Visual attention features for computer-assisted medical diagnosis

Anonymous CVPR submission

Paper ID ****

Abstract

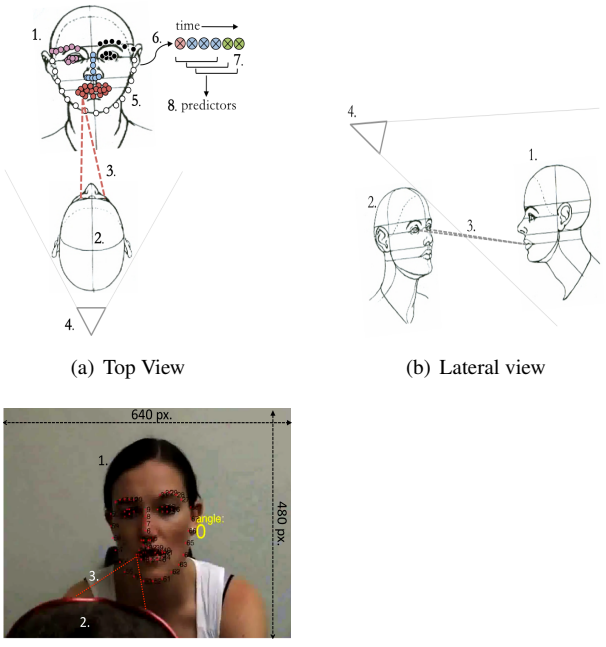
This paper proposes a method to quantitatively distinguish between fine-grained categories of cognitive mental impairments extracting attentional patterns in the eye movements of individuals. We show that there exist short-term visual features that inherently distinguish between known mental conditions. Whereas psychologists typically rely on hours spent with patients and batteries of tests to determine which condition an individual is afflicted by, we demonstrate that an initial screening is possible strictly from multi-modal visual data and standard machine learning classifiers. Our predictors highlight the existence of detectable visual patterns that underlie these diseases. We build a system that can be used to assist medical practitioners in diagnosing these diseases.

1. Introduction

For medical experts, identifying between a patient with and without a developmental disability is relatively simple. However, finer classification between disabilities is a challenging endeavor which typically requires a battery of tests and hours spent on clinical evaluations. Here, we show that using multi-modal video data, sophisticated eye-tracking, and standard classifiers, we can rapidly distinguish between Autism Spectrum Disorder (ASD), more commonly known simply as autism, and a general, non-autistic developmental disorder (DD), on the timescale of seconds. We demonstrate the existence of visual cues that are phenotypically expressed in different ways between these disorders and we develop quantitative evaluation metrics for their assessment.

Autism is a behavioral diagnosis caused by a number of genetic and non-genetic factors. Much remains unknown about the condition. The range of symptoms in autism is wide, but generally a common factor is an impaired ability to communicate and interact socially with other people. Patients may be diagnosed either with autism or autistic-like features, and are typically done so by a pediatrician, neurologist, psychologist or psychiatrist. The diagnosis is made after evaluating the patient using a number of different behavioral tools and tests which suffer from being highly subjective and non-quantitative. When a child is diagnosed with autism, blood tests are often ordered to rule out the known genetic causes. Fragile X Syndrome (FXS) is the most common known genetic (single gene) cause of autism [16], affecting approximately 1 in 3,000 individuals in the United States (approx. 100,000 people).

Whereas autism is a behavioral diagnosis, FXS is a medical or more accurately, a genetic diagnosis. When associated with FXS, the autism is caused by the genetic change or mutation in the Fragile X gene. This is similar to other conditions such as Down syndrome. If a child is diagnosed with autism and then diagnosed with FXS, he or she still has autism, it is just that the cause of their autism is known.



(a) Top View (b) Lateral view

(c) Real View

Figure 1. Environment configuration: 1. Interviewer 2. Participant 3. Eye tracker on participant 4. Monocular color camera. The distance between the participant and the interviewer is $\sim$ 1.5 mts. The camera is located behind and above the participant head to avoid occlusions.

Individuals diagnosed with genetic syndromes associated with developmental disability (e.g., fragile X syndrome , Williams syndrome) often engage in specific forms of aberrant social behavior that can interfere with everyday functioning. For example, individuals with Williams syndrome show a particular form of hypersociability [14, 19]. In the case of FXS-autism children often show prominent eye gaze deficits during social encounters in which they actively seek to avoid social interaction [4, 5]. These social-attentional behavioral phenotypes are considered as important modeling features which drive investigations to better understand the mental impairment [20].

In this paper we present a computer-vision based method capable of assisting health professionals in fine-grained, rapid distinction between FXS and DD without the need of blood tests. This method may be easily extended to distinguish between any number of mental disorders that exhibit visual behavioral phenotypes, and serves as a proof of concept of the role that computer vision can play in medically assistive technologies.

