Vision-Based Classification of Developmental Disorders Using Eye-Movements

Guido Pusiol[1], Andre Esteva[2], Scott Hall[3], Michael Frank[4],
Arnold Milstein[5], and Li Fei-Fei[1]

[1] Dept. of Computer Science, Stanford University, USA
[2] Dept. of Electrical Engineering, Stanford University, USA
[3] Dept. of Psychiatry, Stanford University, USA
[4] Dept. of Psychology, Stanford University, USA
[5] Dept. of Medicine, Stanford University, USA

Abstract. This paper proposes a system for fine-grained classification of developmental disorders via measurements of individuals' eye-movements using multi-modal visual data. While the system is engineered to solve a psychiatric problem, we believe the underlying principles and general methodology will be of interest not only to psychiatrists but to researchers and engineers in medical machine vision. The idea is to build features from different visual sources that capture information not contained in either modality. Using an eye-tracker and a camera in a setup involving two individuals speaking, we build temporal attention features that describe the semantic location that one person is focused on relative to the other person's face. In our clinical context, these temporal attention features describe a patient's gaze on finely discretized regions of an interviewing clinician's face, and are used to classify their particular developmental disorder.

1 Introduction

Autism Spectrum Disorders (ASD) is an important developmental disorder with both increasing prevalence and substantial social impact. Significant effort is spent on early diagnosis, which is critical for proper treatment. In addition, ASD is also a highly heterogeneous disorder, making diagnosis especially problematic. Today, identification of ASD requires a set of cognitive tests and hours of clinical evaluations that involve extensively testing participants and observing their behavioral patterns (e.g. their social engagement with others). Computer-assisted technologies to identify ASD are thus an important goal, potentially decreasing diagnostic costs and increasing standardization.

In this work, we focus on Fragile-X-Syndrome (FXS). FXS is the most common known genetic cause of autism [5], affecting approximately 100,000 people in the United States. Individuals with FXS exhibit a set of developmental and cognitive deficits including impairments in executive functioning, visual memory and perception, social avoidance, communication impairments and repetitive behaviors [14]. In particular, as in ASD more generally, eye-gaze avoidance during social interactions with others is a salient behavioral feature of individuals with FXS. FXS is an important case study for ASD because it can be diagnosed easily as a single-gene mutation. For our purposes, the focus on FXS means that ground-truth diagnoses are available and heterogeneity of symptoms in the affected group is reduced.

Maintaining appropriate social gaze is critical for language development, emotion recognition, social engagement, and general learning through shared attention [3]. Previous studies [10,4] suggest that gaze fluctuations play an important role in the characterization of individuals in the autism spectrum. In this work, we study the underlying patterns of visual fixations during dyadic interactions. In particular we use those patterns to characterize different developmental disorders.

We address two problems. The first challenge is to build new features to characterize fine behaviors of participants with developmental disorders. We do this by exploiting computer vision and multi-modal data to capture detailed visual fixations during dyadic interactions. The second challenge is to use these features to build a system capable of discriminating between developmental disorders. The remainder of the paper is structured as follows: In section 2, we discuss prior work. In section 3, we describe the raw data: its collection and the sensors used. In section 4, we describe the built features and analyze them. In section 5, describe our classification techniques. In section 5, we describe the experiments and results. In section 6 we discuss the results.
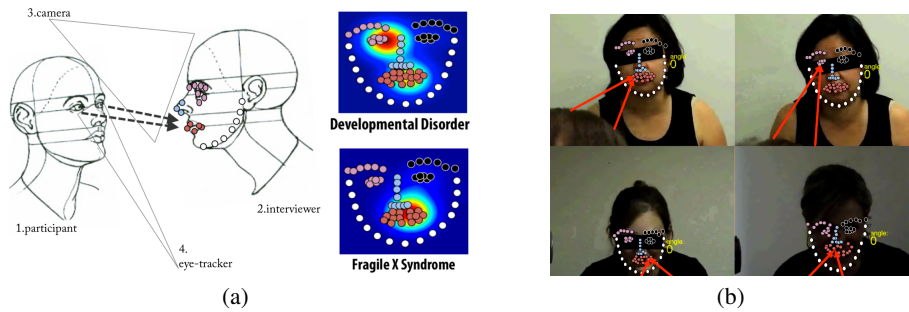

Fig. 1. (a) We study social interactions between a participant with a mental impairment and an interviewer, using multi-modal data from a remote eye-tracker and camera. The goal of the system is to achieve fine-grained classification of developmental disorders using this data. (b) A frame from videos showing the participant's view (participant's head is visible in the bottom of the frame). Eye-movements were tracked with a remote eye-tracker and mapped into the coordinate space of this video.

