Vision-based classification of developmental disorders using eye-movements

Anonymous ICCV submission

Paper ID 2441

Abstract

This paper proposes a system for fine-grained classification of developmental disorders via measurements of individuals' eye-movements using multi-modality 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 computer vision.

The main idea is to build multi-modal features from two pixel-based 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 setup, the temporal attention features refer to a patient's gaze on finely discretized regions of an interviewing clinician's face, and are used to classify their particular developmental disorder. We show that our best method for classification is a state-vector based convolutional neural network (CNN) that achieves 90% accuracy.

1. Introduction

Autism Spectrum Disorder (ASD), is an important developmental disorder to account for and understand in our society today. Significant efforts are spent in early diagnosis, which is a key component to proper treatment. Today, identification of a ASD requires a set of cognitive tests and hours spent on clinical evaluations which involve extensively testing the participant and observing their behavioral patterns (i.e. their social engagement with others). Physicians go through extensive training and require substantial experience to properly assess ASD. The problem with this procedure is that it is both laborious and imprecise due to variability in the clinician's subjective judgement. With the advent of computers and machine learning techniques, people are beginning to look for computer-assisted technology to identify ASD.

In this work, we build multi-modal temporal features that describe the location of an interviewer's face that a patient is paying attention to, and employ state-vector based CNNs to create an automated system to assist physicians in the diagnosis of ASD (see Figure 1).

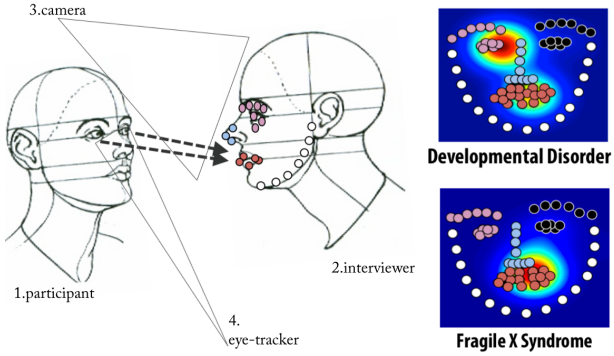


Figure 1. We study social interactions between a participant (1) with a mental impairment and an interviewer (2), using multi-modal data from a remote eye-tracker and camera (3-4). The goal of the system is to achieve fine-grained classification between developmental disorders using this data.

Specifically, we focus on Fragile-X-Syndrome (FXS). FXS is the most common known genetic cause of autism [17], affecting approximately 1 in 3,000 individuals in the United States (approximately 100,000 people nationally). Individuals with FXS often show prominent eye gaze deficits during social encounters in which they actively seek to avoid social interaction [6, 7]. These social-attentional behavioral phenotypes are considered to be salient cues which drive investigations to better asses mental impairment [22]. We build descriptive features from this data that capture social engagement between participant and experimenter. We then use these features to train classifiers to discern between different participants' disorder.

In the rest of this section we motivate our contribution and approach. In section 2, we discuss prior work. In section 3, we describe the data: its collection and the sensors used. In section 4, we present an overview of the proposed system. In section 5, we describe the built features and analyze them. In section 6, we apply classification techniques. In section 7, we describe our approach for classifying the mental condition of an unknown participant. In section 8

we discuss the results and future work.

Our contribution and approach. We build a system capable of performing automatic assessment of a participant's developmental condition using features derived from multi-modal data and classifiers adapted to our problem. We perform quantitative analysis of our data in order to provide insight into the differences and variance in the behavioral patterns of these developmental conditions. Our results show that this system works well above chance, in some cases attaining 90% accuracy. In addition, our data analysis reveals that patterns in eye gaze considered at the fine granularity of facial regions (i.e. nose, mouth, etc.) are strong cues in diagnosing a participant (Figure 1).

Note that our contribution is not based on novel algorithms but in the observation that multi-modal data can be crafted to reveal high-level behavioral characteristics of an individual based on finely discretized actions - in our case revealing a developmental disorder based on regional attention to another's face during conversation.

We believe this will interest both computer vision researchers and engineers working on similar problems, such as emotion classification or activity understanding based on the interaction of people (i.e. understanding social roles).

2. Previous Work

Pioneering work by [30] shows the potential of using eye-tracking information to measure relevant behavior in children with ASD. However, it did not address the issue of fine-grained classification between ASD/FXS and other disorders in an automated way. Our work extends this to develop a means for FXS classification via multi-modal data.

2.1. Classification of disorders

Previous efforts in the classification of developmental disorders such as epilepsy and schizophrenia have relied on using electroencephalogram (EEG) and neurophysiological signals [25, 33]. These methods are accurate, but they suffer from long recording times and the use of EEG probes positioned all over a participants scalp and face. Meanwhile, eye-tracking has long been used to study autism and characterize the disease [3, 20, 9], but a rigorous automated system complete with visual feature extraction and classification has yet to be proposed. Here we propose a step towards an integrated approach to solving this problem.