2. Previous Work

2.1. Medical

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 [17, 22, 26, 25]. In particular [18] shows that eye-gaze avoidance during social interactions with others is a particularly prominent behavioral feature of individuals with FXS. On the basis of these studies and others, several authors have suggested that social impairments observed in children with FXS may be caused by different underlying mechanisms than those diagnosed with ASD [4, 5, 16]. For example, children with FXS are able to discriminate between familiar and unfamiliar persons, and exhibit symptoms of social anxiety during social encounters with unfamiliar people. Maintaining appropriate social gaze is a critical prerequisite for language development, emotion recognition, social engagement, and general learning through joint attention [7, 21, 9]. For example, several studies have indicated that high levels of gaze avoidance can negatively impact social interaction skills and communication flow, given that crucial non-verbal gestures and facial expressions that usually aid social interaction will be missed [8, 24]. Given these factors, it seems important to examine and quantify social gaze behaviors in FXS in a naturalistic social setting (i.e., while individuals are actually engaged in a real-life social interaction). Surprisingly few studies have attempted to employ eye-tracking methodology to quantify social gaze behavior during a real-life social setting. In a study conducted by Farzin and colleagues [10, 11] , for example, photographs of faces that depicted various emotions were presented to 16 individuals with FXS (13 male, 3 female) and 16 typically developing controls. Across the 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 age-matched typically developing individuals. Limitations of both studies include the small number of individuals studied, the lack of ecological validity of the experimental setting, and the fact that the controls were not matched on level of functioning to the participants with FXS. In the present study, we therefore recruited a larger sample of individuals with FXS (males and females) and improved the ecological validity of the experimental setting by employing a social partner (a research assistant) who interacted directly with the participants during the experiment.

2.2. Computer vision

It is commonly believed that visual attention is driven by two mechanisms: 1) A bottom-up (BU), task independent, and image-based mechanism that instinctively guides the human eyes into image salient scene regions such as discontinuities, color, texture, motion, etc. [1], 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 [28, 3]. Our work focuses on this second category in the sense that the goal we are studying is the underlying structures behind social interaction engagement. Despite its importance in understanding human behaviors, only a few approaches have addressed this TD mechanism. This believe is summarized in the quote of [27]: "Eye movement reflects the human thought processes; so the observer's thought may be followed to some extent from records of eye movement". Previous TD work tackles the problem of predicting the eye tracking position using a head mounted monocular camera [12]. Here the goal is to mimic eye-tracking information with egocentric video information and not to understand human behavior. Egocentric activity detection by understanding hand-object interactions from a head-mounted camera [13] has been addressed, as well as social event detection by analyzing someone's gaze [15], but none of them are really trying to understand the goal behind the gaze movement of the person wearing the eye-tracker. In this work we build unique features that describe moments of interactions between individuals. The coarseness of our features are more indicative of true person-to-person engagement in social interactions. We show the potential of these features by performing the automatic classification among mental impaired individuals with promising results.

3. Dataset

3.1. Participants

For our fine-grained study we use two groups of professionally-diagnosed participants with either DD or FXS. There are known gender-related behavioral differences between FXS participants, so we further sub-divide them into FXS-Male and FXS-Female.

The participants were recruited over a 3-year period as part of an investigation of brain function and development. Participants with FXS are between 12 and 28 years old and have a genetic diagnosis of FXS. Participants with DD have been diagnosed with a developmental disability, but have been further tested to confirm that they did not have FXS. Exclusion criteria for both groups includes the presence of sensory impairments, or any other serious medical or neurological condition that affected growth or development (e.g., seizure disorder, diabetes, congenital heart disease).

Finally the group of FXS participants is composed of fifty-one participants with FXS (32 male, 19 female) and 19 individuals diagnosed with developmental disability (10 female, 9 male). The two groups were well matched in terms of chronological age and developmental age, as evidenced by similar mean scores obtained on the subscales and composite score of the Vineland Adaptive Behavior Scales (VABS), a well-established measure of developmental functioning. The VABS adaptive behavior composite standard score was 58.47 (SD = 23.47) for individuals with FXS and 57.68 (SD = 16.78) for controls, indicating that the level of functioning in the two groups was almost 3 standard deviations below the mean.

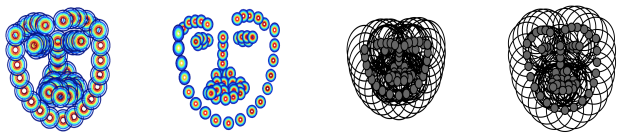
3.2. Experimental Set-Up

The participants were each interviewed for a period of $\sim$10 minutes by a health professional interviewer. The interviews were conducted in a room with minimal visual distractions to control for the participants' attention. The distance between the participants and the interviewer was $\sim$1.5 mts. Figure 1 depicts the physical arrangement of the sensors. The eye-tracker and the camera are synchronized in time. A linear transformation maps the coordinate values of the eye-tracker to the coordinate space of the camera.