2 Previous Work

Pioneering work by Rehg et al. [12] shows the potential of using coarse gaze information to measure relevant behavior in children with ASD. However, this work does not address the issue of fine-grained classification between ASD and other disorders in an automated way. Our work thus extends this work to develop a means for disorder classification via multi-modal data. In addition, some previous efforts in the classification of developmental disorders such as epilepsy and schizophrenia have relied on using electroencephalogram (EEG) recordings [11]. These methods are accurate, but they require long recording times; in addition, the use of EEG probes positioned over a participant's scalp and face can limit applicability to developmental populations. Meanwhile, eye-tracking has long been used to study autism [1,7], but we are not aware of

an automated system for inter-disorder assessment using eye-tracking such as the one proposed here.

3 Dataset

Our dataset consists of 70 videos of an clinician interviewing a participant, overlaid with the participant's point of gaze (as measure by a remote eye-tracker), first reported in [6].

The participants were diagnosed with either an idiopathic developmental disorder (DD) or Fragile-X-Syndrome (FXS). DD presents similar autistic symptoms to FXS, but does not have FXS or any other known genetic syndrome. There are known gender-related behavioral differences between FXS participants, so we further subdivided this group by gender into males (FXS-M) and females (FXS-F). There were no gender-related behavioral differences in the DD group, and genetic testing confirmed that DD participants did not have FXS.

Participants were between 12 and 28 years old, with 51 FXS participants (32 male, 19 female) and 19 DD participants. The two groups were well-matched on chronological and developmental age, and had similar mean scores on the Vineland Adaptive Behavior Scales (VABS), a well-established measure of developmental functioning. The average score was 58.5 ($SD = 23.47$) for individuals with FXS and 57.7 ($SD = 16.78$) for controls, indicating that the level of cognitive functioning in both groups was $2 - 3$ SDs below the typical mean.

Participants were each interviewed by a clinically-trained experimenter. In our setup the camera was placed behind the patient and facing the interviewer. Figure 1 depicts the configuration of the interview, and of the physical environment. Eye-movements were recorded using a Tobii X120 remote corneal reflection eye-tracker, with time-synchronized input from the scene camera. The eye-tracker was spatially calibrated to the remote camera via the patient looking at a known set of locations prior to the interview.

4 Visual Fixation Features

A goal of our work is to design features that simultaneously provide insight into these disorders and allow for accurate classification between them. These features are the building blocks of our system, and the key challenge is engineering them to properly distill the most meaningful parts out of the raw eye-tracker and video footage. We capture the participant's point of gaze and its distribution over the interviewer's face, 5 times per second during the whole interview. There are 6 relevant regions of interest: *nose, left eye, right eye, mouth, jaw, outside face*. The precise detection of these fine-grained features enables us to study small changes in participants' fixations at scale.

For each video frame, we detected a set of 69 landmarks on the interviewer's face using a part-based model [16]. Figure 1 shows examples of landmark detections. In total, we processed 14,414,790 landmarks. We computed 59K, 56K and 156K frames for DD, FXS-Female, and FXS-Male groups respectively. We evaluated a sample of 1K
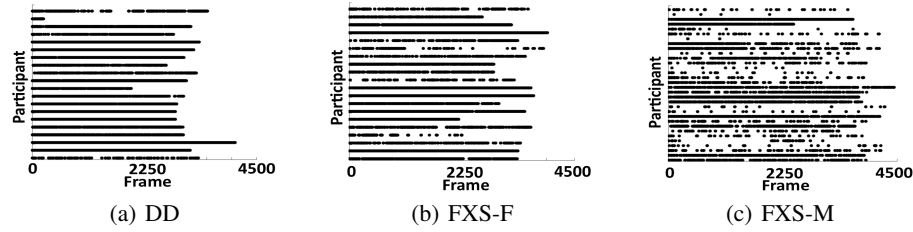
Fig. 2. Temporal analysis of attention to face. X axis represents time in frames (in increments of 0.2 seconds). Y axis represents each participant. Black dot represent time points when the participant was looking at the interviewer's face. White space signifies that they were not.

randomly selected frames, out of which only a single frame was incorrectly annotated. We mapped the eye-tracking coordinates to the facial landmark coordinates with a linear transformation. Our features take the label of the cluster (e.g. jaw) holding the closest landmark to the participant point of gaze. We next present some descriptive analyses of these data.