2.2. Cognitive Impairment Studies

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 [18, 29, 36, 35]. In particular [19] shows that eye-gaze avoidance during social interactions with others is a salient behavioral feature of individuals with FXS. Maintaining appropriate social gaze is critical for language development, emotion recognition, social engagement, and general learning through shared attention [8, 28, 11]. Studies have indicated that high levels of gaze avoidance are characteristic of poor social interaction skills [10, 32]. For instance, when shown images of faces depicting various emotions, participants with FXS looked significantly less at the eye region of the faces, and were more likely to look at the nose region compared to healthy individuals. Despite this, few researchers have attempted to employ eye-tracking methodology to quantify social gaze behavior in real-life social settings [12, 13]. Limitations of previous studies include a small sample of participants, lack of automation & scalability, and the fact that the controls were not matched on level of cognitive functioning to the participants with FXS. As such, in the present study we utilize data in which a larger group of individuals with FXS (both males and females) were studied and in which controls were cognitively matched.

2.3. The mechanism for human visual attention

It is commonly believed that visual attention is driven by two mechanisms: 1) A bottom-up (BU), task independent, image-based mechanism that instinctively guides the human eyes into salient image regions such as discontinuities, color, texture, motion, etc. [26], 2) A top-down (TD) mechanism that guides attention and gaze in a task-dependent and goal-directed fashion, that is able to manage the sequential acquisition of information from the visual environment [37, 4]. We focus on this second category and investigate the underlying visual structures behind social interaction engagement. Despite its importance in understanding human behaviors, only a few approaches have addressed this TD mechanism. Previous TD work tackles the problem of predicting gaze position using a head mounted monocular camera [14, 34]. There the goal was to mimic eye-tracking information with egocentric video information and not to understand human behavior. Egocentric activity detection by predicting hand-object interactions from a head-mounted camera [15] has been addressed, as has social event detecting by facial direction analysis [16]. Previous work has addressed the problem of understanding cognitive and motivational factors of the person wearing the eye tracker using manual analysis of the data [38]. To the best of our knowledge, automated classification of cognitive impairments using eye-tracking data remains an unaddressed problem. In this work we build unique features that capture the location of attention of an individual for the automatic screening and diagnosis of mental conditions.

3. Dataset

Our data are 70 10-minute videos of an experimenter interviewing a patient, overlaid with the patient's point of gaze (as measure by a remote eye-tracker). These participants have either FXS or some other developmental disorder (DD).

3.1. Participants

Participants were patients diagnosed with either an idiopathic developmental disorder or Fragile-X-Syndrome (FXS). 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 patients did not have FXS. Patients were between 12 and 28 years old, with 51 FXS patients (32 male, 19 female) and 19 DD patients. 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 (std = 16.78) for controls, indicating that the level of cognitive functioning in both groups was 2 – 3 SDs below the typical mean.

3.2. Data Collection Methods

Patients were each interviewed for ∼10 minutes by a clinically-trained experimenter. In our setup the camera was placed behind the patient and facing the interviewer. Figure 2 depicts the configuration of the interview. 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 calibrated to the remote camera via the patient looking at a calibration board (60cm x 60cm) prior to the interview. The eye-tracker has an accuracy of 0.5°, its data rate is 120Hz, its latency is 35 ms. The eye-tracker allows a freedom-of-head movement of 30 x 22 x 30 cm with a FOV of 22 x 22 x 30 cm.

4. System Architecture

The assistive system takes as input multi-sensorial eye-tracking and video data and outputs a prediction on the developmental disorder of the participant. Figure 4 depicts the architecture of the system. It is composed of the following three parts, elaborated further in sections 5, 6, 7:

Feature Extraction. We build our features by mapping the eye-tracking information onto facial regions of the interviewer.

Classifier Training. We define and build classification models based on these features to label sequences of features as belonging to a particular mental condition.

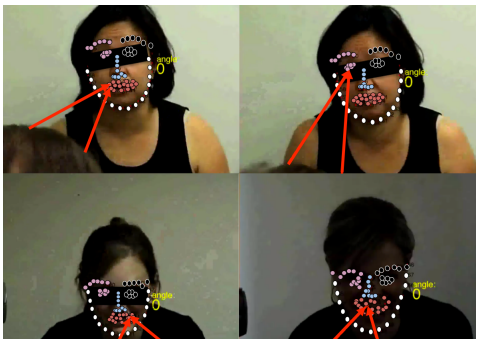
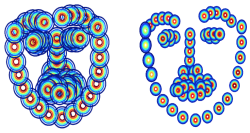


Figure 2. A frame from videos showing the participant's view (participant's head is visible in the bottom of the frame). Eye-movements were tracked with a remote eye-tracker and mapped into the coordinate space of this video.



(a) (b)

Figure 3. Face Motion Analysis. Size of the circles represent the spatial variability of a keypoint over the length of an interview. (a) The interviewer with the greatest facial movement; (b) the interviewer with the least;

Participant Classification. We define a method to predict the condition of a new participant using these models.