4. Feature Extraction & Data Analysis

4.1. Face-mark Extraction

For each video frame we compute a set of face marks in the face of the interviewer. We use a deformable-parts-model based approach [29]. To capture possible dyadic events, we reduce the bias of the interviewer by filtering out the frames where the interviewer is not facing the participant. The frames that are filtered out represent $<$1% of the total number of frames.



(a) (b) (c) (d)

Figure 2. Examples: (a) The worst case of high motion of the interviewer. (b) The most steady interviewer. In all cases the interviewers instructed to limit as much as possible their head movement during the interview. The figures (c) and (d) represent the 20 px threshold radius surrounding the face key points. The threshold is enough to cover all face parts, the same happens for all interviews.

We have computed 58,587 frames for the 19 developmental disorder participants, 56,109 frames for the 19 FXS-female participants, 94,214 frames for the 51 FXS-Male participants. Our frame-rate is 5 frames per second. At each frame we compute 69 face-marks which total 14,414,790 marks over all frames. The quality of the face-marks is done by evaluating a randomly sampled group of 500 frames, where a total of X frames are incorrectly localized. The face-mark extractor is almost perfect. The figure 2 -(a)(b) shows examples of the spacial translation of the face-mark on 2 interviews. We can see that the interviewer's face is spatially stable during the interview and that these results are consistent across the interviews. For additional details, see the supplementary material.

4.2. Time Series Attention Features

The eye-tracking system is calibrated to the camera coordinate system. We map the eye-tracking coordinates and the face marks using a linear transformation at each instant of time t. The binding between the eye-tracking and the face-mark happens when the eye-tracking is within a 20 px. radius distance to the closest face-mark. This value is manually defined such that it covers the entire face of the interviewer. Figure 2(c)(d) show examples of the 20 px. coverage of the face.

In the interest of using more meaningful facial features, we mapped the 69 facemarks into 6 semantic regions: *not looking at the face (0), nose (1), left-eye (2), right-eye (3), mouth (4), jaw (5)*. Figure 3 depicts the average face marks of one of the interviewers grouped by regions. We can thus consider each individual participant to be characterized by a time-series of integer value *0,1,...5* indicating what semantic region they are paying attention to. To convert this time-series into temporal features for learning algorithms we select a time-window of length n and a time-step s to form a feature matrix as follows:

1. Form a temporal feature by selecting n time-consecutive semantic tokens

2. Slide the window across the time-series by a length s

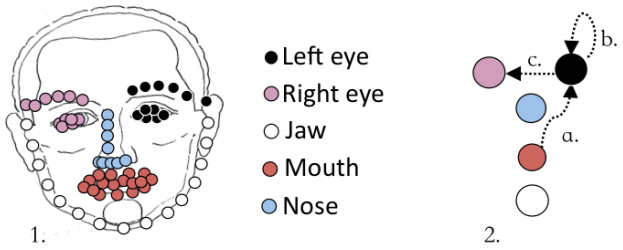


Figure 3. Translation of the Interviewer key marks over time

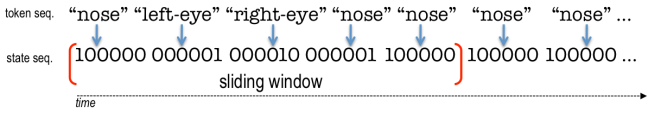


Figure 4. Facial Region State Vectors

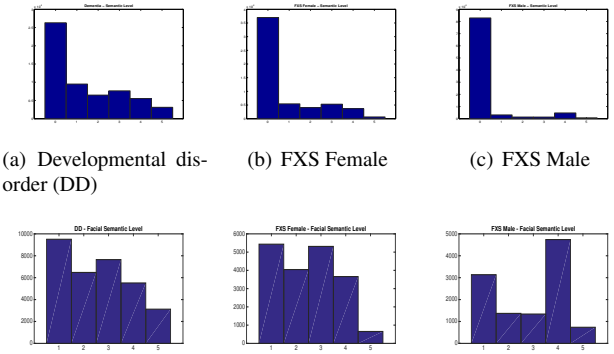
Additionally, we transform these tokens to vectorized representations of states, as shown in Figure 4, to avoid introducing an artificial numerical relationship between the semantic regions. For example the token for nose is $< 0,1,0,0,0,0 >$ and the token "jaw" is $< 0,0,0,0,0,1 >$.

4.3. Fourier Analysis of Features

In the interest of characterizing any attentional oscillations of the patient (say, a constant back-and-forth switching of gaze between nose and mouth), we consider the fourier transform of the time-series of the various mental disorders, and use it in classification. What we find is that in the frequency domain the time-series signals of these patients are characterized by an extremely strong DC component, and very weak and noisy non-DC components. To control for the possibility that this DC component arises from patients gazing away from the face for extended periods of time, we further take the fourier transform of the signal with all non-face frames removed. The same pattern is seen, indicating that patients tend to either gaze at a singular semantic region or flicker their eyes around. As shown in Table 6, when classification is attempted with Fourier components, chance accuracy is attained.