Feature granularity. We want to analyze the relevance of our fine grained attention features. Participants—especially those with FXS—spent only a fraction of the time looking at the interviewer's face. Analyzing the time series data of when individuals are glancing at the face of their interviewer (see Figure 2), we observe high inter-group participant's variance. For example, most of FXS-F individual sequences could be easily confused with the other groups.

Clinicians often express the opinion that the *distribution* of fixations, not just the sheer lack of face fixations—seem related to the general autism phenotype [10,8]. This opinion is supported by the distributions in Figure 3: DD and FXS-F are quite similar, whereas FXS-M is distinct. FXS-M focuses primarily on mouth (4) and nose (1) areas.
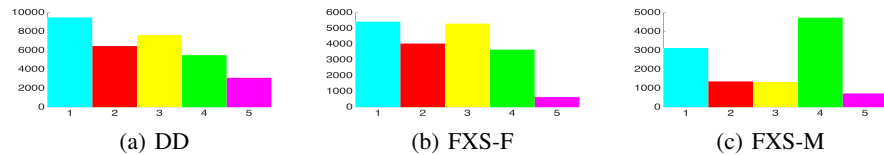

Fig. 3. Histograms of visual fixation for the various disorders. X-axis represents fixations, from left to right: nose (1), eye-left (2), eye-right (3), mouth (4), and jaw (5). The histograms are computed with the data of all participants. The non-face fixation is removed for visualization convenience.

Attentional transitions. In addition to the distribution of fixations, clinicians also believe that the *sequence* of fixations describe underlying behavior. In particular, FXS participants often glance to the face quickly and then look away, or scan between non-eye regions. Figure 4 shows region-to-region transitions in a heatmap. There is a marked
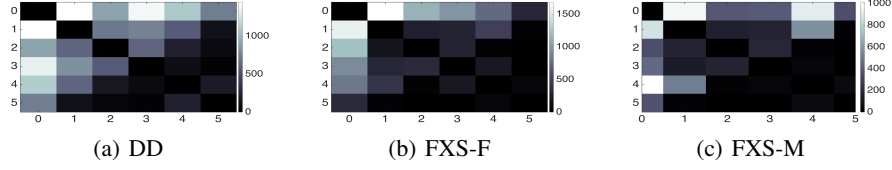
(a) DD (b) FXS-F (c) FXS-M

Fig. 4. Matrix of attentional transitions for each disorder. Each square $[ij]$ represents the aggregated number of times participants of each group transitioned attention from state i to state j. The axes represent the different states: non-face (0), nose (1), eye-left (2), eye-right (3), mouth (4), and jaw (5).

difference between the different disorders: Individuals with DD make more transitions, while those with FXS exhibit significantly less—congruent with the clinical intuition. The transitions between facial regions better identify the three groups than the transitions from non-face to face regions. FXS-M participants tend to swap their gaze quite frequently between mouth and nose, while the other two do not. DD participants exhibit much more movement between facial regions, without any clear preference. FXS-F patterns resemble DD, though the pattern is less pronounced.

Approximate Entropy. We next estimate Approximate Entropy ($ApEn$) analysis to provide a measure of how predictable a sequence is [13] . A lower entropy value indicates a higher degree of regularity in the signal. For each group (DD, FXS-Female, FXS-Male), we selected 15 random participants sequences. We compute $ApEn$ by varying w (sliding window length). Figure 5 depicts this analysis. We can see that there is great variance amongst individuals of each population, many sharing similar entropy with participants of other groups. The high variability of the data sequences makes them harder to classify.
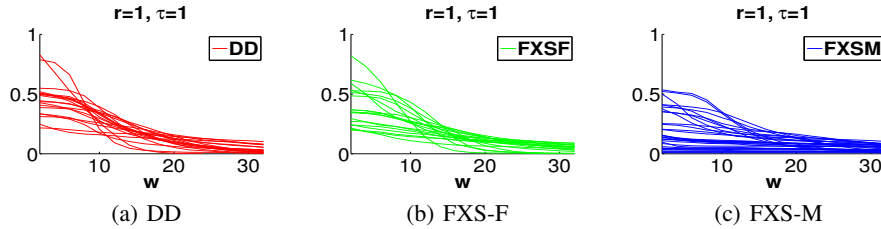

(a) DD (b) FXS-F (c) FXS-M

Fig. 5. (a) - (c) Analysis of the $ApEn$ of the data per individual varying the window length parameter w. Y-axis is $ApEn$ and X-axis varies w. Each line represents one participant's data. We observe great variance among individuals.