5. Feature Extraction

A goal of our work is to build proper descriptive 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. [1] e

5.1. Face Landmark Extraction

As shown in Figure 2 for each video frame we compute a set of landmarks on the face of the interviewer using a deformable-parts-model based approach [39]. We reduce the bias of the interviewer by filtering out the frames where the interviewer is not facing the participant. This represents <1% of all frames.

[1] The camera and eye tracker calibration matrix is computed using a board (60cmx60cm) with marks (20 cm) at the start of each interview. The eye-tracker is Tobii X120 having a data rate of 120 Hz, a latency of 35 ms, a mean time to tracking recovery of 100 ms, an accuracy of 0.5, a spatial resolution of .2 deg., a drift of 0.3 deg. a freedom-of-head movement of 30x22x30 cm, a tracker eld of view of 22x22x30 cm and a top head-motion speed of 35 cm/sec.

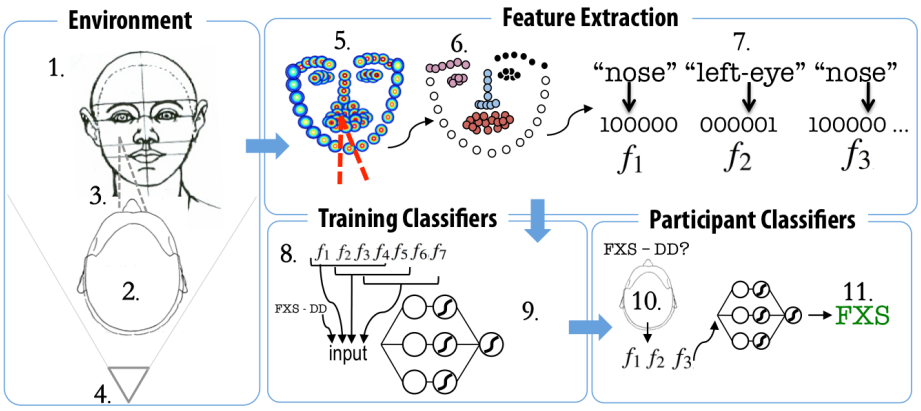


Figure 4. System Architecture. The environment is composed of an interviewer (1) facing a participant (2) who wears a remote eye-tracker (3) while a fixed-camera (4) faces the interviewer. Facial landmarks (5) on the interviewer's face are extracted from the raw video and we map the eye-tracking data onto these landmarks. This mapping is then clustered into semantic facial regions (6) and we represent these facial regions using state vectors (7). Sub-sequences of features (8) are used to train classification algorithms (9). Finally, the full feature sequences of unseen participants (10) are used to predict their developmental disorder (11).

We computed 58,587 frames for the 19 developmental disorder participants, 56,109 frames for the 19 FXS-female participants, and 94,214 frames for the 51 FXS-Male participants, at a frame rate of 5 fps (down-sampled for efficiency). On each frame we computed 69 landmarks, for a total of 14,414,790 over all frames, using 500 cores over 96 hours. Landmark quality was evaluated on a sample of 1000 randomly selected frames, out of which only a single frame was incorrectly annotated. Figure 2 shows an example of detected facial landmarks, and Figure 3(a)(b) shows the spatial range of the landmarks across two different interviewers.

5.2. Feature Construction

The eye-tracking system is calibrated to the camera coordinate system. We mapped the eye-tracking coordinates to the facial landmark coordinates with a linear transformation. The participant's point of gaze was considered to be on a particular facial landmark when the reported point of gaze was within a 20 pixel radius.

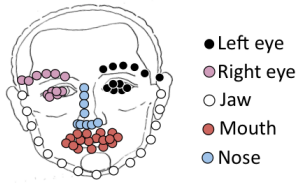


Figure 5. Clustering of the 69 key landmarks of the interviewer's face into 5 semantic facial regions. The colors represent the clustering of these landmarks into 5 facial regions. A 6th region, not shown, is reserved for when the participants look away from the face. These clusters represents our features.

We clustered the 69 landmarks into 6 regions of psychiatric interest: *not looking at the face, nose, left eye, right eye, mouth, jaw* (Figure 5). To capture the independent value of these regions (which play substantially different social roles for individuals on the autism spectrum [23]), we represent these regions as state vectors (Figure 6).

5.3. Notation

We define a feature f to be a 6-element state vector (i.e. $f =< 1,0,0,0,0,0 >$) representing the 6 facial regions and $\bar{f} \in \{0,1,2,3,4,5\}$ to be its corresponding integer-valued representation. For participant p^{α}, represented by the sequence of features f_i as

$$p^{\alpha} = [f_1^{\alpha}, f_2^{\alpha}, ...f_{T_{\alpha}}^{\alpha}] \quad (1)$$

where α indexes the participants and T_{α} represents the length of the time series, we consider the set of all individuals of a given class c to be

$$S_c = \{p^1, p^2, ...p^N\} \quad (2)$$

where c indexes the diagnostic labels {DD, FXS-female, FXS-male}.

