2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Towards Autism Screening through Emotion-guided Eye Gaze Response

Surjya Ghosh[1] and Tanaya Guha[2]

Abstract— **Individuals with Autism Spectrum Disorder (ASD) are known to have significantly limited social interaction abilities, which are often manifested in different non-verbal cues of communication such as facial expression, atypical eye gaze response. While prior works leveraged the role of pupil response for screening ASD, limited works have been carried out to find the influence of emotion stimuli on pupil response for ASD screening. We, in this paper, design, develop, and evaluate a light-weight LSTM (Long-short Term Memory) model that captures pupil responses (pupil diameter, fixation duration, and fixation location) based on the social interaction with a virtual agent and detects ASD sessions based on short interactions. Our findings demonstrate that all the pupil responses vary significantly in the ASD sessions in response to the different emotion (*angry, happy, neutral*) stimuli applied. These findings reinforce the ASD screening with an average accuracy of 77%, while the accuracy improves further (>80%) with respect to *angry* and *happy* emotion stimuli.**

I. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by significantly impaired social interaction and communication abilities [1], [2]. Such impairments include deficits in perceiving, using and responding to various non-verbal cues of communication, such as emotion-related facial expressions [3], [4], [5]. This is often attributed to the atypical eye gaze in deriving these non-verbal cues from [6], [7], [8], [9]. Research has shown that individuals with ASD, while watching videos of social scenes (dynamic stimuli), fixate less to the eye and face region and more to the human body [10], [11], [12] as compared to their typical counterparts. Therefore, there is clear evidence that individuals with ASD have atypical gaze pattern which is prominently observed in the context of processing affective expressions from a communicator's face.

In this paper, we develop and investigate the effectiveness of an *automated screening* system for ASD based on the subjects' eye gaze response as they interact with a virtual reality (VR)-based social communication system. VR-based social communication systems have emerged as an effective alternative to the traditional assessment, intervention and education programs in Autism [13], [14], [15], [16] due to their lower cost and higher accessibility. For this work, we have used a recently developed VR platform designed particularly to help individuals with ASD to improve their emotion recognition skills and performance in social tasks [17], [14]. This platform lets us collect individual's eye physiological index and looking pattern as they interact with virtual agents demonstrating basic, context-relevant emotional expressions. The agents are capable of displaying three basic emotions: neutral, happy and angry. During the course of interaction with the agents, 16 participants' eye gaze data were collected using a commercially available eye-tracking device in terms of (i) pupil response measured in terms of *pupil diameter* changes, and (ii) *fixation coordinate and duration*. This data is used to develop a deep recurrent model that identifies subjects with ASD based on their eye gaze behavior.

There are several past works that analyzed and reported atypicality in eye gaze patterns among subjects with ASD [6], [3], but limited work on using emotion-guided eye gaze responses for ASD screening. DiCriscio and Troiani demonstrated that pupil dilation is correlated with the clinical ASD measure of the social responsiveness [18]. Ahuja et al. developed an ASD screening mechanism based on eye gaze responses to multiple prosaic videos [19]. However, none of the two works [18], [19] have investigated eye gaze as a response to emotional stimuli.

Our automated screening system employs a recurrent model to infer whether or not a subject-agent interaction session is *atypical* (suggesting ASD). We used our own database of eye-gaze patterns collected from 16 adolescents (8 ASD, 8 typically developing (TD)) across more than 1300 small sessions of interaction with virtual agents showing three basic emotions. We first perform a thorough analysis of the participants' eye-gaze data (pupil diameter, fixation duration and location). We observe the sessions involving subjects with ASD (i) exhibit significantly ($p < 0.001$) larger pupil diameter, (ii) have significantly ($p < 0.001$) shorter fixation duration, and (iii) fixate less on the faces of the agents. These observations are consistent with past work on eye-gaze responses in Autism [19], [6], [20], [21]. Motivated by these observable atypicalities, we train a Long-Short Term Memory network (LSTM) using the participants' eye-gaze data (pupil diameter, fixation duration and location) to identify a agent-subject interaction session as ASD/TD. We show that eye-gaze patterns can be used to identify ASD sessions with a reasonably high accuracy of 77% overall, and even higher accuracy of more than 80% when using only emotional stimuli (*angry, happy*). This indicates that emotion-guided eye-gaze response is a promising approach to ASD screening within an interactive VR system.