4.4. Coarseness of the features

Here we analyze where patients tend to focus their attention. As was previously stated, patients with developmental disorder tend to look away from the face of the person they are speaking with, due largely in part from a lack of engagement in social interactions. Figure 5 shows a histogram of the six facial regions for all the time-series data for each disorder. For clarity, (a)-(c) shows the histogram including the time-points when individuals are looking away from the face, and (d)-(f) show the same histograms with those time points removed. As can be seen, patients spend the majority of the interview looking away from the interviewer, and only a fraction of the time looking at the interviewers face. Nevertheless, they each exhibit distinct distributions over their facial regions, which can be leveraged in distinguishing between them. Notice that FXS male distinguishes itself quite clearly from the other two not only by its low face counts but from its focus primarily on the mouth (region 4) and nose (region 1). FXS female and DD appear far more similar. In clinical settings gender is always known, so the most important classifications to consider will be between DD and FXS female, and DD and FXS male. These distributions justify the use of face regions in our learning algorithms, as opposed to simply classifying based on whether or not individuals are looking at the face.



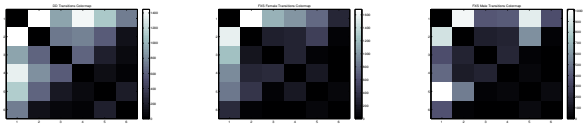
(a) Developmental disorder (DD) (b) FXS Female (c) FXS Male

(d) Developmental disorder (DD) (e) FXS Female (f) FXS Male

Figure 5. Histograms of attention tokens for the various disorders, as well as histograms with token=0 removed.

4.5. Attentional transitions

Another question that arises is the importance of attentional transitions as descriptors of a condition. We consider a transition to be any point in time where the patient changes the facial region that they are paying attention to, and we analyze the transitions that the patients exhibit in order to better understand their visual patterns. Figure 6 shows these transitions in a color map. Each [ij] squares intensity denotes the number of times that patients with a given disorder and gender transition from face region i to face region j. Note that despite the appearance of these plots, they are in fact non-symmetrical and a transition from i to j does not imply a transition from j to i. Again, we see a marked difference between the different disorders and genders. Individuals with DD exhibit greater counts of transitions, whereas those with FXS exhibit significantly less. The real differences between them lie in the inter-facial transitions. Since all disorders and genders spend most of their time looking away from the face, the transitions from non-face to face appear rather similar amongst all of them. However, inter-face transitions are more unique. DD and FXS female both



(a) Developmental disorder (DD) (b) FXS Female (c) FXS Male

Figure 6. Matrix of attentional transitions with respect to the participant condition . Here the numeric values map to: *nose (1), eye-left (2), eye-right (3), mouth (4), jaw (5) and not looking at the face (6)*. While the matrix looks symmetric they are not. In the matrices we can see that the transitions can help to differentiate the condition class.

tend to swap their gaze quite frequently between mouth and nose, whereas FXS male patients do not.

We note that there is significant variance amongst individuals within each group. For example, Figure 7 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 patients also spend noticeable amounts of time looking away from it. We notice that DD, FXS-female, and FXS-male form a spectrum of average frequency, with FXS male looking least frequently at the face, and DD the most. Already these analysis are hinting at the likelihood of higher classification accuracy between FXS males and DD, and harder distinguishment between FXS females and the other two groups.

4.6. Approximate Entropy

The Approximate Entropy ($ApEn$) [23, 2] measures the logarithmic likelihood that if two vectors (x_i^m, x_j^m) representing windows of length m are close (within a distance R) in an m-dimensional space, then they remain close in an $(m+1)$-dimensional space when their window-length is extended. Greater (lesser) likelihood of remaining close produces smaller (larger) $ApEn$ values. To estimate $ApEn(R,m,N)$ for a length N time series $Q = u_1, u_2, ..., u_N$, given the parameters m (window length), $\tau \in \mathbb{N}$ (subsampling coefficient - set here to 1), and $r \in \mathbb{R}^+$, we use the m-dimensional embedded vectors $x_i^m = [u_i, u_{i+\tau}, ui+2\tau, ..., u_{i+(m-1)\tau}]$, with $1 \leq i \leq N-(m-1)\tau$. The embedded vectors are representing ordered subsequences of length m of the time series.

To better understand the complexity of our data, we have measured the $ApEn$ of the our 3 groups (DD, FXS-Female, FXS-Male). For each group we extract 30,000 points from the time series Q of each group. Q contains the information of $\sim$10 participants per group. The participants are chosen randomly. We parametrize $\tau = 1$ to capture the entropy of the shortest period of time allowed by the data resolution (0.2 seconds). Figure 8(a) depicts the $ApEn$ for $m = 3$ and varying the parameter r, where r take the values $r = 0.1i$ for $i = 1...30$, and $R = \text{std}(Q)r$. The FXS-Male (FXSM) data is more stable than DD and FXS-Female (FXSF) data, which share a similar entropy. This makes the data of these last two groups the most challenging to disambiguate in classification.