5.4. Descriptive Analyses

We next briefly present a number of descriptive analyses of the dataset that explore the complexity of this rich data

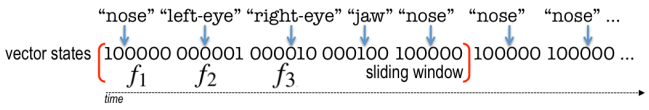
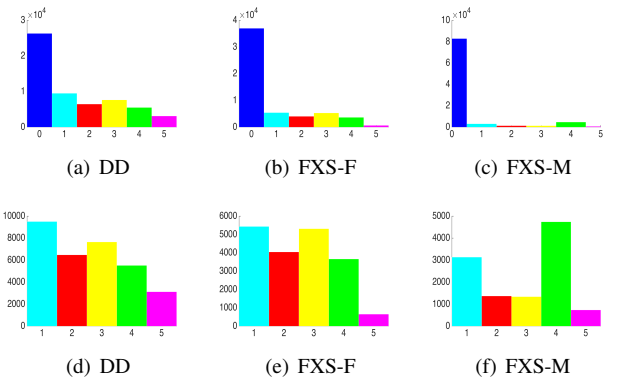


Figure 6. Facial Region State Vectors. The 5 facial regions and the non-face region are represented as state-vector features f_i. A sliding window technique is used to extract the input into our classifiers.



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

(d) DD (e) FXS-F (f) FXS-M

Figure 7. Feature histograms for the various disorders. X-axis represents different states: non-face (0), nose (1), eye-left (2), eye-right (3), mouth (4), and jaw (5). (a)-(c) feature histograms for all participants of each class. (d)-(f) equivalent histograms with the non-face state vector removed.

and justify the importance of using temporal information in classification.

5.4.1 Fourier Analysis of Features

We first consider the spectra of the time-series for each group (DD, FXS-male, FXS-female) by concatenating the integer-representation time-series of each patient $\bar{p}^{\alpha} = [\bar{f}_1^{\alpha}, \bar{f}_2^{\alpha}, ...\bar{f}_{T_{\alpha}}^{\alpha}]$ of the group into one time-series $\bar{S}_c = \{\bar{p}^1, \bar{p}^2, ...\bar{p}^N\}$ and taking the Fourier transform. Each group exhibits strong near-zero-frequency components with weak and noisy higher order components. Further, the spectra of the signal—after eliminating the non-face state vectors—has the same pattern. This finding implies that there exist no characteristic oscillatory movements in the data: Participants tend either to fixate to single region of the face or scan across multiple regions.

5.4.2 Feature Granularity

Patients—especially those with FXS—spent the majority of the interview looking away from the interviewer, and only a fraction of the time on the interviewer's face (Figure 7). Yet clinicians often express the intuition that the *distribution* of fixations, not just the sheer lack of face fixations—seem related to the general autism phenotype [23, 21]. This intuition is supported by the distributions in Figure 7(d)-(f): DD and FXS-F are quite similar, whereas FXS-M is distinct. FXS-M focuses primarily on mouth (4) and nose (1) areas.

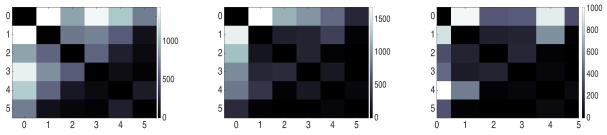
5.4.3 Attentional transitions

In addition to the distribution of fixations, the *sequence* of fixations is also remarked upon by clinicians. In particular, FXS patients often glance to the face quickly and then look away, or scan between non-eye regions. Figure 8 shows region-to-region transitions in a heatmap. There is a marked difference between the different disorders: Individuals with DD make more transitions, while those with FXS exhibit significantly less—congruent with the clinical intuiton. 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.

There is significant variance among individuals in the same group, making it challenging to classify between them. For example, Figure 9 shows time-series data of when individuals are glancing at the face of their interviewer or away from it. Note the varying sparsity of FXS males—some individuals glance at the face very frequently, whereas others may spend minutes without looking at it. FXS females exhibit similar sparsity amongst some individuals and much more consistent face-glancing amongst others. DD participants, on the other hand, tend to have a much higher frequency of glancing at the face, though a few participants also spend noticeable amounts of time looking away from it.