II. Dataset and Analysis

In this section, we describe the dataset collected using a VR-based social intervention system, and analyze its suitability for autism screening.

*Thanks to Prof. Lahiri, IIT Gandhinagar for sharing the data.

[1]S. Ghosh is with the Distributed and Interactive Systems Group, CWI Amsterdam, The Netherlands.

[2]T. Guha is with the Department of Computer Science, University of Warwick, UK CV4 7AL.

Fig. 1: Example of a social situation in the VR-based communication system we employed for data collection. The scene is annotated with two regions of interest (ROI): *face* and *others*. The virtual agent displays *neutral* expression in the scene.

A. Data Description

The VR platform we used for data acquisition is realistic 3D environment of social situations developed by Kuriakose et al. [14], [17]. In this platform a virtual agent narrates stories related to different social situations to a participant, and displays context-relevant emotional expressions. As the participants listen to the social stories narrated by the agents their eye gaze data are collected. More information about data acquisition can be found in our previous work [6]. The experiment was approved by the Institutional Review Board.

Our database (see Table I for a summary) contains eye-gaze data (pupil diameter, fixation duration and location) from 16 subjects (8 ASD, 8 TD) recorded in 48 long sessions. This yields 1,305 short sessions created by breaking down the long sessions into smaller chunks. Each short session is associated and labeled with a single emotion displayed by the agent i.e. neutral, happy or angry. Based on the recorded fixation coordinates, we labeled the participants' fixation patterns corresponding to the two regions of interest (ROI): 1 for *face* and 0 for *others* ROI (see Fig. 1). Every session thus consists of three (synchronized) sequences of 100 data points corresponding to pupil diameter, fixation duration and fixation location (binary sequence). The sessions are balanced across emotion and subject classes. Also note that the ASD and TD groups have no significant difference in age or gender distribution.

B. Data Analysis

We now analyze the eye gaze behavior in ASD and TD sessions to discover any key differences between the two groups. Fig. 2 compares the three components of eye gaze response for the two groups. We observe that the pupil diameters recorded in TD sessions are significantly higher than ASD sessions (Fig. 2a). The median pupil diameter for ASD and TD sessions are 0.497 and 0.649. Since pupil diameter values in our database are not normally distributed

TABLE I: Details of the eye gaze dataset used

Number of subjects	16 (8 ASD, 8 TD)
Total (short) sessions	1,305
Average session duration	16.2 seconds
ASD sessions	625 (218 Angry, 208 Happy, 199 Neutral)
TD sessions	680 (235 Angry, 231 Happy, 214 Neutral)

($p < 0.05$ with Shapiro-Wilk test)[1], we perform the unpaired Mann-Whitney U-test and observe a significant effect of session types on pupil diameters ($U = 148952$, $Z = 9.344$, $p < 0.001$, $r = 0.259$). Fig. 2b also shows significant effect of session types on fixation duration ($U = 174734$, $Z = 5.553$, $p < 0.001$, $r = 0.154$). This indicates that subjects with ASD find it difficult to fixate on the objects as they have significantly lower fixation duration. Similar observations are also reported in earlier studies [22], [21]. To compare the fixation locations (*face, other*), we compute the frequency (mode) of fixation location from each type of sessions (see Fig. 2c). We note that subjects with ASD fixate less on the agents' *face* (denoted by 1) exhibiting significant effect of session type on fixation location ($U = 315220$, $Z = 17.571$, $p < 0.001$, $r = 0.486$).

In summary, all three eye gaze components exhibit significant differences between ASD and TD sessions. Note that the differences are significant even within such short duration as ~ 16.2 seconds. These observations motivate us to develop an automated ASD screening system based on the subjects' eye gaze behavior.

