, 2018.

[51] L. Zwaigenbaum et al. Early identification of autism spectrum disorder: Recommendations for practice and research. *Pediatrics*, 136(Supplement_1):S10–S40, 2015.