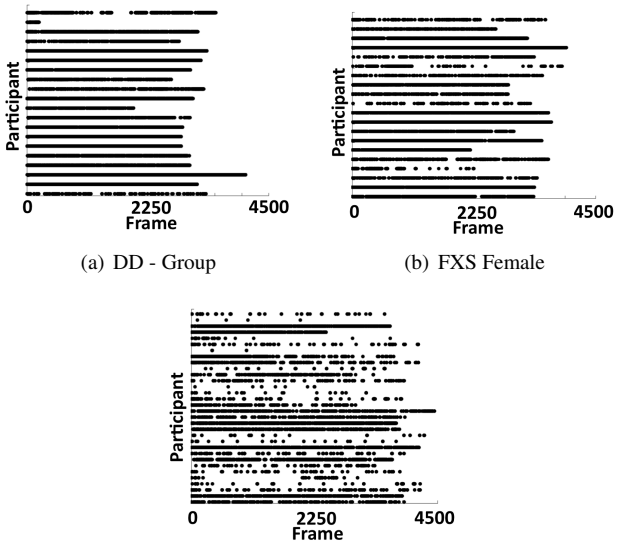
5.4.4 Approximate Entropy

We next use Approximate Entropy ($ApEn$) analysis to provide a measure of how predictable a sequence is. The $ApEN$ of a signal is a value in $[0,1]$, where a higher entropy value indicates unpredictability of fluctuations in the signal, and a low value indicates a high degree of regularity. We employ ($ApEn$) analysis [31, 2] to further emphasize the similarity between DD and FXS-F signals and the difference between FXS-M and the other two. $ApEn$ measures the logarithmic likelihood that if two vectors (q_i^w, q_j^w) representing feature sequences of length w are within a distance R, called the tolerance, in a w-dimensional space, then they remain within R in a $(w+1)$-dimensional space when their



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

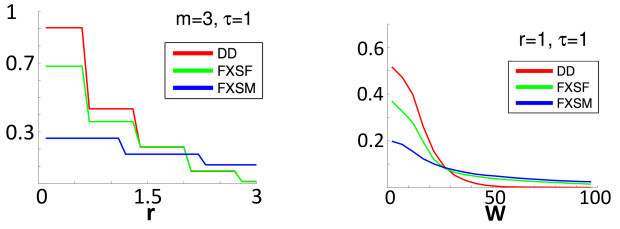
Figure 8. Matrix of attentional transitions for each disorder. Each square $[ij]$ represents the aggregated number of times participants of each group transitioned attention from state i to state j. The axes represent the different states: non-face (0), nose (1), eye-left (2), eye-right (3), mouth (4), and jaw (5).



(a) DD - Group (b) FXS Female

(c) FXS Male

Figure 9. 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.



(a) Tolerance Parameter (b) Dimension

Figure 10. Analysis of the average *ApEN* of the data for each participant class. X-axis represents (a) the tolerance parameter $r = R/std(Q)$, (b) the dimension parameter w.

length is extended by an extra feature. Greater (lesser) likelihood of remaining close produces smaller (larger) $ApEn$ values. To estimate $ApEn(R,m,N)$ for a length N feature sequence $Q = \bar{f}_1, \bar{f}_2, ..., \bar{f}_N$, given the parameters w (window length), $\tau \in \mathbb{N}$ (subsampling coefficient), and $r \in \mathbb{R}^+$, we define the w-dimensional embedded vectors $q_i^m = [\bar{f}_i, \bar{f}_{i+\tau}, \bar{f}_{i+2\tau}, ..., \bar{f}_{i+(m-1)\tau}]$, with $1 \leq i \leq N-(m-1)\tau$.

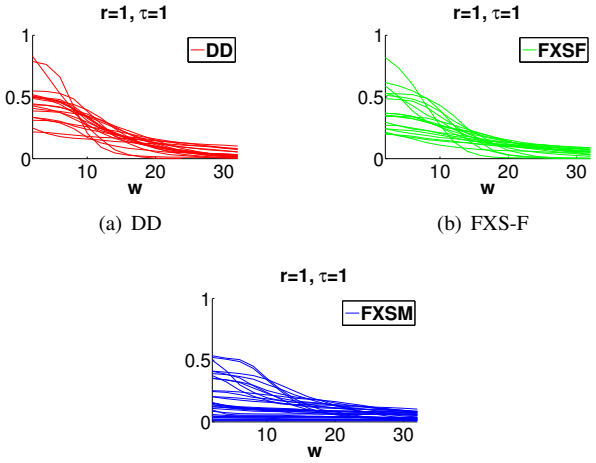
We measured the $ApEn$ of our 3 groups (DD, FXS-Female, FXS-Male). For each group S_c, we selected 10 random participants (about 30,000 features $\bar{f}_i$). We parametrize $\tau = 1$ to capture the entropy of the shortest period of time allowed by the data resolution (0.2 seconds). The tolerance parameter $r = 0.1k$, for $k = 1...30$ defines the tolerance $R = \text{std}(Q)r$. Figure 10(a) depicts the $ApEn$ for $m = 3$ while varying r. The FXS-M data is more stable than DD and FXS-F data, which share a similar entropy.

We vary w in order to understand the volatility of the signals from frame to frame. Figure 10(b) depicts this analysis. Here the parameter R is fixed to $R = \text{std}(Q)$. We see that FXS-M participants, on average, show the most stable decay for small m while FXS-F and DD share a similar decay rate. Simultaneously we see in Figure 11 that there is great variance amongst individuals of each population. This is taken using about 6.6 minutes of data (2000 features) from each participant. We compute the $ApEn$ by fixing $\tau = 1$, $r = 1$ and varying w.