III. Autism Screening

A. Recurrent model

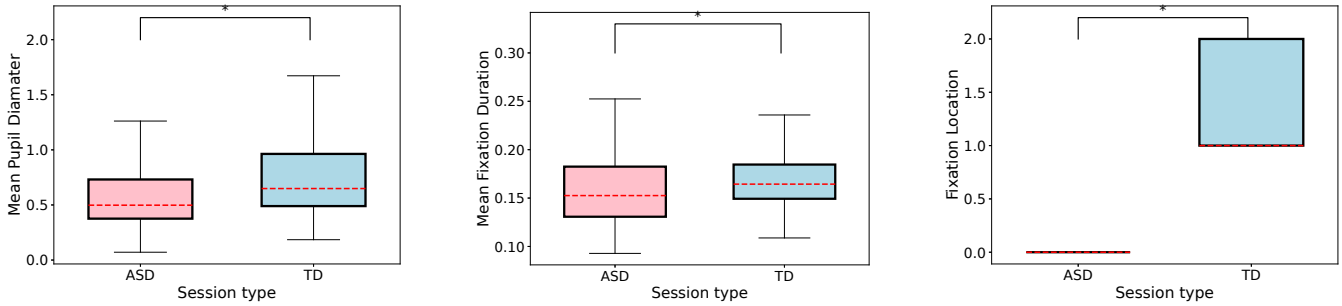
We develop an LSTM-based architecture to identify if a given subject-agent interaction session characterized by eye gaze data is ASD or TD (see Fig. 3).

Given eye gaze sequence $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_T\}$ for a session, the LSTM model takes an input $\mathbf{x}_t = [p_t, d_t, f_t]^T$ at each step t, where p_t is the pupil diameter, d_t is the fixation duration and f_t indicates fixation location at step t. Our architecture consists of T LSTM cells. The LSTM embedding is input to a dropout layer, followed by a dense (fully connected) layer. The output is finally connected to a sigmoid function to perform a binary classification i.e., ASD or TD. We use binary crossentropy loss for training this architecture.

B. Baseline Model

As a baseline, we used a Random Forest (RF) classifier to identify the ASD sessions. We compute the following functionals from every sequence X in a session: mean and median of pupil diameter and fixation duration, and mode of fixation location. The model is constructed using 100 decision trees with the maximum depth of the tree set to unlimited.

[1]Same observation and analysis protocol followed for for fixation duration and location.



(a) Pupil diameter comparison (b) Fixation duration comparison (c) Fixation location comparison

Fig. 2: Comparison of eye gaze patterns between ASD and TD sessions. All three measures show significant differences ($p < 0.001$) between ASD and TD.

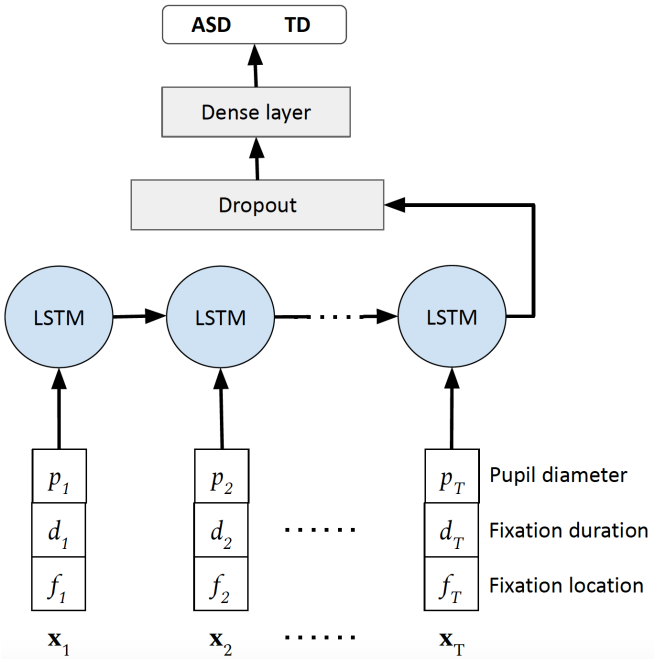
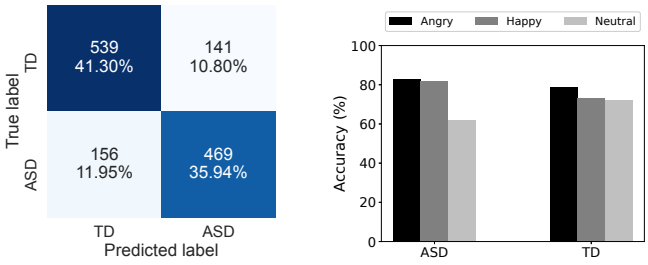


Fig. 3: Our LSTM-based architecture for ASD screening

TABLE II: ASD screening results

Model	**Accuracy** (in %)
RF Baseline	67.8± 8.7
LSTM (proposed)	**77.3**± **3.5**
Ablation	
LSTM-pupil dia	59.1± 6.3
LSTM-fix dur	54.8± 12.5
LSTM-fix loc	69.9± 7.9
LSTM w/o pupil dia	72.2± 6.5
LSTM w/o fix dur	76.4± 6.4
LSTM w/o fix loc	61.9± 9.7



(a) (b)

Fig. 4: Performance of our LSTM-based ASD screening system. (a) Confusion matrix (b) ASD detection accuracy broken down by different emotion stimuli. Note that emotional stimuli elicit more discriminating eye gaze response as compared to the neutral stimuli.