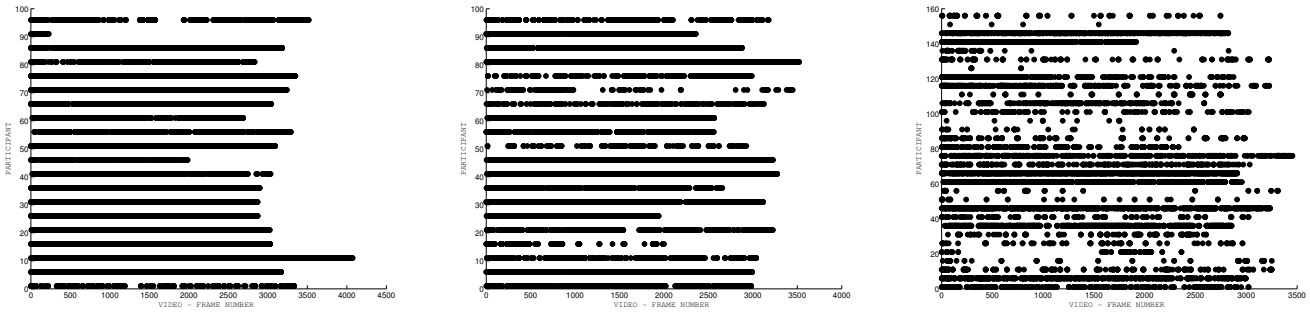
Furthermore, we vary m in order to understand the volatility of the signals from frame to frame. Figure 8(b) depicts such an analysis. Here the parameter R is fixed to $R = \text{std}(Q)r$ which is a reasonable tolerance value according to the results observed in Figure 8(a). The value m varies from $m = 2..100$ with a stride of 5. Once again we see that FXS-Male (FXSM) participants, on average, show the most stable decay for small m while FXS-Female(FXSF) and DD share a similar decay rate, making the last two harder to differentiate automatically.

Simultaneously we see in Figure 9 that there is great variance amongst individuals of each population, making them difficult to classify on an individual basis. This is taken using sequences of 2000 tokens of each participant (6.6 minutes of data). We compute the $ApEn$ by fixing $\tau = 1$, $r = 1$ and vary the dimension of the embedded vectors ($m = 1...35$). We observe that the FXS-Females(FXSF) and DD share similar entropy. The result implies that some of the individuals of the first group will behave in a similar manner as the ones of the second group and vice versa. A similar confusion will happen with some out-liar DD participants but the data seems to be more structured. These results further demonstrate how challenging it is to classify an unknown participant into one of the 3 categories.

5. Classification and Evaluation

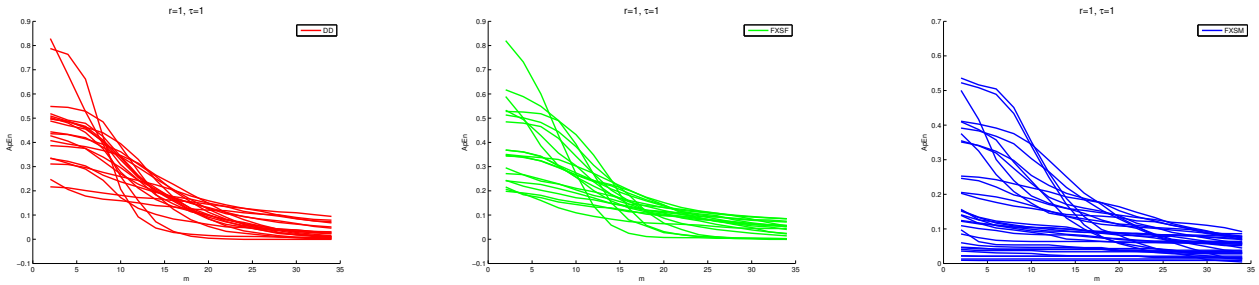
The accumulation of attentional features across the different developmental conditions show clear differences among the groups (Figure 5). We now aim to understand the capability of these visual features at classifying new individuals. We choose to analyze the following pair-wise classifications: DD vs FXS, DD vs FXS male, DD vs FXS female, and FXS male vs FXS female. We consider DD vs FXS male, and DD vs FXS female for their clinical relevance, and we choose the other two in order to better understand the differences in visual phenotypic expression between males and females with FXS, and between all FXS individuals and those with generic DD.

Given the nature of the data, we use classifiers to distinguish between the different disorders. In the construction of the feature matrix for any given experiment, we vary both the step size and window length as hyperparameters of the



(a) DD - Group (b) FXS Female (c) FXS Male

Figure 7. Temporal analysis of attention to face. X axis represents time. Y axis represents each participant. Black dot represent time points where the patient was looking at the interviewer's face. White space signifies that they were not.