6. Training of 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. Given the novelty of this dataset, the key challenge is to find methods that can tackle the intrinsic structure of the data and discern between participants. We design a one-dimensional convolutional neural network to work with state-vectors and find that it outperforms other baseline techniques.

We form an input data matrix and label vector by choosing a window length w and a step size s and converting the feature sequence of our entire dataset, $Q =$



(a) DD (b) FXS-F

(c) FXS-M

Figure 11. Analysis of the $ApEn$ of the data per individual varying the dimension parameter w. Y-axis is $ApEn$ and X-axis varies w. Each line represents one participant's data.

$[f_1^1, f_2^1, ..., f_{T_N-1}^N, f_{T_N}^N]$ into the $m \times w$ matrix

$$X = \begin{bmatrix} f_1^1 & f_2^1 & \cdots & f_w^1 \\ f_{1+s}^1 & f_{2+s}^1 & \cdots & f_{w+s}^1 \\ \vdots & \ddots & & \vdots \\ f_{T_N-w}^N & f_{T_N-w+1}^N & \cdots & f_{T_N}^N \end{bmatrix} \quad (3)$$

With corresponding labels vector $Y = [Y_1, ...Y_m]$ where the label Y_i is either 0 or 1 depending on the disorder of the patient whose features were used to form row i.

6.1. State Vector CNN

Convolutional neural networks are good models for representing data with highly correlated features (e.g. speech and image [24, 1, 5, 27]). Our feature sequences fit this data profile. We engineer a one-dimensional convolutional neural network approach that can exploit the local-temporal relationship between state vectors. The state vector CNN (S-CNN) works with states represented by sparse vectors.

Figure 12 depicts the architecture of the S-CNN. It is composed of one hidden layer of $U = 6$ 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 followed by another sigmoid function.

Recalling the state-vector representation of the features $f =< x_1, x_2, ...x_6 >$ where $x_i \in \{0,1\}$ we now represent the input rows q of X to the CNN in this state-vector form $q = [x_i, ..., x_{i+l}]$ where l is the new dimensionality of the input data under this extension.

We choose the kernel size k of the hidden units and stride s to be a multiple of $\tau = 6$, the length of a feature f, to avoid splitting state-vectors between filter steps. We parametrize $s = 6$ and $k = 24$ (i.e. 0.8 sec. of video). The S-CNN has an output layer composed of a single unit which outputs a continuous value in the range $[0,1]$. We threshold this value at 0.5 for classification.

In the convolutional layer we have that the activation map of a unit is given by a vector $\mathbf{a}$ whose j^{th} entry is given by:

$$a_j = \sigma\left(\sum_{i=1}^{k} w_i^{(1)} x_{i+(j-1)s}\right) \quad (4)$$

Where σ is the sigmoid function and $w_i^{(1)}$ are input weights. The U activation maps are concatenated into a single vector $\mathbf{v}$. The output of the S-CNN is then given by:

$$y = \sigma\left(\mathbf{v}^T \mathbf{w}^{(2)}\right) \quad (5)$$

The S-CNN is trained with backpropagation using stochastic gradient descent with momentum and a learning rate of 0.05.

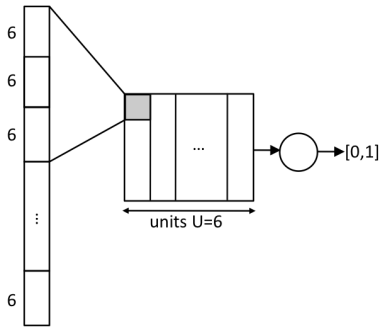


Figure 12. Convolutional Neural Network Design. The input is a binary vector representing a sequence of features. The CNN is composed of a single hidden layer and an output layer. The filters are chosen to be an integer multiple of 6, the length of the state vectors. The output is a sigmoid, thresholded to 0.5 for classification

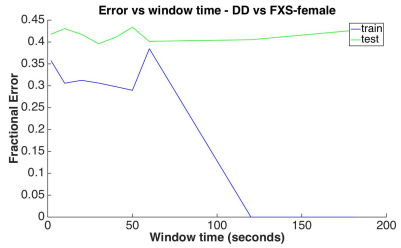
6.2. Other Classifiers

As baselines, we train support vector machines (SVMs), Naive Bayes (NB) classifiers, and Hidden Markov Models (HMMs). SVMs are trained using a chi-Squared kernel function to account for long sub-sequences of features with little variation (i.e. sequences with many 0's). We also train a two-hidden-state HMM with six possible emissions (corresponding to binary classification with vectors of six states). For the HMM classifier we take $Q = [f_1^1, f_2^1, ..., f_{T_N-1}^N, f_{T_N}^N]$ and $Y = [b^1, b^1, ..., b^N]$, where $b^i \in 0, 1$ are the labels of the two mental disorders considered and correspond to the two states of the HMM. We divide both Q and Y into non-overlapping, contiguous subsequences of length w and shuffle them to form the HMM emission sequence Q^s and state sequence Y^s. We take a training set of this data and first estimate the transmission and emission probabilities of an HMM, and then use these probabilities to predict the HMM state of a testing set.