IV. Evaluation

We perform a *leave-one-subject-out* cross-validation to evaluate our model. At each iteration, we randomly select one ASD subject and one TD subject, and hold all their sessions aside for testing. All sessions pertaining to the remaining subjects are used for training both the models. This is repeated 8 times, and the average recognition accuracy across *sessions* is reported. For the recurrent model, we used $T = 100$ LSTM units, a batch size of 64, dropout rate of 0.5, and the Adam optimizer for training.

A. Results

Table II presents the performances the proposed LSTM-based model, the baseline, and the ablation results. We observe that LSTM outperforms the RF baseline by almost 10%. The ablation study shows that fixation location alone has the most discriminative power when compared with fixation duration or pupil diameter. Combining all three measures improve the overall accuracy significantly.

Fig. 4a shows the confusion matrix indicating that the accuracy in identifying ASD sessions is lower than that of TD sessions. This can be attributed to the higher variability in ASD measurements noted in several past works due to the spectrum nature of the disorder itself.

Influence of emotion stimuli: We also investigate the role of emotion stimuli in ASD screening. When comparing performance in the context of emotion labels of the sessions, we observe that screening performance is significantly better for the *angry* and *happy* sessions than the *neutral* sessions

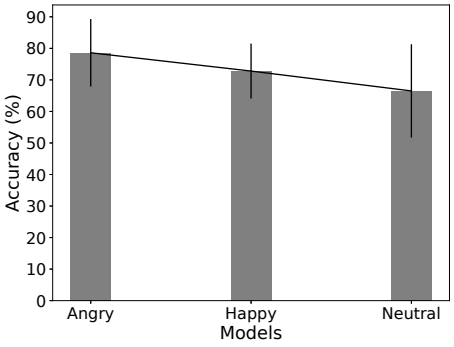


Fig. 5: Classification performance comparison for different models trained with a specific type of emotion stimuli one time. The findings reveal with *angry, happy* stimuli, the ASD detection performance is better.

(Fig. 4b). This emphasizes the importance of the emotion stimuli while detecting ASD.

To investigate further the influence of a specific type of emotion stimuli, we carry out the following experiment. In this setup, we train and evaluate three separate models, each trained with one specific emotion stimuli (*angry* or *happy* or *neutral*) at a time. The models were trained using the identical architecture and validated adopting the same *leave-one-subject-out* approach as as described earlier. We report the findings from the study in Fig.5. In this case also, we observe that the model trained with *angry* and *happy* emotion stimuli return comparatively better performance than the model trained with only *neutral* sessions. This further reinforces the earlier findings that using a specific type of emotion stimuli (i.e., *angry* or *happy*) may help to detect ASD sessions more accurately than using *neutral* stimuli.

V. Conclusion

We investigated the feasibility of emotion-guided gaze response in an interactive VR environment for ASD screening. To this end, we developed an LSTM-based classification model that leverages three key eye gaze response parameters (pupil diameter, fixation duration, and fixation location) to identify a subject-agent interaction session as ASD or TD. Our observations are as follows: (i) Our recurrent model can distinguish between ASD and sessions with an average accuracy of 77%. (ii) ASD screening accuracy is higher ($>$80%) when the subjects' eye gaze signals are in response to emotional stimuli. Our light-weight deep model thus can be an inexpensive yet effective option for ASD screening from short segments of eye gaze data recorded in response to emotional stimuli.

References

[1] P Bolton, H Macdonald, A Pickles, P al Rios, S Goode, M Crowson, A Bailey, and M Rutter, "A case-control family history study of autism," *Journal of child Psychology and Psychiatry*, vol. 35, no. 5, pp. 877–900, 1994.

[2] G Dawson, A N Meltzoff, J Osterling, J Rinaldi, and E Brown, "Children with autism fail to orient to naturally occurring social stimuli," *Journal of autism and developmental disorders*, vol. 28, no. 6, pp. 479–485, 1998.