(a) $ApEn$ (b) $ApEn$ (c) $ApEn$

Figure 9. Analysis of the Approximate Entropy of the data per individual

data. We train support vector machines (SVMs [6]), naive bayes classifiers, HMMs and our 1-D convolutional neural network described in section 5.2. By using 10-fold cross validation and optimizing over the window length and step size, we find classification accuracies up to 90% for pairwise comparisons. We set chance to be 50% by including the same number of positive and negative patients for each comparison, and we verify that the total time-series data concatenated across patients of any given class is roughly equal to the other class in the comparison.

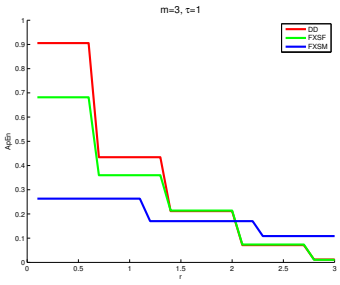
5.1. SVM, NB and HMM classifiers

Our SVM, NB classifiers are trained using a sliding window on the attention features described in section 4.2. On the other hand, our HMM models are trained in the following way. Given N training patients, we concatenate their time-series attention tokens together to form one long time-series. We then select a window length w, as before, and divide up the series into blocks of length w. To inject randomness in the training, these blocks are randomly shuffled before being used to calculate an estimate of the transmission and emission probabilities of an HMM that could generate this data. This HMM has two states, representing the two disorders that are being considered, and each state has 5 possible emissions corresponding to the semantic face regions. These model estimates are then used to classify a testing dataset of patients that were not part of the original N.

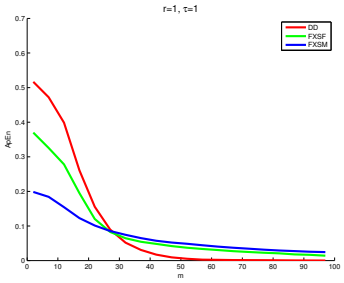
We first train classifiers on small temporal windows of the attentional feature sequences, and we then aggregate the partial window classification results in a voting system that could describe the fit of an unseen participant to one of the developmental conditions.

5.2. State-based CNN: (this section needs to be cleaned up)

We propose a convolutional neural network approach that can exploit the local-temporal relationship between tokens considering their lack of correlation in their representation (i.e. vectorized states). This state-vector based CNN (S-CNN) works with states represented by sparse vectors whose lengths correspond to the number of states. In this case the number of states is the number of facial regions (e.g."nose"). Figure 10(a) depicts this mapping. The input to the S-CNN are sub-sequences of states extracted extracted with a sliding window where $stride = \#states$. We aim at performing binary classification among the groups (DD, FXS-Femal and FXS-Male). Hence, each input sequence is labeled (i.e. $1,0$) corresponding to the individual's condition.

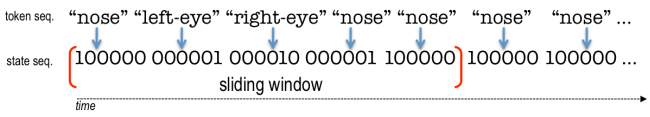


(a) *ApEn* varying the tolerance parameter $R = r * std(Q)$

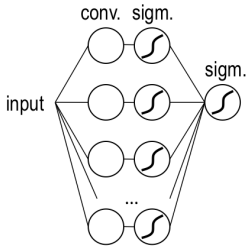


(b) *ApEn* varying the dimension parameter m

Figure 8. Analysis of the Average Approximate Entropy of the data for each participant class.



(a)



(b)

Figure 10. Neural network Design

5.2.1 Setup

Our S-CNN is composed of one hidden layer of 6 convolutional units and a non-linear transformation. The feature map computed at each unit (i.e. a_i) is ordered in a row and feed to an output layer composed of a single unit. Figure 10(b) depicts the architecture of a S-CNN.

The input of the S-CNN is a sub-sequence of states of length n. The length of our states is $\tau = 6$. We configure the S-CNN *stride* value to 12 which corresponds to 2 frames in the video. The kernel size of each unit is set to 24, corresponding to 4 frames of video.

5.2.2 Forward propagation

To compute the pre-nonlinearity input to a unit in our hidden layer, we need to sum up the contributions (weighted by the filter components) from the previous layer cells (a^{l-1}). In our case, each a_i^{l-1} corresponds to a value in the input sequence. The forward pass is computed as follows:

$$x_i^l = \sum_{t=0}^{n-1} w_i^l \, a_{i+\tau t}^{l-1} \tag{1}$$

Then the convolutional layer applies its non linearity.

$$a_i^l = \sigma(x_i^l) \tag{2}$$

Where, σ is the sigmoid function and $\tau = 6$.