5   Classifiers

The goal of this work is to create an end-to-end system for classification of developmental disorders from raw visual information. So far we have introduced features that

capture social attentional information and analyzed their temporal structure. We next need to construct methods capable of utilizing these features to predict the specific disorder of the patient.

Model (RNN). The Recurrent Neural Network (RNN) is a generalization of feedforward neural networks to sequences. Our deep learning model is an adaptation of the attention-enhanced RNN architecture proposed by Hinton et al. [15] (LSTM+A). The model has produced impressive results in other domains such as language modeling and speech processing. Our feature sequences fit this data profile. In addition, an encoder-decoder RNN architecture allows us to experiment with sequences of varying lengths in a cost-effective manner. Our actual models differ from LSTM+A in two ways. First, we have replaced the LSTM cells with GRU cells [2], which are are memory-efficient and could provide a better fit to our data [9]. Second, our decoder produces a single output value (i.e. class). The decoder is a single-unit multi-layered RNN (without unfolding) and with a soft-max output layer. Conceptually it could be seen as a many-to-one RNN, but we present it as a configuration of [15] given its proximity and our adoption of the attention mechanism.

For our experiments, we used 3 RNN configurations: RNN_128: 3 layers of 128 units; RNN_256: 3 layers of 256 units; RNN_512: 2 layers of 512 units. These parameters were selected considering our GPU memory allocation limitation.
We trained our models for a total of 1000 epochs. We used batches of sequences, SGD with momentum and max gradient normalization (0.5).

Other Classifiers. We also trained shallow baseline classifiers. We engineer a convolutional neural network approach (CNN) that can exploit the local-temporal relationship of our data. It is composed of one hidden layer of 6 convolutional units followed by point-wise sigmoidal nonlinearities. The feature vectors computed across the units are concatenated and fed to an output layer composed of an affinity transformation followed by another sigmoid function. We also trained support vector machines (SVMs), Naive Bayes (NB) classifiers, and Hidden Markov Models (HMMs).

6   Experiments and Results

By varying the classification methods described in Section 5 we perform a quantitative evaluation of the overall system. We assume the gender of the patient is known, and select the clinically-relevant pair-wise classification experiments DD vs FXS-F and DD vs FXS-M. For the experiments we use 32 FXS-male, 19 FXS-female and 19 DD participants. To maintain equal data distribution in training and testing we build S_{train} and and S_{test} randomly shuffling participants of each class ensuring a 50%/50% distribution of the two participant classes over the sets. At each new training/testing fold the process is repeated so that the average classification results will represent the entire set of participants. We classify the developmental disorder of the participants, given their individual time-series feature data p, to evaluate the precision of our system. For N total participants, we create an 80%/20% training/testing dataset such that no participant's data is shared between the two datasets. For each experiment, we performed 10-fold cross validation where each fold was defined by a new random 80/20 split of the participants –about 80 participant's were tested per experiment.

	window length	DD vs FXS-female (precision)	DD vs FXS-male (precision)
SVM	3	0.65	0.83
	10	0.65	0.80
	50	0.55	0.85
N.B	3	0.60	0.85
	10	0.60	0.87
	50	0.60	0.75
HMM	3	0.67	0.81
	10	0.66	0.82
	50	0.68	0.74
CNN	3	0.68	0.82
	10	0.68	0.90
	50	0.55	0.77
RNN_128	3	0.69	0.79
RNN_250	10	0.79	0.81
RNN_512	50	**0.86**	**0.91**

Table 1. Comparison of precision of our system against other classifiers. Columns denote pairwise classification precision of participants for DD vs FXS-female and DD vs FXS-male binary classification. Classifiers are run on 3,10, and 50 seconds time windows. We compare the system classifier, RNN to CNN, SVM, NB, and HMM algorithms.

Metric. We consider the binary classification of an unknown participant as having DD or FXS. We adopt a voting strategy where, given a patient's data $p = [f_1, f_2, ...f_T]$, we classify all sub-sequences s of p of fixed length w using a sliding-window approach. In our experiments, w correspond to 3, 10, and 50 seconds of video footage. To predict the participant's disorder, we employ a max-voting scheme over each class. The predicted class C of the participant is given by:

$$C = \underset{c \in \{C_1, C_2\}}{\operatorname{argmax}} \sum_{\text{sub-seq. } s} \mathbf{1}(\text{Class}(s) = c) \tag{1}$$

Where $C_1, C_2 \in \{DD, FXS\text{-}F, FXS\text{-}M\}$, $\text{Class}(s)$ is the output of a classifier given input s. We use 10 cross validation folds to compute the average classification precision.