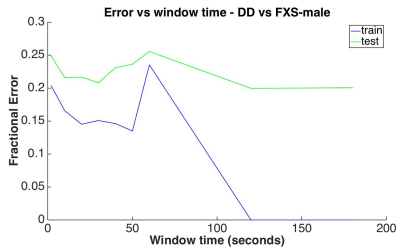
To find an optimal window length for classification, we vary the window size from 1 feature (0.2 sec) up to 900 (3 minutes) for a number of classifiers (Figure 13). We find small variance in classification error for windows up to 50 sec (= 250 features), and observe a similar result when trying other classifiers.

7. Participant Classification

To make the system transferable to clinical settings, we consider the classification an unknown participant as having DD or FXS. We adopt a divide and conquer 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. 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:



(a) DD vs FXS Female



(b) DD vs FXS Male

Figure 13. Analysis of SVM classifier training and testing error as a function of time window (in seconds) for pair-wise classifiers. One second of time corresponds to 5 features.

$$C = \underset{c \in \{C_1, C_2\}}{\operatorname{argmax}} \sum_{\text{sub-seq. } s} \mathbf{1}(\text{Class}(s) = c) \qquad (6)$$

Where $C_1, C_2 \in \{\text{DD}, \text{FXS-F}, \text{FXS-M}\}$, Class$(s)$ is the output of a classifier given input s. This simple strategy improves classification accuracy as the length of the interview increases, yet keeps the system independent of interview length.

7.1. Experiments and Results

By varying the classification methods described in Section 6 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 $S_{train} = \{p^1, p^2, ...p^{0.8N}\}$ and $S_{test} = \{p^{0.8N+1}, ...p^N\}$ such that no patient's data is shared between the two datasets. We perform 10-fold cross validation where each fold is defined by a new random 80/20 split of the participants. At each new training/testing fold the process is repeated so that the average classification results will represent the entire set of participants. We compare the S-CNN to SVM, NB, and HMM models using feature windows corresponding to 3, 10, and 50 seconds of video footage. These window lengths correspond to the observations presented in section 6.

	sec.	DD vs FXS-female	DD vs FXS-male
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
S-CNN	3	0.68	0.82
	10	**0.68**	**0.90**
	50	0.55	0.77

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 second time windows. We compare the system classifier, S-CNN, to SVM, NB, and HMM algorithms.

The system is evaluated on the average precision with which it discerns developmental disorders. The results are reported in Table 1. We find that the highest average precision is attained using the S-CNN model with a 10 second time window. It classifies DD versus FXS-F with 0.68 precision and DD versus FXS-M with 0.90 precision, a remarkably high pair of accuracies given the challenges of working with such a noisy dataset.

8. Discussion

We hereby demonstrate the use of computer vision and machine learning techniques in a rapid system for assistive diagnosis of developmental disorders that exhibit visual phenotypic expression in social interactions.

Data of experimenters interviewing patients with developmental disorders was collected using video and a remote eye-tracker. We built visual features corresponding to joint attention between interviewer and participant, and trained a state-vector based convolutional neural network model using temporal sequences of these features to discern between FXS and idiopathic developmental disorder, achieving accuracies of up to 90%. Despite finding a high degree of variance and noise in the underlying signals used, our high accuracies imply the existence of latent 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, within minutes, a high-probability prediction of the individual being afflicted with one disorder or another. 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] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn. Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. *Acoustics Speech Signal Processing ICASSP 2012 IEEE International Conference.* 7

[2] K. H. Approximate entropy for all signals Chon, C. Scully, and S. L. Medicine. Approximate entropy for all signals. *Engineering in and Biology Magazine, IEEE*, 28(6). 5

[3] Z. Boraston and S. J. Blakemore. The application of eye-tracking technology in the study of autism. *The Journal of Physiology*, 581(3):893–898, June 2007. 2

[4] A. Borji, D. N. Sihite, and L. Itti. 2012 IEEE Conference on Computer Vision and Pattern Recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 470–477. IEEE, 2012. 2

[5] J. Bouvrie. Notes on Convolutional Neural Networks. 2006. 7

[6] I. L. Cohen, G. S. Fisch, V. Sudhalter, E. G. Wolf-Schein, D. Hanson, R. Hagerman, E. C. Jenkins, and W. T. Brown. Social gaze, social avoidance, and repetitive behavior in fragile X males: a controlled study. *American Journal on Mental Retardation*, 92(5):436–446, 1988. 1

[7] I. L. Cohen, P. M. Vietze, V. Sudhalter, E. C. Jenkins, and W. T. Brown. Parent-child dyadic gaze patterns in fragile X males and in non-fragile X males with autistic disorder. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 30(6):845–856, 1989. 1

[8] G. Csibra and G. Gergely. Social learning and social cognition: The case for pedagogy. *In Process of change in brain and cognitive development, attention and performance XXI*, 2006. 2