5.2.3 Backward propagation

Since our S-CNN is has a single hidden layer the gradients (deltas) are computed output layer values. The delta i serving as input to the unit j of the hidden layer is computed as follows:

$$\delta_{j,i}^l = a_{j,i}^{l-1} \left(1 - a_{j,i}^{l-1}\right) (y - \hat{y}) \tag{3}$$

Where y is the output of the last unit and $\hat{y}$ is the expected value.

The weight gradients computation in the hidden unit needs to consider the inherent structure of the vectorized states. For a unit j, the weight i is updated as follows:

$$\frac{\partial E}{\partial w_{i,j}^l} = \sum_{t=0}^{n-1} \delta_t^{l+1} \, a_{i+\tau t}^l \tag{4}$$

Where $\tau = 6$. We use stochastic gradient descent to update the weights, where the learning ratio is 0.05.

5.2.4 Classification

The S-CNN has an output layer composed of a single unit. The network output value y is a continuous value $[0,1]$, we discretize this value to obtain a discrete a classifier, by thresholding y to 0.5.

5.3. Temporal Window Classification

As can be seen from the plots of Figure 11, SVMs consistently decrease in average training error (over multiple folds) from small time windows of 2 seconds up to mid-size windows of 50 seconds. For longer time windows we see a characteristic spike in the training error followed by significant over-fitting for larger windows of time. In the case of

DD vs FXS classification, generalization error slightly decreases as we increase the time window from 2 to 50. DD vs FXS-female and male vs female both fluctuate in error, while DD vs FXS-male sees first a decrease then an increase in error starting around 30 seconds. They all attain above-chance prediction, even for very short window lengths of only a few seconds, pointing to the fact that visual phenotypic expression of these disorders happens on the timescale of seconds.

DD vs FXS-female classification results exhibit the highest errors at around 0.42, pointing to the similarity in phenotypic expression between female patients afflicted with FXS and general DD patients. On the other hand, DD vs FXS-male classification results are the most accurate, hovering around a generalization error of 0.24 and alluding to the pronounced phenotypical visual differences between FXS males and general DD patients. Males and females exhibit eye patterns which can be distinguished with this classifier with roughly 0.33 error, a similar error to classifying DD vs FXS.

These results are consistent with the intra-group analysis of Figures 11 and 5. There, we saw strongest distinction between DD and FXS male, with FXS female appearing to be a hybrid blend of the two.

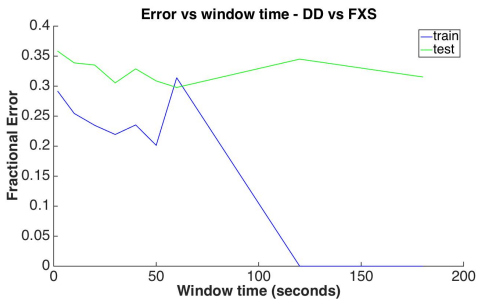
6. Patient Classification

To make these results transferable to clinical settings, we consider the profiling of individuals as having either DD or FXS using these classifiers. To do this we take their time-series data and classify each attentional feature (as described above), and compute a maximum over the count of one class vs the other. As such, we have that the predicted class C_p of an individual patient p is given by:

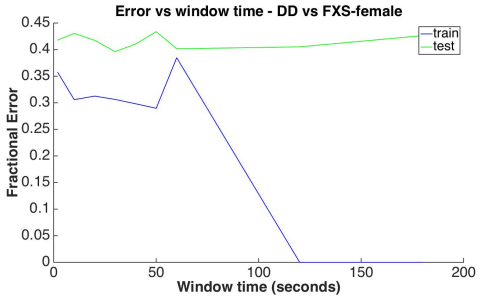
$$C_p = argmax_{C \in \{C_1, C_2\}} \sum_{i=1}^{m} \mathbf{1}(t_i = C) \qquad (5)$$

Where C_1 and C_2 generically refer to one of DD, FXS-female, FXS-male in the pair-wise classifiers, and $\mathbf{1}(x) = \{1 \text{ if x is true; } 0 \text{ otherwise}\}$ is the indicator function.

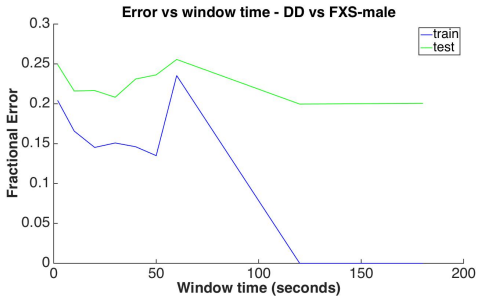
The pairwise profiler accuracy can be find in Table6. We have run profilers based on Naive-Bayes (NB) SVMs, HMMs and S-CNNs window classifications. As expected, the results show better accuracies for profiling DD vs FXS-male, and worse accuracies for DD vs FXS-female. As we can see, NB does a bit worse than SVMs in DD vs FXS-female for time windows of 3 and 10 seconds, whereas its results are actually slightly better than SVMs in DD vs FXS-male over the same time windows. The best performing profilers are based on classifiers that model the temporal correlation of the data: HMMs and S-CNNs. In particular S-CNNs performs slightly better in most of the experiments.