Results. The results are reported in Table 1. We find that the highest average precision is attained using the RNN.512 model with a 50 second time window. It classifies DD versus FXS-F with 0.86 precision and DD versus FXS-M with 0.91 precision. We suspect that the salient results produced by the RNN_512 are related to its high capacity and its capability of representing complex temporal structures.

7 Conclusion

We hereby demonstrate the use of computer vision and machine learning techniques in a cost-effective system for assistive diagnosis of developmental disorders that exhibit visual phenotypic expression in social interactions. Data of experimenters interviewing participants with developmental disorders was collected using video and a remote eye-tracker. We built visual features corresponding to fine grained attentional fixations, and developed classification models using these features to discern between FXS and idiopathic developmental disorder. Despite finding a high degree of variance and noise

in the signals used, our high accuracies imply the existence of temporal structures in the data.

This work serves as a proof of concept of the power of modern computer vision systems in assistive development disorder diagnosis. We are able to provide a high-probability prediction about specific developmental diagnoses based on a short eye-movement recording. This system, along with similar ones, could be leveraged for remarkably faster screening of individuals. Future work will consider extending this capability to a greater range of disorders and improving the classification accuracy.

References

1. Boraston, Z., Blakemore, S.J.: The application of eye-tracking technology in the study of autism. The Journal of Physiology 581(3), 893–898 (Jun 2007) 2
2. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches arXiv:1409.1259 (2014) 6
3. Csibra, G., Gergely, G.: Social learning and social cognition: The case for pedagogy. In Process of change in brain and cognitive development, attention and performance (2006) 2
4. Golarai, G., Grill-Spector, K., Reiss, A.: Autism and the development of face processing. Clinical neuroscience research p. 145160 (2006) 2
5. Hagerman, P.J.: The fragile X prevalence paradox. Journal of medical genetics (2008) 1
6. Hall, S.S., Frank, M.C., Pusiol, G.T., Farzin, F., Lightbody, A.A., Reiss, A.L.: Quantifying naturalistic social gaze in fragile x syndrome using a novel eye tracking paradigm. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics (2015) 3
7. Hashemi, J., Spina, T.V., Tepper, M., Esler, A., Morellas, V., Papanikolopoulos, N., Sapiro, G.: A computer vision approach for the assessment of autism-related behavioral markers. In: 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL). pp. 1–7. IEEE (2012) 2
8. Jones, W., Klin, A.: Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism. Nature (2013) 4
9. Jzefowicz, R., Zaremba, W., Sutskever, I.: An empirical exploration of recurrent network architectures. In: Bach, F.R., Blei, D.M. (eds.) ICML. JMLR Proceedings, vol. 37, pp. 2342–2350. JMLR.org (2015) 6
10. Klin, A., Jones, W., Schultz, R., Volkmar, F., Cohen, D.: Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. Archives of general psychiatry 59(9), 809–816 (2002) 2, 4
11. Kumar, Y., Dewal, M.L., Anand, R.S.: Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine. Neurocomputing 133 (Jun 2014) 2
12. Rehg, J.M., Rozga, A., Abowd, G.D., Goodwin, M.S.: Behavioral imaging and autism. IEEE Pervasive Computing 13(2), 84–87 (2014) 2
13. Restrepo, J.F., Schlotthauer, G., Torres, M.E.: Maximum approximate entropy and r threshold: A new approach for regularity changes detection arXiv:1405.7637(nlin.CD) (2014) 5
14. Sullivan, K., Hatton, D.D., Hammer, J., Sideris, J., Hooper, S., Ornstein, P.A., Bailey, D.B.: Sustained attention and response inhibition in boys with fragile X syndrome: measures of continuous performance. American Journal of Medical Genetics. Part B: Neuropsychiatric Genetics 144B(4), 517–532 (2007) 1
15. Vinyals, O., Kaiser, L.u., Koo, T., Petrov, S., Sutskever, I., Hinton, G.: In: Advances in Neural Information Processing Systems 28. pp. 2773–2781 6
16. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR. pp. 2879–2886. IEEE (2012) 3