[9] 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*, 8(4):519–526, Apr. 2005. 2

[10] G. Doherty-Sneddon, L. Whittle, and D. M. Riby. Gaze aversion during social style interactions in autism spectrum disorder and Williams syndrome. *Research in Developmental Disabilities*, 34(1):616–626, 2013. 2

[11] N. J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, 2000. 2

[12] F. Farzin, S. M. Rivera, and D. Hessl. Brief report: Visual processing of faces in individuals with fragile X syndrome: an eye tracking study. *Journal of Autism and Developmental Disorders*, 39(6):946–952, 2009. 2

[13] F. Farzin, F. Scaggs, C. Hervey, E. Berry-Kravis, and D. Hessl. Reliability of eye tracking and pupillometry measures in individuals with fragile X syndrome. *Journal of Autism and Developmental Disorders*, 41(11):1515–1522, 2011. 2

[14] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV'12: Proceedings of the 12th European conference on Computer Vision*. Springer-Verlag, 2012. 2

[15] A. Fathi and J. M. Rehg. 2013 IEEE Conference on Computer Vision and Pattern Recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2579–2586. IEEE, 2013. 2

[16] S. From Ego to Nos-Vision Detecting Social Relationships in First-Person Views Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara. From Ego to Nos-Vision: Detecting Social Relationships in First-Person Views. 2

[17] P. J. Hagerman. The fragile X prevalence paradox. *Journal of medical genetics*, 2008. 1

[18] S. Hall, M. DeBernardis, and A. Reiss. Social escape behaviors in children with fragile X syndrome. *Journal of Autism and Developmental Disorders*, 36(7):935–947, 2006. 2

[19] S. S. Hall, A. A. Lightbody, B. E. McCarthy, K. J. Parker, and A. L. Reiss. Effects of intranasal oxytocin on social anxiety in males

with fragile X syndrome. *Psychoneuroendocrinology*, 37(4):509–518, 2012. 2

[20] J. Hashemi, T. V. Spina, M. Tepper, A. Esler, V. Morellas, N. Papanikolopoulos, and G. Sapiro. 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)*, pages 1–7. IEEE, 2012. 2

[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*, 2013. 5

[22] C. H. Kennedy, M. Caruso, and T. Thompson. Experimental analyses of gene-brain-behavior relations: some notes on their application. *Journal of Applied Behavior Analysis (Abstracts)*, 34(4):539–549, 2001. 1

[23] 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*, 59(9):809–816, 2002. 4, 5

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012. 7

[25] Y. Kumar, M. L. Dewal, and R. S. Anand. Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine. *Neurocomputing*, 133, June 2014. 2

[26] L. L Itti, C. Koch, E. P. A. Niebur, and M. I. I. T. on. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume:20 , Issue: 11)*, (11). 2

[27] B. B. Le Cun, J. S. Denker, and D. Henderson. Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems*, 1990. 7

[28] M. Morales, P. Mundy, C. E. Delgado, M. Yale, R. Neal, and H. K. Schwartz. Gaze following, temperament, and language development in 6-month-olds: A replication and extension. *Infant Behavior and Development*, 23(2):231–236, 2000. 2

[29] M. M. M. Pimentel. Fragile X syndrome (review). *International Journal of Molecular Medicine*, 3(6):639–645, 1999. 2

[30] J. M. Rehg. Behavior Imaging: Using Computer Vision to Study Autism. In *MVA2011 IAPR, Conference on Machine Vision Applications*, 2011. 2

[31] J. F. Restrepo, G. Schlotthauer, and M. E. Torres. Maximum approximate entropy and r threshold: A new approach for regularity changes detection. *arXiv.org*, nlin.CD, 2014. 5

[32] D. M. Riby, G. Doherty-Sneddon, and L. Whittle. Face-to-face interference in typical and atypical development. *Developmental Science*, 15(2):281–291, 2012. 2

[33] M. Sabeti, S. Katebi, and R. Boostani. Entropy and complexity measures for EEG signal classification of schizophrenic and control participants. *Artificial Intelligence in Medicine*, 47(3), Nov. 2009. 2

[34] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *Neural Networks, IEEE Transactions on*, 13(4):928–938, 2002. 2

[35] K. Sullivan, D. Hatton, J. Hammer, J. Sideris, S. Hooper, P. Ornstein, and D. Bailey. ADHD symptoms in children with FXS. *American Journal of Medical Genetics Part A*, 140(21):2275–2288, 2006. 2

[36] K. Sullivan, D. D. Hatton, J. Hammer, J. Sideris, S. Hooper, P. A. Ornstein, and D. B. Bailey. 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. 2

[37] W. YI and D. Ballard. Recognizing Behavior in hand-eye coordination patterns. *International Journal of Humanoid Robotics*, 06(03):337–359, 2009. 2

[38] C. Yu and L. B. Smith. What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, 14(2):165–180, 2011. 2

[39] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886. IEEE, 2012. 3