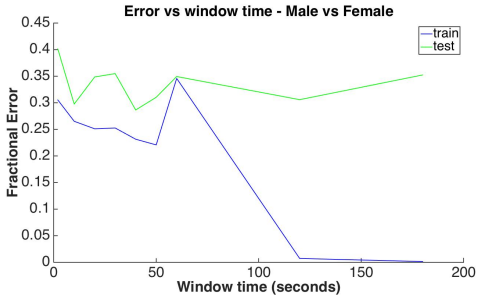
(a) DD vs FXS



(b) DD vs FXS Female



(c) DD vs FXS Male



(d) Male vs Female

Figure 11. Analysis of classifier accuracy as a function of window length for pair-wise classifiers.

We thus see that our temporal window analysis generalizes to the clinical case of assistive patient classification, with comparable accuracy. This further strengthens the result that visual phenotypical expression of these disorders happens on second time-scales and is expressed through

	sec.	DD vs FXS-female	DD vs FXS-male
SVM + FFT	10	0.5	0.5
SVM + TF	3	0.65	0.83
	10	0.65	0.80
	50	0.55	0.85
N.B + TF	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
S-CNN	10	**0.67**	**0.90**
S-CNN	50	0.55	0.77

Table 1. Pairwise Profiling Accuracy of Patients for DD vs FXS-female and DD vs FXS-male

trackable movements of the eyes.

7. Discussion

We hereby demonstrate the use of computer vision and machine learning techniques in a rapid system for assistive diagnosis of mental disorders that exhibit visual phenotypic expression in social interactions. By employing multimodal data, eye-tracking algorithms, semantic region compositionally, and classifiers, we are able to profile individuals as having one mental disorder vs another on a time-scale of a few seconds by analyzing the way they interact with someone else. In particular HMMs and S-CNNs perform the best for profiling implying that the temporal structures in the data need to be considered. Given that individuals with mental disorders of these types exhibit social impairment in the way they interact with others, in particular when being interviewed and asked questions, we are able to leverage this and make a statement about how strongly these phenotypes are exhibited, as well as how similar these disorders are in terms of their impact on patients' social-situation eye patterns.

This work serves as a proof of concept of the power of computer vision systems in assistive medical diagnosis. Trained professionals take hours of effort and batteries of tests to accurately diagnose these disorders, whereas we are able to provide, within a few seconds, 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 visual symptoms.

References

[1] L. A model of saliency-based visual attention for rapid scene analysis Itti, C. Koch, E. P. A. Niebur, and M. I. I. T. on. A model of saliency-based visual attention for rapid scene analysis. *Fathi, Alilreza*, 20(11). 2

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

[3] 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

[4] 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, Mar. 1988. 2

[5] 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, Nov. 1989. 2

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 1995. 6

[7] G. Csibra and G. Gergely. Social learning and social cognition: The case for pedagogy. ... *in brain and cognitive development Attention and ...*, 2006. 2

[8] 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, Jan. 2013. 2

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

[10] 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, June 2009. 2

[11] 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, Nov. 2011. 2

[12] 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, Oct. 2012. 2

[13] 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

[14] E. Frigerio, D. M. Burt, C. Gagliardi, G. Cioffi, S. Martelli, D. I. Perrett, and R. Borgatti. Is everybody always my friend? Perception of approachability in Williams syndrome. *Neuropsychologia*, 44(2):254–259, Jan. 2006. 2

[15] 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. *Fathi, Alilreza*. 2

[16] P. J. Hagerman. The fragile X prevalence paradox. *Journal of medical genetics*, 2008. 1, 2

[17] 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, Oct. 2006. 2

[18] 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, Apr. 2012. 2

[19] W. Jones, U. Bellugi, Z. Lai, M. Chiles, J. Reilly, A. Lincoln, and R. Adolphs. II. Hypersociability in Williams Syndrome. *Journal of Cognitive Neuroscience*, 12(supplement 1):30–46, Mar. 2000. 2

[20] 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, Jan. 2001. 2

[21] 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, Jan. 2000. 2

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

[23] 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, May 2014. 5

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

[25] 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, Nov. 2006. 2

[26] 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, June 2007. 2

[27] A. L. Yarbus, B. Haigh, and L. A. Rigss. Eye movements and vision — Clc. 1967. 2

[28] W. YI and D. BALLARD. RECOGNIZING BEHAVIOR IN HAND-EYE COORDINATION PATTERNS. *International Journal of Humanoid Robotics*, 06(03):337–359, Sept. 2009. 2

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