[3] A Klin, W Jones, R Schultz, F Volkmar, and D Cohen, "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism," *Archives of General Psychiatry*, vol. 59(9), pp. 809–816, 2002.

[4] R B Grossman, R Edelson, and H Tager-Flusberg, "Emotional facial and vocal expressions during story retelling by children and adolescents with high-functioning autism," *J Speech, Language, and Hearing Research*, vol. 56(3), pp. 1035–1044, 2013.

[5] T Guha, Z Yang, R B Grossman, and S S Narayanan, "A computational study of expressive facial dynamics in children with autism," *IEEE Trans. Affective Computing*, vol. 9(1), pp. 14–20, 2016.

[6] Z Akhtar and T Guha, "Computational analysis of gaze behavior in autism during interaction with virtual agents," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1075–1079.

[7] C Trepagnier, M M Sebrechts, and R Peterson, "Atypical face gaze in autism," *Cyberpsychology & Behavior*, vol. 5(3), pp. 213–217, 2002.

[8] K A Pelphrey, J P Morris, and G McCarthy, "Neural basis of eye gaze processing deficits in autism," *Brain*, vol. 128(5), pp. 1038–1048, 2005.

[9] E Bal, E Harden, D Lamb, Amy V Van H, J W Denver, and S W Porges, "Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state," *J Autism and Developmental Disorders*, vol. 40, no. 3, pp. 358–370, 2010.

[10] L L Speer, A E Cook, W M McMahon, and E Clark, "Face processing in children with autism: Effects of stimulus contents and type," *Autism*, vol. 11, no. 3, pp. 265–277, 2007.

[11] K M Dalton, B M Nacewicz, T Johnstone, H S Schaefer, M A Gernsbacher, H H Goldsmith, A L Alexander, and R J Davidson, "Gaze fixation and the neural circuitry of face processing in autism," *Nature neuroscience*, vol. 8, no. 4, pp. 519, 2005.

[12] K A Pelphrey, N J Sasson, J S Reznick, G Paul, B D Goldman, and J Piven, "Visual scanning of faces in autism," *Journal of autism and developmental disorders*, vol. 32, no. 4, pp. 249–261, 2002.

[13] M Bellani, L Fornasari, L Chittaro, and P Brambilla, "Virtual reality in autism: state of the art," *Epidemiology and Psychiatric Sciences*, vol. 20(3), pp. 235–238, 2011.

[14] S Kuriakose and U Lahiri, "Understanding the psycho-physiological implications of interaction with a virtual reality-based system in adolescents with autism: a feasibility study," *IEEE Trans Neural Systems and Rehab Engg*, vol. 23, no. 4, pp. 665–675, 2015.

[15] S Alves, A Marques, C Queirós, and V Orvalho, "Lifeisgame prototype: A serious game about emotions for children with autism spectrum disorders.," *PsychNology Journal*, vol. 11, no. 3, 2013.

[16] P Mitchell, S Parsons, and A Leonard, "Using virtual environments for teaching social understanding to 6 adolescents with autistic spectrum disorders," *Journal of autism and developmental disorders*, vol. 37, no. 3, pp. 589–600, 2007.

[17] P R KB, P Oza, and U Lahiri, "Gaze-sensitive virtual reality based social communication platform for individuals with autism," *IEEE Trans Affective Computing*, 2017.

[18] A S DiCriscio and V Troiani, "Pupil adaptation corresponds to quantitative measures of autism traits in children," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[19] K Ahuja, A Bose, M Jain, K Dey, A Joshi, K Achary, B Varkey, C Harrison, and M Goel, "Gaze-based screening of autistic traits for adolescents and young adults using prosaic videos," in *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2020, pp. 324–324.

[20] D Alie, M H Mahoor, W I Mattson, D R Anderson, and D S Messinger, "Analysis of eye gaze pattern of infants at risk of autism spectrum disorder using markov models," in *2011 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 2011, pp. 282–287.

[21] W Jones and A Klin, "Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism," *Nature*, vol. 504, no. 7480, pp. 427–431, 2013.

[22] A Klin, W Jones, R Schultz, F Volkmar, and D Cohen, "Defining and quantifying the social phenotype in autism," *American Journal of Psychiatry*, vol. 159, no. 6, pp. 895–908, 2002.